Machine learning dalam Program Chatting untuk Merespon Emosi Teks Berbahasa Indonesia Menggunakan Text Mining dan Naïve Bayes

Entin Martiana, Rizky Yuniar Hakkun, Nur Rosyid M, Miftahul Firodh
Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh Nopember Surabaya
Kampus PENS-ITS Keputih Sukolilo Surabaya 60111
Telp (+62)31-5947280, 5946114, Fax. (+62)31-5946114
entin@eepis-its.edu ,risky@eepis-its.edu, rosyid@eepis-its.edu, tahool_ff@yahoo.co.id

Abstract

Komunikasi Verbal tidak hanya terjadi secara lisan, tetapi juga melalui tulisan. Program chatting, surat, atau email merupakan contoh media komunikasi berbentuk tulisan yang sering digunakan. Namun bahasa tulisan mempunyai kelemahan dalam menyampaikan faktor emosi kepada penerima pesan. Kekurangan ini bisa diatasi dengan penggunaan emoticon pada email, chatting, dan media elektronik untuk komunikasi bahasa tulisan. Sehingga sebuah teks tidak hanya menyampaikan keterangan dari suatu informasi, tetapi juga berisi informasi tentang perilaku manusia termasuk emosi. Jenis emosi seperti senang, sedih, marah, takut dan sebagainya telah dikenal sejak lama dan menjadi aspek yang penting dari perilaku manusia. Pada penelitian ini dibahas tentang pembuatan model machine learning dalam aplikasi chatting untuk merespon emosi dari kalimat teks berbahasa Indonesia dengan menggunakan algoritma text mining dengan proses naïve bayes untuk mencari keterhubungan kalimat dengan kelas emosi.

Keywords: komunikasi verbal, factor emosi, text mining, naïve bayes.

1. Pendahuluan

Komunikasi Verbal tidak hanya terjadi secara lisan, tetapi juga melalui tulisan. Program chatting, surat, atau email merupakan contoh media komunikasi berbentuk tulisan yang sering digunakan. Komunikasi verbal melalui tulisan menjadi jawaban untuk masalah jarak dengan biaya yang relative rendah. Bahasa tulisan merupakan suatu penerjemahan dari bentuk pikiran yang dituangkan ke dalam tulisan yang melalui proses dalam pembuatan setingkat lebih tinggi dari bahasa lisan. Itu sebabnya banyak dalam presentasi, orang lebih banyak menuangkannya dalam bahasa tulisan, agar dapat mudah dimengerti. Karena orang paham sebesar 30% dari pendengaran dan 70% dari penglihatan. Namun bahasa tulisan mempunyai kelemahan dalam menyampaikan faktor emosi kapada penerima pesan. Kekurangan ini bisa diatasi dengan penggunaan emoticon pada email, chatting, dan media elektronik untuk komunikasi bahasa tulisan. Sehingga sebuah teks tidak hanya menyampaikan keterangan dari suatu informasi, tetapi juga berisi informasi tentang perilaku manusia termasuk emosi. Jenis emosi seperti senang, sedih, marah, takut dan sebagainya telah dikenal sejak lama dan menjadi aspek yang penting dari perilaku manusia. Penerapan emosi belum banyak digunakan dalam interaksi manusia dan komputer, padahal emosi cenderung berperan dalam komunikasi antar manusia di kehidupan sehari-hari. Oleh karena itu dibutuhkan sistem interaksi manusia dan komputer yang baik yang dapat mengenali, menginterpretasikan dan memproses emosi manusia, dalam hal ini emosi yang berasal dari teks.

Penelitian di bidang emosi merupakan sebuah proses yang kompleks karena dapat berubah secara dinamis. Penelitian emosi yang berbasis teks biasanya dilakukan karena bentuk teks relatif lebih sederhana dibandingkan bentuk lain seperti visual atau suara. Penelitian yang telah dilakukan sebagian besar masih menggunakan teks bahasa Inggris, sedangkan untuk teks berbahasa Indonesia masih jarang dilakukan. Destuardi telah melakukan pemodelan Machine Learning menggunakan metode Bayesian untuk klasifikasi teks ini [1]. Pada penelitian ini dilakukan pemodelan metode Bayesian dalam aplikasi chatting untuk merespon emosi dari kalimat teks berbahasa Indonesia dengan menggunakan algoritma text mining dengan proses naïve bayes untuk mencari keterhubungan kalimat dengan kelas emosi.

