Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)
Edisi. .. Volume. ... Bulan 20.. ISSN : 2089-9033

KLASIFIKASI EMOSI PADA TEKS BAHASA INDONESIA MENGGUNAKAN METODE K-NEAREST NEIGHBOR DENGAN PEMBOBOTAN WIDF

Wildan Abdul Gani

Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
wildan12gani@gmail.com

ABSTRAK

Emosi merupakan sebuah ekspresi dari suatu keadaaan yang berkaitan dengan perasaan seseorang yang dipengaruhi berbagai faktor baik *internal* maupun *eksternal*. Dalam mengekpresikan sebuah emosi seseorang biasanya menggunakan berbagai cara salah satunya dalam sebuah tulisan. Sebuah tulisan yang mengandung emosi sangat sulit untuk ditafsirkan oleh pembaca karena dipengaruhi oleh sudut pandang yang berbeda antara penulis dan pembaca. Maka dari itu untuk memudahkan pembaca mengetahui emosi dari sebuah tulisan bisa menggunakan cara klasifikasi.

Pada penelitian ini metode yang digunakan dalam mengklasifikasikan emosi pada teks yaitu *K-Nearest Neighbor*, pemilihan metode ini didasarkan pada penelitian yang dilakukan yang menjelaskan bahwa metode ini memiliki kinerja yang baik dengan hasil akurasi sampai 86%. Pada penelitian sebelumnya yang telah dilakukan dengan tahapan proses dari mulai *preprocessing*, pembobotan TF.IDF dan pengklasifikasian hasil akurasi yang didapat masih kurang sekitar 60%, hal ini dipengaruhi oleh beberapa kelemahan salah satunya dalam metode pembobotan yang digunakan. Oleh karena itu dalam penelitian ini metode pembobotan yang akan digunakan adalah WIDF yang memiliki kinerja yang baik dan pengembangan dari TF.IDF.

Berdasarkan hasil pengujian dari penelitian ini yang mengimplementasikan lirik lagu bahasa Indonesia dalam mengklasifikasikan emosi dengan metode *K-Nearest Neighbor* dan pembobotan WIDF ke dalam lima kategori yaitu senang, takut, marah, sedih dan bersalah, menghasilkan akurasi sebesar 73.3% pada nilai K = 5. Untuk nilai relevansi dari setiap kategori emosi dengan menggunakan *precision* dan *recall* menghasilkan nilai rata-rata terbaik pada kategori "Takut" untuk *precision*, dan kategori "Marah" untuk *recall*.

Kata kunci : Emosi, Klasifikasi, *K-Nearest Neighbor*, Pembobotan WIDF

1. PENDAHULUAN

1.1 Latar Belakang

Bahasa dalam sebuah tulisan merupakan alat komunikasi untuk menerjemahkan sebuah pikiran yang ingin diungkapkan oleh seorang penulis. Dalam sebuah tulisan seorang penulis tidak hanya menyampaikan sebuah keterangan dari suatu informasi, tetapi juga berisi informasi tentang perilaku manusia termasuk emosi [1]. Emosi yang tersembunyi dibalik tulisan sangat sulit untuk ditafsirkan oleh pembaca dikarenakan pengaruh oleh sudut pandang yang berbeda antara penulis dan pembaca [2]. Maka dari itu untuk memudahkan pembaca mengetahui emosi dari sebuah tulisan bisa menggunakan cara klasifikasi.

Banyak metode yang bisa digunakan dalam klasifikasi teks diantaranya yaitu metode *naive bayes*, *k-nearest neighbor* (KNN), *support vector machine* (SVM) *decision tree, neural network* dan sebagainya. Masing – masing metode mempunyai kelebihan dan kekurangan masing – masing dan tingkat akurasi yang berbeda-beda. Seperti pada penelitian yang berjudul "Review on Comporation Between Text Classification Algorithms" dari hasil perbandingan algoritma *naive bayes*, *k-nearest neighbor* (KNN), dan *support vector machine* (SVM), diketahui bahwa algoritma *K-nearest neighbor* (KNN) memiliki kinerja yang bagus dengan nilai akurasi sampai 86% [3].

