Klasifikasi Teks Emosi Bahasa Aceh Menggunakan Metode Term Frekuensi / Invers Dokument Frekuensi

, Mursyidah[1], Hari Toha Hidayat[2] ,Desy Meutia Sari[3]

Program Studi Teknik Multimedia dan Jaringan, Jurusan Teknologi Informasi dan Komputer, Politeknik Negeri Lhokseumawe, Jalan Banda Aceh-Medan Km. 280,3 Buketrata, Lhokseumawe, 24301 PO.BOX 90 Telpon (0645) 42670, 42785 Fax 42785, Indonesia

Abstrak-Dalam bahasa Aceh teks emosi marah, senang, sedih, jijik dan normal dapat dinyatakan dalam bentuk verbal (menulis kata-kata). Emosi marah, senang, sedih, jijik dan normal juga dapat ditunjukkan dengan teks, akan tetapi tingkatan emosinya agak sulit ditebak disebabkan teks yang tidak dikenali suatu teks itu berupa emosi marah, senang, sedih, jijik dan bahkan emosi normal. Tingkatan emosinya agak sulit ditebak karena dalam bahasa aceh teks emosi belum tentu dapat digambarkan dengan pasti perasaan emosi yang dirasakan oleh teks itu sendiri. Oleh karena itu, dibuat sebuah program aplikasi yang dapat mengetahui tingkat emosi seseorang. Penelitian ini bertujuan untuk membahas teks emosi marah melalui kalimat atau teks marah, senang, sedih, jijik dan normal dalam bahasa Aceh untuk memperoleh tampilan persentase tingkat emosi dari suatu teks. Program aplikasi ini menggunakan metode *Term Frekuensi / Invers Dokument Frekuensi.* Penelitian sistem aplikasi ini menghasilkan akurasi kebenaran prediksi sebesar 80%.

Kata Kunci : Teks Emosi, *TF-IDF* , Klasifikasi teks

Abstract- In the language of the text Aceh angry emotions, happy, sad, disgusted and normal can be expressed in the form of verbal (written words). Emotions angry, happy, sad, disgusted and normal can also be shown with the text, but the level of emotions somewhat unpredictable because the text does not recognize the text in the form of angry emotions, happy, sad, disgusted and even a normal emotion. Depth emotions somewhat unpredictable because the Acehnese language is not necessarily emotional text can be described by a definite feeling emotions felt by the text itself. Therefore, made an application program that can determine the level of a person's emotions. This study aims to discuss the text of the sentence or angry emotions through texts angry, happy, sad, disgusted and normal in the Acehnese language to get to see the percentage level of emotion of a text. This application program using Term Frequency / Inverse Frequency Documents. This application systems research resulted in the predicted accuracy by 80%.

Key words:Text Emotions, TF-IDF, text classification

I. PENDAHULUAN

Sebuah teks dapat berisi suatu informasi tentang perilaku manusia terhadap emosi. Jenis emosi seperti normal, senang, sedih, marah dan jijik telah di kenal sejak lama dan menjadi peranan dalam komunikasi antar manusia di kehidupan sehari hari dan menjadi aspek penting dari perilaku manusia pada umumnya. Akan tetapi penerapan emosi belum banyak di gunakan untuk interaksi antara manusia dan komputer.

Teks emosi tidak hanya ditunjukkan melalui lisan, tetapi juga dapat melalui tulisan. Emosi seseorang lebih mudah ditebak karena tergambar melalui ekspresi wajah dan nada suara seseorang pada saat berbicara.

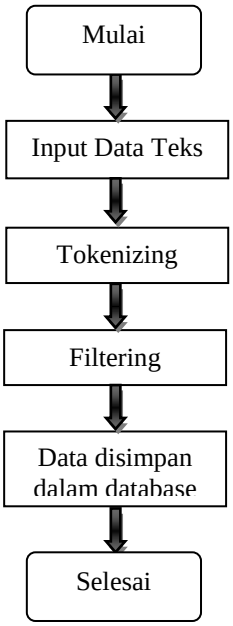
Seiring dengan perkembangan teknologi yang semakin canggih, dilakukan penelitian di bidang emosi antara manusia dan komputer. Penelitian yang telah di lakuakan sebagian besar masih menggunakan bahasa inggris dan bahasa Indonesia, sedangkan untuk teks bahasa aceh belum dilakukan. Oleh karena itu
pada tugas akhir ini akan dibuat penelitian klasifikasi teks emosi menggunakan bahasa aceh .