Data yang digunakan selain diambil dari ISEAR juga dikuatkan dengan survey dari masyarakat Indonesia.

2. Teori

2.1 ISEAR (International Survey On Emotion Antecedents And Reaction)

Pada penelitian ini, data yang diolah diambil dari ISEAR databank dikuatkan dengan survey ke masyarakat Surabaya. Selama bertahun-tahun selama tahun 1990-an, sebuah kelompok besar psikolog di seluruh dunia mengumpulkan data dalam proyek ISEAR, dipimpin oleh Klaus R. Scherer dan Harald Wallbott. Mahasiswa responden, baik psikolog dan non-psikolog, diminta untuk melaporkan situasi di mana mereka mengalami semua 7 emosi utama (senang, takut, marah, sedih, jijik, malu, dan rasa bersalah). Dalam setiap kasus, pertanyaan-pertanyaan meliputi cara mereka telah mengenali situasi dan bagaimana mereka bereaksi. Data akhir ini memuat laporan tentang tujuh emosi masing-masing sekitar 3000 responden di 37 negara di 5 benua. Penggunaan survey dilakukan untuk memastikan karena data ISEAR berasal dari bahasa Inggris yang memungkinkan ekspresi emosinya berbeda dengan bahasa Indonesia.

2.2 Emosi

Emosi adalah pengalaman afektif yang disertai penyesuaian dari dalam diri individu tentang keadaan mental dan fisik dan berwujud suatu tingkah laku yang tampak.

Adapun beberapa fungsi dari emosi adalah sebagai berikut :

1. Survival atau untuk mempertahankan hidup, seperti pada hewan.
2. Energizer atau pembangkit energi yang memberikan kegairahan dalam kehidupan .
3. Messenger atau pembawa pesan (Martin dalam Khodijah, 2006).

Dalam Penelitian ini adapun jenis emosi yang dibahas adalah senang, takut, marah, sedih, jijik, malu, dan bersalah.

2.3 Text Mining

Text mining adalah proses menambang data berupa teks dengan sumber data biasanya dari dokumen dan tujuannya adalah mencari kata - kata yang mewakili dalam dokumen sehingga dapat dilakukan analisa keterhubungan dalam dokumen. Data teks akan diproses menjadi data numerik agar dapat dilakukan proses lebih lanjut. Sehingga dalam *text mining* ada istilah *preprocessing data*, yaitu proses pendahulu yang diterapkan terhadap data teks yang bertujuan untuk menghasilkan data numerik.

Adapun proses preprocessing yang dilakukan antara lain :

- Tokenizing

Teks dalam bentuk mentah mereka, bagaimanapun, hanya rangkaian karakter tanpa informasi eksplisit tentang batas kata dan kalimat. Sebelum diproses lebih lanjut dapat dilakukan, teks perlu tersegmentasi ke dalam kata-kata dan kalimat. Proses ini disebut tokenization. Tokenization membagi urutan karakter menjadi kalimat dan kalimat ke dalam token. Tidak hanya kata-kata dianggap sebagai bukti, tetapi juga angka, tanda baca, tanda kurung dan tanda kutip.

- Filtering

Tahap *filtering* adalah tahap mengambil kata - kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist / stopword* adalah katakata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah “yang”, “dan”, “di”, “dari” dan seterusnya.

- Penanganan kata ‘Tidak’

Kata ‘tidak’ dapat sangat mempengaruhi makna dari sebuah kalimat. Sebuah kalimat yang awalny menunjukkan emosi senang dapat berbuah menjadi sedih ketika disisipi dengan kata tidak. Misalnya kata berhasil menjadi tidak berhasil. Untuk diperlukan penanganan khusus untuk kata ‘tidak’ ini, yaitu dengan mencari antonym dari kata yang muncul setelah ‘tidak’.

- Stemming

Tahap *stemming* adalah tahap mencari *root* kata dari tiap kata hasil *filtering.* Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tahap ini kebanyakan dipakai untuk teks berbahasa inggris dan lebih sulit diterapkan pada teks berbahasa Indonesia. Hal ini dikarenakan bahasa Indonesia tidak memiliki rumus bentuk baku yang permanen.

- Counting

Dimulai dengan perhitungan jumlah kata dalam setiap dokumen, yang kemudian akan dihitung menggunakan skema pembobotan yang dikehendaki.

2.4 Naïve Bayes

Naïve Bayes merupakan salah satu metode machine learning yang menggunakan perhitungan probabilitas. Untuk klasifikasi Bayes sederhana yang lebih dikenal sebagai naïve Bayesian Classifier dapat diasumsikan

bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain. Asumsi ini disebut *class conditional independence* yang dibuat untuk memudahkan perhitungan-perhitungan pengertian ini dianggap "naive", dalam bahasa lebih sederhana naïve itu mengasumsikan bahwa kemunculan suatu term kata dalam suatu kalimat tidak dipengaruhi kemungkinan kata-kata yang lain dalam kalimat padahal dalam kenyataanya bahwa kemungkinan kata dalam kalimat sangat dipengaruhi kemungkinan keberadaan kata-kata yang dalam kalimat. Naive Bayes classifier dapat dilatih untuk mengklasifikasi pola-pola yang melibatkan ribuan atribut dan diterapkan untuk ribuan pola. Akibatnya, Naive Bayes merupakan algoritma yang banyak digunakan untuk text mining dan masalah lain klasifikasi besar. Teori dari *Naïve Bayes* sendiri adalah sebagai berikut:

$$P(A|B) = (P(B|A) * P(A))/P(B) \quad (1)$$

Rumus di atas dapat dibaca sebagai peluang kejadian A sebagai B ditentukan dari peluang B saat A, Peluang A, dan Peluang B. Agar lebih jelas, dalam kasus klasifikasi emosi dari teks maka rumus di atas akan dirubah menjadi:

$$P(Ki|T) = \Sigma \log (P(T|Ki) + \log (P(Ki)) / P(T)) \quad (2)$$

Dimana P(Ki|T) adalah peluang dokumen teks T pada Kategori Ki. Untuk mewakili kata yang tidak terdapat dalam suatu kelas maka diberikan nilai yang kecil yaitu 10^{-13}.

3 Desain Sistem

Sebelum pembuatan akan dilakukan perencanaan untuk alur pada sistem yang terdiri dari Deskripsi Umum, Text Mining, Naïve Bayes, Rancangan Proses.

3.1 Deskripsi Umum Sistem

Pada aplikasi chatting server client ini, proses text mining dibagi menjadi 2 jenis, yaitu proses pertama adalah text mining untuk mencari kata kunci dari tiap kelas emosi dengan inputan data ISEAR, dan proses kedua adalah text mining untuk mengolah teks inputan(obrolan) dari client untuk dikirimkan hasil text mining nya ke client lain. Proses text mining yang pertama terjadi sekali hanya saat aplikasi dijalankan di sisi server. Sedangkan proses text mining yang kedua dilakukan setiap kali client I mengirimkan teks obrolan ke client II melewati sisi server. Proses text mining selalu terjadi di sisi server.

3.1 Proses Text Mining

Proses text mining digunakan untuk mencari kata kunci/keyword. Proses ini sangat dibutuhkan untuk mencari perwakilan dari sebuah dokumen. Untuk mendapatkan kata kunci dari sebuah dokumen, maka harus melewati tahapan – tahapan proses dibawah ini. Khusus untuk dokumen berbahasa Indonesia tahapan prosesnya adalah Tokenizing, Filtering, Stemming, dan Counting.

3.2.1 Tokenizing

Proses tokenizing dilakukan dengan tahapan algoritma sebagai berikut :

- Membuat object StringTokenizer dengan alias st, untuk melakukan proses tokenizing. Karena sebenarnya source untuk proses ini sudah ada pada class di java dengan nama StringTokenizer. Deklarasi object tersebut disertakan dengan pattern – pattern sebagai penanda token. Misalnya : – …&?/.,;:\'\"“”-() \r\n, dsb.
- Masukkan kedalam hasil token kedalam variable temp dengan metod nextToken() dari object st.
- Selama kondisi st.hasMoreTokens true, dalam artian text masih ada yang di token. Maka kembali ke tahap 2.

3.2.2 Filtering

Tahap *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 lebih jelasnya, berikut adalah algoritma proses Filtering :

- Insialisasi flag=0, inisialisasi content_filter = "", string_input=hasil token.
- Lakukan proses Loop dengan kondisi while(!feof(string_input)), kata_string=fgets(string_input)
- Lakukan nested Loop dengan kondisi while(!feof(stopdic))
- Cek jika string_input==stopdic, maka flag="sama". Keluar dari loop stopdic. Jika tidak, maka flag="beda". Ulangi langkah ini sampai kondisi pada langkah nomor 3 false.
- Setelah keluar Loop stopdic, jika flag="beda", content_filter=kata_string. Ulangi langkah 3-5 sampai kondisi pada langkah nomor 2 false.
- content_filter adalah hasil dari proses filtering.

3.2.2.1 Penanganan Kata ‘Tidak’

Pada tahap filtering, ketika ditemukan kata ‘tidak’, maka akan dicari antonim kata dari kata setelah kata ‘tidak’.

3.2.3 Stemming

Berikut merupakan algoritma dari proses stemming :

- Masukkan hasil filter kedalam arrayWord.
- Inisialisasi word=arrayWord[i].
- Cek awalan dengan memasukkan ke dalam fungsi cekAwalan(word). Kemudian cek resultWord.
- Apabila resultWord is empty maka word = ArrayWord[i], jika tidak maka word = resultWord().
- Cek akhiran dengan memasukkan ke dalam fungsi cek Akhiran(word). Kemudan cek resultWord.
- Apabila resultWord is empty maka word = ArrayWord[i], jika tidak maka word = resultWord().
- Cek KPTS dengan memasukkan ke dalam fungsi cek KPTS(word). Kemudan cek resultWord.
- Apabila resultWord is empty maka word = ArrayWord[i], jika tidak maka word = resultWord().

3.2.4 Counting

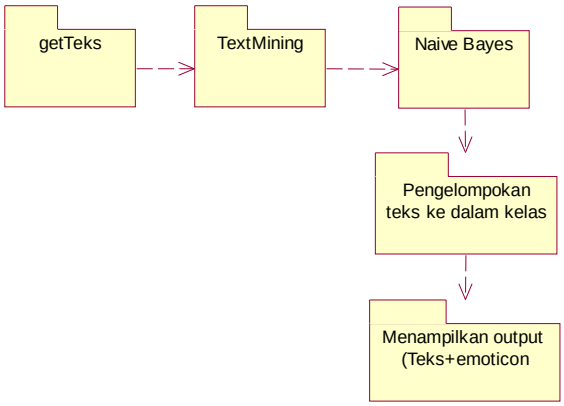
Proses counting atau penghitungan kata kunci dilakukan pada proses preprocessing. Proses ini ditujukan untuk melakukan proses korelasi, dimana korelasi tersebut membandingkan nilai kata kunci antar kedua dokumen. Dokumen yang dibandingkan adalah dokumen berita dan dokumen kamus kategori. Nilai dari korelasi tersebut menunjukkan seberapa sering sebuah kata keluar di kelas tertentu.

3.3 Proses Text Mining

Algoritma Naïve Bayes digunakan untuk mencari probabilitas tertinggi suatu teks termasuk kelas yang mana. Algoritma Naïve Bayes diawali terlebih dahulu oleh proses text mining namun proses counting digunakan untuk mencari probabilitas kata dalam tiap kelas.

3.4 Rancangan Proses Sistem

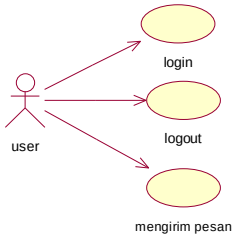
Use Case Utama di atas menggambarkan fungsionalitas yang diharapkan dari pembuatan aplikasi ini. Adapun yang ditekankan pada pembuatan aplikasi adalah apakah system dapat mengenali emosi yang terdapat dalam sebuah teks.



