CENTROID BASED CLASSIFIER DENGAN FITUR TF-IDF-ICF UNTUK KLASIFIKASI KELUHAN MAHASISWA PADA APLIKASI E-COMPLAINT DI UNIVERSITAS MUHAMMADIYAH SIDOARJO

(Centroid Based Classifier With TF – IDF – ICF for Classfication of Student's Complaint at Appliation E-Complaint in Muhammadiyah University of Sidoarjo).

Mochamad Alfan Rosid[1], Gunawan[2], Edwin Pramana[3]
Universitas Muhammadiyah Sidoarjo[1]
Sekolah Tinggi Teknik Surabaya, Surabaya [2,3]
alfanrosid@umsida.ac.id[1], gunawan@stts.edu[2], epramana@stts.com[3]

Abstrak
Text mining mengacu pada proses mengambil informasi berkualitas tinggi dari teks. Informasi berkualitas tinggi biasanya diperoleh melalui peramalan pola dan kecenderungan melalui sarana seperti pembelajaran pola statistik. Salah satu kegiatan penting dalam *text mining* adalah klasifikasi atau kategorisasi teks. Kategorisasi teks sendiri saat ini memiliki berbagai metode antara lain metode *K-Nearest Neighbor*, *Naïve Bayes*, dan *Centroid Base Classifier*, atau *decision tree classification*.

Pada penelitian ini, klasifikasi keluhan mahasiswa dilakukan dengan metode *centroid based classifier* dan dengan fitur TF-IDF-ICF, Ada lima tahap yang dilakukan untuk mendapatkan hasil klasifikasi. Tahap pengambilan data keluhan kemudian dilanjutkan dengan tahap preprosesing yaitu mempersiapkan data yang tidak terstruktur sehingga siap digunakan untuk proses selanjutnya, kemudian dilanjutkan dengan proses pembagian data, data dibagi menjadi dua macam yaitu data latih dan data uji, tahap selanjutnya yaitu tahap pelatihan untuk menghasilkan model klasifikasi dan tahap terakhir adalah tahap pengujian yaitu menguji model klasifikasi yang telah dibuat pada tahap pelatihan terhadap data uji. Keluhan untuk pengujian akan diambilkan dari database aplikasi e-complaint Universitas Muhammadiyah Sidoarjo. Adapun hasil uji coba menunjukkan bahwa klasifikasi keluhan dengan algoritma *centroid based classifier* dan dengan fitur TF-IDF-ICF memiliki rata-rata akurasi yang cukup tinggi yaitu 79.5%. Nilai akurasi akan meningkat dengan meningkatnya data latih dan efesiensi sistem semakin menurun dengan meningkatnya data latih.

Kata Kunci: Text Mining, Klasifikasi, Keluhan, Centroid Based Classifier

1. Pendahuluan

Pada saat ini proses klasifikasi dilakukan secara manual oleh staf BPM, namun dengan semakin banyaknya keluhan yang harus diproses maka waktu yang dibutuhkan akan semakin lama dan tinggat akurasi juga semakin berkurang dikarenakan keterbatasan kemampuan manusia dalam memproses data. Ide untuk membuat aplikasi keluhan secara elektronikpun sudah dilakukan yakni aplikasi e-complaint dengan menyediakan pilihan unit kerja tujuan pada form keluhan, namun ide ini masih belum efektif dan belum akurat dikarenakan mahasiswa terkadang salah memilih unit

kerja tujuan yang sesuai dengan isi keluhan yang disampaikan.

Dari permasalahan diatas maka disiplin ilmu text mining menjadi titik tolak metode penting yang dimanfaatkan pada penelitian ini. Teks mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Metode yang digunakan dalam Pengklasifikasian ini adalah Centroid Based Classifier.

Metode Centroid Based merepresentasikan dokumen kedalam bentuk vektor. Metode ini membentuk vektor centroid pada sekumpulan dokumen yang termasuk pada suatu kelas tertentu. Adapun pembentukan vektor centroid nantinya menggunakan metode CFC(Class Feature Centroid), vektor centroid tersebut yang akan digunakan sebagai model untuk mengelompokan suatu dokumen dengan menggunakan kesamaan kosinus[7]. Dalam metode ini digunakan fitur TF-IDF-ICF. Metode TF-IDF-ICF (Term Frequency-Inverse Document Frequency-Inverse Class Frequency) adalah metode untuk melakukan pembobotan hubungan suatu kata (term) terhadap dokumen. Metode ini menggabungkan tiga konsep untuk perhitungan bobot yaitu Term Frequency (TF) merupakan frekuensi kemunculan kata pada kalimat, document frequency (DF) adalah banyaknya dokumen dimana suatu kata muncul dan inverse class frequency (ICF) adalah banyaknya kelas dimana suatu kata muncul.

