doi: 10.47002/metik.v5i1.213

METIK VOLUME. 5 NOMOR. 1 2021 ISSN-P 2442-9562 ISSN-E 2580-1503

Analisa Perbandingan Algoritma ID3 dan KNN pada Klasifikasi Emosi Teks Berita Berbahasa Indonesia

Pramudya P.Insan[1*], Kusrini[2]

[1,2] Teknik Informatika, Magister Teknik Informatika, AMIKOM

[1]STMIK Balikpapan

[*]pamudyapi@stmikbpn.ac.id

Abstract

The use of algorithms in proving the text-based classification process or text mining is very rarely compared, especially for an emotion classification. Many do research in classification without an element of comparison in it and there is no use of a system that was built independently. In this study, a comparison was made to measure the ability of the algorithm in obtaining the level of accuracy in the classification process using ID3 and KNN. The data used are 220 news text-based data taken on the online media news site viva.co.id, the data training process is carried out with different weighting processes for each algorithm, namely with term weighting tf-idf for ID3 while KNN with similarity and vector space models. Classification is carried out to obtain data in the category of emotions with accuracy results obtained from testing classification with various comparison data, the highest accuracy is 71.25, namely by comparison of training data with test data of 75% - 25%. Thus the use of the ID3 algorithm is better in classifying emotions in Indonesian, which is a very efficient method of grouping data by category either manually or by system.

Keywords: data mining, emotion classification, text mining

Abstrak

Penggunaan algoritma pada pembuktian proses klasifikasi berbasis teks atau text mining sangat jarang dilakukan perbandingan khususnya untuk sebuah klasifikasi emosi. Banyak yang melakukan penelitian dalam klasifikasi tanpa unsur perbandingan didalamnya serta tidak terdapat penggunaan sistem yang dibangun secara mandiri. Pada penelitian ini perbandingan dilakukan untuk mengukur kemampuan algoritma dalam perolehan tingkat akurasi pada proses klasifikasi menggunana ID3 dan KNN. Data yang digunakan sebanyak 220 data berbasis teks berita yang diambil pada sistus warta media online yaitu viva.co.id, proses pelatihan data dilakukan dengan perbedaan proses pembobotan pada masing-masing algoritma yaitu dengan term weighting tf-idf untuk ID3 sedangkan KNN dengan similarity dan vector space model. Klasifikasi yang dilakukan untuk memperoleh data berkategori emosi dengan hasil akurasi yang didapatkan dari klasifikasi testing dengan data perbandingan yang beragam didapatkan akurasi paling tinggi yaitu 71.25 yaitu dengan perbandingan data latih dengan data uji 75%-25%. Demikian penggunaan algoritma ID3 lebih baik dalam pengklasifikasian emosi berbahasa Indonesia dimana sebuah metode yang sangat efisien dalam pengelompokkan data berdasarkan kategori baik secara manual ataupun sistem.

Kata kunci: *data mining*, klasifikasi emosi, teks *mining*

1. Pendahuluan

Teks adalah media yang paling sederhana akan tetapi teks juga memiliki peranan penting dalam komunikasi karena sebuah teks tidak hanya mempunyai pesan atau sebuah informasi belaka, namun juga menginformasikan tentang perilaku atau sifat manusia termasuk emosi. Media teks juga lebih mudah diperoleh dari bermacam-macam sumber misalnya pada buku, koran, majalah, situs, serta media elektronik lainnya seperti yang tersedia di internet lainnya. Tetapi emosi cenderung sangat berperan dalam komunikasi antar manusia di kehidupan sehari-hari. Sebagai contoh pada jejaring social yang semakin marak dan digemari para manusia untuk mengekspresikan emosi ke dalam bentuk teks yaitu update status pada akun facebook atau twiter. Oleh karena itu sistem interaksi

doi: 10.47002/metik.v5i1.213

manusia dan komputer yang baik harus dapat mengenali, representasikan dan memproses emosi manusia. Dalam perkembangan teknologi informasi yang semakin pesat, saat ini telah ditemui untuk melakukan pemrosesan teks yang bisa digunakan bermacam-macam metode. Diantaranya Naïve bayes, SVM, VCM, KNN, juga KNN-improve. Dari metode-metode tersebut digunakan untuk mengenali emosi pada teks.