Gambar 1. Diagram Sistem

3.4.1 Use Case Diagram Utama Sistem

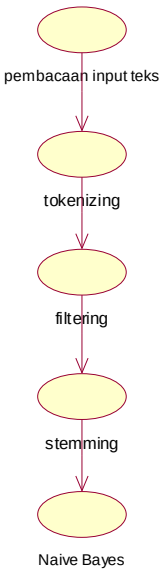
Perlakuan user yang bisa dilakukan kepada ystem antara lain adalah login, logout, dan mengirim pesan. User sebagai actor dapat melakukan aksi-aksi misalnya login, logout dan mengirim pesan. User melakukan login ketika pertama kali membuka aplikasi. User harus memasukkan id dan password yang telah didaftarkan. User melakukan logout ketika ingin mberhenti melakukan chatting dan mengakhiri koneksi dengan server. Mengirim pesan merupakan aksi utama karena merupakan fungsional dari aplikasi ini.



Gambar 2. Diagram Use Case

3.4.2 Use Case Diagram Text Mining

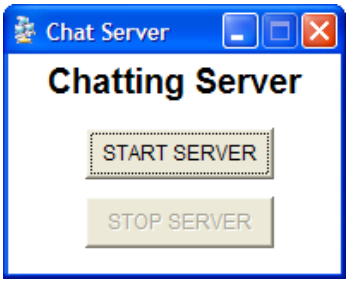
Pertama-tama input text (obrolan) dari user diterima oleh server untuk dijadikan input dalam proses text mining. Selanjutnya text inputan mengalami proses tokenizing, filtering, dan stemming untuk mendapatkan vektor kata-kata pembentuk text. Selanjutnya vector kata hasil proses sebelumnya diproses dengan algoritma naïve bayes untuk mendapatkan klasifikasi terhadap text.



Gambar 3. Diagram Use Case Text Mining

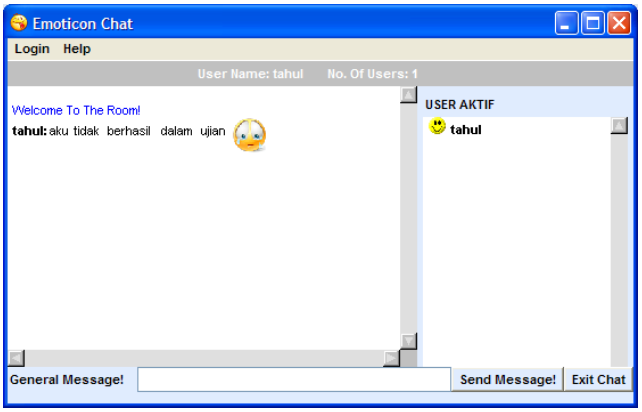
4. Uji Coba dan Analisa

Aplikasi ini dibagi menjadi dua bagian yaitu aplikasi server dan aplikasi client chatting. Hasil akhir dari tujuan penelitian ini adalah bagaimana mengenali emosi yang terdapat pada percakapan chatting antara client. Pada aplikasi server, tugas server hanya menekan tombol 'start' untuk menjalankan server socket dan melakukan proses preprocessing text mining. Pada aplikasi chatting client ('Emoticon Chat'), user harus melakukan login untuk dapat melakukan chatting dengan user lain yang sedang aktif.



Gambar 4. Aplikasi Desktop Server

Ketika client mengirim pesan, sebelum pesan dikirim ke client lain, pesan akan diproses dulu oleh server untuk menentukan klasifikasi teks. Kemudian teks pesan ditambahkan dengan emoticon sesuai hasil klasifikasi teks berdasarkan emosi nya seperti yang ditunjukkan pada gambar 4.2.



Gambar 4.2 Halaman Utama dengan output pesan

4.1 Uji Coba Text Mining

Proses Text Mining disini dilakukan dua kali, yaitu pada proses preprocessing dan proses klasifikasi teks. Proses preprocessing digunakan untuk menambang kata dari database disertai nilai kemunculan kata pada kelas. Kelas yang digunakan antara lain : senang, takut, marah, sedih, jijik, malu, dan bersalah. Yang mana pada setiap kelas memiliki kamus sebagai nilai bobot yang akan digunakan pada perhitungan probabilitas kelas. Berikut merupakan contoh sebagian kamus kelas yang tersedia.