Diharapkan dengan adanya penelitian ini, akan dihasilkan sebuah sistem pengklasifikasian keluhan mahasiswa yang dapat membagi keluhan yang masuk berdasarkan unit kerja yang ada yaitu Kemahasiswaan, Sarana dan Prasarana, Kebersihan/PRT, BAA, BAK, PUSDAKOM,Fakultas, Perpustakaan, Keamanan, LPPM, BPM, Lain-Lain dengan otomatis dan cepat untuk menambah performa dari Aplikasi E-Complaint yang sudah ada.

2. Centroid Based Classifier

Dalam metode centroid-based, vektor centroid dihitung untuk mewakili dokumen dari masing-masing kelas, dan dokumen baru diberikan ke kelas yang sesuai dengan vektor centroid yang paling mirip, diukur dengan fungsi kosinus. Untuk dapat mengimplementasikan metode ini, diperlukan representasi dokumen menggunakan *vector space model* , dalam model ini, setiap dokumen *dj* ditransformasikan menjadi suatu vektor $dj = (w_{1j}, w_{2j}, \ldots w_{ij},)$ dimana w_{ij}, adalah bobot *term* ke-i pada dokumen j.

Bobot setiap *term* dapat direprisentasikan secara binari (*true* atau *false* atau dengan frekuensi dan frekuensi invers dokumennya (TF-IDF), metode TF-IDF dinyatakan sebagai berikut:

$$w_{ij} = tf \cdot \log (N/df_i) \quad (1)$$

dimana w_{ij} adalah bobot term i pada dokumen j, N adalah jumlah dokumen yang diproses dan df_i adalah jumlah dokumen yang memiliki *term* I didalamnya

Pada penelitian ini, diusulkan model representasi dokumen dengan metode TF-IDF-ICF dan menggunakan metode CFC pada saat pembentukan centroid tiap kelas. Rumus metode TF-IDF-ICF adalah sebagai berikut:

$$wij = TFij * \left(log \frac{N}{DFi}\right) * \left(log \frac{|C|}{Cfi}\right) \quad (2)$$

dimana w*ij* adalah bobot *term* i pada dokumen j, TF*ij* adalah frekuensi *term* i pada dokumen j, N adalah jumlah dokumen yang diproses, DF*i* adalah jumlah dokumen yang memiliki *term* i, |C| adalah total kelas dan Cf*i* adalah jumlah kelas yang mengandung *term* i. Sedangkan rumus CFC adalah sebagai berikut:

$$wij = \left(b \wedge \frac{DF_{ti}^{j}}{|Cj|} \times \log(\frac{|C|}{CFi}\right) \quad (3)$$

dimana wij adalah bobot *term* i pada dokumen vektor centroid kelas j, b adalah konstanta (b>1), DF_{ti}^{j} adalah jumlah frekuensi term i pada kelas Cj, |Cj| adalah jumlah dokumen dikelas Cj, |C| adalah total kelas dan CF*i* adalah jumlah kelas yang mengandung *term* i.

Setelah didapat bobot masing-masing *term*, dibentuk centroid masing-masing kelas dengan rumus:

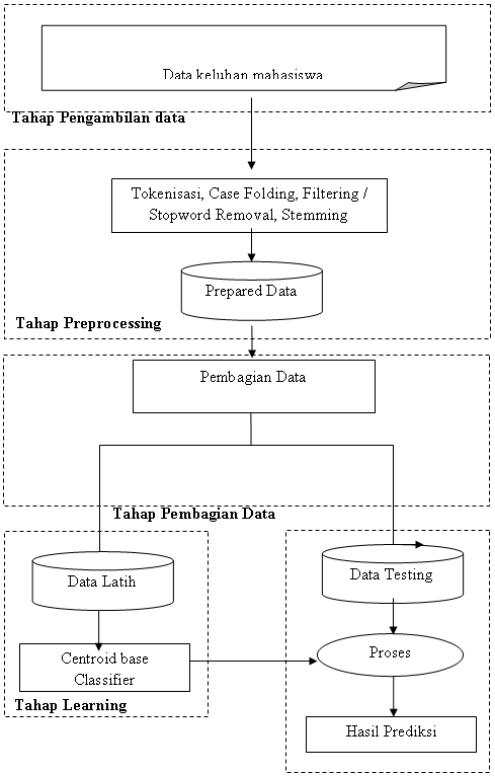
$$Ci = \frac{1}{|Si|} \sum_{d \in S} d \qquad (4)$$