Pada penelitian sebelumnya dalam pengolahan teks dalam metode pembobotan *term* sering menggunakan metode TF.IDF. metode ini memiliki kelemahan seperti yang disebutkan dalam penelitian yang berjudul "Text Categorization Based on Weighted Inverse Document Frequency" menjelaskan yaitu semua *text* yang berisi *term* tertentu diperlakukan sama, sehingga IDF tidak memperhitungkan jumlah kemunculan suatu *term* pada suatu dokumen. Sehingga untuk metode pembobotan term yang akan digunakan dalam penelitian ini adalah metode WIDF (*Weighted Inverse Document Frequency*), metode ini memiliki kinerja yang lebih baik dibanding dengan TF.IDF dengan memberikan peningkatan akurasi pada IDF sebesar 7.4% Maksimum [4].

Berdasarkan gambaran yang telah dijelaskan, dalam penelitian ini digunakan metode *K-nearest neighbor* (KNN) dengan metode pembobotan *term* WIDF untuk mengklasifikasikan emosi pada teks bahasa Indonesia dan mengetahui besar akurasi yang didapatkan.

1.2 Maksud dan Tujuan

Berdasarkan permasalahan yang akan diteliti, maka maksud penelitian ini adalah mengklasifikasikan emosi pada teks bahasa Indonesia menggunakan metode *K-nearest neighbor* dengan pembobotan WIDF

Tujuan yang ingin dicapai dalam simulasi pengenalan emosi pada teks bahasa Indonesia adalah :

1. Untuk mengklasifikasikan emosi pada teks bahasa Indonesia
2. Untuk mengetahui besar akurasi yang didapatkan dari pengklasifikasian emosi pada teks bahasa Indonesia menggunakan metode *K-nearest neighbor* dengan pembobotan WIDF.

1.3 Tahapan Pembangunan Sistem

Sistem klasifikasi emosi yang akan dibangun akan melewati tahapan-tahapan sebagai berikut :

a. Analisis Sumber Data

Data latih yang digunakan berasal dari data ISEAR (*International Survey On Emotion Antecedents and Reaction)* yang berbahasa inggris kemudian diterjemahkan ke dalam bahasa indonesia tanpa mengurangi maksud dari kalimat – kalimat dalam ISEAR.

Sedangkan untuk data uji yang akan digunakan adalah data lirik lagu berbahasa Indonesia dengan format (.txt)

b. Preprocessing

Pada tahap ini, data yang terkumpul akan diproses sehingga data yang didapat menjadi lebih terstruktur dan mudah untuk diolah. Langkah - langkah preprocessing terdiri dari case folding, convert negation, tokenizing, stopword removal dan stemming.

c. Pembobotan Kata

Pada tahap ini akan dilakukan proses pengekstrakan keyword menggunakan nilai WIDF (*Weighted Inverse Document Frequency*). Term (kata) di ambil dari hasil proses prepocessing terakhir yaitu stemming. Nilai dari hasil pembobotan akan digunakan sebagai tahapan pengklasifikasian menggunakan metode k-nearest neigbor.

d. Klasifikasi

Langkah selanjutnya adalah proses pengklasifikasian yang akan diproses menggunakan metode k-nearest neighbor untuk menentukan kandungan emosi dari data uji yang dimasukan.