Pada penelitian sebelumnya disimpulkan bahwa teknologi media digital telah banyak dilakukan untuk proses mengenali sebuah pattern yang terdapat pada teks. Penelitian "*Klasifikasi Teks Bahasa Indonesia pada Corpus Tak Seimbang Menggunakan NWKNN*" sebuah penerapan metode pengembangan dari KNN untuk dokumen teks berbahasa Indonesia [1]. Kesamaan dalam proses klasifikasi adalah data yang diolah dan diujikan untuk proses tahapan awal sebuah klasifikasi yaitu preprocessing untuk metode KNN. Penelitian ini masih belum adanya perbandingan dengan metode lain khususnya yang mempunyai perhitungan *entropy*, Information Gain dan lainnya. Proses pengenalan emosi tidak dapat dipisahkan dari proses kalsifikasi teks yang digunakan untuk menentukan jenis emosinya. Namun dalam penelitian klasifikasi emosi tidak selalu berhubungan dengan kata dan kalimat teks semata, seperti apa yang telah dilakukan [2] "*User Emotion Identification In Twitter Using Specific Features Hastag, Emoji, Emoticon And Adjective Term*", identifikasi emosidilakukan menggunakan beberapa *character emoji* dan *emoticon* untuk mendapatkan informasi tentang jenis emosi secara otomatis. Meskipun emoji dan emoticon tersebut tidak dapat terfilter untuk perumusan preprocessing dan ekstraksi indeksnya, hasil yang diperoleh untuk identifikasi emosi mencapai 92%.

Dengan beberapa tahapan pada setiap metode yang digunakan dalam penelitian ini memiliki luasan lingkup yang tak terbatas, oleh karena itu perlu adanya sebuah batasan masalah dalam penelitian ini antara lain:

1 Evaluasi perbandingan yang dilakukan terhadap ke-dua algoritma dalam penelitian ini hanya mengevaluasi hasil akurasi dari proses klasifikasi.
2 Proses *Stemming* pada setiap dokumen digunakan Porter Stemmer algorithm.
3 Kategori emosi yang diklasifikasi hanya yaitu marah, sedih, senang, takut.
4 Data latih dan data uji dikoreksi keakuratannya oleh pakar Bahasa Indonesia di kampus STMIK Balikpapan atau dari PTS untuk validasi data sesuai dengan kategori masing-masing.
5 Data yang akan digunakan untuk penelitian ini hanya format teks (.txt), dan total data yang akan digunakan adalah 200 dengan perbandingan data latih dan data uji adalah 70:30.

2. Metoda Penelitian

Penelitian ini menerapkan metode kulitatif mengenai metode dan langkah yang dilakukan dalam penelitian ini hingga tujuan penelitian melakukan perbandingan penggunaan metode ID3 dan KNN untuk hasil klasifikasi emosi pada teks bahasa Indonesia. Adapun tahapan-tahapan pada penelitian dapat dilihat pada Gambar 1 yang meliputi (1) Observasi, (2) Studi Pustaka, (3) Perancangan, (4) Implementasi Sistem, (5) Pengujian dan Evaluasi, dan (6) Analisa Hasil.

Langkah penelitian mengenai pengklasifikasian emosi pada bertia berbahasa Indonesia dijelaskan pada Gambar 1 berikut ini:

