Computation of Scientific References Using Vector Space Model over Cosine Similarity and Hamming Distance (Case Study: Department of Informatics UNS)

Lydia Permata Sari
Program Studi Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36A Surakarta
lydia.ps@student.uns.ac.id

Ristu Saptono S.Si., M.T
Program Studi Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36A Surakarta
ristu.saptono@staff.uns.ac.id

Esti Suryani S.Si., M.Kom
Program Studi Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36A Surakarta
estisuryani@staff.uns.ac.id

ABSTRACT

Bibliography on final project contains references that are used by writer. References cited in the form of books, journals, magazines, newspapers or internet. Scientific works in the form of journals and proceedings used as references also are listed in the bibliography. Calculation of scientific works that have been noted in the bibliography can be used to calculate the frequency of the use of scientific works used by undergraduate students of Informatics UNS in the form of Vector Space Model using Cosine Similarity and Hamming Distance.

Calculation of the frequency of use of the work of the scientific method using Vector Space Model to find the value of the weight of the title scientific papers where the results will be used to identify the title similarity search papers by using the method of Cosine Similarity. Hamming Distance method used to calculate the distance of the similarity of author names and year. The data used for this research is 100 final project documents. There are 47% of eligible documents Harvard and APA.

The result of the calculation has been done i.e. There are 27 same scientifi works, and produces 12 scientific papers that have a frequency of occurrence of twice or more. There are 5 scientific papers originating from within the country, and 7 scientific papers which came from international.

Keywords

Cosine Similarity, Frequency of Scientific Papers, Hamming Distance, Vector Space Model.

1. PENDAHULUAN

Syarat akhir untuk memperoleh gelar sarjana seorang mahasiswa adalah menyelesaikan skripsi yang melalui proses pembimbingan yang mengikuti syarat penulisan karya ilmiah berdasarkan hasil penelitian sesuai dengan bidang yang diambil. Tugas akhir memiliki tiga bagian yaitu bagian awal, utama dan akhir. Bagian awal meliputi halaman sampul, judul, persetujuan, pengesahan, persembahan, kata pengantar, daftar isi, daftar tabel, daftar gambar, daftar lampiran. Bagian utama terdiri atas lima bagian yaitu, bagian pendahuluan, landasan teori, metodologi, pembahasan dan penutup. Bagian akhir terdiri atas daftar pustaka dan daftar lampiran**.** Daftar pustaka berisi semua sumber pustaka yang digunakan pada laporan tugas akhir. Daftar pustaka dapat digunakan untuk membantu pembaca dalam mencari sumber kutipan yang digunakan dalam skripsi. Isi daftar pustaka diurutkan berdasarkan nama penulis secara alfabetis tanpa gelar kesarjanaan. Sumber yang dikutip berupa buku, jurnal, majalah, surat kabar atau internet. Daftar pustaka memiliki minimal 2 jurnal terbaru [1].

Perhitungan karya ilmiah yang telah dicantumkan di daftar pustaka dapat digunakan untuk menghitung frekuensi penggunaan karya ilmiah yang digunakan oleh mahasiswa S1 Informatika UNS. Dengan mengidentifikasi karya ilmiah yang telah digunakan dalam skripsi dapat mengetahui karya ilmiah mana saja yang berasal dari dalam institusi UNS. Pemanfaatan referensi karya ilmiah yang berasal dari dalam institusi mahasiswa tersebut akan mempermudah untuk menyelesaikan penelitiannya [2].

Penelitian yang bertujuan untuk menghitung frekuensi dapat dimanfaatkan untuk mengetahui trending topik sesuai dengan obyek penelitian yang digunakan. Analisis mengenai perhitungan frekuensi yang tinggi dapat menjadi topik yang trending telah dilakukan dalam beberapa penelitian sebelumnya, misalnya [3], [4].

Proses pencarian kesamaan karya ilmiah yang digunakan menggunakan metode *Vector Space Model* (VSM) terlebih dahulu. *Vector Space Model* (VSM) adalah metode yang digunakan untuk menghitung kemiripan suatu dokumen dan suatu query dengan mewakili setiap dokumen dalam sebuah koleksi sebagai sebuah titik dalam ruang. Metode VSM berguna untuk menghitung jarak antar dokumen, kemudian diurutkan berdasarkan tingkat kedekatannya, sehingga semakin kecil jarak antar dokumen, maka semakin tinggi tingkat kemiripannya [6].

Proses pencarian kemiripan dapat dipercepat dengan menggunakan algoritma *Hamming Distance*. *Hamming Distance* merupakan metode pengujian untuk mencari seberapa mirip sebuah vector terhadap vector lainnya berdasarkan nilai kedekatannya. Jika nilai kedekatan semakin kecil maka artinya kemiripan kedua vector semakin besar sebaliknya jika nilai kedekatan semakin besar artinya kemiripan kedua vector semakin kecil. Algoritma *Hamming Distance* sangat membantu dalam mempercepat eksekusi sistem [7].

Berdasarkan penelitian yang telah dilakukan sebelumnya, penulis ingin melakukan perhitungan penggunaan referensi karya ilmiah pada daftar pustaka tugas akhir mahasiswa dalam bentuk metode *Vector Space Model* dengan menggunakan metode *Cosine Similarity* untuk mencari kesamaan judul karya ilmiah yang telah digunakan. Selain itu juga menggunakan metode *Hamming Distance* yang dapat mempercepat pencarian kemiripan nama pengarang karya ilmiah.

2. DASAR TEORI

2.1 Tata Cara Penulisan Daftar Pustaka

Daftar pustaka adalah bagian laporan tugas akhir yang berisi semua sumber pustaka yang digunakan pada kutipan laporan tugas akhir. Semua kutipan pustaka yang digunakan dalam laporan tugas akhir harus ada dalam daftar pustaka. Daftar pustaka memiliki minimal dua jurnal terbaru [1].