2. ISI PENELITIAN

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 artibut, artibut dapat berupa kontinyu ataupun kategoris, salah satunya artibut menunjukan kelas untuk *record.*

2.2 Preprocessing

Preprocessing merupakan proses menggali, mengolah, mengatur informasi dengan cara menganalisis hubungannya, aturan-aturan yang ada di data tekstual semi terstruktur atau tidak terstruktur. bagian – bagian tahapan dari preprocessing, dapat dijelaskan dengan menggunakan flowchart berjenis Flowchart Sistem, yaitu bagan yang menunjukkan alur kerja atau apa yang sedang dikerjakan di dalam sistem secara keseluruhan dan menjelaskan urutan dari prosedur-prosedur yang ada di dalam sistem.

1. Case Folding

Pada tahap ini, semua huruf akan diubah menjadi huruf kecil.

2. Convert Negation

Convert Negation merupakan proses konversi kata-kata negasi yang terdapat pada suatu pernyataan, karena kata negasi mempunyai pengaruh dalam merubah nilai sentimen pada suatu tweet. Jika terdapat kata negasi maka akan disatukan dengan kata setelahnya. Kata - kata negasi tersebut meliputi kata “bukan”, “tidak”, “tak”, “ga”,”gak”, “enggak”, “jangan”, dan ”nggak”.

3. Tokenizing

Tokenizing merupakan tahap pemotongan kalimat berdasarkan tiap kata yang menyusunnya. Proses ini melakukan penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan simbol seperti titik(.), tanda seru(!), tanda tanya (?), koma(,), spasi, emoticon.

4. Stopword Removal

Stopword didefinisikan sebagai term yang tidak berhubungan dengan subyek utama dari database meskipun kata tersebut sering kali hadir di dalam dokumen dan kata yang dianggap tidak dapat memberikan pengaruh dalam menentukan suatu kategori sentimen..

5. Stemming

Stemming merupakan tahap untuk mentransformasi kata-kata yang terdapat dalam suatu dokumen ke kata-kata akarnya (root word) dengan menggunakan aturan-aturan tertentu. Dengan menggunakan stemming dapat mengurangi variasi kata yang sebenarnya memiliki kata dasar yang sama. Salah satu algoritma stemming yaitu Algoritma Nazief dan Adriani.

2.3 Pembobotan WIDF

Metode pembobotan kata yang digunakan dalam sistem ini adalah Weighted Inverse Document Frequency (WIDF). WIDF merupakan sebuah metode pengembangan dari metode IDF. Inverse Document Frequency (IDF) adalah proses mengukur term yang jarang muncul pada corpus. Penilaiannya menggunakan seluruh dokumen latih yang digunakan. Jika sebuah term sering muncul, maka term tersebut tidak bisa dianggap sebagai term yang mewakili dokumen tersebut. Sebaliknya, jika term yang jarang muncul pada corpus, maka term tersebut dikatakan memiliki hubungan dengan dokumen. Inverse Document Frequency merupakan perhitungan dari jumlah seluruh dokumen (D) dibagi dengan Frekuensi Dokumen (DF) dari kata (ti), dapat dituliskan dengan persamaan sebagai berikut.

$$\text{IDF} = \log\left(\frac{N}{Dfi}\right)$$

Keterangan :
IDF = *inverse document frequency*
N = Jumlah kalimat yang berisi term(t)
Dfi = Jumlah kemunculan *term* terhadap D

Kelemahan dari metode IDF adalah semua teks yang berisi *term* tertentu diperlakukan sama. Artinya, IDF tidak memperhitungkan jumlah kemunculan suatu *term* pada suatu dokumen. Contohnya pada tabel 2.1 berikut, dimana d (kolom) adalah text dan t (row) adalah term yang dicari, pertemuan antara di dan tj adalah *termfrequency* dari term tj di dalam text di

	d_1	d_2	d_3	d_4	d_5
⋮	⋮	⋮	⋮	⋮	⋮
t_x	2	50	3	2	4
t_y	3	2	3	2	3
⋮	⋮	⋮	⋮	⋮	⋮

Gambar 1. Contoh Perhitungan Kemunculan Term

Dapat dilihat dari persamaan (2.1) bahwa IDF (tx) dan IDF (ty) masing-masing bernilai 0, karena *term* yang dicari muncul pada semua text. Sedangkan, distribusi frekuensi tx dan ty terhadap d1 sampai d5 berbeda, IDF tidak dapat mengidentifikasi perbedaannya, karena bersifat binary counting. Sehingga, tidak dapat menentukan jumlah frekuensi dari term.

IDF mengasumsikan bahwa *term* berbanding terbalik dengan jumlah text yang mengandung *term* tersebut. Oleh karena itu, bagian penting dari persamaan (1) di atas adalah faktor 1/df(t). Metode WIDF menghitung factor ini dengan *term* frekuensi. Sebagai contoh, 1/df(t) dari tabel 2.1 di atas menjadi

$$\frac{1}{1+1+1+1+1}$$

untuk semua text. Kemudian angka "1" diganti dengan frekuensi dari masing-masing term, menjadi

$$\frac{50}{2+50+3+2+4}$$

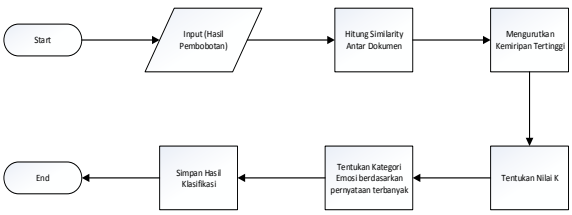
pada kasus d2. Faktor inilah yang yang disebut WIDF. Tidak seperti IDF, WIDF dapat membedakan masing-masing text d1…d5. Sehingga, WIDF dengan term t dalam text d dapat dituliskan sebagai persamaan (2) berikut :

$$WIDF(d,t) = \frac{TF(d,t)}{\sum_{i \epsilon D} TF(i,t)}$$

dimana TF(d,t) adalah *term frequency* dari term t di dalam text d dan i menyatakan jumlah text. WIDF dari term t menjumlahkan semua *term frequency* dari dari semua kumpulan text. Dengan kata lain WIDF adalah bentuk normalisasi term frequency dari semua kumpulan text [4].

2.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah suatu metode yang menggunakan algoritma supervised dimana hasil dari query instance yang baru diklasifikan berdasarkan mayoritas dari kategori pada KNN. KNN memiliki beberapa kelebihan yaitu ketangguhan terhadap training data yang memiliki banyak noise dan efektif apabila training data-nya besar. Sedangkan, kelemahan KNN adalah KNN perlu menentukan nilai dari parameter k (jumlah dari tetangga terdekat), training berdasarkan jarak tidak jelas mengenai jenis jarak apa yang harus digunakan dan atribut mana yang harus digunakan untuk mendapatkan hasil terbaik, dan biaya komputasi cukup tinggi karena diperlukan perhitungan jarak dari tiap query instance pada keseluruhan training sample.



Gambar 2. Diagram Alur K-Nearest Neighbor

Langkah – langkah metode KNN :
1. Hitung jarak antara data sampel (data uji) dengan data latih yang telah dibangun. Salah satu persamaan dalam menghitung jarak kedekatan Cosine Similarity.

$$Cos(i,k) = \frac{\sum_k (d_i d_k)}{\sqrt{\sum_k d^2_{ik}} \sqrt{\sum_k d^2_{jk}}}$$

Keterangan :

$\sum_k (d_i d_k)$: Vector dot produk dari i dan k

$\sqrt{\sum_k d^2_{ik}}$: Panjang vector i

$\sqrt{\sum_k d^2_{jk}}$: Panjang vector k
i : data uji ke-i
k : data latih ke-j

2. Menentukan parameter nilai k = jumlah tetanggaan terdekat bebas.
3. Mengurutkan jarak terkecil dari data uji
4. Pasangkan kategori sesuai dengan kesesuaian
Cari jumlah terbanyak dari tetanggaan terdekat Kemudian tetapkan kategori.