Pada Penelitian ini dapat dideteksi suatu teks emosi seseorang melalui teks bahasa aceh menggunakan metode *Term Frekuensi / Invers Dokument Frekuensi* (TF-IDF). *Term Frekuensi / Invers Dokument Frekuensi* (TF-IDF) adalah suatu metode yang digunakan untuk menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*. Dalam proses menghitung bobot setiap kata teks emosi menggunakan bahasa aceh disimpan

dalam format .txt di dalam suatu folder tersenderi berdasarkan kelasnya masing-masing. Setelah data teks disimpan tahap selanjutnya Indexing untuk memberikan nomor pada setiap kata, tahap *tokenizing* untuk memecah kata berdasarkan spasi, tahap *filtering* dalam penelitian ini untuk membuang *stopword* yaitu kata penghubung akan dihapus karena termasuk kata yang tidak mempunyai makna atau arti

II. METODOLOGI PENELITIAN

Untuk gambaran proses data *training* sistem klasifikasi teks emosi dalam bahasa aceh ditunjukkan pada gambar 1 berikut ini.



Gambar 1. Blog diagram data *training*

2.1Klasifikasi

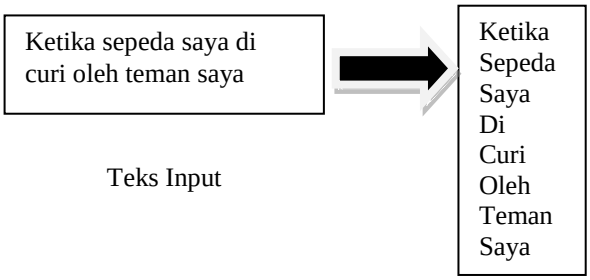
Klasifikasi adalah proses menentukan suatu obyek kedalam suatu kelas atau kategori yang telah ditentukan. Penentuan obyek dapat menggunakan suatu model tertentu bebeapa model yang bisa digunakan antara lain: *classification (IF-THEN)rules*, *decision trees*, formula matematika atau *neural networks*. Klasifikasi data atau dokumen dimulai dengan membangun aturan klasifikasi dengan algoritma klasifikasi tertentu menggunakan data training (tahapan ini sering disebut dengan tahapan pembelajaran) dan tahap pengujian algoritma dengan data testing.

2.2 Teks Mining

Teks mining merupakan penerapan konsep dan teknik data mining untuk mencari pola dalam teks proses penganalisisan teks untuk menyarikan informasiyang bermanfaat untuk tujuan tertentu. Proses data mining untuk data dokumen atau teks memerlukan lebih banyak tahapan, mengingat data teks memiliki karakteristik yang lebih kompleks.. Kolom menunjuk dokumen dan baris menunjuk kata, sedangkan selnya menunjuk frekuensi kata dalam dokumen. Teks Mining mencoba untuk mengekstrak informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi dari suatu pola menarik. Sumber data berupa sekumpulan dokumen dan pola menarik yang tidak ditemukan dalam bentuk database tetapi dalam data teks yang tidak terstruktur

2.3 *Tokenizing*

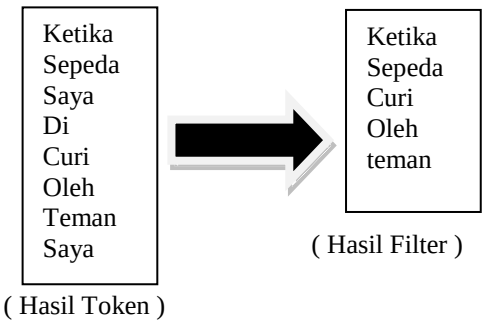
Tokenizing adalah proses memecah teks menjadi kalimat dan kata/token. Fitur ini terdiri dari tipe kapitalisasi, keberadaan digit, tanda baca, karakter spesial dan lain sebagainya. Hasil keluaran dari proses *tokenizing*akan dipergunakan sebagai masukan dalam tahap transformasi teks.Untuk gambaran proses tokenizing ditunjukkan pada gambar 2 berikut ini.



