# KLASIFIKASI EMOSI UNTUK TEKS BERBAHASA INDONESIA DENGAN MENGGUNAKAN *K-NEAREST NEIGHBOR*

Lailatus Sofiyana[1], Zainal Abidin[2], Hani Nurhayati[3]

Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang[1,2,3]
Jl. Gajayana 50 Malang 65144, Indonesia
Email: Sofy.ana.165@gmail.com[1], zainal@ti.uin-malang.ac.id[2], hani.hayati@gmail.com[3]

**ABSTRAK**

Penggunaan emosi yang tepat dalam situasi yang tepat dapat mempengaruhi terhadap hasil dari aktifitas yang dilakukan oleh manusia. Maka dari itu perlu adanya pengklasifikasian emosi guna membantu seseorang mengatur dan mengendalikan emosi pada dirinya. Emosi dapat diklasifikasikan berdasarkan sifatnya yaiu emosi positif dan negatif. Dalam penelitian ini pengklasifikasian emosi diimplementasikan pada lirik lagu yang berbahasa Indonesia, karena lagu mengandung emosi dan merupakan ekspresi hati. Salah satu pemanfaatan aplikasi klasifikasi emosi ini yaitu dalam dunia penyiaran yang berhubungan dengan program acara musik yang perlu adanya aplikasi penunjang dalam pemilihan lagu untuk penyajiannya yang berhubungan dengan karakter emosi dari segmentasi pendengar. Proses pengklasifikasiannya menggunakan data latih yang telah diketahui kelas emosinya yaitu senang, sedih, marah, bersalah dan takut dengan menggunakan metode *text mining* serta menggunakan algoritma *K-Nearest Neighbor*. Sistem yang dikembangkan telah berhasil melakukan pengklasifikasian teks berbahasa Indonesia sesuai dengan kategori emosi dengan prosentase tertinggi dengan nilai 60 % pada k=5.

Kata kunci: *emosi, text mining, k-nearet neighbor*

## 1. PENDAHULUAN

Bahasa merupakan alat komunikasi yang digunakan oleh manusia. Bahasa mempunyai dua bentuk, yaitu bahasa lisan dan bahasa tulis. Bahasa lisan berurusan dengan tata bahasa, kosakata, dan lafal. Bahasa tulis merupakan suatu alat untuk menerjemahkan pikiran manusia. Dalam penelitiaanya Firodh mengatakan sebuah tulisan tidak hanya menyampaikan keterangan dari suatu informasi, tetapi juga berisi informasi tentang perilaku manusia termasuk emosi. Sedangkan emosi memberikan informasi penting mengenai pemahaman terhadap dunia sekitar.

Emosi dapat diklasifikasikan berdasarkan sifatnya yaiu emosi positif dan negatif. Emosi-emosi positif seperti rasa gembira dan rasa syukur mengekspresikan sebuah evaluasi atau perasaan menguntungkan, sedangkan emosi-emosi negatif seperti rasa marah atau rasa bersalah mengekspresikan sebaliknya (Hude, 2006).

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 (Firodh, 2011).

Metode pengklasifikasian salah satunya yaitu *K-nearest neighbor* (KNN). KNN merupakan metode klasifikasi yang dapat memprediksikan suatu kelas berdasarkan titik terdekat dari suatu dokumen latih. Untuk cara kerjanya sendiri dipilih dokumen latih yang memiliki kategori terdekat dengan dokumen uji, sehingga akan diperoleh dokumen uji tersebut akan masuk dalam satu kategori.

## 2. KAJIAN PUSTAKA

### 2.1 Klasifikasi

Klasifikasi adalah sebuah proses untuk menemukan model yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui. Di dalam klasifikasi diberikan sejumlah *record* yang dinamakan *training set*, yang terdiri dari beberapa atribut, atribut dapat berupa kontinyu ataupun kategoris, salah satu atribut menunjukkan kelas untuk *record*.

### 2.2 Emosi

Emosi merupakan suatu aspek psikis yang berkaitan dengan perasaan dan merasakan. Misalnya, merasa senang, sedih, kesal, marah, tegang, dan lain sebagainya. Emosi pada diri seseorang erat kaitannya dengan suatu keadaan psikis tertentu yang distimulasi, baik oleh faktor *internal* (dari dalam) maupun *eksternal* (dari luar) (Al-firdaus, 2011).

### 2.3 *Text mining*

*Text Mining* adalah subjek riset yang tergolong baru pada bidang data mining. *Text Mining* dapat memberikan solusi baru dalam hal pemrosesan, pengelompokan atau pengorganisasian dan analisis teks dalam jumlah yang banyak atau besar. Pada umumnya permasalahan yang terdapat dalam *text mining* adalah jumlah data yang besar, *high dimensional*, struktur yang berubah-ubah, ambigu, *dependency* dan data *nois*.

