

Klasifikasi Berita *Online* dengan menggunakan Pembobotan *TF-IDF* dan *Cosine Similarity*

Bening Herwijayanti¹, Dian Eka Ratnawati², Lailil Muflikhah³

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: ¹bening.herwijayanti@yahoo.co.id, ²dian_ilkom@ub.ac.id, ³lailil@ub.ac.id

Abstrak

Dalam klasifikasi berita *online* dengan menggunakan pembobotan *tf-idf* dan *cosine similarity* ini mendapatkan referensi penelitian sebelumnya mengenai klasifikasi berita *online* menggunakan algoritma *single pass clustering*, dimana data yang akan digunakan berasal dari *website* berita *online* yaitu *kompas.com*. Karena banyaknya berita yang dimasukkan ke dalam *website*, sehingga terkadang berita tersebut terposting tidak sesuai dengan kategorinya. *Human error* akan menjadi masalah berita yang salah *posting*. Selain kesalahan *posting* pengelompokan berita *online* juga penting untuk kenyamanan *user* untuk mencari berita sesuai dengan kategorinya. Menerapkan klasifikasi berita *online* dengan menggunakan *tf-idf* dan *cosine similarity*, memerlukan proses *preprocessing* yaitu *tokenizing*, *stopword* dan *stemming* dapat memperkecil term sehingga mempercepat proses perhitungan pembobotan term menggunakan *tf-idf* dan mempercepat proses *cosine similarity*. Tujuannya adalah untuk mempermudah *human error* serta mengurangi terjadinya kesalahan pengkategorian. klasifikasi mampu mengelompokkan berita dengan tingkat akurasi sebesar 91.25%.

Kata Kunci: klasifikasi berita *online*, *TF-IDF*, *Cosine Similarity*

Abstract

In discussing the online news by using the weighting of tf-idf and cosine of this similarity the previous research reference on online news information using single pass clustering algorithm, where the data to be used comes from the online news website that is kompas.com. Because of the many news that is on the website, so sometimes the news is posted not in accordance with the category. Human error will be the problem of wrong news posting. In addition to posting errors online news groupings are also important for the convenience of users to search for news according to their category. Implementing online news stories using tf-idf and cosine similarities, preprocessing processes ie tokenizing, stopwords and stemming can reduce the term process of speeding the weighting of terms using tf-idf and accelerating the cosine process of similarity. The goal is to facilitate human error as well as reduce caution categorization. The value is able to classify news with accreditation rate of 91.25%.

Keywords: online news classification, *TF-IDF*, *Cosine Similarity*

1. PENDAHULUAN

Dalam klasifikasi berita *online* dengan menggunakan pembobotan *tf-idf* dan *cosine similarity* ini mendapatkan referensi penelitian mengenai klasifikasi berita *online* menggunakan algoritma *single pass clustering*, dimana data yang akan digunakan diambil dari salah satu *website* berita *online* yaitu *kompas.com*. Pada penelitian sebelumnya yang dilakukan oleh Agus Zaenal Arifin dan Ari Novan Setiono yang menjelaskan bahwa berita yang mempunyai event yang sama cenderung mengelompok

menjadi satu *cluster*, nilai batas atau yang disebut dengan *threshold* yang paling bagus digunakan adalah 0,0175 dengan nilai *recall* sebesar 76% dan nilai *precision* sebesar 87% sehingga disimpulkan bahwa *single pass clustering* cukup handal untuk digunakan dalam mengklasifikasikan event. Referensi berikutnya yaitu jurnal dengan judul “pembuatan web portal sindikasi berita Indonesia dengan klasifikasi metode *single pass clustering*” menjelaskan bahwa *single pass clustering* sangat tepat untuk klasifikasi dengan tingkat kemiripan berita dengan *recall precision* 79% dengan nilai

recall rata-rata 76% dan *precision* rata - rata 87%.

Dari pengertian *tf-idf* dan *cosine similarity* dianggap cocok untuk klasifikasi berita *online*. Karena banyaknya berita yang dimasukkan ke dalam *website*, sehingga terkadang berita tersebut terposting tidak sesuai dengan kategorinya. *Human error* akan menjadi masalah berita yang salah *posting*. Selain kesalahan *posting* klasifikasi berita *online* juga penting untuk kenyamanan *user* untuk mencari berita sesuai dengan kategorinya. Seperti yang terjadi pada postingan yang ada di *kompas.com* dengan judul berita “*Ridwan Kamil Berikan "Kadeudeuh" untuk Persib*” yang diposting tanggal 21 Maret 2017 jam 21.31 WIB terdapat kesalahan kategori *posting* karena judul beserta isi dari artikel tersebut tersebut seharusnya masuk dalam kategori *news* olahraga sedangkan yang terjadi kategori tersebut masuk kategori *news* regional. Harapannya dengan dibuat sistem ini bertujuan untuk meminimallisir kesalahan dalam klasifikasi pada berita *online*.

Oleh karena itu, agar tidak terjadi kesalahan dalam pengkategorian terhadap berita *online* yang ada, klasifikasi pada skripsi ini akan menerapkan suatu pengkategorian suatu berita *online* dengan menggunakan pembobotan *tf-idf* dan *cosine similarity*.

2. LANDASAN TEORI

2.1 Text mining

Text mining merupakan proses analisis dalam data yang berupa teks dimana sumber data didapatkan dari dokumen (Ronen Feldman, 2007). Konsep *text mining* biasanya digunakan dalam klasifikasi dokumen tekstual dimana dokumen-dokumen tersebut akan diklasifikasikan sesuai dengan topik dokumen tersebut. Dengan bantuan *text mining* suatu artikel dapat diketahui jenis kategorinya melalui kata-kata yang terdapat pada artikel tersebut. Kata-kata yang dapat mewakili isi dari artikel tersebut dianalisa dan dicocokkan pada basis data kata kunci yang telah ditentukan sebelumnya. Sehingga dengan adanya *text mining* dapat membantu melakukan pengelompokkan suatu dokumen dengan waktu yang singkat.

Tahapan dalam melakukan analisa pada *text mining* yaitu melakukan pengumpulan data kemudian melakukan ekstraksi terhadap fitur yang akan digunakan (Ronen Feldman, 2007).

Adapun teknik yang digunakan dalam ekstraksi fitur yaitu melakukan pembersihan data mulai dari *tokenizing*, *stop words removal* dan *stemming*. Selanjutnya yaitu melakukan *transform* data dengan pembobotan terhadap *term* yang telah dibersihkan. Kemudian dilanjutkan dengan reduksi data. Tahap terakhir yaitu melakukan analisis terhadap proses klasifikasi untuk merepresentasikan hasil informasi yang ditemukan.

2.2 Processing data

Processing bertujuan untuk mendapatkan dataset yang dapat diolah dengan cepat dan menghasilkan kesimpulan yang tepat. Salah satu proses processing data yang dapat dilakukan adalah pemilihan fitur (*feature selection*). Ada beberapa tahapan dalam pemilihan fitur, antara lain:

Tokenizing, merupakan tahap pemotongan string input untuk memisah kalimat menjadi kata. *Stopword*, Pada tahap ini dilakukan proses menghilangkan kata yang tidak penting dalam teks. Untuk proses ini diperlukan suatu kamus kata-kata yang menyimpan kata-kata yang bisa dihilangkan. *Stemming*, merupakan tahapan yang melakukan proses untuk mengubah kata turunan menjadi kata dasar.

2.3 Pembobotan Term Frequency Inverse Document Frequency (Tf-Idf)

Data yang telah melalui tahap *preprocessing* harus berbentuk numerik. Untuk mengubah data tersebut menjadi numerik yaitu menggunakan metode pembobotan TF-IDF. Metode *Term Frequency Invers Document Frequency* (TF-IDF) merupakan metode yang digunakan menentukan seberapa jauh keterhubungan kata (*term*) terhadap dokumen dengan memberikan bobot setiap kata [1]. Metode TF-IDF ini menggabungkan dua konsep yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen dan inverse frekuensi dokumen yang mengandung kata tersebut (Fitri, 2013).

Dalam perhitungan bobot menggunakan TF-IDF, dihitung terlebih dahulu nilai TF perkata dengan bobot masing-masing kata adalah 1. Sedangkan nilai IDF diformulasikan pada Persamaan (1).

$$IDF(word) = \log \frac{td}{df} \quad (1)$$

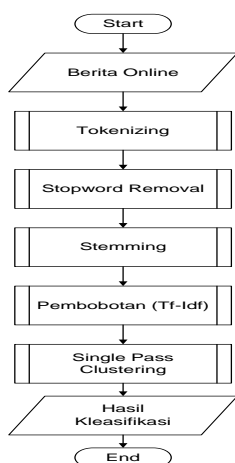
$IDF(word)$ adalah nilai IDF dari setiap kata yang akan di cari, td adalah jumlah keseluruhan dokumen yang ada, df jumlah kemuculan kata pada semua dokumen.

2.4 Cosine Similarity

Model ruang vektor dan pembobotan $tf-idf$ digunakan untuk merepresentasikan nilai numerik dokumen sehingga kemudian dapat dihitung kedekatan antar dokumen. Kemiripan antar dokumen dihitung menggunakan suatu fungsi ukuran kemiripan (*similarity measure*). Semakin besar hasil fungsi *similarity*, maka kedua objek yang dievaluasi semakin mirip, demikian pula sebaliknya. Ukuran ini memungkinkan perankingan dokumen sesuai dengan kemiripan (relevansi)nya terhadap *query*. Kualitas hasil dari dokumen yang didapatkan sangat tergantung pada fungsi *similarity* yang digunakan.

3. PERANCANGAN SISTEM

Pada bab ini akan dibahas mengenai perancangan sistem, yaitu formulasi penyelesaian masalah secara sederhana, perancangan antarmuka pengguna, serta perancangan pengujian sistem. Gambar merupakan diagram alir proses klasifikasi dokumen berita *online* secara umum menggunakan pembobotan $tf-idf$ dan *cosine similarity*.

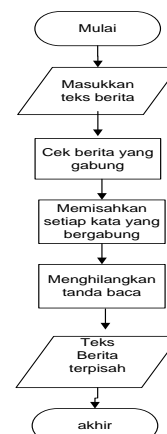


Gambar 1 Diagram Alir Klasifikasi Berita Online

3.1 Tokenizing

Tokenizing merupakan tahap pemotongan string input untuk memisah kalimat menjadi kata dan penghapusan karakter tanda baca, berikut

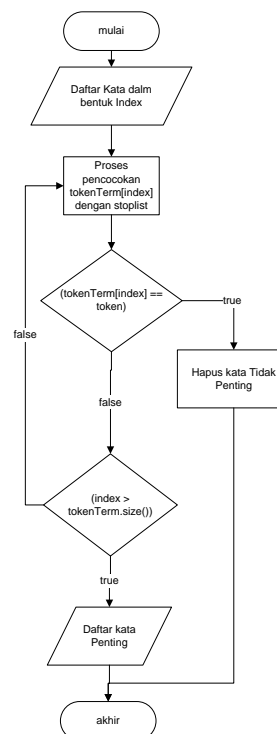
merupakan asumsi dari proses tokenizing yang diambil dari salah satu data studi.



Gambar 2 Diagram alir Tokenizing

3.2 Stopword Removal

Dalam proses ini dilakukan proses menghilangkan kata yang tidak penting dalam teks. Untuk proses ini diperlukan suatu kamus kata-kata yang menyimpan kata-kata yang bisa dihilangkan atau dengan kata lain kata –kata yang tidak penting.

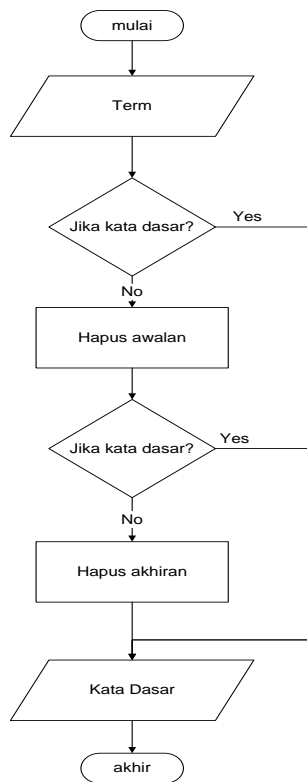


Gambar 3 Diagram Alir Stopword Removal

3.3 Stemming

Proses *stemming* merupakan proses yang berfungsi untuk menghilangkan imbuhan seperti

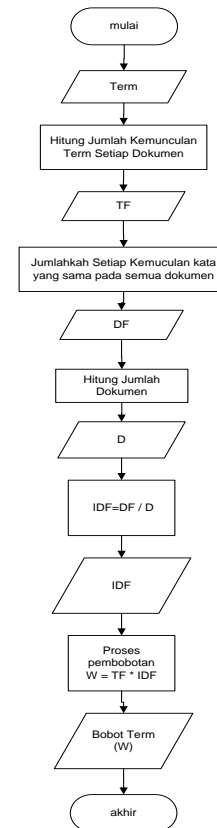
awalan dan akhiran proses stemming akan di jelaskan dalam gambar alur seperti pada gambar 4 berikut:



Gambar 4 Diagram Alir Stemming

3.4 Proses Pembobotan (TF-IDF)

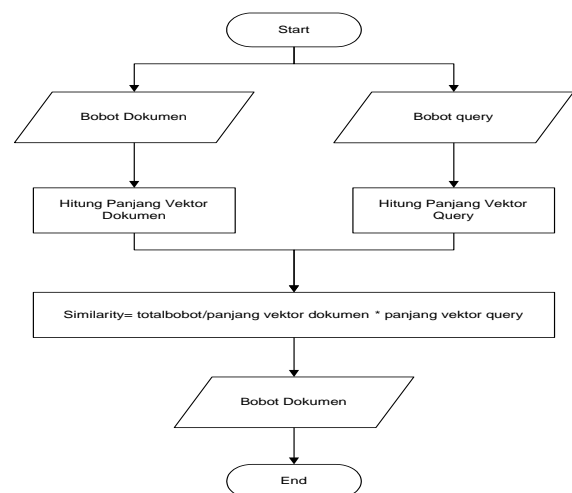
Dalam perhitungan bobot menggunakan TF-IDF, dihitung terlebih dahulu nilai TF perkata dengan bobot masing-masing kata adalah 1. Sedangkan nilai IDF diformulasikan pada persamaan 2. dimana $IDF(word)$ adalah nilai IDF dari setiap kata yang akan di cari, td adalah jumlah keseluruhan dokumen yang ada, df jumlah kemuculan kata pada semua dokumen. Setelah mendapat nilai TF dan IDF, maka untuk mendapatkan bobot akhir dari TF-IDF diformulasikan pada persamaan 1 dimana w (wordi) adalah nilai bobot dari setiap kata, TF (wordi) adalah hasil perhitungan dari TF. $IDFi$ adalah hasil dari perhitungan IDF. Berikut merupakan proses perhitung bobot menggunakan metode TF-IDF:



Gambar 5 Diagram Alir TF-IDF

3.5 Cosine similarity

Dalam proses *cosine similarity* yang menjadi masukan adalah bobot dari term setiap data, bobot *term* tersebut di gunakan dala proses perhitungan jarak kemiripan dengan kata kluster, kemudian dari setiap nilai akan menentukan *centroid* setiap kluster. Berikut merupakan proses perhitungan cosine similarity:



Gambar 6 diagram alir cosine similarity

4. IMPLEMENTASI

4.1 Implementasi Halaman Utama

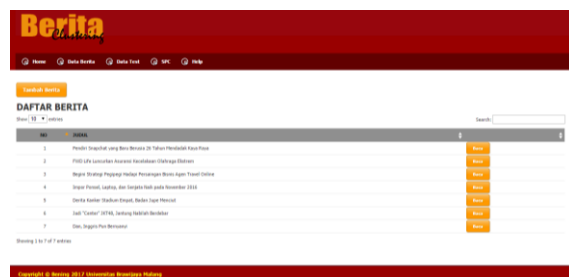
Halaman utama merupakan halaman yang berfungsi untuk menampilkan judul aplikasi, dan menu utama aplikasi, gambar 7 ini merupakan penjelasan dari implementasi halaman utama:



Gambar 7 Implementasi Antarmuka Halaman Utama

4.2 Implementasi Halaman Data Berita

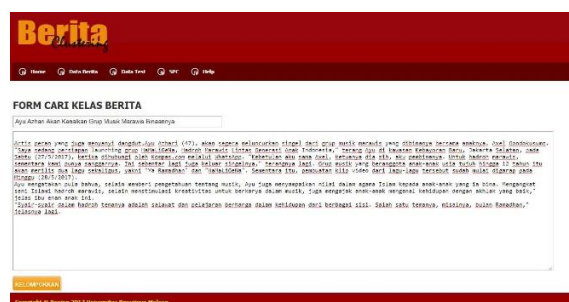
Halaman data berita berfungsi untuk menampilkan data berita yang menjadi data training, implementasi data berita dapat di lihat pada gambar 8 berikut.