2.5 Pengujian

Pengujian dalam penelitian ini dilakukan dengan menguji sebanyak 30 lirik lagu bahasan Indonesia dengan nilai K yang digunakan adalah 2, 3, 4, 5 dan kategori klasifikasi yang ditentukan adalah emosi senang, sedih, tahut, marah, dan bersalah.

Berdasarkan pengujian yang telah dilakukan maka akurasi yang didapatkan sebagi berikut :

Tabel 1. Hasil Akurasi

No	Nilai K	Akurasi
1	2	60%
2	3	63.3%
3	4	70%
4	5	73.3%

Sedangkan untuk nilai relevansi dari setiap kategori emosi dengan menggunakan pengujian *precision* dan *recall*. Berikut hasil pengujian relevansi dari setiap kategori :

Tabel 2. Hasil Nilai Precesion

	2	3	4	5	Rata-rata
Senang	0.4	0.45	0.56	0.56	0.49
Takut	1	1	1	1	1
Marah	0.5	0.5	0.6	0.6	0.55
Sedih	0.75	0.72	0.75	0.81	0.75
Bersalah	0.6	1	1	1	0.9

Tabel 3. Hasil Nilai Recall

	2	3	4	5	Rata-rata
Senang	1	0.8	1	1	0.95
Takut	0.5	0.5	0.75	0.75	0.62
Marah	1	1	1	1	1
Sedih	0.42	0.57	0.64	0.64	0.56
Bersalah	0.5	0.5	0.25	0.5	0.43

3 PENUTUP

3.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan mengenai klasifikasi emosi pada teks bahasa Indonesia menggunakan metode *k-nearest neighbor* dengan pembobotan WIDF dapat disimpulkan sebagai berikut ;

1. Simulator yang dibuat untuk pengklasifikasian emosi pada teks bahasa Indonesia menggunakan metode k-nearest neighbor dengan mengimplementasikan lirik lagu berbahasa Indonesia sebagai data ujinya berhasil dilakukan, dengan memberikan kategori emosi yang sesuai dengan kandungan lirik lagu tersebut.
2. Keberhasilan mengkalsifikasikan emosi pada lirik lagu dipengaruhi oleh nilai K dalam proses k-nearest neighbor dan data latih yang digunakan.
3. Penggunaan metode pembobotan WIDF dalam proses pengklasifikasian emosi memberikan nilai akurasi yang meningkat dari penelitian sebelumnya dengan besar akurasi yang didapatkan sebesar 73.3%.
4. Untuk nilai relevansi dari setiap kategori emosi dengan menggunakan precision dan recall menghasilkan nilai maksimum sebesar 100%. Nilai tersebut didapatkan di kategori “Takut” untuk precision, dan kategori “Marah” untuk recallnya.

3.2 Saran

Berdasarkan penelitian yang telah dilakukan ada kekurangan yang didapat yaitu proses pengklasifikasian yang memerlukan waktu lama, sehingga penulis memberikan saran untuk penelitian selanjutnya untuk bisa menyempurnakan lagi dan data latih yang digunakan untuk bisa menggunakan data asli berbahasa Indonesia agar hasil yang didapatkan lebih relevan.

DAFTAR PUSTAKA

[1] C. Z.-j. H.W, "Multimodal Emotion Recognation From Speech," p. 1, 2004.

[2] S. M. Kim, "Recognising Emotions and Sentiments in Text," Sydney, The University of Sydney, 2011, p. 26.

[3] J. A. Vaibhav C.Gandhi, "Review on Comparison between Text Classification Algorithms," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) , vol. 1, no. 3, 2012.

[4] M. I. Takenobu Tokunage, "Text Categorization Based on Weighted Inverse Document Frequency," in ISSN 0918-2802, Ookayama, Meguro Tokyo Japang, Department Of Computer Science Tokyo Institute Of Technology , 1994.