Gambar 2. Tahapan Proses Tokenizing

2.4 Filtering

Filtering adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma stop list (membuang kata yang kurang penting) atau word list (menyimpan kata penting).Untuk gambaran proses filtering pada pengujian sistem klasifikasi teks emosi ditunjukkan pada gambar 3 berikut ini.



Gambar 3. Proses Teks Filtering

2.5 Metode TF-IDF

Metode TF-IDF merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini akan menghitung nilai *Term Frequency* (*TF*) dan *Inverse Document Frequency* (*IDF*) pada setiap token (kata) di setiap dokumen.
Metode ini akan menghitung bobot setiap *token* t di dokumen d. Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah cara pemberian bobot hubungan suatu kata (*term*) terhadap dokumen. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu *Term frequency* (TF) merupakan frekuensi kemunculan kata (t) pada kalimat (d). TF adalah algoritma pembobotan heuristik yang menentukan bobot dokumen berdasarkan kemunculan *term* (istilah). Untuk rumus TF-IDF dapat dihitung dengan rumus sebagai berikut :

$$IDF = \log(\frac{D}{df})$$

Dengan keterangan sebagai berikut:
D = jumlah dokumen
df = jumlah kemunculan (frekuensi) *term*terhadap *D*.

Adapun persamaan yang digunakan untuk menghitung bobot (*W*) masing-masing dokumen terhadap kata kunci (*query*)dapat dihitung dengan rumus sebagai berikut :

$$W_{d.t} = tf_{d.t} * IDF_t$$

dengan keterangan sebagai berikut :
d = dokumen ke–d
t = *term* ke–t dari kata kunci
tf = *term* frekuensi/frekuensi kata
Wd.t = bobot dokumen ke–d terhadap term ke–t.

2.6 *Cosine Similarity*
Cosine Similarity merupakan metode yang digunakan untuk menghitung *similarity* (tingkat kesamaan) antar dua buah objek. Secara umum penghitungan metode ini didasarkan pada *vector space similarity measure.* Metode *cosine similarity* ini menghitung *similarity* antara dua buah objek (misalkan D1 dan D2) yang dinyatakan dalam dua buah vektor dengan menggunakan *keywords* (kata kunci) dari sebuah dokumen sebagai ukuran. Hal ini memperkecil pengaruh panjang dokumen. Jarak *euclidean* (panjang) kedua vektor digunakan sebagai faktor normalisasi. Hal ini diperlukan karena dokumen yang panjang cenderung mendapatkan nilai yang besar dibandingkan dengan dokumen yang lebih pendek. Perhitungan *cosine similarity* yang memperhitungkan perhitungan pembobotan kata pada suatu dokumen dapat dinyatakan dengan rumus sebagai berikut ini

$$\cos(\theta_{ij}) = \frac{\sum_{k=1}^{t}(d_{ik}d_{jk})}{\sqrt{\sum_{k=1}^{t}(d_{ij})^2}.\sqrt{\sum_{j=1}^{t}(d_{ij})^2}}$$

Diketahui:
d_{ik} = Nilai Pembobotan kata *training*
d_{jk} = Nilai Pembobotan katauji

III. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

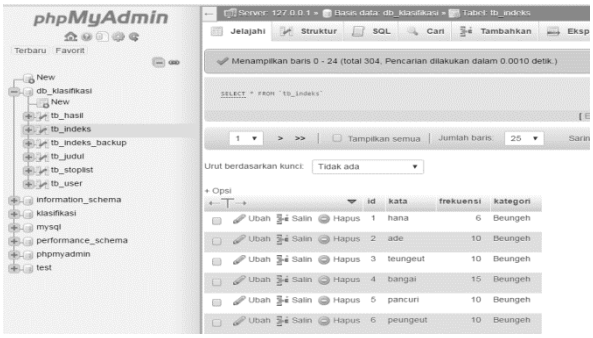
Pengumpulan data dilakukan dengan menggunakan data sampel dari teks emosi dari penelitian sebelumnya yang pernah dilakukan dengan menggunakan bahasa inggris. Data yang didapat akan di translate ke bahasa Indonesia dan di terjemahkan ke dalam bahasa aceh. Emosi yang digunakan ada 5 yaitu normal, senang, sedih, marah dan jijik. Masing-masing teks emosi dilakukan 5 pengujian dengan teks emosi yang berbeda beda . Jumlah seluruh sample yang akan diuji sebanyak 25 teks untuk mengetahui keakuratan program yang telah dibuat.Sampel teks emosi bahasa acehdapat dilihat pada table1 sebagai berikut :