Gambar 1. Alur Penelitian

doi: 10.47002/metik.v5i1.213

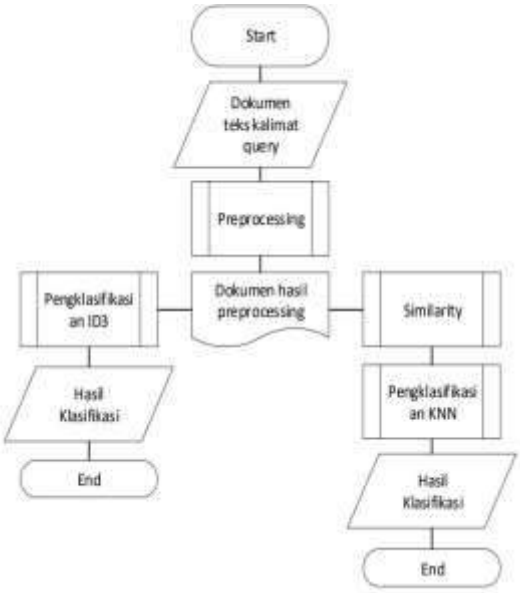
2.1. Alur Proses Data Latih

Berikut ini adalah bentuk pemodelan penelitian untuk proses data latih berdasarkan teknik pengumpulan data, dengan gambar berikut ini dijelaskan bagaimana proses data latih sesuai dengan tahapan preprocessing hingga diperoleh transformasi data dalam kelas interval setelah pengumpulan data sebayak 100 data yaitu 25 data setiap kategori emosinya [3].

Gambar 2. Alur Proses Data Latih

2.2. Perancangan Proses Klasifikasi

Sesuai dengan pemodelan alur pada proses data latih sebelumnya pengklasifikasikan kedalam algoritma baik secara manual atau menggunakan sistem memperoleh tingkat akurasi yang lebih baik perlu dibagi menjadi dua, dikarenakan penggunaan algoritma yang berbeda tahapannya untuk sebuah proses klasifikasi. Berikut alur proses klasifikasi emosi pada teks berita berbahasa Indonesia [4].



Gambar 3. Alur Proses Klasifikasi

2.3. Analisis Kebutuhan Sistem

Pada penelitian ini dibutuhkan sebuah data yang telah dilakukan proses data latih sebelumnya *preprocessing*, pembobotan, dan generalisasi data [5]. Sehingga sistem dapat melakukan perhitungan secara otomatis dan tepat dalam agar proses pembobotan dokumen interval dapat ditentukan dengan benar dan sesuai. Sistem dibagi dalam 3 tahap, yaitu *pre-processing*, pembelajaran, dan pengklasifikasian. Sistem akan dapat bekerja jika pengguna memasukkan data latih yang berupa dokumen teks. fungsi dari data latih itu sendiri adalah untuk membentuk *classifier* yang diperlukan dalam pengkategorian dokumen baru.

2.4. Perhitungan Manual

Pada perhitungan manual ini membahas contoh perhitungan manual untuk *data training*, *data testing*, pembobotan tf-idf, proses generalisasi dengan distribusi frekuensi dan pengklasifikasian dengan algoritma ID3 dan KNN Perhitungan manual diawali setelah kata melalui tahap *preprocessing*.

2.5. Pengujian Sistem

Dalam penerapan kedua algoritma kedalam sistem, diharapkan sebuah sistem dapat melakukan proses klasifikasi sesuai dengan tahapan dan juga hasil yang lebih baik dikarenakan perlakuan secara otomatis. Dalam pengujian terdapat scenario pengujian pada setiap masing-masing algoritma. Untuk algoritma ID3 di skenariokan sebuah pengujian dalam perbandingan 70:30 yang dimana terdapat beberapa perbedaan prosesntase dari data latih dan data uji yang akan digunakan dalam proses pengujian. Sedangkan dalam proses pengujian KNN akan dilakukan perbedaan dalam penggunaan sistem, nantinya file yang akan digunakan adalah file hasil ekstrasi dari proses klasifikasi ID3 akan dilakukan proses pengujian dengan menggunakan data latih yang sama dengan data yang digunakan dalam pengklasifikasian sebelumnya yaitu pengujian ID3.

doi: 10.47002/metik.v5i1.213

METIK VOLUME. 5 NOMOR. 1 2021 ISSN-P 2442-9562 ISSN-E 2580-1503

2.6. Proses Pembobotan ID3

Dalam penerapan kedalam sisitem distribusi frekuensi dilakukan pada setiap atribut karna setiap atribut memiliki data bobot berbeda. Pada perhitungan manual ini distribusi frekuensi tidak dilakukan ke seluruh atribut karna jumlah dokumen terbatas, maka dilakukan distribusi frekuensi keseluruh data hasil pembobotan dari rata training. Berikut langkah2 untuk mebuat distribusi frekuensi dalam tahapan pembobotan algoritma ID3.

Tabel 1. Hasil Pengurutan Data Pembobotan

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0.30102
0.30102	0.30102	0.54018	0.60206	0.60206	0.60206
0.60206	0.90031	0.90031	1.08036	1.20412	1.44048
1.80618	1.80618	2.53733	5.41854		

Dari hasil yang telah diketahui melalui tabel pengurutan diatas, dapat dilihat nilai maksimum yang didapat adalah 5.41854 sedangkan nilai miminum dari data adalah 0 dari jumlah data sampel 66. Perhitungan rentang nilai minimum dengan maksimum sesuai dengan persamaan.

R=5.41854 – 0 =5.41854

1 Untuk menentukan banyak kelas yang akan digunakan dengan menggunakan persamaan.

$K = 1 + 3.3 \log 66$

2 Hasil dari perhitungan banyak kelas tersebut dapat dibuat 6 kelas.

3 Menentukan Panjang interval setiap kelas menggunakan persamaan.

$$Interval = \frac{5.41854}{7} = 0.9$$

4 Memilih ujung paling bawah di kelas interval pertama, dengan begitu dapat diambil sama dengan data terkecil atau data lebih kecil, namun selisih harus kurang dari panjang kelasnya. Diambil data terkecil adalah 0, maka Batasan kelas atas dengan menambahkan interval kemudian kurangi batas atas kelas dengan skala terkecil dari data. Maka kelas pertama 0 hingga 0.9 dan interval kedua 0,9-1,9 dan seterusnya.

5 Setiap range interval diberikan nama atau indeks agar mudah bilamana dilakukan pemanggilan data. Dalam permsalahan mengatasi nilai bobot lebih dari Xmax dalam distribusi frekuensi, data maksimum diproses berbeda perlakuannya dengan kondisi data x > max. hasilnya pada tabel 2.

Tabel 2. Hasil Pengurutan Interval Kelas

Data interval	Nilai interval	Frekuensi
Int1	0 – 0.9	56
Int2	0.9 – 1.9	7
Int3	1.9 – 2.9	1
Int4	2.9 – 3.9	1
Int5	3.9 – 4.9	0
Int6	X >5	1
Jumlah data		66

Berikut hasil transformasi data untuk data training dan label kategori setiap dokumen.

Tabel 3. Hasil Transformasi Data

	korban	paksa	buruk	bunuh	marah	takut	kesal	Kat
D1	Int2	Int1	Int1	Int3	Int1	Int1	Int1	Marah
D2	Int2	Int1	Int1	Int1	Int1	Int1	Int1	Marah
D3	Int1	Int1	Int6	Int1	Int1	Int2	Int1	Marah
D4	Int2	Int1	Int1	Int1	Int1	Int1	Int1	Marah
D5	Int1	Int3	Int1	Int1	Int1	Int1	Int1	Marah
D6	Int1	Int1	Int1	Int1	Int1	Int1	Int2	Senang
D7	Int1	Int1	Int1	Int1	Int1	Int1	Int1	Senang
D8	Int2	Int1	Int1	Int1	Int1	Int1	Int1	Senang

3. Hasil Penelitian

Penelitian mengenai klasifikasi emosi menggunakan data berita ini dengan format yang digunakan adalah dalam format teks (.txt). Data berita dikelompokkan berdasarkan emosinya setelah dianalisis dan dipelajari oleh ahli Bahasa Indonesia setempat, dan data tersebut akan divalidasi oleh ahli Bahasa Indonesia tersebut menjadi beberapa kelompok emosi sesuai dengan kategori emosi senang, marah,takut, dan sedih. Setelah itu dilakukan pengolahan data sesuai dengan tahapan-tahapan teks *mining* untuk menggunakan data tersebut kedalam sistem agar dapat diolah menjadi data latih dan dapat dilakukannya pengujian pada setiap algoritma ID3 dan KNN.

doi: 10.47002/metik.v5i1.213

Hasil yang didapatkan dari klasifikasi emosi berbasis teks berita berbahasa Indonesia dapat dilihat pada tabel berikut ini

Tabel 4. Hasil Pengujian

Prosentase Data (%)		Tingkat Akurasi (%)	
Data Latih	Data Uji	Percobaan Ke-1	Percobaan Ke-2
40%	**60%**	40.25	44
50%	**50%**	60.22	60.13
60%	**40%**	56.13	55
75%	**25%**	71.75	70
80%	**20%**	50	46
85%	**15%**	66.85	72.13
90%	**10%**	25	22.55

Sedangkan pengujian untuk algoritma KNN memperoleh hasil dalam bentuk probabilitas antar dokumen latih dengan dokumen uji yang dapat dilihat pada gambar berikut ini.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7
D_1	1.08036	0.60206	0	1.80618	0	0.60206	0
D_2	0.90031	0.90031	0	0.60206	0.30102	0	0
D_3	0	0	5.41854	0	0	1.20412	0
D_4	1.44048	1.44048	0	0	0	0	0
D_5	0	0	0	0	0	0	0
D_6	0.54018	0.54018	0	0	0.30102	0	1.80618
D_7	0	0	0	0	0.30102	0	0
D_8	0.90031	0.90031	0	0	0	0	0
Uji_1	3.01031	0.60206	0	0.60206	0	0.30103	0
Uji_2	0	0	0	0.30103	0	0	0
Uji_3	0	0	0	0	0.60206	0	0.60206
Uji_4	0	0	0	0	0	0.30103	0.60206

Gambar 4. Matriks *Term Document*

Dari data matriks diatas dihitung tingkat kemiripan antar dokumen dan salah satu contoh dokumen dalam penyelesaian perhitungan similarity dapat dilihat pada sebagai berikut.

Dok	*Term*						
	T1	T2	T3	T4	T5	T6	T7
D_1	1.076	0	0.416	0	0.444	1.267	0.717
Uji	1.155	0	0	0.416	0.653	1.893	0

Dengan menggunakan persamaan berikut ini,

$$Sim(Q,D) = \frac{Uji.D_1}{|Uji||D_1|}$$

untuk melakukan perhitungan dari kemiripan pada dokumen 1 dengan dokumen uji

$$Uji \cdot D_1 = (1.076 * 0) + (0 * 0) + (0.416 * 0) + (0 * 0.416) + (0.444 * 0.653) + (1.267 * 1.893) + (0.717 * 0) = 0 + 0 + 0 + 0 + 0.389 + 2.398 + 0 = 2.87$$

$$|Uji| = \sqrt{1.155^2 + 0^2 + 0^2 + 0.416^2 + 0.653^2 + 1.893^2 + 0^2} = 5.267$$

$$|D1| = \sqrt{1.076^2 + 0^2 + 0.416^2 + 0^2 + 0.444^2 + 1.267^2 + 0.717^2} = 3.647$$

$$Sim(Q,D) = \frac{Uji \cdot D_1}{|Uji|\,|D_1|} = \frac{2.87}{5.267*5.568} = 0.098|$$

Dari perhitungan di atas maka dapat diketahui bahwa kemiripan antara dokumen 1 dan dokumen uji sebesar 0.240. Kemudian semua dokumen latih akan di hitung kemiripan nya dengan dokumen uji, sehingga dapat disimpulkan dokumen uji masuk ke dalam kategori yang telah ditentukan pada dokumen latih. Pada perhitungan di atas hanya menggunakan k=1, sehingga di dapat dokumen uji masuk ke dalam kategori Marah. Perhitungan kemiripan dokumen uji dengan dokumen latih akan dijelaskan pada tabel hasil perhitungan kemiripan dokumen training dengan dokumen *testing*.