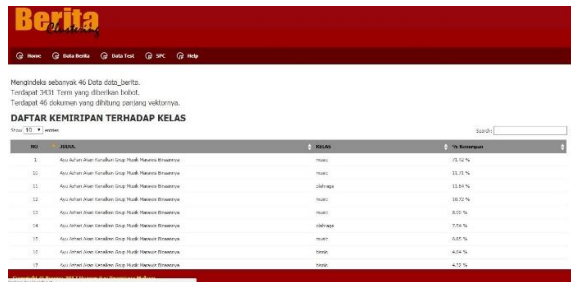
Gambar 8 Implementasi Antarmuka Halaman Data Berita

4.3 Implementasi Halaman Data Test

Halaman data test merupakan halaman yang berfungsi untuk menampilkan form masukan data test berupa berita, implementasi data test dapat di lihat pada gambar 9 dan hasil data test dapat dilihat pada gambar 10 berikut:



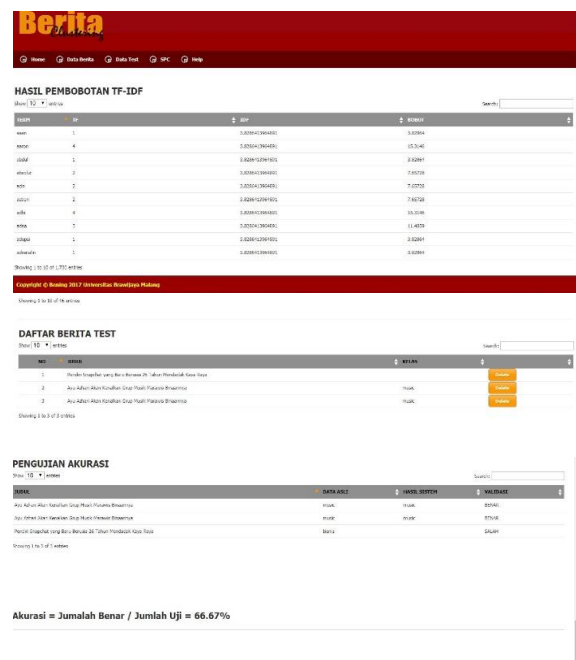
Gambar 9 Implementasi Antarmuka Halaman Data Test



Gambar 10 Implementasi Antarmuka hasil data test

4.4 Implementasi Halaman kluster

Halaman kluster merupakan halaman yang akan menjelaskan tentang proses perhitung kluster sampai dengan proses perhitung akurasi, implementasi halaman kluster dapat di lihat pada gambar 11 berikut.



Gambar 11 implementasi antarmuka kluster

5. PENGUJIAN

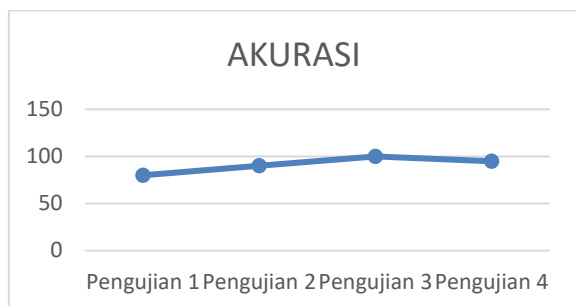
Hasil pengujian akurasi seperti pada tabel 1 berikut:

Tabel 1 hasil pengujian

Pengujian	Data latih	Data uji	Akurasi
Pengujian 1	90%	10%	80%
Pengujian 2	80%	20%	90%

Pengujian 3	70%	30%	100%
Pengujian 4	60%	40%	95%
Rata-rata			91.25%

Grafik hasil pengujian akurasi dapat dilihat pada gambar 12 berikut:



Gambar 12 Grafik hasil pengujian

5.1 Analisis Hasil Pengujian

Berdasarkan pengujian yang telah dilakukan sebanyak 4 kali percobaan seperti yang sudah dilakukan pada tabel pengujian. Adapun penjelasan dari setiap percobaan adalah sebagai berikut:

1. Pada percobaan I, peneliti mencoba jumlah data training pada datababse sebanyak 46 data. Dengan pembagian data latih 90% dan data uji 10% Didapatkan hasil akurasi sebesar 80%.
2. Pada percobaan 2, peneliti mencoba jumlah data training pada datababse sebanyak 41 data. Dengan pembagian data latih 80% dan data uji 20% Didapatkan hasil akurasi sebanyak 90%
3. Pada percobaan 3, peneliti mencoba jumlah data training pada datababse sebanyak 36 data. Dengan pembagian data latih 70% dan data uji 30% Didapatkan hasil akurasi sebanyak 100%
4. Pada percobaan 4, peneliti mencoba jumlah data training pada datababse sebanyak 31 data. Dengan pembagian data latih 60% dan data uji 40% Didapatkan hasil akurasi sebesar 95%

Dari pengujian I, II, III, dan IV diketahui bahwa semakin banyak jumlah data uji maka semakin tinggi tingkat akurasi. Oleh karena itu hasil dari pengujian pada tabel 6.6 mempunyai rata-rata 91.25% dan mempunyai tingkat akurasi paling tinggi pada pengujian ke 3.

6. KESIMPULAN

Berikut adalah kesimpulan yang dapat diambil dari hasil yang telah didapatkan dari perancangan, implementasi dan pengujian yang telah dilakukan yaitu :

1. Untuk Menerapkan pengelompokkan berita online dengan menggunakan algoritma Single Pass Clustering memerlukan proses preprocessing yaitu tokenizing, stopword dan stemming. Preprocessing tersebut dapat memperkecil term sehingga bisa mempercepat proses perhitungan pembobotan term menggunakan tf-idf dan mempercepat proses cosine similarity.
2. Dari pengujian I, II, III, dan IV diketahui bahwa semakin banyak jumlah data uji maka semakin tinggi tingkat akurasi. hasil rata-rata pengujian 91.25% dan mempunyai tingkat akurasi paling tinggi pada pengujian ke 3 dengan akurasi 100 %. Hal tersebut dapat terjadi karena hasil pada kategori data hasil sistem sesuai dengandata asli.

DAFTAR PUSTAKA

- Arifin Agus Zainal , Setiono Ari novan (2002). Klasifikasi dokumen berita kejadian berbahasa Indonesia dengan algoritma single pass clustering, Proceeding of seminar on intelegent Technology and Its Applications (SITIA), Teknik Elektro, Institut Teknologi Sepuluh.
- Februariyanti Henry, Zuliarso Eri (2013.)Klastering dokumen berita dari web menggunakan Algoritma *single pass clustering*, teknologi Informasi, Universitas Stikubank.
- Februariyanti Henry, Zuliarso Eri (2012.)klastering berita online tentang bencana dengan algoritma *single pass clustering*, teknologi Informasi, Universitas Stikubank.
- Feldman, Ronen , Sanger, dkk. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York.
- Fitri, Meisya. (2013). Perancangan Sistem Temu Balik Informasi Dengan Metode Pembobotan Kombinasi Tf-Idf Untuk Pencarian Dokumen Berbahasa Indonesia. Universitas Tanjungpura : Semarang.
- Ifada Noor, Husni, Liyantanto Rahmady (2011). Pembuatan Web Portal Sindikasi Berita

Indonesia Dengan Klasifikasi Metode *Single Pass Clustering*, teknik Informatika, Universitas Trujoyo.

Klampanos I A., Joemon M. J, C. J. Keithvan Rijsbergen, (2006). *Single Pass Clustering for peer-to-peer Information Retrieval: The Effenct Of Document Ordering, Processings of the firs international conference on scalable information systems.*

Lentera Kecil, (2015). pengertian media online. LenteraKecil.com/pengertian_media_online/. Diakses pada tanggal 20 Februari 2017.

Mahayasa I Nyoman , Jasa Lie, (2016) Sistem Pendukung Keputusan Perekrutan Pegawai Menggunakan Perangkingan Madm Topsis Dan Klasifikasi *Naive Bayes*, Manajemen Sistem Informasi Dan Komputer Universitas Udayana, Denpasar.

W.B.Frakes, and Yates Baeza, R. (1992) *Information Retrieval , Data Structures and Algorithm, Prentice Hall, Englewood New Jersey.*

Zhang J., Jianfeng G., Ming Z., Jiaying W., (2001). *Improving the Effectiveness of Information Retrieval with Clustering and Fusion, Computational Linguistics and Chinese Language Processing*, Vol. 6, No. 1, February 2001, pp. 109-125.