Table 1 Teks Jumlah Data Teks Emosi

NO	KATAGORI	JUMLAH
1	Emosi Luwat	26
2	Emosi Beungeh	34
3	Emosi Normal	16
4	Emosi Seunang	35
5	Emosi Seudeh	14

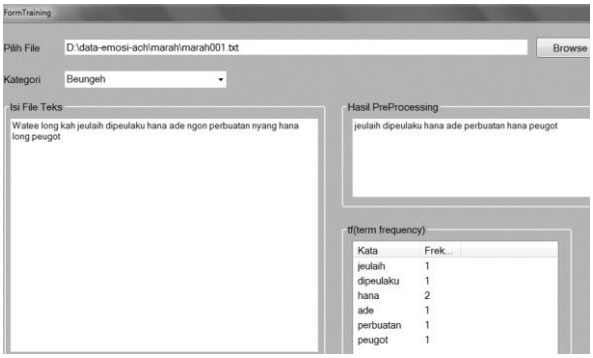
3.2 Pengolahan Teks

Perancangan table dilakukan untuk menyimpan data teks emosi kedalam data base. Untuk merancang tabel pada phpMyAdmin dapat dilihat pada gambar 4 sebagai berikut :



Gambar 4 Perancangan Table

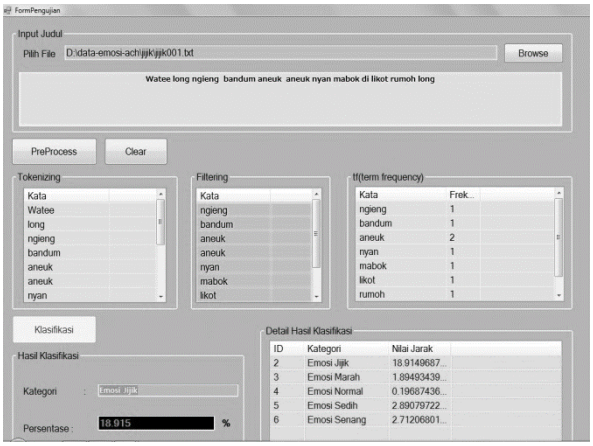
Selanjutnya akan dilakukan data training untuk menghitung bobot kata dan banyaknya frekuensi kemunculan kata dalam setiap teks yang disimpan ke dalam database. Proses training data dapat dilihat pada gambar 5 sebagai berikut :



Gambar 5 Training Data

Setelah dilakukan training data selanjutnya kita akan lakukan pengujian teks emosi.

Hasil pengujian sistem aplikasi teks emosi jijik dapat dilihat gambar 6.berikut ini :



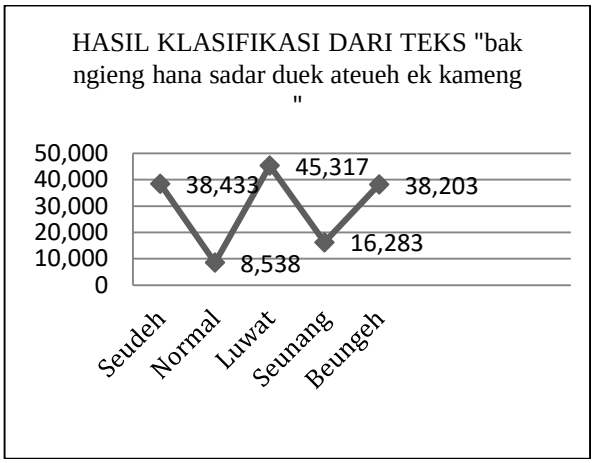
Gambar 6 Pengujian Teks Emosi

Pada penelitian ini, dilakukan 5 tahap pengujian menggunakan data sampel dari 5 kalimat jijik bahasa Aceh dengan 5 teks yang berbedaMasing-masing Untuk masing-masing kalimat jijik bahasa Aceh diinput 1 kali jumlah sampel teks yang diuji adalah 1 x 26 = 26 data sampel.