Tabel 4. Hasil Perhitungan Kemiripan

	Duji	Kategori
D1	0.847	Sedih
D2	0.654	Marah
D3	0.898	Marah
D4	1.076	Senang
D5	0.158	Marah
D6	0.355	Senang
D7	1.027	Senang
D8	1.084	Senang

4. Kesimpulan

Berdasarkan analisa hasil pengujian baik secara sistem maupun manual yang telah dijabarkan pembahasan dan evaluasi yang dilakukan pada bab sebelumnya, maka kesimpulan yang dapat diambil sebagai berikut.

1 Telah diketahui perbedaan untuk proses pembobotan pada masing-masing algoritma proses yang berbeda tersebut yaitu pada proses transformasi data latih, jika ID3 mengelompokkan beberapa dokumen berdasarkan interval dan KNN dengan membentuk *knn classifier* dengan membentuk panjang *vector* berdasarkan hasil dari pembobotan *tf-idf*.

2 Pengaruh proses stemming dalam perlakuan data latih untuk klasifikasi

doi: 10.47002/metik.v5i1.213

emosi berdasarkan teks, pada penelitian ini disimpulkan bahwa proses tersebut tidak terlalu berpengaruh dalam tingkatan akurasi sebuah penerapan algoritma. Meski hanya dilakukan percobaan pada salah satu algoritma yaitu ID3.

3 Perolehan tingkat akurasi ini hanya diperoleh dari klasifikasi algoritma ID3 yang paling tinggi yaitu pada skenario perbandingan data latih dengan data uji sebesar 60% - 40% dengan tingkat akurasi mencapai 87.11% . Akurasi yang didapat sangat tidak konsisten karena adanya penggunaan beberapa dokumen uji yang tidak selalu sama dengan pengujian pada setiap skenario pengujian, sedangkan untul algoritma KNN telah berhasil dilakukan perhitungan probabilitas namun perlu adanya pengujian untuk proses klasifikasi dengan menggunakan lebih banyak dokumen uji untuk proses pengujian agar diketahui tingkat keakuratan sebuah proses algoritma klasifikasi

.

5. Saran

Adapun saran untuk penelitian ini adalah sebagai berikut:

1. Perlu dilakukan peringkasan temu kembali informasi pada setiap dokumen latih agar dapat mendapatkan nilai bobot yang sesuai sehingga perolehan tingkat akurasi yang lebih baik untuk proses klasifikasi.
2. Penelitian ini masih sangat diperlukan perlakuan tambahan yaitu dengan pengujian menggunakan aplikasi/sistem yang lainnya
3. Penelitian ini dapat menggunakan metode tambahan untuk menguji ketepatan emosi pada data latih seperti ontology emosi thrayer model, dan lainnya.
4. Penelitian ini masih harus dilakukan perbaikan yaitu perlunya penggunaan aplikasi yang sama untuk setiap algoritma agar lebih mudah dalam menganalisis perbandingan algoritmanya.

6. Daftar Pustaka

[1] A. Ridok and R. Latifah, “Klasifikasi Teks Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan NWKNN,” *Konf. Nas. Sist. dan Inform. 2015*, no. Oktober, pp. 222–227, 2015.

[2] Y. A. Sari, E. K. Ratnasari, S. Mutrofin, and A. Z. Arifin, “USER EMOTION IDENTIFICATION IN TWITTER USING SPECIFIC FEATURES: HASHTAG, EMOJI, EMOTICON, AND ADJECTIVE TERM,” *J. Ilmu Komput. dan Inf. (Journal Comput. Sci. Information).*, vol. 7, no. 1, pp. 18–23, 2014.

[3] B. Santoso, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, 1st ed. Yogyakarta: Graha Ilmu, 2007.

[4] H. Kamber, *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman, 2006.

[5] A. Rohman, “Model Algoritma K-Nearest Neighbor (K-Nn) Untuk Prediksi Kelulusan Mahasiswa,” *Neo Tek.*, vol. 1, no. 1, 2015.