dimana C*i* adalah vektor centroid kelas i, |S| adalah jumlah dokumen dikelas i dan *d* adalah bobot term. Klasifikasi dokumen dicari dengan mencari nilai terbesar kesamaan kosinus antara dokumen uji dengan centroid masing-masing kelas, dengan menggunakan rumus:

$$\cos(di, Cj) = \frac{d.C}{||d||_2 * ||C||_2} \qquad (5)$$

3. Arsitektur Sistem

Pada gambar 1 dijelaskan secara rinci proses utama dalam penelitian ini, yaitu terdiri dari 5 tahap, tahap pengambilan data, tahap *preprocessing*, tahap pelatihan dan tahap *prediction*.



Gambar 1. Arsitektur Sistem

1. Tahap Pengambilan Data

Pada proses ini, data langsung diambil dari database aplikasi e-complaint Universitas Muhammadiyah Sidoarjo pada tabel keluhan mahasiswa tanpa melalui proses crowling.

2. Tahap Preprocessing

Data keluhan yang digunakan pada penelitian ini berupa data teks yang tidak terstruktur (*unstructured data*). Sehingga diperlukan langkah – langkah untuk mempersiapkan data tersebut sehingga siap digunakan untuk proses selanjutnya. Tahap preprocessing terdiri atas tahap *Case Folding*, Tokenisasi, *Stopword Removal*, dan stemming. Pada tahap stemming, stemmer yang digunakan adalah stemmer for bahasa Indonesia yang telah diteliti oleh Fadillah Z. Talla[12].

3. Tahap Pembagian Data

Setelah melakukan tahap preprocessing, tahap selanjutnya adalah tahap pembagian data menjadi 2 (dua) bagian, yaitu data latih dan data uji. Proses pembagian data dilakukan dengan cara mengambil berapa prosen dari data praproses untuk data latih dan berapa prosen untuk data uji, misalnya 70% untuk data latih dan 30% untuk data uji.

4. Tahap Pelatihan

Tahap ini menjelaskan bagaimana sistem mampu membuat basis pengetahuan melalui proses pembelajaran menggunakan data – data yang telah dipersiapkan pada proses sebelumnya. Metode *Centroid Base Classifier* digunakan pada proses ini untuk menghasilkan model pembelajaran, sehingga model ini dapat digunakan untuk tahap pengujian atau testing nantinya.

5. Tahap Prediction (Pengujian)

Pada tahap ini dilakukan proses testing menggunakan data testing/uji untuk menguji keakuratan prediksi dari model basis pengetahuan telah dihasilkan menggunakan pendekatan algoritma *Centroid Base Classifier.*

4. Model Klasifikasi

Model klasifikasi dengan *Centroid Based Classifier* adalah bagaimana menentukan nilai *centroid* dan mencari nilai cosine similarity terbesar antara *centroid* masing-masing kelas

dengan jumlah bobot dokumen yang akan diklasifikasi. Berikut ini disajikan contoh perolehan nilai cosine similarity untuk dokumen D1, D15 dan D45 terhadap centroid K1, K2, K3, K4 dan K5 pada tabel 1.

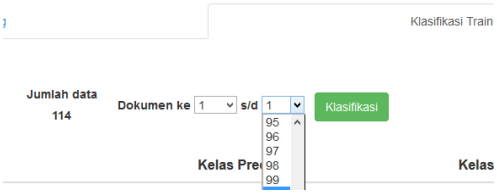
Pada tabel 1 diperoleh nilai cosine terbesar untuk dokumen D1 adalah sebesar 0.0367 dengan atau lebih dekat dengan *centroid* K1 sehingga dapat diketahui bahwa dokumen D1 terklasifikasi kedalam kelas K1. Diketahui juga bahwa tidak ada perbedaan kelas antara pemberian kelas secara manual dengan yang diperoleh dengan melalui *centroid based classifier* yaitu sama-sama termasuk kedalam kelas K1.