Pengujian yang dilakukan dalam penelitian ini, diantaranya adalah srbagai berikut :

- Pengujian dengan teks emosi jijik

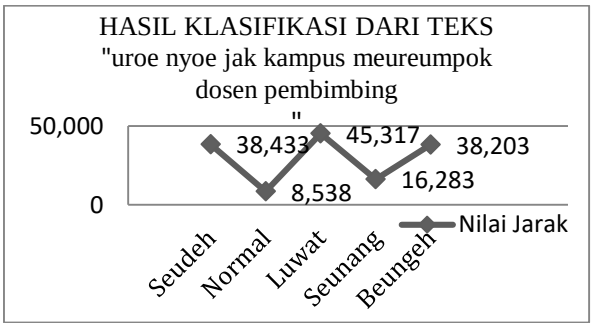
Hasil pengujian sistem aplikasi teks emosi jijik dapat dilihat dalam bentuk grafik pada gambar 7 berikut ini :



Gambar 7 Grafik Teks Emosi Jijik

- Pengujian dengan teks emosi normal

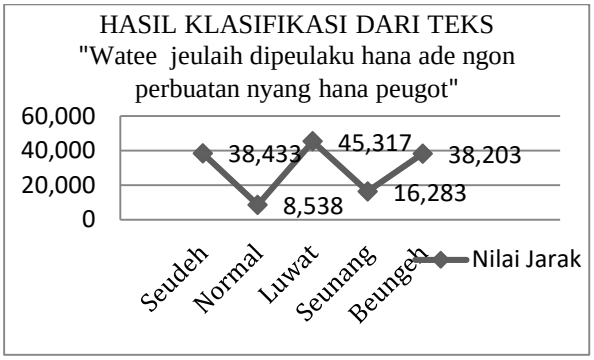
Hasil pengujian sistem aplikasi teks emosi normal dapat dilihat dalam bentuk grafik pada gambar 8 berikut ini :



Gambar 8 Teks Emosi Normal

- Pengujian dengan teks emosi marah

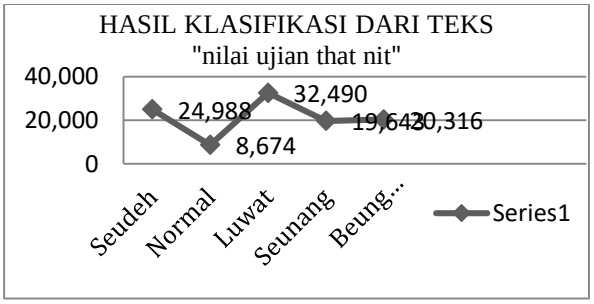
Hasil pengujian sistem aplikasi teks emosi marah dapat dilihat dalam bentuk grafik pada gambar 9 berikut ini :



Gambar 9 Teks Emosi Marah

- Pengujian dengan teks emosi sedih

Hasil pengujian sistem aplikasi teks emosi sedih dapat dilihat dalam bentuk grafik pada gambar 10 berikut ini :

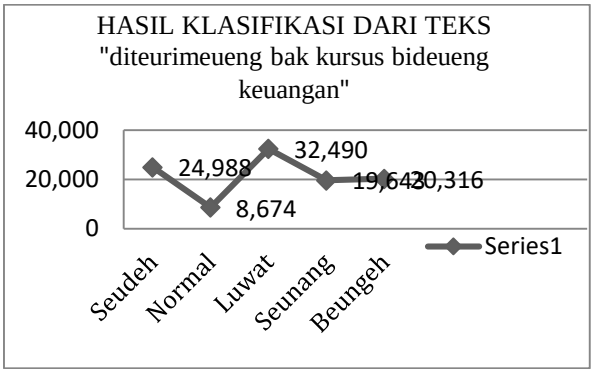


Gambar 10 Teks Emosi Sedih

Dari hasil pengujian kategori teks emosi sedih tingkat akurasi keberhasilan yang dihasilkan ialah 80% berhasil, hal ini dikarenakan teks yang diinput tidak sesuai dengan hasil klasifikasi. Tetapi terdapat pendekatan dengan kategori yang dituju dan nilai yang dihasilkan.

