jurnalteknik.unisla.ac.id/index.php/informatika **ISSN: 2503-071X**
JOUTICA Volume 6 No.2 2021 **E-ISSN: 2621-511X**

PENCARIAN TEMA SEJENIS SINOPSIS NOVEL BAHASA INDONESIA DENGAN MENGGUNAKAN GVSM

Munif[1], Endang Setyati[2], Yosi Kristian[3]
[1]*Jurusan Teknik Informatika, Fakultas Teknik, Universitas Islam Lamongan*
[2]*Jurusan Teknologi Informasi, Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terpadu Surabaya*
[3]*Jurusan Teknologi Informasi, Fakultas Sains dan Teknologi, Institut Sains dan Teknologi Terpadu Surabaya*
E-mail: munif@unisla.ac.id, endang@stts.edu, yosi@stts.edu

ABSTRAKS
Tujuan utama dalam penelitian ini adalah untuk membantu memberikan informasi dokumen yang relevan sehingga mempermudah dalam melakukan pencarian novel dan lebih mengetahui isi dari sebuah novel. Output dari sistem yang dibangun ini adalah daftar sinopsis novel bahasa Indonesia beserta judul novel yang relevan sesuai dengan keyword yang diinputkan. Dataset yang digunakan dalam penelitian ini adalah synopsis novel bahasa indonesia yang didapatkan dengan cara crawling dan sudah dipisahkan dalam bentuk teks dokumen. Dataset berjumlah lebih dari 1500 record. Selajutnya dataset tersebut dilakukan pre-prosesing yang meliputi case folding, tokenizing, filtering dan stemming. Proses pencarian dilakukan dengan metode Generalized Vector Space Model sedangkan proses pengklasifikasian dilakukan dengan metode Naïve Bayes Classifier. Uji coba sistem dilakukan dengan menggunakan 300 data uji, dengan rincian 150 data uji untuk pencarian dan 150 untuk uji klasifikasi. Hasil dari sistem yang dibuat berupa ranking dokumen relevan berdasarkan urutan nilai cosine similarity dan juga menampilkan hasil klasifikasi sinopsis novel. Uji coba sistem pencarian yang dilakukan menghasilkan nilai recall yaitu 90% dan precision 85% sedangkan uji coba klasifikasi menghasilkan nilai akurasi sampai dengan 86%.

Kata Kunci: Sinopsis, Generalized Vector Space Model, Naïve Bayes Classifier, Cosine Similarity.

ABSTRACT
The main purpose of this research is to help provide relevant document information to make it easier to find novels and find out more about the contents of a novel. The output of the system being built is a list of synopsis of Indonesian novels along with relevant novel titles according to the entered keywords. The data set used in this study is a synopsis of Indonesian novels obtained by crawling and is already in the form of document text. The dataset may be more than 1500 records. The next dataset is pre-processed which includes case folding, tokenizing, filtering and stemming. The weighting process is carried out using the Generalized Vector Space Model method, while the classification process is carried out using the Naïve Bayes Classifier method. System testing is carried out using 300 test data, with details of 150 test data for searching and 150 for classification testing. The results of the system are made in the form of ranking the relevant documents based on the order of the cosine similarity values and display the classification results of the novel synopsis. The search system testing performed resulted in a recall value of 90% and accuracy of 85%, while the classification test resulted in an accuracy value of up to 86%.

Kata Kunci: Synopsis, Generalized Vector Space Model, Naïve Bayes Classifier, Cosine Similarity.

1. PENDAHULUAN

1.1 Latar Belakang

Pencarian informasi atau yang dikenal dengan istilah temu balik informasi (Information Retrieval) merupakan sistem yang dapat digunakan untuk menemukan informasi yang relevan dengan kebutuhan dari penggunanya secara otomatis dari suatu koleksi informasi (Mandala dan Setiawan, 2002). Sistem temu kembali menerima masukan (input) berupa kata-kata kunci (keyword) dari informasi yang dicari, dan dalam waktu yang relatif singkat sistem akan menampilkan daftar dokumen yang sesuai dengan kebutuhan informasi pengguna (Hendra Bunyamin dan Chathalea, 2008).

Generalized Vector Space Model merupakan perluasan dari *vector space model* (vsm) yaitu dengan menambahkan jenis informasi tambahan, disamping term, dalam merepresentasikan dokumen (Jasman Pardede dkk, 2013). Sistem temu kembali dengan *generelized vector space model* (gvsm) merepresentasikan dokumen dengan similaritas vector terhadap semua dokumen yang ada. deskripsi ringkas mengenai gvsm adalah *linierly independent. Generalized Vector Space Model* (gvsm) menghindari pengasumsian dengan penggunaan dokumen-dokumen sebagai dasar ruang vector dari pada term. Dalam *"dual space"* suatu dokumen direpresentasikan oleh suatu vector dimana dimensinya merujuk terhadap dokumen (Baeza, 1999).

Naïve bayes classifier merupakan pengklasifikasian dengan metode probabilitas dan statistik yang memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema bayes. Secara sederhana, naïve bayes menggunakan kemiripan fitur antara data training dan data testing dimana nantinya akan diambil class yang paling mirip dari data training tersebut (Junaedi widjojo, 2012). Kelebihan dari naïve bayes adalah sederhana namun memiliki tingkat akurasi yang cukup tinggi.

ISSN: 2503-071X
E-ISSN: 2621-511X

Sinopsis novel merupakan bagian penting dari sebuah novel, setiap novel memiliki sinopsis yang dapat memudahkan pembaca untuk lebih memahami karakter dan isi cerita yang ada dalam novel itu sendiri. Dalam penelitian kali ini penulis membuat sistem aplikasi yang dapat memberikan informasi teks sinopsis novel beserta klasifikasinya berdasarkan keyword yang dimasukkan pengguna pada system.

1.2 Tujuan dan Manfaat

Adapun dari rumusan masalah diatas, maka penelitian ini memiliki tujuan dan manfaat sebagai berikut.

a. Mengimplementasikan dan mengevaluasi hasil temu kembali dalam dokumen.
b. Membuat system information retrieval yang dapat memberikan informasi teks sinopsis novel berdasarkan keyword yang diinputkan oleh pengguna.
c. Mengklasifikasikan teks sinopsis novel dalam 5 kategori yaitu : novel anak, novel remaja, novel islami, novel fantasi dan novel misteri berdasarkan keyword.