Tabel 1 Nilai Cosine untuk masing-masing dokumen

. Awal	entroid	lai Cosine	ermasuk Klasifikasi	asifikasi Manual
D1	*troid* K1	0.0367	K1	K1
	troid K2	0.0000		
	troid K3	0.0000		
	troid K4	0.0000		
	troid K5	0.0000		
D15	*troid* K1	0.0000	K2	K2
	troid K2	0.0360		
	troid K3	0.0000		
	troid K4	0.0000		
	troid K5	0.0000		
D45	*troid* K1	0.0008	K5	K5
	troid K2	0.0000		
	troid K3	0.0000		
	troid K4	0.0000		
	troid K5	0.0230		

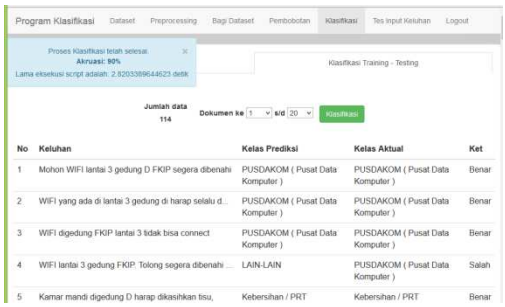
Setelah model klasifikasi diperoleh dari data latih, langkah berikutnya adalah menggunakan model klasifikasi tersebut terhadap data uji. Proses penentuan jenis kelas terhadap data uji dengan menggunakan model klasifikasi yang dihasilkan oleh data latih, digunakan untuk menghitung akurasi data uji terhadap jenis kelas yang dihasilkan oleh Centroid Based Classifier dengan jenis kelas manual.

Proses penentuan jenis kelas untuk data uji dilakukan dengan memilih jumlah data uji yang akan diklasifikasi, ditunjukkan pada gambar 1.



Gambar 1 Form Klasifikasi untuk Data Uji

Pada gambar 1 admin memilih range dokumen data uji yang akan diklasifikasi dan memilih tombol Klasifikasi. Proses penentuan jenis kelas untuk data uji sama seperti proses pada data latih yaitu melalui beberapa tahap. Dengan menggunakan model klasifikasi yang telah diperoleh pada data latih, diperoleh hasil berupa jenis kelas terhadap data uji. Tampilan penentuan jenis kelas untuk data uji dapat dilihat pada gambar 2.



Gambar 2 Tampilan Jenis Kelas Untuk Data Uji

5. Uji Coba

Pengujian dilakukan untuk mengklasifikasi data keluhan mahasiswa kedalam klasifikasi unit kerja yang telah ada. Dataset menggunakan data keluhan mahasiswa yang sudah ada yaitu sebanyak 500 data keluhan yang telah terklasifikasi secara manual. Data keluhan mahasiswa sebanyak 500 data dibagi menjadi dua bagian, sejumlah 350 data untuk data latih, dan 150 data untuk data uji. Contoh data latih dapat dilihat pada tabel 2 dan contoh data uji dapat dilihat pada tabel 3.