- Pengujian dengan teks emosi Senang

Hasil pengujian sistem aplikasi teks emosi Senangdapat dilihat dalam bentuk grafik pada gambar 11 berikut ini :



Gambar 11 Teks Emosi Senang

Dapat dianalisa dari hasil pengujian kategori teks emosi marah tingkat akurasi keberhasilan yang dihasilkan ialah 100% berhasil, hal ini dikarenakan teks yang diinput pada saat pengujian semua sesuai dengan hasil klasifikasi kategori yangdiinginkan.

IV. KESIMPULAN

Berdasarkan pembahasan yang telah diuraikan pada bab-bab sebelumnya, maka dapat disimpulkan beberapa hal sebagai berikut.

1. Klasifikasi dokumen memiliki tingkat persentase keakuratan yang cukup baik dalam melakukan proses pengelompokan berdasarkan data uji dengan data yang berasal dari *database*, yaitu tingkat keakuratan mencapai 80%.
2. Untuk mendapatkan hasil klasifikasi teks yang sempurna, sebaiknya menggunakan dua metode, metode TF-IDF sebagai metode untuk pembobotan dan metode *Cosine Similarity* untuk menghitung jarak dari klasifikasikan teks.
3. Kesalahan pada saat tidak sesuai pengklasifikasi dokumen ialah karena terdapat kata yang sama dan nilai pembobotan kata dari kategori lain lebih besar. Hal ini mengakibatkan klasifikasi teks yang diinput lebih mendekati kategori lain bukan kategori yang dimaksud.

REFERENSI

[1] Abdul Aziz Maarif. " Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah". Skripsi Mahasiswa Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Dian Nurwantoro), Semarang, 2011.

[2] Agustinus Widiantoro. " Peringkasan Teks Otomatis Pada Dokumen Berbahasa Jawa Menggunakan Metode TF-IDF". (Skripsi Mahasiswa Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Sanata Dharma), Yogyakarta, 2014.

[3] Darujati Cahyo, Gumelar Bimo Agustinus. 2012 " Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia". Surabaya

[4] Destuardi I dan Sumpeno Surya. 2009. " Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naïve Bayes". Surabaya

[5] Fitri Meisya. " Perancangan Sistem Temu Balik Informasi Dengan Metode Pembobotan Kombinasi TF-IDF Untuk Pencarian Dokumen Berbahasa Indonesia". Skripsi Mahasiswa Jurusan Teknik Elektro Fakultas Teknik Universitas Tanjung Pura)

[6] Hamzah Amir. 2012. " Klasifikasi Teks Dengan Naïve Bayes Clasifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis". Yogyakarta

[7] Indah Kadek. 2012. " Pengembangan Aplikasi Text Mining Dengaan Metode Associations Rule Analysis Untuk Pencarian Dokumen". Karmapati

[8] Indranandita amalia. 2008. " Sistem Klasifikasi dan Pencarian Jurnal Dengan Menggunakan Metode Naïve Bayes Dan Vector Space Model ". Skripsi mahasiswa jurusan Teknik Informatika.

[9] Raharjo Suwanto, Edi Winarko. 2014. " Klasterisasi, Klasifikasi dan Peringkasan Teks Bahasa Indonesia. Depok

[10] Silfia Andini. 2013. "Klasifikasi Dokument Teks Menggunakan Algoritma Naïve Bayes Dengan Bahasa Pemograman Java. Padang

[11] Sumpeno Surya. 2010. " Text mining For Fuzzy Based Emotion Exspresions". Journal For Technology and Science.

[12] Wijaya Chandra Marvin, Tjiharjadi Semuil. 2010. " Aplikasi Klasifikasi Dokumen Menggunakan Metode Naïve Bayes". Yogyakarta