Pada *text mining* tedapat dua teknik pembelajaran, *unsupervised learning* dan *supervised learning*. *Clustering* adalah contoh dari *unsupervised learning*, yaitu sekelompok data kemudian langsung dikelompokkan berdasarkan tingkat kemiripan data tanpa dilakukan supervisi. Sedangkan klasifikasi merupakan bentuk dari *supervised learning* yang merupakan salah satu teknik dalam pembelajaran mesin untuk membentuk model yang merupakan fungsi dari data latihan (*trainingset*). Pada *supervised learning*, data latih yang digunakan terdiri dari beberapa pasangan nilai-nilai masukan dan nilai keluaran (nilai dari atribut tujuan). Model yang terbentuk dari data latih digunakan sebagai dasar pengetahuan untuk mengklasifikasikan data-data yang baru (*testing set*) (Even.Yahir dan Zohar, 2002).

#### 2.3.1 *Case Folding*

*Case folding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.

#### 2.3.2 *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.

#### 2.3.3 *Filtering*

Tahap *filtering* adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* atau *stopwords* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting).

*Stopwords* didefinisikan sebagai hal yang tidak relevan sehubungan dengan subjek utama dari database, meskipun mereka mungkin sering terjadi dalam dokumen. Mereka termasuk penentu, konjungsi, preposisi, dan sejenisnya (Cios, 2007).

#### 2.3.4 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.

#### 2.3.5 Pembobotan TFIDF

Pada tahap ini dilakukan proses perhitungan bobot (w) dokumen agar diketahui seberapa jauh tingkat similaritas antara kata kunci yang dimasukkan dengan dokumen. Rumus dari algoritma TF/IDF bisa dilihat pada **persamaan 1** dan **persamaan 2**.

)

$$IDF = \log\left(\frac{D}{df}\right) \quad W_{d,t} = tf_{d,t}.IDF_t \quad (1)$$

(2)

dimana:

d : dokumen ke-d

t : kata ke-t dari kata kunci

W : bobot dokumen ke-d terhadap kata ke-t

D : total dokumen

df : banyak dokumen yang mengandung kata yang dicari

tf : banyak kata yang dicari pada sebuah dokumen

### 2.4 *K-Nearest Neighbor*

*K-Nearest Neighbor* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data *learning* yang jaraknya paling dekat dekat objek tersebut. KNN termasuk algoritma *supervised learning* dimana *query instance* yang baru diklasifikan berdasarkan mayoritas dari kategori pada KNN. Kelas yang paling banyak muncullah yang akan menjadi kelas hasil klasifikasi (Nasution, 2011).

Selanjutnya langkah-langkah penerpan metode ini adalah sebagai berikut (Ridok):

1. Membuat dokumen X dari semua sampel latih menjadi bentuk vektor fitur yang sama ($X_1$, $X_2$, ..., $X_m$).

2. Menghitung kesamaan antara semua sampel latih dan dokumen X. Mengambil dokumen ke-i $d_i$ ($d_{i1}$, $d_{i2}$, ..., $d_{im}$) sebagai contoh, kesamaan SIM (X, $d_i$) adalah sebagai berikut:

$$sim(X, d_i) = \frac{\sum_{j=1}^{m} X_j.d_y}{\sqrt{(\sum_{j=1}^{m} X_j)^2}.\sqrt{(\sum_{j=1}^{m} d_y)^2}} \quad (3)$$

3. Memilih k sampel yang lebih besar dari kesamaan N dari SIM (X, $d_i$), (i=1,2,..., N). Dan memperlakukan mereka sebagai kumpulan K-NN dari

X. Kemudian menghitung probabilitas X ke masing-masing kategori menggunakan rumus berikut:

$$P(X,C_j) = \sum_{d_i \in KNN} SIM(X,d_i).y(d_i,C_j) \quad (4)$$

Dimana, $y(d_i, C_j)$ adalah fungsi atribut kategori yang memenuhi persamaan berikut:

$$y(d_i,C_j) = \begin{cases} 1, d_i \in C_j \\ 0, d_i \notin C_j \end{cases} \quad (5)$$

Uji dokumen X untuk mengetahui kategorinya dengan melihat $P(X, C_j)$ terbesar.