Tabel 2 Contoh Data Latih

No	Keluhan	Klasifikasi Manual

1	Kasih alat musik untuk IKABAMA agar kita bisa belajar alat music dengan nyaman Trims	Kemahasiswaan
2	Lebih memfasilitasi tiap UKM (trutama IKABAMA) dengan fasilitas yang layak selain secretariat	Kemahasiswaan
3	Papan sudah tidak layak pakai ruang 301-308	Sarana dan Prasarana
4	Lampu sering mati mengganggu proses belajar mengajar.	Sarana dan Prasarana
5	Kamar mandinya fak agama islam KOTOR. Mohon tiap minggu di bersihkan	Kebersihan / PRT
6	Mohon Kurangnya penerangan dan kebersihan dikampus terutama ditoilet.	Kebersihan / PRT
7	Untuk BAA Mohon perhatianya untuk absen anak tarbiyah kok belum ada. Padahal dah beberapa pertemuan kuliah. Apakah harus menunggu ujian tiba	BAA (Biro Administrasi Akademik)
8	Mohon diberikan pelayanan pada sore hari misalnya BAA dan yang lain supaya mahasiswa kelas sore tidak harus bolak-balik ke kampus. Karena mahasiswa kelas sore/malam tidak selalu punya waktu untuk kekampus pada pagi/siang hari yang alasannya karena terhambat pekerjaan (bekerja) terimakasih	BAA (Biro Administrasi Akademik)
9	Petugas BAK kampus ini tidak bisa sopan dan senyum yaa..?	BAK (Biro Administrasi Keuangan)
10	Kalau bisa SPP/bayar ujian jangan terlalu mahal	BAK (Biro Administrasi Keuangan)
11	Tolong WIFI dihidupkan di lantai 3 gedung FKIP	PUSDAKOM (Pusat Data Komputer)
12	WIFI lantai 3 tolong dinyalakan untuk mempermudah mengerjakan tugas	PUSDAKOM (Pusat Data Komputer)
13	Admin terlalu lamban dan dipersulit. Misal surat keterangan PKL.	Fakultas
14	Dipermudah dalam peminjaman buku diperpus UMSIDA	Perpustakaan
15	Buku diperpus kurang lengkap. Terlalu acak-acakan	Perpustakaan

Pada percobaan klasifikasi pertama jumlah record data yang akan dipakai untuk training set atau data uji sejumlah 350 data keluhan, klasifikasi awal dilakukan secara manual oleh petugas Badan Penjaminan Mutu.

Tabel 3 Contoh Data Uji

No	Keluhan	Klasifikasi Manual
1	Mohon WIFI lantai 3 gedung D FKIP segera dibenahi	PUSDAKOM (Pusat Data Komputer)
2	Kuliah tarbiyah gak jelas dosennya, tidak profesional.Sudah bayar mahal, dosennya jarang masuk	BPM (BADAN PENJAMINAN MUTU)
3	Tidak ada tanggung jawab saat kehilangan barang dikendaraan yang sedang diparkir.	Keamanan
4	Tempat area parkir mohon diberi atap tidak hanya tempat parkir dosen aja yang ada atapnya dan biar sepeda kami tidak kepanasan, mbulak ntar	Sarana dan Prasarana
5	Tolong kelas fisip AN kebersihannya harap diperhatikan karena banyak sarang laba-laba dan masih berdebu disekitar jendela	Kebersihan / PRT
6	UKM agar difasilitasi computer dan printer	Kemahasiswaan
7	Masalah denda jangan terlalu over dosis	BAK (Biro Administrasi Keuangan)
8	proses pembayaran terlalu banyak meja	BAK (Biro Administrasi Keuangan)
9	staf fakultas terkadang kurang ramah	Fakultas
10	BUKU LITERATUR HARUS DI TAMBAH	Perpustakaan
11	Tolong diperjelas alur pengurusan PKM kami butuh	LPPM

6. Hasil Pengujian

Berdasarkan hasil pengujian, diperoleh hasil klasifikasi keluhan mahasiswa yang dilakukan dengan menggunakan metode centroid based classifier, ada beberapa

keluhan yang terklasifikasi secara tidak tepat dan ada yang tidak tepat. Tingkat akurasi yang dapat diperoleh adalah:

Jumlah klasifikasi yang benar : 119
Jumlah klasifikasi yang salah : 31
Jumlah dokumen keseluruhan : 150

Maka nilai akurasi dari aplikasi yang telah dibangun adalah:

Akurasi=

$$\frac{Jumlah\ Klasifikasi\ Benar}{Jumlah\ Dokumen\ Keseluruhan} X\ 100$$

Akurasi = $\frac{119}{150} X\ 100 = 79.3$

Untuk melakukan verifikasi hasil percobaan, dilakukan cross validation, percobaan akan dilakukan sebanyak 10 kali, dataset dari percobaan ini dibagi menjadi dua buah yakni data latih dan data testing, penentuan data data latih dan data testing ini dilakukan secara acak dan merata agar tidak terjadi pengelompokan dokumen-dokumen yang berasal dari satu kategori tertentu pada sebuah percobaan. Data latih diambil sebanyak 70% dari dataset secara acak dan merata untuk setiap kategori, sedangkan sisanya dipakai sebagai data testing, pembagian data ini akan dilakukan sebanyak 10 kali untuk sepuluh percobaan. Hasil dari percobaan nantinya dicari nilai akurasi rata-ratanya. Hasil cross validation dari 10 percobaan ditunjukkan oleh tabel 4.