Sedangkan manfaat dari penelitian ini adalah dapat mempermudah bagi user untuk menemukan judul dan sinposis novel dengan hasil akurat dan cepat serta mengetahui klasifikasi novel berdasarkan sinopsisnya.

1.3 Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini adalah sebagai berikut :

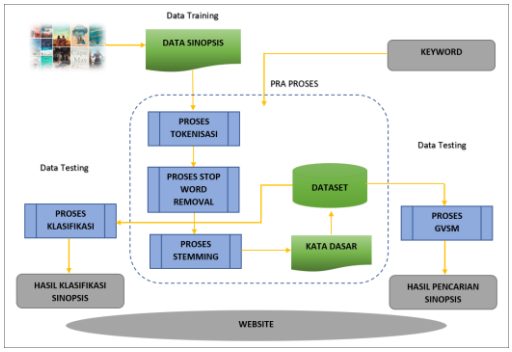
a. Pengumpulan Data
   Proses pengumpulan data sinopsis dilakukan dengan menggunakan metode *crawling* dari website penjulaan buku *online*. Crawling dibangun sendiri dengan menggunakan bahasa pemrograman PHP.
b. Studi Literatur
   Dengan mempelajari buku-buku referensi dan jurnal yang berkaitan dengan permasalahan penelitian yang diangkat serta mencari solusi yang terbaik. Topik bahasan utama yang dibutuhkan diantaranya adalah *Information Retrival*, GVSM (*Generalized Vector Space Model*) serta *Naïve Bayes Classifier* (NBC).
c. Analisa
   Melakukan uji coba secara teoritis terhadap masalah yang diangkat guna menganalisa apakah rancangan algoritma yang digunakan dapat menghasilkan solusi yang sesuai dengan tujuan penelitian.
d. Implementasi
   Membuat program dari hasil rancangan algoritma yang telah dibuat untuk mengimplementasikan serta membuktikan bahwa hasil analisa secara teoritis yang telah dilakukan benar-benar sesuai yang diharapkan.
e. Pengujian
   Pengujian dilakukan untuk melihat apakah data yang telah menjadi input akan diproses sesuai dengan output yang diharapkan. Hal ini juga dilakukan untuk mengevaluasi apakah metode yang diusulkan mampu menjawab tujuan yang telah diusulkan.
f. Dokumentasi
   Merupakan langkah akhir, penyusunan laporan mulai dari latar belakang permasalahan sampai dengan pengambilan kesimpulan akan dijelaskan dalam tahap dokumentasi ini.

2. ANALISA DAN DESAIN SISTEM

2.1 Analisa Desain Sistem

Sistem yang dibangun untuk melakukan penelitian ini adalah sebuah sistem yang dapat digunakan sebagai mesin pencarian sinopsis novel bahasa indonesia serta dapat mengklasifikasikan sinopsis berdasarkan kategori yang sudah ditentukan.

Arsitektur sistem digunakan untuk menggambarkan sistem kerja yang digunakan dalam proses analisa dan implementasi. Dengan arsitektur sistem dapat dilihat alur sistem secara lengkap, adapun arsitektur sistem dari keseluruhan sistem dapat ditunjukkan pada Gambar 1.



Gambar 1. Desain arsitektur sistem pencarian sinopsis novel dan klasifikasi sinopsis novel

Berikut ini adalah penjelasan dari arsitektur sistem pada Gambar 1:

1. Proses awal dari sistem ini adalah proses pengambilan data dari website bukukita.com atau crawling yang dibangun dengan menggunakan Bahasa pemrograman php sehingga disimpanlah menjadi sebuah data mentah pada database Mysql.
2. Preprosessing dilakukan untuk membentuk dataset dalam basis data. Diantaranya proses tersebut adalah :
   a) Proses tokenisasi, yaitu memisahkan setiap kata dalam satu dokumen sinopsis sehingga terbentuk menjadi sebuah kumpulan kata.
   b) Proses berikutnya adalah Stopword removal, yaitu menghilangkan kata-kata yang dianggap tidak perlu dan sering muncul di setiap dokumen.

ISSN: 2503-071X
E-ISSN: 2621-511X

c) Berikutnya dilakukan proses stemming, yaitu proses pembentukan kata dasar dengan cara menghilangkan imbuhan disetiap kata.
d) Setelah terbentuk menjadi kata dasar, barulah data tersebut disimpan kedalam database untuk dijadikan sebagai dataset yang akan diproses oleh sistem.

3. Dari dataset yang sudah terbentuk, selanjutnya proses pencarian sinopsis dilakukan dengan menggunakan metode *Generalized Vector Space Model*.
4. Jika terdapat data sinopsis baru yang belum memiliki kategori maka sistem akan mengklasifikasikan data tersebut menggunakan metode *naïve bayes*.

2.2 Analisa Kebutuhan Data

Data sinopsis novel yang digunakan dalam penelitian ini terdapat lebih dari 1500 sinopsis, hasil *crawling* yang dibangun sendiri dengan bahasa pemrograman php. Data synopsis diambil dari website penjualan buku online. Sinopsis novel terdiri dari beberapa jenis novel diantaranya, novel anak, islami, remaja, fantasi dan misteri. Adapun rincian data yang digunakan dalam penelitian ini dijelaskan pada tabel 1.

Tabel 1. Rincian data penelitian

No	Kategori Novel	Jumlah Sinopsis
1	Novel Anak	350
2	Novel Remaja	350
3	Novel Fantasi	300
4	Novel Islami	200
5	Novel Misteri	300
Jumlah Sinopsis		1500

Adapun contoh sinopsis novel yang telah berhasil dikumpulkan menggunakan crawling dijelaskan pada gambar 2.

judul_buku	alamat_url	sinopsis
SIRRUHART ? Rahasia Hati	Buku-Novel/Fantasi/100076-SIRRUHART-?-Rahasia-Hati--(Buk	BUKU INI TERSEDIA MULAI TANGGAL 30 JANUARI 2012. Dengarkan Mimpimu... Ikuti Kat
Kkpk : Violettes Secret	Buku-Novel/Anak-Anak/100126-Kkpk--:--Violettes-Secret.html	Violette harus mengikuti ayahnya pindah tugas ke Sydney. Di quintta School--sekolahba
Kkpk : Two Of Friends	Buku-Novel/Anak-Anak/100127-Kkpk--:--Two-Of-Friends.html	Chrystal dan Anette adalah saudara kembar. Mereka bersahabat dengan sepupunya, Pa
Kkpk : Holiday In Japan	Buku-Novel/Anak-Anak/100131-Kkpk--:--Holiday-In-Japan.html	Marsya berilubur ke rumah neneknya--Nenek Izumi--Jepang. Dia boleh mengajak teman
Kkpk : Diari Kocak Sulthan	Buku-Novel/Anak-Anak/100132-Kkpk--:--Diari-Kocak-Sulthan.ht	Hahaha ... hihihi ... aduh, aduh, ini diari anak SD yang kocaaak banget! Hari-harinya ken
Kcpk : Planet Zombiered	Buku-Novel/Anak-Anak/100134-Kcpk--:--Planet-Zombiered.htm	Di manakah sahabatku, Ziyya? Adakah kehidupan lain selain di Bumi? Pertanyaan itu sel
The Shadows	Buku-Novel/Fantasi/100247-The-Shadows.html	Saat Olive, 11 tahun, pindah ke rumah Victoria peninggalan Ms. McMartin, dia merasa r
Kkpk : I Can Fly (Petualanga	Buku-Novel/Anak-Anak/100324-Kkpk--:--I-Can-Fly--(Petualanga	Pernah Berkhayal menjadi peri? Hati-hati, Ifho, nanti bisa dijemput oleh peri beneran da
The Invention Of Hugo Cab	Buku-Novel/Fantasi/100502-The-Invention-Of-Hugo-Cabret.ht	Kehadirannya bagaikan hantu. Hugo menyelinap dan satu bilik ke bilik lain, menyusun j
Something Stupid (Buku_Re	Buku-Novel/Remaja/100629-Something-Stupid-(Buku_Rekom-Ini	dia kisah cerdas nan romantis yang dituliskan dengan alur penuh kejutan dan sangat me
Prom And Prejudice : Pesta	Buku-Novel/Remaja/100706-Prom-And-Prejudice--:-Pesta-Dan	Status anak beasiswa dianggap aib luar biasa di sekolah elite Akademi Longbourn. Ini bi
Kkpk : Love Music (Perjalanu	Buku-Novel/Anak-Anak/100808-Kkpk--:--Love-Music-(Perjalana	Asyiiik! Bikin band di sekolah, pasti seru banget. Apalagi, bisa tampil dan bertemu deng
Kcpk : Me And Bangka Islan	Buku-Novel/Anak-Anak/100811-Kcpk--:--Me-And-Bangka-Islan	Berprestasi dan dikenal orang banyak tentu menjadi impi setiap orang, termasuk Ade
40 Kisah Di Lorong Rahasia	Buku-Novel/Anak-Anak/100812-40-Kisah-Di-Lorong-Rahasia.h	Sejak kematian Raja, Ratu Bielna tampak kerepotan memimpin Kerajaan Vickhara. Seda
Koin : Xenopet Concurso #1	Buku-Novel/Fantasi/100813-Koin--:-Xenopet-Concurso--1.html	Pratinjau ... .???? xenopet adalah mahluk hidup hasil rekayasa genetik yang dibiakkan s
Kkpk : Kapok, Deh!	Buku-Novel/Anak-Anak/100814-Kkpk--:-Kapok,-Deh!.html	Anya dan teman-temannya sedang berdiskusi. Mereka memutuskan pergi ke rumah A
Kkpk : Me & My Sister	Buku-Novel/Anak-Anak/100815-Kkpk--:--Me-&-My-Sister	Liburan ke Bali, yeaaa ... pasti asyik. Tapi, kok, adikku ngambek, sih? Jadi garing, deh. Pa
Pcpk : Chocolate Girl	Buku-Novel/Anak-Anak/100818-Pcpk--:--Chocolate-Girl.html	Hari itu Arkha ulang tahun. Dia akan mengadakan pesta coklat di rumahnya. Kue ulan
BoyBand Story : Cinta Cekat	Buku-Novel/Remaja/100992-BoyBand-Story--:-Cinta-Cekat-Cel	Boyband...oh...Boyband. Fenomena Boyband yang makin menggila ternyata butuh pej
Kkpk : My Lovely Friends	Buku-Novel/Anak-Anak/101106-Kkpk--:-My-Lovely-Friends.htr	Bingung enggak, sih, bila suatu hari ada beberapa orang yang namanya mirip kita mens
Smash Cut - Dramatis (Fres	Buku-Novel/Mistery-Thriller/101586-Smash-Cut---Dramatis-(F	BUKU INI TERSEDIA MULAI TANGGAL 15 MARET 2012. Ingin membaca romantic suspen
Ini Dia Si Paling Badung (Gra	Buku-Novel/Remaja/101595-Ini-Dia-Si-Paling-Badung-(Grame	Ada murid cowok baru di kelas Elizabeth. Ia pemarah dan tidak ramah, dan dia mati-ma

Gambar 2. Data sinopsis hasil crawling

2.3 Analisa Kebutuhan User

Dilihat dari sisi user maka sistem yang dibangun ini bertujuan untuk memudahkan user atau masyarakat umum dalam mencari judul novel beserta sinopsisnya, dan seperti yang telah dijelaskan sebelumnya dalam mencari judul novel serta sinopsis novel dapat dicari dari Queri sesuai yang diinputkan oleh user.

User hanya memerlukan sistem untuk melakukan pencarian saja akan tetapi sistem ini juga memiliki sub sistem untuk pre-prosessing. Dimana sub sistem ini hanya dapat digunakan oleh admin sistem.

Sebelum melakukan pre-prosessing admin perlu melakukan analisa lebih lanjut terhadap setiap tahapan pre-prosessing. Seperti pada tahap pemilihan stopwords, maka admin perlu melakukan analisa lanjutan untuk memilih kata- kata yang dianggap stopword dengan melihat objek penelitian yang dibangun ini. Dalam pre-prosessing ini admin hanya perlu melakukan satu kali pre-prosessing.

2.4 Analisa Kebutuhan Software

Untuk membangun sistem ini digunakan beberapa software, yaitu :

a) Windows XP/7/W10
   Sistem operasi atau sistem dasar untuk mengoperasikan sebuah alat elektronik dalam hal ini adalah komputer. Karena merupakan sistem dasar maka dapat ditambahkan program aplikasi seperti, Microsoft Word termasuk Xampp dan lainnya.
b) PHP
   PHP adalah sebuah skrip pemrograman yang terletak dan dieksekusi di server. Salah satunya adalah untuk menerima, mengolah dan menampilkan data dari dan ke sebuah situs. Data yang diterima akan diolah disebuah program database server (program database yang terletak di sisi server, MySql contohnya) untuk kemudian hasilnya ditampilkan kembali dalam layar browser sebagai situs. Dengan demikian PHP dapat membuat sebuah sits menjadi dinamis karena data situs tersebut dapat selalu berubah sesuai dengan permintaan.
c) Database MySql
   MySql merupakan sebuah perangkat lunak sistem manajemen basis data SQL (Database Manajemen Sistem) atau DBMS yang multithread dan multi user.

3. PEMBAHASAN

3.1 Generalized Vector Space Model (GVSM)

Generalized Vector Space Model (GVSM) merupakan pengembangan dari metode Vector Space Model (VSM). Algoritma GVSM termasuk kedalam Algebraic model, dimana algebraic model merupakan salah satu dari 3 model besar information retrieval. Pada information retrieval yang menggunakan algebraic model, seluruh dokumen dan query direpresentasikan menjadi vektor, untuk menemukan similarity antara dokumen dan query maka nilai skalar antara vektor query dan vektor dokumen akan dikaitkan sehingga muncul nilai skalar sebagai acuan pengurutan dokumen (Tsatsaronis, G. dan Panagiotopoulou, 2009 Hal. 70-78).

Algoritma Generalized Vector Space Model yang dibahas menggunakan konsep ruang vektor.

ISSN: 2503-071X
E-ISSN: 2621-511X

Masukan dari user dan kumpulan dokumen diterjemahkan menjadi vektor-vektor. Kemudian vektor-vektor tersebut dikenakan operasi perkalian titik dan hasilnya menjadi acuan dalam menentukan relevansi masukkan pengguna (query) terhadap kumpulan dokumen.

Menurut Baeza terdapat beberapa langkah atau proses untuk mendapatkan hasil pencarian sesuai dengan query yang dimasukkan, yang disebut sebagai algoritma Generalized Vector Space Model.

a) Membuang kata depan dan kata penghubung
b) Menggunakan stemmer pada kumpulan dokumen dan query. Tahap ini biasa disebut sebagai tahap preprocessing
c) Menentukan pola minterm untuk mengetahui pola frekuensi kata. Panjang minterm ini didasarkan pada banyak kata yang diinput pada query. Kemudian diubah menjadi vector orthogonal sesuai dengan pola minterm yang muncul
d) Melakukan perhitungan kemunculan kata yang terdapat pada dokumen berdasarkan kata kunci. Pada penelitian ini menggunakan rumus TF-IDF.
e) Menghitung index term menggunakan persamaan 1

$$\overrightarrow{K_i} = \frac{\sum_{\forall r, gi(Mr)=1} Ci, r.\overrightarrow{Mr}}{\sqrt{\sum_{\forall r, gi(Mr)=1} Ci, r2}} \quad (1)$$

Dimana :

$\overrightarrow{K_i}$: Index term
$\overrightarrow{M_r}$: Vektor Orthogonal
$C_{i,r}$: Faktor Korelasi antara indeks term ke-i dengan minterm r
Sedangkan faktor korelasi dapat dinyatakan pada persamaan 2.

$$C_{i,r} = \sum_{dj \mid gi(\overrightarrow{dj})=gi(Mr)} W_{i,j} \quad (2)$$

Dimana :

$W_{i,j}$: Vektor dokumen ke-i
$gi(Mr)$: Bobot index term K_i dalam minterm M_r

f) Mengubah dokumen dan query menjadi vektor. Untuk menghitung nilai similaritas atau kemripan dokumen dan query, maka dokumen dan query harus diubah menjadi vektor terlebih dahulu. Untuk merubah dokumen dan query menjadi vektor menggunakan persamaan 3 dan 4.

$$\overrightarrow{d_j} = \sum_{i=1}^{n} W_{i,j} \; x \; \overrightarrow{K_i} \quad (3)$$

$$\vec{q} = \sum_{i=1}^{n} q_i \; x \; \overrightarrow{K_i} \quad (4)$$

Dimana :

$\overrightarrow{d_j}$: Vektor dokumen ke-j
$\vec{q}$: Vektor query
$W_{i,j}$: Berat indeks term i pada dokumen j
q_i : Berat indeks term pada query i
n : Jumlah indeks term

g) Menghitung nilai similaritas dokumen menggunakan perkalian vektor menggunakan persamaan 5

$$Sim(\overrightarrow{d_j}.\vec{q}) = \frac{\overrightarrow{d_j}.\vec{q}}{|\overrightarrow{d_j}||\vec{q}|} \quad (5)$$

Dimana :

$\overrightarrow{d_j}$: Vektor dokumen ke-j
$\vec{q}$: Vektor query

3.2 Naïve Bayes Classifier

Selain proses pencarian sinopsis pada penelitian ini juga terdapat proses klasifikasi. Proses klasifikasi dilakukan jika terdapat sinopsis baru yang belum mempunyai kelas. Untuk proses klasifikasi digunakan metode naïve bayes sebagai metode klasifikasi.

Sebuah *bayes classifier* adalah *classifier* probabilistik sederhana berdasarkan penerapan *teorema Bayes* (dari statistik Bayesian) dengan asumsi independen (naif) yang kuat. Sebuah istilah yang lebih deskriptif untuk model probabilitas yang digaris bawahi adalah "model fitur independent".

Dari proses ekstraksi data selanjutnya dilakukan pemodelan untuk mengelompokkan berdasarkan kategori. Untuk implementasi data tersebut kedalam model algoritma naïve bayes adalah dengan membuat tabel *term document matrix* dan selanjutnya akan ditentukan nilai probabilitas untuk masing–masing kategori seperti ditunjukkan pada persamaan 6.

$$P(W_{jk}|C_i) = \frac{f(W_{jk}.C_i) + 1}{f(C_i) + |W|} \quad (6)$$

Dimana :

$P(W_{jk}|C_i)$= Nilai kemunculan kata Wjk pada kategori Ci
$f(C_i)$ = Jumlah keseluruhan kata pada kategori Ci
|W| = Jumlah keseluruhan kata atau fitur yang digunakan.

Kemudian dilakukan perhitungan TF-IDF dengan persamaan 7.

$$P(C_i) = \frac{fd(C_i)}{|D|} \quad (7)$$

Dimana :

$fd(C_i)$ = Jumlah dokumen yang memiliki kategori Ci
|D| = Jumlah seluruh training dokumen

Sedangkan untuk klasifikasi bentuk naïve bayes digunakan persamaan 8.

$$C* = argmax\; p(C_i|d_j) = argmax \prod p(W_{ki}|C_i) \; x \; p(C_i) \quad (8)$$

ISSN: 2503-071X
E-ISSN: 2621-511X

Dimana W_{kj} merupakan fitur atau kata dari dokumen d_j yang ingin diketahui kategorinya. Nilai $p(C_i|d_j)$ dipelajari dari data training yang dimiliki dengan menggunakan informasi jenis fitur yang berbeda.

3.3 Recall dan Precision

Recall dan precision digunakan untuk mengukur efektifitas kinerja dari suatu *information retrieval*. Recall adalah perbandingan antara jumlah dokumen relevan yang ditampilkan terhadap jumlah seluruh dokumen yang relevan (Nisa Putri, 2016). Perhitungan recall menggunakan persamaan 9.

$$recall = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)} \quad (9)$$

Precision adalah perbandingan antara jumlah dokumen relevan yang ditampilkan terhadap jumlah seluruh dokumen yang ditampilkan. Perhitungan precision dinyatakan dengan persamaan 10.

$$precision = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)} \quad (10)$$

Penotasian precision dan recall dalam bentuk contingency table ditunjukkan pada tabel 2.

Tabel 2. Contingency table precision dan recall

	Relevant	*Non Relevant*
Retrieved	True Positives (tp)	False Positives (fp)
Non Retrieved	False Negatives (fn)	True Negatives (tn)

Dari tabel 2 diperoleh sebuah persamaan untuk menghitung precision dan recall, seperti ditunjukkan pada persamaan 11 dan 12.

$$Precision = \frac{tp}{(tp + fp)} \quad (11)$$

$$Precision = \frac{tp}{(tp + fn)} \quad (12)$$

4. UJI COBA DAN HASIL

Berikut ini merupakan hasil uji coba yang telah dilakukan pada penelitian ini.

4.1 Pencarian dengan Metode Generalized Vector Space Model (GVSM)

Dengan menggunakan perhitungan algoritma generalized vector space model, terdapat sebuah *query* serta 5 dokumen seperti terlihat pada tabel 3.

Tabel 3. Query dan dokumen

Query	*Libur Rumah Nenek Teman*
D1	nenek rumah grettel marrie marsya kejut izumi senyum bagaiman temu ada apa libur teman ajak dekat nenek sampai izumi-jepang
D2	aduh diari sedih malu-maluin senang teri pokok unik benar-benar lugu lucu sulthan seru kocaaak banget sd anak hihihi hari-harinya kerap hahaha gokil konyol isi alam
D3	amelia hutan teman ambar sentanu meneganggkan baru petualangan angker genderuwo tengah mati takut warga percaya benar nenek dikenalnya ski sepupunya temu maya berhasil mama izin kakek dunia desa rumah bersikeras libur sana
D4	marina renata benar baca jessica kamu tarik yuk cerita-cerita tunggu buku temu tindas artis banget senang idola sahabat tivi benci sebentar
D5	dunia siapa si mimpi nyata untung panggil dewasa sweets artinya ikut ingin magic ingin orang-orang pemilik sack biasa nama sempurna indah kadang hani kalista hitung hadir sma siswi pukau

Dari tabel 3 diatas telah dilakukan tahap *preprocessing* yang meliputi *case folding* dan *tokenizing, filtering* serta *stemming.*

Setelah tahap preprocessing maka selanjutnya dilakukan proses untuk menghitung bobot setiap term menggunakan TF-IDF. Hasil perhitungan TF-IDF dapat dijelaskan pada tabel 4.

Tabel 4. Perhitungan nilai idf

Term	TF						df	d/df	Idf
	Q	D1	D2	D3	D4	D5			
Libur	1	1	0	1	0	0	2	2,5	0,3979
Rumah	1	4	0	1	0	0	5	1,0	0
Nenek	1	4	0	1	0	0	5	1,0	0
Teman	1	1	0	2	0	0	3	1,7	0,2218

Tabel 4 merupakan hasil hasil perhitungan IDF dari 4 term yang ada. Kolom df pada term "libur" diperoleh dari jumlah dokumen yang mengandung term "libur" yaitu sebanyak 2 dokumen. Setelah dilakukan perhitungan nilai IDF, selanjutnya dilakukan perhitungan nilai bobot (W). Hasil perhitungan bobot dapat dilihat pada tabel 5.

Tabel 5. Perhitungan nilai w

Term	W					
	Q	D1	D2	D3	D4	D5
Libur	1	1	0	1	0	0
Rumah	1	4	0	1	0	0
Nenek	1	4	0	1	0	0
Teman	1	1	0	2	0	0

Tabel 5 merupakan hasil perhitungan nilai W yang diperoleh dari jumlah term pada setiap dokumen dikalikan dengan nilai IDF. Untuk term

ISSN: 2503-071X
E-ISSN: 2621-511X

"libur" pada D1 terdapat 1 term, diperoleh nilai W = 1 x 0,3979 dan seterusnya untuk term yang lain.

Sesuai dengan perhitungan algoritma generalized vector space model terdapat 6 langkah dalam menyelesaikan proses pencarian sebagai berikut

a) Melakukan *preprocessing* berupa *case folding, tokenizing, filtering, dan stemming*
b) Menentukan pola *minterm* yang muncul sesuai dengan *query* yang diinputkan oleh *user*. Dalam penelitian ini penulis menggunakan 4 kata kunci yaitu Libur, Rumah, Nenek, Teman. Sehingga pola minterm yang terbentuk dijelaskan pada tabel 6.

Tabel 6. Pola minterm

Pola minterm	Term query
M_w	Libur
M_x	Rumah
M_y	Nenek
M_z	Teman

c) Hasil perhitungan bobot (W) pada tabel 5 digunakan nilainya kedalam bentuk vektor orthogonal berdasarkan pola minterm yang sudah terbentuk. Penentuan bentuk vektor orthogonal dijelaskan pada tabel 7

Tabel 7. Penentuan vektor orthogonal

Dok	Libur	Rumah	Nenek	Teman	Vektor Orthogonal
D1	0,3979	0	0	0,2218	$\overrightarrow{M_1}$
D2	0	0	0	0	$\overrightarrow{M_2}$
D3	0,3979	0	0	0,4437	$\overrightarrow{M_3}$
D4	0	0	0	0	$\overrightarrow{M_4}$
D5	0	0	0	0	$\overrightarrow{M_5}$
Q	0,3979	0	0	0,2218	

d) Setelah vektor orthogonal tebentuk selanjutnya adalah menghitung nilai *index term.* Untuk menghitung index term menggunakan persamaan 1. Dalam menghitung index term terlebih dahulu harus diketahui nilai faktor korelasi setiap term. Faktor korelasi dapat dihitung menggunaka persamaan 2. Nilai faktor korelasi yang digunakan adalah korelasi antara query dengan dokumen. Hasil dari faktor korelasi dapat dilihat pada tabel 8.

Tabel 8. Perhitungan faktor korelasi

$C_{1,1} = 0,3979$	$C_{2,1} = 0$	$C_{3,1} = 0$	$C_{4,1} = 0,2218$
$C_{1,2} = 0$	$C_{2,2} = 0$	$C_{3,2} = 0$	$C_{4,2} = 0$
$C_{1,3} = 0,3979$	$C_{2,3} = 0$	$C_{3,3} = 0$	$C_{4,3} = 0,4437$
$C_{1,4} = 0$	$C_{2,4} = 0$	$C_{3,4} = 0$	$C_{4,4} = 0$
$C_{1,5} = 0$	$C_{2,5} = 0$	$C_{3,5} = 0$	$C_{4,5} = 0$

Dari faktor korelasi tabel 8 diatas, selanjutnya digunakan untuk menghitung nilai *index term* menggunakan persamaan 1. Hasil perhitungan index term dijelaskan pada tabel 9

Tabel 9. Nilai index term

	Index Term
K1	$\overrightarrow{K_1} = \frac{0,3979\overrightarrow{M_1} + 0,3979\overrightarrow{M_3}}{\sqrt{0,3166}}$
K2	$\overrightarrow{K_2} = \frac{0}{\sqrt{0}}$
K3	$\overrightarrow{K_3} = \frac{0}{\sqrt{0}}$
K4	$\overrightarrow{K_4} = \frac{0,2218\overrightarrow{M_1} + 0,4437\overrightarrow{M_3}}{\sqrt{0,2461}}$

e) Merubah dokumen dan query kedalam bentuk vektor menggunakan persamaan 3 dan 4. Hasil perhitungan vektor dokumen dan vektor query dijelaskan pada tabel 10

Tabel 10. Perhitungan vector dokumen dan query

Dok	Vektor Dokumen					Q
	D1	D2	D3	D4	D5	
D1	0,3806	0,0000	0,4798	0,0000	0,0000	0,3806
D2	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
D3	0,4798	0,0000	0,6782	0,0000	0,0000	0,4798
D4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
D5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

f) Setelah dokumen dan query diubah menjadi vektor, selanjutnya adalah menghitung nilai similaritas dokumen dan query menggunakan persamaan 5. Hasil perhitungan nilai similaritas dokumen dan query dijelaskan pada tabel 11

Tabel 11. Nilai similaritas dokumen

Dokumen	Nilai Similaritas
D1	1,000
D2	0,000
D3	0,998
D4	0,000
D5	0,000

Dari nilai *Cosine Similarity* dan Peringkat yang diperoleh pada tabel 11, ada tiga dokumen yang mempunyai nilai 0, sehingga dokumen tersebut tidak ditampilkan dalam hasil. Sedangkan dokumen yang ditampilkan oleh IR sistem kepada *user* (pengguna) urut berdasarkan pringkatnya adalah dokumen D1 dan Dokumen D3.

ISSN: 2503-071X
E-ISSN: 2621-511X

4.2 Uji Coba pada Sistem

a) Unit Testing

Terdapat tiga komponen utama dalam pengujian pencarian, yaitu uji coba input keyword, proses, dan hasil. Dalam setiap pengujian blackbox, penguji akan melihat perubahan yang ada dalam database dengan bantuan software database management.

1. Uji Coba Input Keyword

Input teks yang akan digunakan untuk uji coba seperti pada table 12 berikut ini

Tabel 12. Keyword pencarian

No	Keyword Text
1	Berlibur kerumah nenek bersama teman
2	99 cahaya dilangit eropa
3	Jutawan tewas terbunuh
4	Fenomena boyband
5	Buku diari gembel
6	Hidung besar seperti pinokio
7	Rahasia hati siti hajar
8	Knife dan ratu peri pemberani
9	Romeo dan juliet
10	Takdirku telah terukir dilangit

Jika user akan melakukan pencarian maka user akan masuk ke menu Input Keyword, kemudian teks keyword akan otomatis masuk kedalam tabel buku. Tampilan input keyword seperti tampak pada gambar 3 berikut :

Gambar 3. Halaman input keyword pencarian

Uji coba tampilan pada tabel buku setelah proses input keyword dilakukan ditunjukkan pada gambar 4 berikut

Gambar 4. Tabel buku update keyword

2. Uji Coba Proses Pencarian

Tampilan pada uji coba proses hanya merupakan keterangan dan hasil perhitungan. Namun untuk uji coba keberhasilan terletak pada tabel–tabel yang digunakan sebagai perhitungan. Berikut merupakan tahap proses dan hasil yang terdapat dalam tabel.

Hasil yang didapatkan dalam tabel hitung seperti pada Gambar 5 dibawah ini

did	term	tf	tfidf
0QUERY	teman	1	1
0QUERY	nenek	1	1
0QUERY	rumah	1	1
0QUERY	libur	1	1
100126	violette	5	0
100126	maker	3	0
100126	trouble	3	0
100126	rahasia	2	0
100126	ganggu	2	0
100126	marah	1	0
100126	heran	1	0
100126	ikut	1	0
100126	tetap	1	0
100126	cari	1	0
100126	yuk	1	0
100126	tenang	1	0

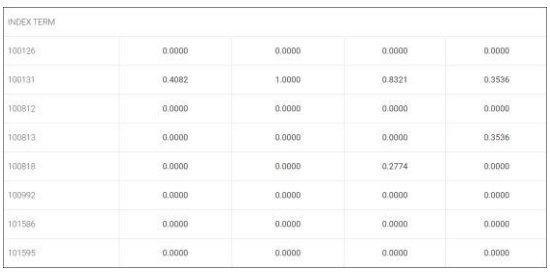
Gambar 5. Hasil Uji Coba Term pada Tabel Hitung

Dari hasil uji coba, terdapat beberapa kelemahan yaitu pada proses pemenggalan kata menjadi kata dasar tidak berjalan sempurna, beberapa kata belum dapat dijadikan sebagai kata dasar. Sehingga akan mempengaruhi jumlah term frequency jika imbuhan berbeda namun memiliki kata dasar sama. Proses pengambilan term pada tiap dokumen terdapat pada pre-process.php. Hasil yang didapatkan dalam tabel hitung seperti pada Gambar 5 dibawah ini

idhitung	did	term	tf
0QUERY_teman	0QUERY	teman	1
0QUERY_nenek	0QUERY	nenek	1
0QUERY_rumah	0QUERY	rumah	1
0QUERY_libur	0QUERY	libur	1
100126_violette	100126	violette	5
100126_maker	100126	maker	3
100126_trouble	100126	trouble	3
100126_rahasia	100126	rahasia	2
100126_ganggu	100126	ganggu	2
100126_marah	100126	marah	1
100126_heran	100126	heran	1
100126_ikut	100126	ikut	1

Gambar 6. Hasil uji coba tf pada tabel hitung

Proses menghitung index term pada setiap dokumen dapat dilihat pada gambar berikut ini

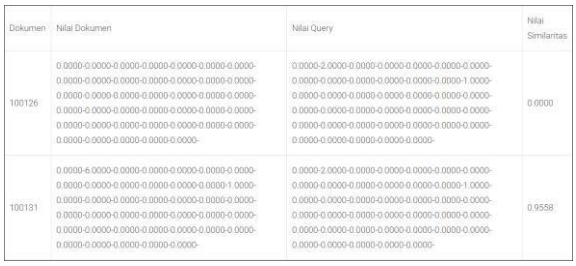


Gambar 7. Hasil perhitungan index term

Proses perhitungan index term digunakan untuk menghitung indeks term pada setiap dokumen untuk masing – masing kata kunci yang digunakan. Nilai indeks term dokumen **100126** terhadap kata kunci **"LIBUR", "NENEK", "RUMAH"** dan

ISSN: 2503-071X
E-ISSN: 2621-511X

"TEMAN". Proses menghitung kemiripan dokumen terhadap query dapat ditampilkan pada Gambar dibawah



Gambar 8. Proses uji coba perhitungan kemiripan dokumen

Proses perangkingkan dokumen digunakan untuk mengurutkan nilai dokumen secara descending, tampilan perangkingan dokumen dapat dilihat pada Gambar dibawah ini

idbuku	similar
100131	0,9558
100818	0,8944
102561	0,4472
112562	0,4472
133990	0,8944
143999	0,8944
144224	0,8944

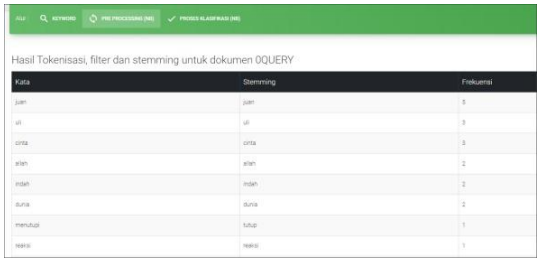
Gambar 9. Proses uji coba perhitungan kemiripan dokumen

3. Uji Coba Proses Klasifikasi

Proses pertama sebelum dilakukan klasifikasi adalah menginputkan keyword berupa sinopsis novel bahasa Indonesia yang akan diklasifikasikan. Proses ini sama dengan proses pada input keyword untuk pencarian. Proses input dapat dilihat pada gambar 10 berikut ini

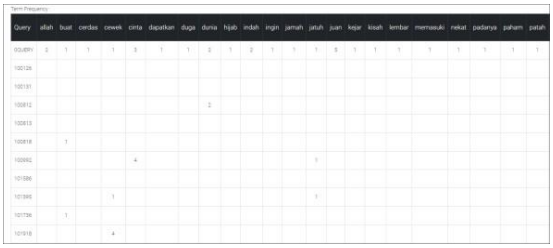
Gambar 10. Input keyword sinopsis klasifikasi

Dalam proses perhitungan selanjutnya, setiap dokumen akan dihitung jumlah kemunculan pada setiap term, proses ini dapat dijelaskan pada gambar 11 berikut ini.



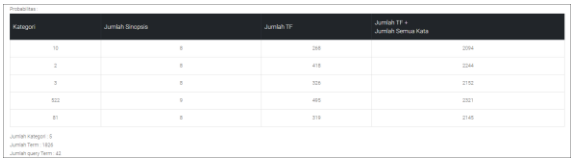
Gambar 11. Perhitungan term frequency

Pada proses ini dilakukan menggunakan metode naïve bayes sesuai dengan rumus perhitungan metode tersebut. Seperti tampak pada gambar 12 berikut ini.



Gambar 12. Perhitungan term frequncy setiap dokumen

Pada proses selanjutnya dihitung probabilitas yang akan menentukan keyword yang telah dimasukkan masuk dalam kategori apa seperti pada gambar 13 dibawah ini



Gambar 13. Jumlah term setiap dokumen

Setelah dilakukan proses perhitungan maka program akan menghasilkan nilai tertinggi dalam setiap kategori seperti tampak pada gambar 14 berikut ini

Gambar 14. Hasil nilai probabilitas setiap kategori

b) Hasil Uji Coba Pencarian

Dari 150 keyword pencarian yang telah diproses dengan menggunakan 700 dokumen didapatkan hasil dokumen yang relevan sesuai dengan query atau keyword seperti pada tabel dibawah ini. Pada tabel 13 dibawah ini berjumlah 30 keyword dari total 150 data yang telah diuji

Tabel 13. Dokumen yang ditampilkan system

No	Keyword	Jml Dok
1	Tadabbur Cinta	24
2	Xenopet adalah makhluk hidup hasil rekayasa	19
3	Perang nuklir yang menghancurkan dunia	62
4	Mereka menjulukinya malaikat kematian	54
5	Pegasus adalah teman yang sempurna	36
6	Takdirku telah terukir dilangit	28
7	Summer hantu perempuan dihotel	58
8	Negara ravka shadow fold alina dan grisha	19
9	Diana putri ratu amazon	46
10	Berlibur kerumah nenek bersama teman	113

Tabel 13. Lanjutan

11	Arkha mengadakan pesta coklat	22
12	Rena punya ibu baru	36
13	Marina bertemu artis idolanya	18
14	Hidung besar seperti pinokio	6
15	Kerajaan vickhara dan prince ghifari	18
16	Cher yang tomboi dan suka dance	19
17	Nindya berkhayal menjadi peri	11
18	Petualangan iza dan teman-teman	45
19	99 cahaya dilangit eropa	14
20	Kisah fatimah az-zahrah	38
21	Rahasia hati siti hajar	58
22	Masa lalu masa kini dan masa depan	0

Diperoleh jumlah dokumen yang sesuai dengan masing-masing query. Sebagai contoh untuk query "Takdirku telah terukir dilangit" diperoleh 28 dokumen. Sedangkan untuk query "Berlibur kerumah nenek bersama teman" diperoleh 113 dokumen. Untuk menghitung precision dan recall dibutuhkan sudut pandang pakar untuk menilai dokumen yang relevan berdasarkan keyword.

c) Precision dan Recall

Dari Uji coba yang telah dilakukan terhadap 700 dokumen sinopsis dengan 150 data keyword, serta telah didapatkan hasil dari sudut pandang pakar, maka perhitungan precision recall dari ke 150 data uji tersebut dapat dijelaskan pada tabel 14.

Tabel 14. Hasil perhitungan precision

No	Kata Kunci	Precision	Recall
1	Tadabbur Cinta	79%	69%
2	Xenopet adalah makhluk hidup hasil rekayasa	79%	95%
3	Perang nuklir yang menghancurkan dunia	85%	81%
4	Mereka menjulukinya malaikat kematian	89%	82%
5	Summer hantu Perempuan dihotel	84%	79%
6	Pegasus adalah teman yang sempurna	78%	95%

Dari pengujian dengan 150 keyword dan terhadap 700 dokumen, pencarian tema sejenis sinopsis menggunakan metode generalized vector space model mendapatkan nilai rata-rata precision sebesar 85% dan rata-rata nilai recall sebesar 90%.

d) Hasil Uji Coba Klasifikasi

Dalam uji coba klasifikasi pada laporan penelitian ini, penulis memproses data latih sebanyak 700 dokumen dan data uji sebanyak 150 dokumen seperti pada tabel berikut ini. Dari uji coba dokumen pada tabel 5.5 diperoleh tingkat akurasi sebesar 86% seperti ditunjukkan pada tabel berikut ini

Tabel 15. Akurasi uji coba klasifikasi

Kategori	Jumlah
Benar	129
Salah	21
Tingkat Akurasi	86%

Secara garis besar proses klasifikasi dengan menggunakan metode naïve bayes pada penelitian ini telah berjalan dengan baik dan hasilnya sesuai dengan hipotesa pada penelitian ini, meski demikian ada beberapa hal yang menjadi catatan terkait dengan kesalahan dalam klasifikasi.

Secara garis besar hasil uji coba pada penelitian ini telah berjalan dengan baik. Dari perhitungan precision dan recall bahwa IR sistem dengan menggunakan metode generalized vector space model memiliki akurasi yaitu rata-rata nilai Recall 90% dan Precision sebesar 85%. Sedangkan untuk uji coba klasifikasi menggunakan metode naïve bayes juga berajalan cukup baik dengan tingkat akurasi mencapai 86% dan telah memenuhi harapan.

5. KESIMPULAN

Dari pembahasan sebelumnya serta berdasarkan uji coba terhadap sistem diambil kesimpulan antara lain :

1. Hasil uji coba pada penelitian ini menunjukkan hasil yang baik, pada proses pencarian dengan menggunakan metode Generalized Vector Space Model dari uji coba sebanyak 150 keyword dengan 700 data sinopsis nilai recall yang didapatkan sebesar 90% dan nilai precision sebesar 85%.
2. Berdasarkan hasil pengujian lama waktu, pencarian dengan menggunakan metode generalized vector space model membutuhkan waktu yang sangat lama dikarenakan proses pencarian dengan metode gvsm membutuhkan waktu untuk mencari kedekatan makna antar term.
3. Dari penelitian yang telah dilakukan, faktor korelasi merupakan faktor yang membedakan antara metode gvsm dengan metode sebelumnya yaitu vector space model.
4. Hasil klasifikasi menggunakan metode naïve bayes pada penelitian ini mendapatkan hasil yang baik. Uji coba yang dilakukan dengan menggunakan 150 data uji dan 700 data training mendapatkan hasil akurasi sebesar 86%.
5. Sinopsis yang didapatkan dengan cara crawling yang sudah terkategori otomatis pada data training sangat mempengaruhi hasil klasifikasi, banyak kategori sinopsis yang tidak sesuai dengan kategori yang semestinya.
6. Pada penelitian ini terdapat dua fitur sistem yang berbeda, fitur pencarian menggunakan metode gvsm dan fitur klasifikasi menggunakan metode naïve bayes.

PUSTAKA

Baeza, Ricardo, B. Ribeiro. 1999. *Modern Information Retrieval*. ACM Press, United States of America.

Bunyamin Hendra, Chatalea Puspa Negara. 2008. *Aplikasi Information Retrieval (IR) CATA dengan metode Generalized Vector Space Model*. Bandung.

Digilib.unila.ac.id, diakses tgl. 3 April 2020 Pukul. 14.05 WIB.

I Made Suwija Putra, Ni Putu Ayu Widiardi, I Wayan Gunaya. 2019. *Implementasi Generalized Vector Space Model dalam Pencarian Buku Diperpustakaan*. Merpati, Vol (7): No 1.

Jasman Pardede, Mira Musrini Barnawi, Wildan Deny Pramono, 2013. *Implementasi metode Generalized Vector Space Model pad Aplikasi Informastion Retrieval*. Institut Teknologi Nasional Bandung. Bandung.

Junaedi Widjojo. 2012. *Prediksi Jenis Kelamin dan Usia untuk Blog Berbahasa Indonesia dengan Metode Klasifikasi Teks yang Dilengkapi dengan Pemilihan Fitur Terbaik*. Jurusan Teknologi Informasi, ISTTS, Surabaya.

Nisaa Putri Lestari. 2016. *Uji Recall dan Precision Sistem Temu Kembali Informasi OPAC Perpustakaan ITS Surabaya*. Surabaya.

Suprianto, Sunardi, Abdul Fadlil, 2018. *Aplikasi Sistem temu Kembali Angket Mahasiswa Menggunakan Metode Generalized Vector Space Model*. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), Vol (6): No 1.

Tsatsaronis, G., Panagiotopoulou, 2009. *A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness*, Proceedings of the EACL Student Research Workshop, Greece.