## 3. Desain Sistem

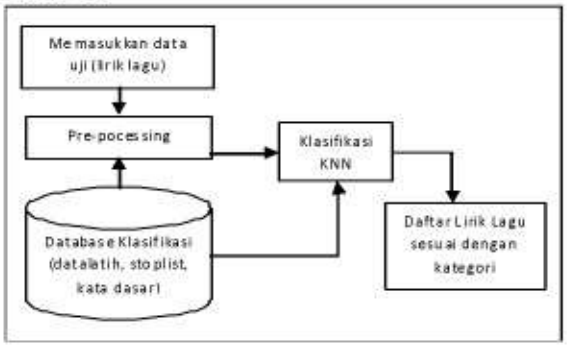
### 3.1 Desain Umum Sistem

Tahapan dalam klasifikasi ini bisa dilihat pada **gambar 1** yaitu:

1. memasukkan data latih yang diambil dari *International Survey On Emotion Antecedents And Reaction* (ISEAR) yang berbahasa Inggris kemudian diterjemahkan ke dalam bahasa Indonesia tanpa mengurangi maksud dari kalimat-kalimat dalam ISEAR.
2. memasukkan data uji dimana data ini berasal dari lirik lagu yang akan diketahui kategori emosinya.
3. *preprocessing* data latih dan data uji, tahapannya yaitu *case folding, tokenizing, filtering* dan *stemming*. *Case folding* yaitu mengubah huruf menjadi huruf kecil semua. *Filtering* yaitu menghilangkan kata-kata yang dianggap tidak penting yaitu kata-kata yang masuk dalam daftar *stopword*. *Stemming* yaitu mengembalikan kata berimbuhan ke dalam bentuk dasar sebuah kata dan menganalisa bobot *term frequency/inseversed document frequency* (TF/IDF) dan membentuk ruang ector.
4. klasifikasi emosi dengan menggunakan *k-nearest neighbor*.

Pada sistem ini hanya terdapat satu aktor yaitu *user*. Ketika pertama kali menjalankan sistem, *user* dapat melakukan klasifikasi lirik lagu dengan memasukkan lirik lagu yang belum diketahui kategorinya dan sistem akan memproses lirik lagu untuk mendapatkan hasil kategori emosinya. *User* juga dapat melihat hasil dari klasifikasi tersebut.

*Use case diagram* dari aplikasi ini dapat dilihat pada **gambar 2.**




Gambar 1 **Alur Pembuatan Sistem Klasifikasi Emosi Lirik Lagu**


Gambar 2 **Diagram *Use Case* Klasifikasi Emosi Lirik Lagu**

### 3.2 Desain Database

Database digunakan untuk menyimpan data-data yang diperlukan selama proses klasifikasi emosi. Tabel kategori emosi digunakan untuk menyimpan kategori emosi. Struktur tabelnya ditampilkan pada **tabel 1.** Tabel data latih digunakan untuk menyimpan data yang sudah ada kategori emosinya, struktur tabelnya bisa dilihat pada **tabel 2.** Kemudian tabel lagu digunakan untuk menyimpan informasi lagu dan *path file* lagu, strukturnya ditampilkan pada **tabel 3.** Tabel *stopword* digunakan untuk menyimpan kata-kata yang dianggap kurang penting. Struktur tabel *stopword* dapat dilihat pada **tabel 4.** Tabel Kata Dasar digunakan untuk menyimpan kata dasar bahasa Indonesia untuk membantu proses *stemming*. Struktur tabel kata dasar dapat dilihat pada **Tabel 5.** Sedangkan tabel klasifikasiemosi digunakan untuk menyimpan hasil pengklasifikasian yang dilakukan system, struktur tabelnya ditampilkan pada **tabel 6.**

**Tabel 1** Struktur Tabel Kategori Emosi

| Nama Field | Tipe Data |
|---|---|
| id_kategori | int(11) |
| kategori | varchar(20) |

**Tabel 2** Struktur Tabel Data Latih

| Nama Field | Tipe Data |
|---|---|
| id_latih | int(11) |
| dokumen | text |
| id_kategori | varchar(5) |

**Tabel 3** Struktur Tabel Lagu

| Nama Field | Tipe Data |
|---|---|
| id_lagu | int(11) |
| judul_lagu | varchar(50) |
| penyanyi | varchar(50) |
| lirik_lagu | text |

**Tabel 4** Struktur Tabel *Stopword*

| Nama Field | Tipe Data |
|---|---|
| id_stopword | int(11) |
| Kata | varchar(20) |

**Tabel 5** Struktur Tabel Kamus Kata Dasar

| Nama Field | Tipe Data |
|---|---|
| id_katadasar | int(11) |
| kataDasar | varchar(20) |