Tabel 4 Hasil Cross Validation

	Fold										**Rata**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Data	78	78	78	75	78	82	80	86	77	83	**79.5**

Pada tabel 4 ditunjukkan nilai akurasi terbesar terjadi pada percobaan ke 8 yaitu sebesar 86%, sedangkan nilai akurasi rata-rata yang ditunjukkan pada tabel 4 sebesar 79.5%

7. Kesimpulan

Dari hasil penelitian dan berdasarkan atas hipotesa penelitian, maka dapat ditarik beberapa kesimpulan sebagai berikut:

1. Dari hasil percobaan yang dilakukan untuk mengklasifikasi data keluhan mahasiswa menggunakan metode Centroid Based Classifier yang diusulkan terbukti mampu melakukan klasifikasi dokumen keluhan mahasiswa.
2. Dengan Menggunakan metode klasifikasi teks yaitu metode Centroid Based Classifier , dari segi kecocokan jenis kelas yang dihasilkan oleh Centroid Based Classifier terhadap penentuan jenis kelas yang dilakukan oleh petugas BPM (Badan Penjaminan Mutu) tergolong baik. Prosentasi kecocokan jenis kelas terhadap 162 data uji rata-rata sebesar 79.5%.
3. Dari hasil percobaan yang dilakukan untuk mengklasifikasi data keluhan mahasiswa dengan menambahkan metode CFC pada proses pembobotan dapat meningkatkan akurasi klasifikasi rata-rata sebesar 0.6%.
4. Dengan pengklasifikasian data keluhan mahasiswa dengan metode Centroid Based Classifier, dapat meningkatkan efektifitas aplikasi e-complaint yang sebelumnya harus memilih jenis kelas secara manual pada saat pengisian form, dan dapat mempercepat pengambilan keputusan oleh pimpinan Universitas Muhammadiyah Sidoarjo.

8. Daftar Pustaka

[1]. Verayuth Lertnattee, Chanisara Leuviphan., "*Using Class Frequency for Improving Centroid-based Text Classification*". Department Of Health-related Informatics, Silpakorn University, Maung, Nakorn Pathom, Thailand, 2012.

[2]. Eui-Hong (Sam) Han, George Karypis., "*Centroid-Based Document Classification: Analysis & Experimental Results*". Department of Computer Science / Army HPC

Research Center, University of Minnesota.

[3]. Songbo Tan *," An improved centroid classifier for text categorization",* Intelligent Software Department, Institute of Computing Technology, Chinese Academy of Sciences, PR China, 2007,

[4]. Hidayet Takci, Tunga Gungor. *"A High Performance Centroid-based Classification Approach for Language Identification".* Department of Computer Enginering, GYTE, Kocaeli, Turkey, 2012.

[5]. Joel W. Reed, Yu Jiao, Thomas E. Potok, Brian A. Klump, Mark T. Elmore, Ali R. Hurson, "*TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams",* Computer Science and Engineering Department, The Pennsylvania State University, University Park, 2006.

[6]. Manning C. D. and H. Shutze: *Foundations of Statistical Natural Language Processing*, Chapter 15. MIT Press. 1999.

[7]. Hu Guan, Jingyu Zhou, Minyi Guo, " *A Class-Feature-Centroid Classifier for Text Categorization",* Computer Science Dept, Shanghai Jiao Tong University, China, 2009.

[8]. Ronen Feldman, James Sanger, 2007. "*The Text Mining Handbook, Advanced Approaches in analyzing Unstructured Data*". Cambridge University Press, Cambridge, England.

[9]. Bambang Kurniawan, Syahril Efendi, dan Opim Salim Sitompul, *Klasifikasi Konten Berita Dengan Metode Text Mining,* Jurnal Dunia Teknologi Informasi vol.1, No.I, 2012

[10].Chakrabarti, Soumen, 2003, Mining the Web: *Discovering knowledge from hypertext data*. San Francisco: Morgan Kaufman.

[11].Porter, M. F. ,1980, *An algorithm for suffix stripping,* Program 14(3), p. 130-137.

[12].Fadillah Z. Talla, A Study of *Stemming Effects on Information* Retrieval in Bahasa, MS Thesis, 2003.