Tabel 1. Tabel Kamus Kelas

Senang			
KATA	NILAI	KATA	NILAI
periode	1	musim	4
jatuh	4	panas	4
cinta	11	harap	6
temu	15	hari	31
utama	3	rasa	50
pisah	14	pasang	4
terima	37	teman	61
surat	19	dama	2
tawarkat	1	riku	1
kerja	14	kalam	4

4.2 Percobaan Bayesian

Berikut akan dilakukan percobaan terhadap satu contoh kalimat input.

Kalimat	Kelas	Probabilitas
Ketika baju dan barang - barang saya dicuri dari lemari saya	Senang	-107.6552
Ketika baju dan barang - barang saya dicuri dari lemari saya	Takut	-59.19255
Ketika baju dan barang - barang saya dicuri dari lemari saya	Marah	-35.75863
Ketika baju dan barang - barang saya dicuri dari lemari saya	Sedih	-61.10425
Ketika baju dan barang - barang saya dicuri dari lemari saya	Jijik	-62.49339
Ketika baju dan barang - barang saya dicuri dari lemari saya	Malu	-59.87458
Ketika baju dan barang - barang saya dicuri dari lemari saya	Bersalah	-57.11864

Probabilitas menunjukkan angka minus dikarenakan hasil penjumlahan dari hasil logaritmik probabilitasi tiap kata dalam kelas yang bersangkutan. Kalimat "Ketika baju dan barang-barang saya dicuri dari lemari saya" menunjukkan kelas yang didapat adalah Marah.

4.3 Analisa Sistem

Dari beberapa percobaan untuk proses di atas, faktor – faktor yang mempengaruhi tingkat ketepatan dalam proses klasifikasi antara lain : kamus kata kunci dan nilai dari kata kunci. Jadi semakin banyak jumlah kata kunci dalam kamus kelas, akan meminimalisir persentase error yang akan muncul. Rata – rata error yang timbul pada percobaan – percobaan yang dilakukan adalah bernilai 20 s.d 30%.

5 Kesimpulan dan Saran

Berdasarkan analisa dari beberapa pengujian yang diterangkan pada bab sebelumnya, kesimpulan yang didapatkan adalah :

1. Semakin banyak jumlah kata kunci dalam kamus kelas, akan meminimalisir persentase error yang akan muncul.
2. Hasil dari ujicoba terhadap data traing didapatkan nilai error sebesar 18%. Nilai error besar karena terdapat missing value serta pengkategorian dari data learning yang sedikit.
3. Proses klasifikasi text mining sangat bergantung kepada banyaknya jumlah dan variasi data training yang digunakan.
4. Survei merupakan hal yang dibutuhkan untuk data training karena emosi juga berhubungan dengan budaya.

Adapun saran untuk pengembangan ke depan adalah sebagai berikut:

1. Pada proses pengkategorian berita, yakni masih terdapat beberapa error yang mungkin terjadi. Diharapkan untuk kedepannya agar kamus di dalam kategori dapat di tambah.
2. Dimaksudkan untuk meminimalisir jumlah error.Untuk klasifikasi text pada tahap analisa dapat di coba dengan menggunakan metode yang lain. Agar dapat dibandingkan seberapa besar tingkat ketepatan.
3. Sistem ini diharapkan dapat dikembangkan untuk aplikasi-aplikasi berbasis suara.

Daftar Pustaka

[1] Destuardi, Surya Sumpeno, Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes, 2009.

[2] Surya Sumpeno, Dwi Cahyono , Junaidillah Fadlil, Mochammad Hariadi, Agen Percakapan Berbasis Pengetahuan Teks Berbahasa Indonesia, 2009.

[3] http://emotion-research.net/toolbox/ waktu akses : 14 Juli 2010, 17:34

[4] Moh Badrullami, Rancang Bangun Aplikasi Server Crawling Berita Online Sebagai Penyedia Berita Up To Date Pada Handphone Yang Mendukung WAP, 2010.

[5] Endhy Pitoyo (2010), Klasifikasi Dokumen Dampak Lumpur Sidoarjo Menggunakan Metode Bayesian.