**Tabel 6** Struktur Tabel Klasifikasi Emosi

| Nama Field | Tipe Data |
|---|---|
| Id_klas | int(11) |
| id_lagu | varchar(5) |
| id_kategori | varchar(5) |

## 1. IMPLEMENTASI

Implentasi di dalam penelitian ini berdasarkan pada desain penelitian yang telah dibuat sebelumnya. Tampilan antar muka ditampilkan pada **gambar 3** . Lagu dimasukkan pada kolom lagu dengan mengklik tombol *browse* akan mucum kotak dialog tempat penyimpanan lagui. Kemudian Lirik lagu dimasukkan dengan mengklik tombol *browse* atau menuliskan langsung pada *text area*. Untuk mengetahui hasil klasifikasi lagu tersebut, maka dengan mengklik tombol analisa. Di dalam *event* klik tombol analisa dilakukan *preprocessing* lagu kemudian dilanjutkan dengan proses pengklasifikasian dengan mencari nilai terdekat antara data latih dengan data uji dengan *k-nearest neighbor*. Hasil pengklasifikasian ditampilkan pada form di bawah lirik lagu.


Gambar 3 **Tampilan Antar Muka Aplikasi Klasifikasi Emosi**

## 2. Hasil dan Pembahasan

Data yang digunakan dalam uji coba ini ada dua yaitu data latih dan data uji. Data latih mengambil dari *International Survey On Emotion Antecedents And Reaction* (ISEAR) yang berjumlah 1000 kalimat dengan masing-masing kategori berjumlah 200 kalimat sedangkan data ujinya yaitu berupa lirik lagu yang berjumlah 30.

Di dalam penelitian ini pengujian keberhasilan dengan membandingkan dengan penentuan kategori emosi lagu yang dilakukan oleh para penyiar radio Mass FM. Sehingga relevan tidaknya suatu jawaban di tentukan oleh hasil penentuan kategori emosi tersebut. Pengukuran keberhasilan dilakukan dengan menghitung prosentase keberhasilan tiap pengujian dengan rumus dalam **persamaan 6**. Untuk hasil semua uji coba ditampilkan pada **tabel 8**.

$$Prosentase = \frac{jumlah\ yg\ benar}{jumlah\ keseluruhan} \times 100\% \quad (6)$$

Hasil prosentasenya ditampilkan pada **tabel 7**.

Tabel 7 **Prosentase Keberhasilan Aplikasi**

| No. | Nilai k | Prosentase |
|---|---|---|
| 1. | 2 | 53.33 % |
| 2. | 3 | 50 % |
| 3. | 4 | 56.67 % |
| 4. | 5 | 60 % |

Tabel 8 **Pengujian Aplikasi**

| No. | Judul Lagu | Kandungan Emosi | Kategori Emosi Hasil dari Aplikasi dengan nilai k | | | |
|---|---|---|---|---|---|---|
| | | | k=2 | k=3 | k=4 | k=5 |
| 1. | Saat Bahagia-Ungu | Senang | Bersalah | Bersalah | Bersalah | Senang |
| 2. | Kesedihanku-Sammy | Sedih | Sedih | Sedih | Sedih | Sedih |
| 3. | Jancok-Sujiwo Tejo | Marah | Senang | Marah | Marah | Marah |
| 4. | Takut-Vierra | Takut | Marah | Marah | Marah | Marah |
| 5. | Surti Tejo-Jamrud | Marah | Marah | Marah | Marah | Marah |
| 6. | Surat Cinta-Vinna | Senang | Bersalah | Bersalah | Bersalah | Bersalah |
| 7. | Emang Dasar-Wali | Marah | Marah | Marah | Marah | Marah |
| 8. | Saat Kau Pergi-Bunga Citra Lestari | Sedih | Sedih | Sedih | Sedih | Sedih |
| 9. | Kesempatan Kedua-Tangga | Bersalah | Bersalah | Bersalah | Bersalah | Bersalah |
| 10. | Hip hip hura-Ruben | Senang | Senang | Senang | Senang | Senang |
| 11. | Bila Waktu Telah Berakhir-Opick | Sedih | Senang | Senang | Sedih | Senang |
| 12. | Kemesraan-Iwan Fals | Senang | Sedih | Senang | Sedih | Senang |
| 13. | Salahkah-Tompi | Bersalah | Sedih | Sedih | Sedih | Sedih |
| 14. | Cinta Mati-Agnes feat Ahmad Dhani | Sedih | Sedih | Sedih | Sedih | Sedih |
| 15. | Taubat-Opick | Bersalah | Bersalah | Bersalah | Bersalah | Bersalah |
| 16. | Rasa kehilangan-Letto | Sedih | Sedih | Sedih | Sedih | Sedih |
| 17. | Pernah Muda-BCL | Senang | Senang | Sedih | Sedih | Sedih |
| 18. | Karma-Cokelat | Marah | Marah | Marah | Marah | Marah |
| 19. | Semua Tentang Kita-Peterpan | Sedih | Senang | Senang | Senang | Senang |
| 20. | Maafkan Aku Mencintai Kekasihmu-Rebecca | Bersalah | Sedih | Sedih | Sedih | Sedih |
| 21. | Saat Terakhir-ST 12 | Sedih | Marah | Marah | Marah | Marah |
| 22. | Arti Sahabat-Nidji | Senang | Sedih | Sedih | Sedih | Sedih |
| 23. | Mata Keranjang-Aura Kasih | Marah | Marah | Marah | Marah | Marah |
| 24. | Keterlaluan-The Potters | Marah | Marah | Marah | Marah | Marah |
| 25. | Hidup Hanya Sementara-Ungu | Takut | Takut | Takut | Takut | Takut |
| 26. | Ku Bersalah-Angkasa | Bersalah | Bersalah | Sedih | Sedih | Sedih |
| 27. | Tak Merasa Bersalah-Five Minutes | Bersalah | Sedih | Sedih | Sedih | Sedih |
| 28. | Takut Jatuh Cinta-Blink | Takut | Takut | Marah | Takut | Takut |
| 29. | Takut-Anggun | Takut | Senang | Takut | Takut | Takut |
| 30. | Takut Patah Hati-Melly | Takut | Takut | Sedih | Sedih | Sedih |
| | | | | | | |
| **Jumlah Yang Benar** | | | **16** | **15** | **17** | **18** |
| **Jumlah Yang Salah** | | | **14** | **15** | **13** | **12** |