Standart format penulisan karya ilmiah yang digunakan di S1 Informatika UNS adalah menggunakan standar penulisan referensi Harvard. Berikut merupakan elemen-elemen yang diberikan untuk penulisan referensi yang berasal dari jurnal:

- Nama Pengarang
- Tahun Publikasi
- Judul Artikel
- Nama jurnal ditulis menggunakan format *italic*
- Nomor Volume
- Nomor isu
- Nomor halaman

Selain menggunakan standar penulisan referensi Harvard, juga digunakan standar referensi penulisan APA *Style.* APA *Style* merupakan salah satu bentuk sitasi yang dikeluarkan oleh organisasi APA terutama untuk bidang psikologi dan social. Beberapa ciri gaya penulisan sitiran dari APA *Style* adalah [8]:

1. Daftar Pustaka diurutkan alfabetis berdasarkan Nama Belakang Penulis atau Judul apabila tidak ada penulis.
2. Nama Depan penulis ditulis sebagai inisial.
3. Apabila ada penulis sama dalam daftar pustaka ditulis berurutan dari tahun yang paling lama.
4. Bisa ditambahkan huruf a,b,c, setelah tahun.

2.2 Preprocessing Process

Preprocessing merupakan proses untuk mengubah suatu format dokumen menjadi format yang sesuai agar dapat diproses. Terdiri dari tiga langkah proses *preprocessing*, yaitu: *tokenization, stopword removal, stemming* [8].

2.2.1 *Tokenization*

Tokenizing adalah tahapan pemotongan *string input* berdasarkan setiap kata yang menyusunnya.

2.2.2 *Stopword Removal*

Stopword Removal merupakan penghapusan kata yang tidak penting dalam sebuah kalimat. Dalam perhitungan frekuensi karya ilmiah ini tidak semua kata dalam sebuah daftar pustaka itu penting. Proses ini disebut disebut *stopword removal* atau *stopword filtration*. Kata yang telah diseleksi akan dicocokan dengan *database stopword*. Jika kata tersebut terdapat pada *database stopword* maka kata itu dihapus.

2.2.3 *Stemming*

Stemming merupakan proses pengurangan varian betuk kata menjadi bentuk kata dasar [9]. Algoritma *stemming* biasanya menghapus sufiks dan prefiks untuk mendapatkan kata dasar. Teknik penghapusan afiks akan menghapus prefiks dan sufiks. Jenis Algoritma Stemming yang digunakan didalam penelitian ini adalah:

2.2.3.1 Algoritma Porter

Proses pada setiap tahap algoritma Porter dilakukan berdasarkan keadaan kata, sufiks dan prefiks. Algoritma porter terdiri atas 5 tahap [10].

2.2.3.2 Algoritma Nazief-Adriani

Algoritma ini menggunakan kamus kata dasar untuk memeriksa apakan kata yang distem sudah mencapai kata dasar [11]. Algoritma Nazief–Adriani terdiri atas 5 langkah yaitu *finding root word,* membuang *inflection suffixes,* hapus hapus *derivation suffixes, derivation prefix, recoding*.

2.3 Naïve Bayes Classifier

Naive Bayes Classifier adalah algoritma klasifikasi probabilitas berdasarkan pada teorema Bayes dengan asumsi yang sangat kuat (naif) akan independesi dari masing-masing kondisi. Naive Bayes Classifier dikenal sebagai algoritma klasifikasi Bayes sederhana [12].

Apabila terdapat kejadian A dan kejadian B, maka teorema Bayes dirumuskan sebagai berikut:

$$P(A|B) = \frac{P(A)}{P(B)} P(B|A) \quad (1)$$

Secara sederhana, rumus ditulis sebagai berikut:

$$Posterior = \frac{prior\ x\ likelihhood}{evidence} \quad (2)$$

Nilai evidence selalu tetap untuk tiap kelas pada satu sampel. Penentuan klasifikasi sampel menjadi kelas yang berbeda dengan cara membandingkan nilai Posterior tiap kelas.

Rumus Bayes dapat dituliskan sebagai berikut:

$$P(C|F_1, \dots, F_n) = \frac{P(c)}{Z} \prod_{i=1}^{n} P(F_1|C) \quad (3)$$

Rumus no 3 di atas adalah rumus teorema *Naive Bayes Classifier* yang akan digunakan untuk proses klasifikasi dokumen. Penentuan kelas yang cocok dilakukan dengan cara membandingkan nilai *posterior* untuk masing-masing kelas dan mengambil kelas dengan nilai *posterior* tertinggi Rumus matematis klasifikasi sebagai berikut :

$$C_{NB} = argmax_{c \in C} P(C) \prod_{i=1}^{n} P(F_i|C) \quad (4)$$

Dengan c yaitu variable kelas yang tergabung dalam suatu himpunan kelas C_i [13].

Agar nilai probabilitas kondisional pada *Naive Bayes Classifier* dapat bernilai 0, maka digunakan teknik *smoothing*. Teknik yang kerap digunakan pada algoritma *Naive Bayes Classifier* adalah *Laplacian Smoothing*. Cara yang digunakan adalah dengan menambahkan angka 1 pada perhitungan *Likelihood* [14].

Perhitungan nilai *Likelihood* seperti berikut ini :

$$P(F_1|C_i) = \frac{1 + n\ (F_1 C_i)}{s_i + n(C_i)} \quad (5)$$

2.4 Vector Space Model

Metode untuk melihat tingkat kedekatan atau kesamaan (*similarity)* term dengan cara pembobotan term disebut dengan *Vector Space Model* (VSM). *Vector Space Model* sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vektor. Relevansi sebuah dokumen ke sebuah *query* didasarkan pada similaritas diantara vektor dokumen dan vektor *query* [15].

2.4.1 Pembobotan Kata

Menurut Robertson (2005), pemberian bobot hubungan suatu term terhadap dokumen dapat menggunakan metode TF-IDF. Metode ini menggabungkan dua konsep perhitungan bobot yaitu frekuensi kemunculan suatu kata dalam suatu dokumen dan *inverse* dari frekuensi yang mengandung kata tersebut. Persamaan dalam perhitungan TF-IDF adalah :

$$W_{(t,d)} = tf_{(t,d)} x\ idf \quad (7)$$

$$W_{(t,d)} = tf_{(t,d)} x Log \frac{d}{df_t} \quad (8)$$

Dimana :

$W_{(t,d)}$: bobot term t pada dokumen d
$tf_{(t,d)}$: jumlah kemunculan term t dalam dokumen d
d : jumlah seluruh dokumen
df_t : jumlah dokumen yang memiliki term t

2.4.2 *Cosine Similarity*

Metode *Cosine Similarity* diperlukan untuk menentukan nilai tingkat hubungan antara dua parameter yang akan menunjukkan apakah kedua parameter tersebut independen (saling bebas) atau tidak. Dalam algoritma *clustering*, pemilihan fungsi similaritas antar objek menjadi kunci keberhasilan algoritma. Untuk tujuan *clustering* dokumen, jarak fungsi yang paling baik adalah fungsi similaritas *Cosine*. (Hamzah, Soesianto, Susanto, & Istiyanto, 2008).

Jika dimisalkan dokumen x direpresentasikan oleh vektor $\vec{x} = \{x_1, x_2, \dots, x_t\}$ dan dokumen y direprensentasikan oleh vektor $\vec{y} = \{y_1, y_2, \dots, y_t\}$. Dengan $|\vec{x}|$ dan $|\vec{y}|$ merupakan panjang x dan y. Rumus untuk Cosine Similarity adalah sebagai berikut:

$$sim(x,y) = \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \times |\vec{y}|} = \frac{\sum_{i=1}^{t}(W_{i,x} \times W_{i,y})}{\sqrt{\sum_{i=1}^{t} W_{i,x}^2} \times \sqrt{\sum_{i=1}^{t} W_{i,y}^2}} \quad (9)$$

Nilai sim(x,y) bervariasi dari 0 sampai 1. Nilai ini menunjukkan bahwa semakin tinggi nilai sim(x,y), maka semakin besar kemiripan dari kedua vektor tersebut [17].

2.5 Hamming Distance

Hamming distance adalah salah satu algoritma mengukur kedekatan *item*. Jika nilai jarak makin kecil, maka kedua *item* itu semakin dekat dan berlaku sebaliknya. Yang biasanya dibandingkan adalah kata dan bilangan *biner* [17].

Selain digunakan untuk membandingkan kata dan bilangan biner, algoritma Hamming Distance juga dapat digunakan pada kalimat. Algoritmanya sebagai berikut :

1. Cek panjang kalimat. Sebagai contoh berikut :
   Kalimat 1 = {(gusi),(berdarah)}
   Kalimat 2 = {(gigi),(berdarah)}
   Analisa panjang kedua kalimat :
   - Kalimat 1 memiliki panjang 2 anggota himpunan, dengan anggota himpunan : gusi dan berdarah.
   - Kalimat 2 memiliki panjang 2 anggota himpunan, dengan anggota himpunan : gigi dan berdarah.
2. Pilih pembanding . Misal Kalimat 1 sebagai pembanding.
3. Cek anggota himpunan yang posisinya sama. Jika setiap anggota anggota himpunan pada kedua himpunan memiliki kata yang sama, maka nilainya 0, dan jika berbeda akan diberi nilai 1.

Perhitungan *Hamming Distance* dapat ditunjukkan pada tabel 1:

Tabel 1. Perhitungan *Hamming Distance* pada Himpunan Kata

Dokumen 1	Dokumen 2	Status	Nilai Hamming
gusi		Beda	1
	gigi	Beda	1
berdarah	berdarah	Sama	0
TOTAL			2

Karena total kedekatan dokumen 1 dan dokumen 2 adalah 2, maka dapat disimpulkan bahwa Kalimat 1 dan Kalimat 2 mirip.

2.6 Reguler Expression

Regular Expression adalah bahasa yang digunakan untuk *parsing* dan memanipulasi data teks. Sering digunakan untuk menyelesaikan masalah pengolahan teks yang kompleks. *Regular Expression* atau sering disingkat menjadi *RegEx* menggunakan beberapa simbol tertentu untuk menentukan pola yang terdiri atas meta-karakter dan repetisi. Meta-karakter adalah simbol yang digunakan untuk penanda *pattern*. Repetisi adalah simbol yang menyatakan pengulangan *pattern* karakter. Berikut merupakan daftar meta-karakter *RegEx* PHP ditunjukkan pada tabel 2.4. (Stubblebine, 2007)

Tabel 2. Meta-karakter *RegEx* pada PHP

Simbol	Fungsi
/	Mengawali dan mengakhiri *pattern*
^	Mencocokan *pattern* yang terletak pada awal subjek
$	Mencocokan *pattern* yang terletak pada akhir subjek
.	Mencocokan dengan karakter apapun, kecuali baris baru
.*?	Mencocokan dengan karakter apapun, termasuk baris baru
[]	Membuka dan menutup definisi *character class*
|	Tanda pemisah dari untuk opsi alternatif
()	Membuka dan menutup sub-*pattern*
\	Karakter *escape*
{x,y}	Pembilang repetisi dengan nilai minimal x dan maksimal y
?	Pembilang repitisi minimal nol dan maksimal satu {0, }
*	Pembilang repitisi minimal nol dan maksimal tidak terbatas {0,1}
+	Pembilang repitisi minimal satu dan maksimal tidak terbatas {1, }

3. METODOLOGI

3.1 Studi Literatur

Studi Literatur digunakan untuk mendapatkan dasar – dasar referensi yang kuat untuk membantu dalam penyusuan laporan penelitian. Dengan mengumpulkan literatur yang berkaitan dengan, Metode *Vector Space Model, Stemming*, *Cosine*

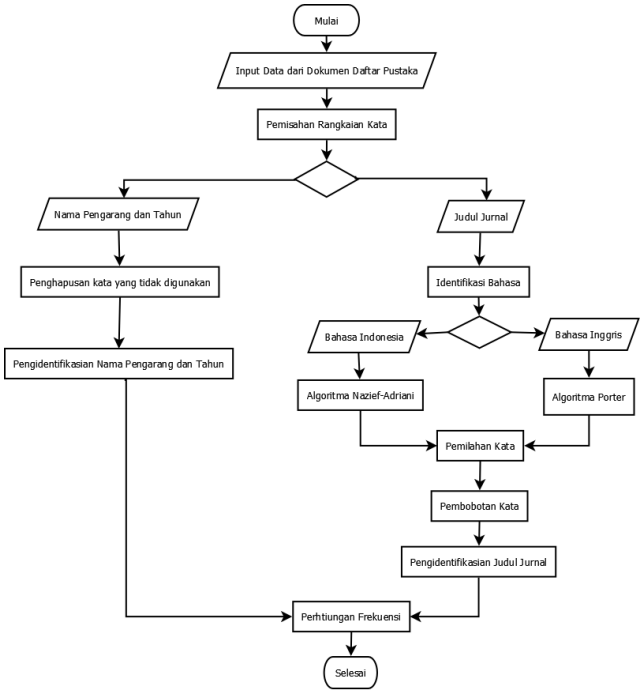
Similarity, Hamming Distance. Sumber yang dapat digunakan untuk literatur berupa jurnal, karya ilmiah, buku, teks, paper, dan situs – situs yang dapat menunjang penyelesaian penelitian.

3.2 Pengumpulan Data

Pengumpulan data digunakan untuk mengumpulkan objek yang akan digunakan untuk penelitian ini. Data yang akan digunakan diambil merupakan Tugas Akhir Mahasiswa S1 Informatika UNS.

3.3 Penerapan Metode

Gambar 1 menunjukkan proses penerapan metode pada system



Gambar 1. Proses Penerapan Metode pada Sistem

3.3.1 *Input* Data

Data yang digunakan dalam proses merupakan dokumen daftar pustaka dari tugas akhir mahasiswa S1 Informatika UNS. Daftar pustaka yang digunakan hanya yang menggunakan sistem penulisan daftar pustaka Harvard. Selain itu dihapus. Karya Ilmiah yang digunakan merupakan jurnal dan *proceeding*. Dokumen sudah dalam *format* .doc.

3.3.2 Pemisahan Rangkaian Kata

Proses ini berguna untuk memotong *string* judul karya ilmiah menjadi rangkaian kata, menghapus tanda baca yang tidak diperlukan dan mengubah huruf besar menjadi huruf kecil

3.3.3 Judul Karya Ilmiah

3.3.3.1 Identifikasi Bahasa

Metode *Naïve Bayes Classifier* digunakan untuk mengidentifikasi bahasa pada judul karya ilmiah. Metode *Naïve Bayes Classifier* berfungsi untuk mengklasiffikasi dokumen menjadi dua klasifikasi, yaitu Bahasa Indonesia dan Bahasa Inggris.

3.3.3.2 Penghapusan kata-kata yang tidak digunakan

Setelah data telah diklasifikasi menjadi 2 yaitu Bahasa Indonesia dan Bahasa Inggris, data yang sudah di-*token* akan melalui proses *stopword removal*. Proses *stopword removal* bertujuan untuk menghapus kata yang tidak penting. Untuk Bahasa Indonesia kata-kata yang tidak digunakan yaitu dan, atau, untuk, dengan, jika, dan lain-lain. Sedangkan untuk Bahasa Inggris, kata-kata yang tidak digunakan adalah *if, or, and, is, are, am, for, with*. Kata-kata tersebut akan dihapus.

3.3.3.3 Pemilahan Kata Dasar

Pemilahan kata dasar ini menggunakan dua algoritma berdasarkan klasifikasinya. Untuk data yang masuk dalam klasifikasi Bahasa Indonesia akan menggun akan Algoritma *Stemming* Nazief-Adriani. Sedangkan data yang masuk dalam klasifikasi Bahasa Inggris akan menggunakan Algoritma *Stemming* Porter. Pencarian kata dasar berdasarkan *library* kata dasar di dalam *database*.

3.3.3.4 Pembobotan Kata

Setiap kata dalam data yang sudah dikelompokkan kata dasarnya akan melalui proses pembobotan kata. Proses ini berguna untuk menghitung nilai kemunculan kata didalam data yang digunakan. Perhitungan menggunakan perhitungan pembobotan TF-IDF.

3.3.3.5 Pengidentifikasian Judul Karya Ilmiah

Pengidentifikasian judul karya ilmiah dengan menggunakan metode *Cosine Similarity*. Judul karya ilmiah yang sama akan memiliki nilai Cos 1.

3.3.4 Nama Pengarang dan Tahun

Proses identifikasi nama pengarang dan tahun akan menggunakan algoritma Hamming Distance. Algoritma ini mengukur kedekatan item. Semakin kecil nilai kedekatannya maka semakin dekat item tersebut.

3.3.5 Perhitungan Frekuensi

Karya Ilmiah yang sama atau mirip akan dihitung frekuensi kemunculannya dalam data tugas akhir mahasiswa S1 Informatika UNS. Semakin banyak frekuensi kemunculan suatu karya ilmiah yang digunakan, maka karya ilmiah tersebut banyak digunakan oleh mahasiswa S1 Informatika UNS.

4. HASIL DAN PEMBAHASAN

4.1 Deskripsi Data

Data diperoleh dari bagian Administrasi Jurusan Informatika UNS. Data yang akan digunakan adalah Tugas Akhir Mahasiswa S1 Informatika UNS. Dari Laporan yang didapat hanya bagian Daftar Pustaka saja yang digunakan untuk penelitian. Data yang diperoleh dari bagian Administrasi dalam bentuk CD. Dimana didalam tersebut terdapat kumpulan Tugas Akhir. Total data yang didapatkan adalah 100 data. Dokumen Tugas Akhir yang diambil berasal dari mahasiswa angkatan tahun 2007 hingga 2010.

Dokumen daftar pustaka yang akan digunakan sudah harus dipisahkan dari dokumen Tugas Akhir. Selain itu untuk mempercepat waktu, dokumen sudah harus dalam *format* *.docx* atau format Microsoft Word tahun 2010 keatas.

4.2 Penerapan Metode

4.2.1 *Input* Data

Data yang harus dimasukkan terlebih dulu adalah informasi dari dokumen daftar pustaka yang akan dimasukkan kedalam sistem. Informasi yang harus dimasukkan merupakan identitas mahasiswa dan informasi Tugas Akhir. Terdiri dari: Nomor Induk, Nama Lengkap, Tahun Skripsi dan Judul Skripsi. Identitas mahasiswa tersebut akan dimasukkan dalam database mahasiswa.

Dokumen daftar pustaka yang sudah sesuai dengan format lalu diupload kedalam sistem. Dokumen diupload sesuai dengan nama mahasiswa yang telah dimasukkan. Judul karya Ilmiah yang sesuai dengan format *Harvard* dan *APA* yang akan dimasukkan kedalam *database* daftar pustaka.

Dari 100 dokumen daftar pustaka tugas akhir, terdapat 47% dokumen yang penulisannya sesuai dengan standar penulisan. Terdapat 264 karya ilmiah yang sesuai dengan format *Harvard dan* APA. Namun, dari total 264 daftar pustaka yang terdeteksi tidak semuanya merupakan karya ilmiah. Data yang bukan karya ilmiah tersebut merupakan daftar pustaka dengan sumber buku, web maupun majalah, namun karena penulisannya yang mirip dengan penulisan karya ilmiah dengan format Harvard dan APA maka terdeteksi sebagai karya ilmiah.

4.2.2 Pemisahan Rangkaian Kata

Karya Ilmiah yang telah sesuai dengan format Harvard lalu melalui proses *tokenizing* dimana akan dipecah menjadi 4 bagian yaitu Nama Pengarang, Tahun Terbit, Judul Karya Ilmiah dan Nama Jurnal Karya Ilmiah.

adaptive probabilities of crossover and mutation in genetic algorithms

4.2.3 Judul Karya Ilmiah

4.2.3.1 Identifikasi Bahasa

Untuk mengetahui bahasa yang digunakan judul karya ilmiah menggunakan metode *Naïve Bayes Classifier.* Sebelum memulai metode NBC, judul karya ilmiah tersebut perlu di token lagi untuk memecah kata-kata. Berikut merupakan hasil *tokenizing* judul karya ilmiah :

Berikut merupakan perhitungan *Prior* dan *Likelihood*:

Tabel 3. Perhitungan *Prior* dan *Likelihood*

Kata	$n(F_i, C)$		*P(C)*		*P(X|C)*	
	Bhs Indo	Bhs Ing	Bhs Indo	Bhs Ing	Bhs Indo	Bhs Ing
adaptive	0	1	0	0,111	0,111	0,111
probabilities	0	1	0	0,111	0,111	0,111
of	0	1	0	0,111	0,111	0,111
crossover	0	1	0	0,111	0,111	0,111
and	0	1	0	0,111	0,111	0,111
mutation	0	1	0	0,111	0,111	0,111
in	0	1	0	0,111	0,111	0,111
genetic	0	1	0	0,111	0,111	0,111
algorithms	0	1	0	0,111	0,111	0,111

Setelah itu menghitung nilai *Posterior:*

Tabel 4. Perhitungan Nilai *Posterior*

Kata	*P(C|X)*	
	Bahasa Indonesia	Bahasa Inggris
adaptive	0	0.00000000258117
probabilities		
of		
crossover		
and		
mutation		
in		
genetic		
algorithms		

Berdasarkan tabel diatas nilai *Posterior* untuk Bahasa Inggris lebih tinggi dibandingkan dengan nilai *Posterior* Bahasa Indonesia dengan nilai 0.00000000258117. Maka dapat ditentukan bahwa kalimat "Adaptive Probabilities of Crossover and Mutation in Genetic Algorithms" adalah kalimat Bahasa Inggris.

4.2.3.2 Penghapusan Kata-kata yang tidak digunakan

Setelah kalimat judul karya ilmiah telah diidentifikasi bahasa yang digunakan, maka langkah selanjutnya adalah penghapusan kata-kata yang tidak digunakan. Daftar kata yang akan dihapus dibagi menjadi 2, yaitu kata Bahasa Indonesia dan kata Bahasa Inggris. Untuk kata Bahasa Indonesia terdapat 357 kata henti, sedangkan untuk kata Bahasa Inggris terdapat 319 kata henti.

Contoh kata henti untuk Bahasa Inggris adalah a, an,is, are, and, the, if, for, of. Sedangkan untuk contoh kata henti Bahasa Indonesia adalah untuk, apabila, jika, dan, dari.

4.2.3.3 Pemilahan Kata Dasar

Kata yang akan diproses ini akan diubah menjadi bentuk kata dasar dari kata tersebut. Proses pemilahan kata dasar menggunakan 2 algoritma tergantung dari klasifikasi bahasa dari kata tersebut. Bahasa Indonesia menggunakan algoritma Nazief-Adriani sedangkan untuk bahasa inggris menggunakan algoritma Porter.

Contoh judul karya ilmiah ini masuk dalam klasifikasi Bahasa Inggris, maka akan diproses dengan menggunakan algoritma Porter. Berikut merupakan proses pemilahan kata dasar Bahasa Inggris yang telah melewati proses penghapusan kata yang tidak digunakan.

Proses *Stemming* algoritma Porter terdapat 5 langkah kerja, yaitu:

1. Menghapus huruf *particle.*
2. Menghapus kata ganti.
3. Menghapus huruf awal. Lanjut langkah 4a apabila tidak ada, Langkah 4b apabila ada
4. a. Hapus awalan kedua, lanjutkan ke langkah 5a.

b. Hapus akhiran, jika tidak ada maka kata tersebut adalah *root word*. Lanjut langkah 5b apabila ditemukan.

5. a. Hapus akhiran. Kemudian kata akhir diasumsikan sebagai *root word.*
   b. Hapus awalan kedua. Kemudian kata akhir diasumsikan sebagai *root word*.

4.2.3.4 Pembobotan Kata

Kata yang telah melalui proses *stemming* diatas dihitung bobotnya dengan menggunakan metode *Vector Space Model* perhitungan TF-IDF.

Contoh hasil *text preprocessing* ditunjukkan pada tabel 5:

Tabel 5. Contoh Hasil *Text Preprocessing*

Dok	Judul Karya Ilmiah	Hasil *Text Preprocessing*
D1	Adaptive Probabilities of Crossover and Mutation in Genetic Algorithms	Adapt probability crossover mutation genetic algorithm
D2	Adaptive Probabilities of Crossover and Mutation in Genetic Algorithms	adapt probability crossover mutation genetic algorithm
D3	Classification of Breast Cancer by Comparing Backpropagation Training Algorithms	classification breast cancer compare backpropagation train algorithm

Hasil perhitungan pada tabel 5 untuk pembobotan kata ditunjukkan pada tabel 6 :

Tabel 6. Perhitungan Pembobotan Kata

Term (t)	tf				idf	W=tf x idf		
	D1	D2	D3	df		D1	D2	D3
adapt	1	1	0	2	0.17	0.17	0.17	0
algorithm	1	1	1	3	0	0	0	0
backpropagation	0	0	1	1	0.47	0	0	0.47
breast	0	0	1	1	0.47	0	0	0.47
cancer	0	0	1	1	0.47	0	0	0.47
classification	0	0	1	1	0.47	0	0	0.47
compare	0	0	1	1	0.47	0	0	0.47
crossover	1	1	0	2	0.17	0.17	0.17	0
genetic	1	1	0	2	0.17	0.17	0.17	0
mutation	1	1	0	2	0.17	0.17	0.17	0
probability	1	1	0	2	0.17	0.17	0.17	0
train	0	0	1	1	0.47	0	0	0.47

4.2.3.5 Pengidentifikasian Judul Karya Ilmiah

Proses mengidentifikasi judul karya ilmiah yang sama dihitung menggunakan metode *Cosine Similarity*. Proses perhitungan membandingkan kemiripan D1, D2, dan D3 kedalam bentuk vector.

Tabel 7. Perhitungan Panjang Vektor dan Bobot Kombinasi

Term (t)	Panjang vektor			Nilai Bobot		
	D1	D2	D3	D1.D2	D1.D3	D2.D3
adapt	0.03	0.03	0	0.03	0	0
algorithm	0	0	0	0	0	0
backpropagation	0	0	0.228	0	0	0
breast	0	0	0.228	0	0	0
cancer	0	0	0.228	0	0	0
classification	0	0	0.228	0	0	0
compare	0	0	0.228	0	0	0
crossover	0.03	0.03	0	0.03	0	0
genetic	0.03	0.03	0	0.03	0	0
mutation	0.03	0.03	0	0.03	0	0
probability	0.03	0.03	0	0.03	0	0
train	0	0	0.228	0	0	0
$\sum {W_{q,j}}^2$	0.15	0.15	1.366			
$\sqrt{\sum {W_{q,j}}^2}$	0.39	0.39	1.169			
Total				0.15	0	0

Setelah mendapatkan total nilai bobot per kombinasi dokumen dan nilai panjang vector tiap bobot, maka nilai *Cosine Similarity* tiap dokumen dapat dihitung sebagai berikut :

$$Sim(D_1, D_2) = \frac{0.155}{0.394\ x\ 0.394} = 1$$

$$Sim(D_1, D_3) = \frac{0}{0.394\ x\ 1.169} = 0$$

$$Sim(D_2, D_3) = \frac{0}{0.394\ x\ 1.169} = 0$$

Hasil dari perhitungan nilai kemiripan antar dokumen diatas dengan metode *Cosine Similarity* menunjukkan bahwa Dokumen 1 dan Dokumen 2 memiliki nilai kemiripan 1 dimana nilai tersebut menunjukkan bahwa kedua dokumen tersebut adalah sama.

4.2.4 Nama Pengarang dan Tahun

Setelah pencarian persamaan untuk judul karya ilmiah selesai, lalu pencarian persamaan untuk Nama Pengarang dan Tahun menggunakan metode *Hamming Distance.* Perhitungan dengan metode ini dengan cara mengukur nilai kedekatan antar kata atau biner. Dalam penelitian ini pengukuran nilai kedekatan dibagi menjadi dua, yaitu perhitungan untuk nama pengarang dengan menggunakan string(kata) dan perhitungan untuk tahun terbit karya ilmiah dengan menggunakan biner.

Tabel 8. Nama Pengarang dan Tahun

Dok	Nama Pengarang	Tahun
D1	Srinivas, M., & Patnaik, L. M.	1994

D2	Srinivas, M. & Patnaik, L.M.,	1994
D3	Paulin, F., & Santhakumaran, A.	2010

Diantara ketiga dokumen tersebut hanya D1 dan D1 yang bisa dibandingkan karena jumlah kata yang sama. Berikut merupakan perhitungan *Hamming Distance* untuk perhitugan Nama Pengarang:

Tabel 9. Perhitungan Nama Pengarang

D1	D2	Status	Nilai Jarak
Srinivas	Srinivas	Sama	0
M	M	Sama	0
Patnaik	Patnaik	Sama	0
L	L	Sama	0
M	M	Sama	0
Total			0

Berikut merupakan perhitungan *Hamming Distance* untuk perhitungan tahun:

Tabel 10. Perhitungan Tahun

D1	D2	Status	Nilai Jarak
1994	1994	sama	0
Total			0

Nilai kedekatan untuk perbandingan nama pengarang dan tahun karya ilmiah untuk D1 dan D2 adalah 0, maka D1 dan D2 bisa dibilang sama.

4.2.5 Perhitungan Frekuensi

Hasil perhitungan yang akan digunakan dari identifikasi judul karya ilmiah adalah yang mempunyai nilai *Cosine Similarity* 1. Dimana nilai tersebut mengindikasikan bahwa perbandingan dari 2 judul karya ilmiah adalah sama. Sedangkan untuk perbandingan yang mempunyai nilai dibawah 1 dimana nilai tersebut mengindikasikan bahwa judul karya ilmiah tesebut adalah mirip ataupun tidak sama, tidak akan ditampilkan.

Untuk hasil perhitungan dari identifikasi nama pengarang dan tahun yang akan ditampilkan hanya yang mempunyai nilai kedekatan 0. Dimana nilai tersebut mengindikasikan bahwa tidak ada jarak antara 2 nama pengarang atau tahun yang dibandingkan.

Dari 41 judul karya ilmiah yang mirip/sama yang memiliki nilai *threshold* 0.65. Terdapat 19 judul karya ilmiah yang sama atau yang mempunyai nilai *Cosine Similarity* 1. Dari 19 judul karya ilmiah yang sama tersebut, menghasilkan 9 karya ilmiah yang sama.

Berikut merupakan frekuensi untuk karya ilmiah yang sama :

Tabel 11. Hasil Perhitungan Frekuensi Karya Ilmiah

No	Karya Ilmiah	Frekuensi	NIM
1	Srinivas, M., & Patnaik, L. 1994. Adaptive Proba-bilities of Crossover and Mutation in Genetic Algorithms. *IEEE Transaction o/n systems,* 656-667	2	M0508031 M0508053
2	Tjhay, F. 2009. Ancaman Penyakit Radang Panggul pada Infeksi Menular Seksual. *Majalah Kedokteran Damaianus, 8*, 105-114	2	M0509013 M0510035
3	Kusumadewi, S. 2008. Aplikasi K-Means Untuk Pengelompokan Mahasis Berdasarkan Nilai Body Mass Index (BMI) dan Ukuran Kerangka. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI),* E-45	3	M0508013 M0508059 M0509018
4	Paulin, F., & Santhakumaran, A. 2011. Classification of Breast Cancer by Comparing Backpropagation Training Algorithms. *International Journal on Computer Science and Engineering (IJCSE)*	2	M0508075 M0509013
5	Yasmin, N., Rahman, I. A., & Eftekh ari, M. 2010. Forecasting Low-Cost Housing Demand in Johor Bahru, Malaysia Using Artificia Neural Networks (ANN). *Journal of Mathematics Research*, vol 2 no 1 , 14-19	2	M0508005 M0508075
6	Hardibroto, B, R. 2005. Mioma Uteri. *Majalah Kedokteran Nusantara, 38*	2	M0509013 M0510035
7	Karypis, G. Kumar, V. 1884. Performance and Scalability of the Parallel Simplex Method for Dense Linier Programming. *University of Minnesota Technical Report*, TR 94-43	2	M0508045 M0508049
8	Widayati, P., Ariyanto, A., & Lestari, W. 2009. Produksi Kit Immunoradiometricassay (IRMA) CA-15 untuk Deteksi DIni Kanker Ovarium. Jurnal Ilmu Kefarmasian, 7	2	M0509013 M0510035
9	Ali, A. H. 2008. Self-Organization Maps for Prediction of Kidney Dysfunction. *Telecommunications Forum TELFOR 2008, 775-778*	3	M0507031 M0508031 M0508053
10	Nawi, N. M., R. Ghazali and M.N.M. Salleh. 2011. Predicting Patients with Heart Disease by Using an Improved Back-propagation Algorithm. Journal of Computing. III(2) : 53-58.	2	M0507031

Tabel 11. Hasil Perhitungan Frekuensi Karya Ilmiah (Lanjutan)

No	Karya Ilmiah	Frekuensi	NIM
11	Deshpande, M. dan Karypis, G. 2004. Item-based Top-N	2	M0508008 M0508065

	Recommendation Algorithms. ACM Transactios on Information Systems 22 No. 1 pp. 143- 177.		
12	Darlis Heru Murti, Nanik Suciati, Daru Jani Nanjaya. (2005). Clustering Data Non-Numerik dengan Pendekatan Algoritma K-Means dan Hamming Distance Studi Kasus Biro Jodoh. Jurnal Ilmiah Teknologi Informasi , 46-53.	3	M0508013 M0508059 M0509018

5. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa menghitung frekuensi penggunaan karya ilmiah yang digunakan pada Tugas Akhir bisa menggunakan metode *Vector Space Model, Cosine Similarity* dan *Hamming Distance.* Dari 100 dokumen terdapat 47% yang sesuai dengan standar penulisan daftar pustaka. Dari penelitian yang telah dilakukan Terdapat 27 judul karya ilmiah yang sama atau yang mempunyai nilai *Cosine Similarity* 1. Dari 27 judul karya ilmiah yang sama tersebut, menghasilkan 12 karya ilmiah yang memiliki frekuensi kemunculan sebanyak 2 kali atau lebih. Penulis dari karya ilmiah yang sama tersebut tidak berasal dari institusi UNS. Terdapat 5 karya ilmiah yang berasal dari dalam negeri, dan 7 karya ilmiah yang berasal dari internasional.

Saran yang dapat dipertimbangkan untuk penelitian lebih lanjut antara lain:

1. Perlu dilakukan evaluasi terhadap penulisan daftar pustaka pada Tugas Akhir yang telah dikerjakan oleh mahasiswa karena banyaknya penulisan daftar pustaka yang tidak sesuai dengan syarat penulisan yang telah ditetapkan oleh bagian divisi Tugas Akhir.
2. Penelitian selanjutnya dapat mengembangkan sistem yang dapat mengklasifikasikan format penulisan daftar pustaka sesuai dengan pedoman penulisan atau salah.

6. DAFTAR PUSTAKA

[1] FMIPA UNS, *Peraturan Fakultas MIPA UNS No. 14 a/H27.1.28/KP/2008 Tentang Pedoman Pelaksanaan Skripsi Program Sarjana Fakultas MIPA UNS,* Surakarta: UNS Surakarta, 2008.

[2] T. Ilmiyah and S. Ati, "Pengaruh Pemanfaatan Koleksi Local Content Terhadap Kegiatan Penelitian Mahasiswa yang sedang Mengerjakan Skripsi/Tugas Akhir di Perpustakaan Fakultas Ilmu Budata Universitas Diponegoro Semarang," *Jurnal Ilmu Perpustakaan,* vol. Vol. 2 No. 2, pp. 1-9, 2013.

[3] E. S. R. Rahayu and Z. A. Hasibuan, "Identification of technology trend on Indonesian patent documents and research reports on chemistry and metallurgy fields," *Proceedings of the Asia-Pacific Conference on Library & Information Education & Practice 2006,* pp. 581-586, 2006.

[4] A. Widodo, I. Budi and R. F. Aji, "Prediksi Topik Penelitian Menggunakan Kombinasi Antara Support Vector Regression dan Kurva Logistik," *Seminar Nasional Aplikasi Teknologi Informasi 2012 (SNATI 2012),* 2012.

[5] P. Jackson and I. Moulinier, Natural Languange Processing for Online Applications: Text Retrieval, Extraction and Categorization, Volume 5 ed., Amsterdam: John Benjamins Publishing CO, 2002.

[6] P. D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space," *Journal of Artificial Intelligence Research,* pp. 141 -188, 2010.

[7] V. B. Wicaksono, R. Saptono and S. W. Sihwi, "Analisis Perbandingan Metode Vector Space Model dan Weighted Tree Similarity dengan Cosine Similarity pada kasus Pencarian Informasi Pedoman Pengobatan Dasar di Puskesmas," *JURNAL ITSMART,* vol. Vol. 4 No. 2. Desember 2015, pp. 73-83, 2015.

[8] A. Surachman, "Panduan Gaya Penulisan Sitiran Karya Ilmiah," 2016. [Online]. Available: lib.ugm.ac.id/data/panduan_sitiran.pdf.

[9] Suprianto, C., Affandy, "Kombinasi Teknik Chi Square dan Singular Value Decomposition Untuk Reduksi Fitur Pada Pengelompokan Dokumen," *Seminar Nasional Teknologi Informasi & Komunikasi Terapan,* pp. Universiti Teknikal Malaysia ISBN 979-26-0255-0, 2011.

[10] J. W. H. d. T. S. Asian, "Stemming Indonesia," *Conferences in Research and Practice in Information Technology,* vol. Vol. 38, 2005.

[11] G. d. M. M. Kowalski, Information Storage and Retrieval Systems: Theory and Implementation. Second Edition, New York: Kluwer Academic Publisher, 2002.

[12] J. Asian, Effective Techniques for Indonesian Text Retrieval, Melbourne: School of Computer Science and Information Technology, 2007.

[13] D. D. Lewis, Representation and Learning In Information Retrieval, USA: MA, 1992.

[14] S. Natalius, Metoda Naive Bayes Classifier dan Penggunaannya Pada Klasifikasi Dokumen, Bandung: Institut Teknologi Bandung, 2010.

[15] W. Dai, G.-R. Xue, Q. Yang and Y. Yu, "Transferring Naive Bayes Classifiers for Text Classification," *Proceedings of the 22nd national conference on Artificial Intelligence,* vol. 1, pp. 540-545, 2007.

[16] R. Y. Baeza and R. Neto, Modern Information Retrieval, Boston: Addison Wesley-Pearson Internatioal Edition, 1999.

[17] S. Susanto and D. I. Sensuse, "Pengklasifikasian Artikel Berita Berbahasa Indonesia secara Otomatis Menggunakan Naive Bayes Classifier," *Jurnal Ilmu Komputer dan Informasi,* vol. Vol. 1 No. 2, 2008.

[18] L. S. Putro, "Penerapan Kombinasi Algoritma Minhash dan Binary Hamming Distance pada Hybrid Perekomendasi

Lagu," Universitas Sebelas Maret, Surakarta, 2014.

[19] R. Sarno and F. Rahutomo, "PENERAPAN ALGORITMA WEIGHTED TREE SIMILARITY PENERAPAN ALGORITMA WEIGHTED TREE SIMILARITY," *JUTI,* pp. 35-42, 2008.