Prosentase tertinggi terdapat pada k=5 dengan nilai 60 %, sedangkan prosentase terendah pada k=3 dengan nilai 50%.

## 3. KESIMPULAN DAN SARAN

### 3.1 Kesimpulan

Berdasarkan uji coba yang telah dilakukan pada aplikasi klasifikasi emosi pada teks lirik lagu berbahasa Indonesia ini maka dapat disimpulkan bahwa keberhasilan sistem mengklasifikasi dokumen tergantung pada penetapan nilai k dalam *k-nearest neighbor*, sedangkan untuk prosentase keberhasilan aplikasi dalam mengklasifikasikan lagu sebanyak 30, lagu yang benar berjumlah 18 dengan nilai k=5 dengan prosentase nilai 60 % dan nilai prosentase terendah pada nilai k=3, jumlah lagu yang benar 15

dan nilai prosentasenya adalah 50 %.

### 3.2 Saran

Saran untuk penelitian selanjutnya untuk menyempurnakan aplikasi klasifikasi emosi agar informasi yang ditemukan lebih relevan, hendaknya dilakukan perbaikan *rule* atau bisa diimplementasikan pada kasus yang berbeda dan lebih modern lagi seperti membahas bahasa simbol emosi atau *emoticon*.

## 4. DAFTAR PUSTAKA


Al-firdaus, Iqra'. 2011. *Dampak Hebat Emosi Bagi Kesehatan*. Yogyakarta: Flashbook.

Cios, Krzysztof J. 2007. *Data Mining A Knowledge Discovery Approach*. Springer.

Even, Yahir dan Zohar. 2002. *Introduction to Text Mining*. Automated Learning Group National Center For Supercomputing Aplications. University of Illionis. http://algdocs.ncsa.uiuc.edu/PR200211162.ppt. Diakses tanggal 17 Juli 2012.

Firodh Fatroni, Miftahul. 2011. *Kecerdasan Buatan Dalam Program Chatting Untuk Merespon Emosi Dari Teks Berbahasa Indonesia Menggunakan Teks Mining Dan Naïve Bayes*. Surabaya: PENS-ITS.

Hude, M. Darwis. 2006. *Emosi Penjelajahan Religio-Psikologis tentang Emosi Manusia dalam Al-Qur'an*. Jakarta: Erlangga.

Nasution, Zulhamsyah Fachrurrazi. 2011. *Penerapan Algoritma Klasifikasi K-Nearest Neighbor Pada Sistem Context Aware File Sharing Berbasis Web Service*. Surabaya: ITS.

Ridok, Achmad dan Furqon, Muhammad Tanzil. *Pengelompokan Dokumen Berbahasa Indonesia Menggunakan Metode K-NN*. Malang: Universitas Brawijaya.

Tala, Fadillah Z. 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Master of Logic Project Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands.