KLASIFIKASI *HELPDESK* UNIVERSITAS JENDERAL ACHMAD YANI MENGGUNAKAN *CONCEPT FREQUENCY-INVERSE DOCUMENT FREQUENCY* (CF-IDF) DAN *K-NEAREST NEIGHBORS* (K-NN)

Taufiq Akbar Herawan[*], Yulison Herry Chrisnanto, Asep Id Hadiana
Jurusan Informatika, Fakultas MIPA, Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Sudirman, Cimahi, Jawa Barat, 40513
[*]Email: tfqkbrhrwn@icloud.com

Abstrak

Universitas Jenderal Achmad Yani (Unjani) memiliki fasilitas Helpdesk pada website sebagai tempat untuk menampung pelayanan berupa pesan yang terdiri dari pertanyaan atau komplain terhadap permasalahan yang berkaitan dengan civitas akademik. Banyaknya jumlah pesan yang diterima setiap harinya serta dengan dibutuhkannya tingkat kesiapan yang tinggi dapat berpotensi menimbulkan kesulitan dalam melakukan klasifikasi isi pesan, dengan demikian distribusi terhadap pesan tersebut menjadi terhambat. Proses klasifikasi memiliki beberapa proses preprocessing yang terdiri dari proses case folding, tokenizing, stemming, dan filtering. Pembobotan yang dilakukan adalah dengan menggunakan Concept Frequency-Inverse Document Frequency (CF-IDF). Cosine similarity salah satu metode yang dapat diterapkan untuk membandingkan kedekatan antara data latih degan data uji. K-Nearest Neighbors (K-NN) merupakan suatu metode yang menggunakan algoritma supervised yang dimana metode ini digunakan untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang tingkat kemiripanya paling dekat dengan objek tersebut. Berdasarkan hasil dari penelitian yang telah dilakukan semakin besar jumlah K yang digunakan maka akurasi dari klasifikasi semakin menurun. Akurasi terbesar didapatkan dengan menggunakan jumlah kedekatan K=1 dengan akurasi sebesar 95%.

***Kata kunci**: CF-IDF, cosine similarity, helpdesk, klasifikasi, K-NN*

1. PENDAHULUAN

Peningkatan kualitas pelayanan merupakan hal yang penting dalam sebuah instansi atau perusahaan. Standar kualitas pelayanan harus diupayakan apabila instansi tersebut ingin memberikan kontribusi yang maksimal pada pemakai jasa layanan. Dengan berkembangnya teknologi dalam bidang sistem informasi mempunyai dampak yang signifikan dalam kehidupan manusia sehari-hari. Internet dapat memudahkan pekerjaan manusia menjadi lebih cepat dan lebih praktis dalam mendapatkan informasi yang dibutuhkan (Nursalim dkk., 2014). Unjani telah memiliki *website* yang salah satunya terdapat fitur yang bernama *Helpdesk*. Fitur tersebut mempunyai fungsi yaitu sebagai penampung berbagai macam jenis pesan berupa kritik, saran ataupun komplain dari berbagai kalangan mulai dari mahasiswa, karyawan dan masyarakat umum terhadap Unjani. untuk dipertimbangkan yang semata-mata sebagai masukkan untuk kemajuan dan perkembangan Unjani. Dengan jumlah rata-rata 7 pesan perhari yang diterima dari tahun 2013 hingga tahun 2015 terdapat jumlah data sebanyak 4919 baris data. *Helpdesk* pada dasarnya adalah sebuah *center point* dimana masalah atau *issue* dilaporkan, diatur secara terurut dan diorganisasikan. *Helpdesk* Universitas Jenderal Achmad Yani (Unjani) dibuat agar perbaikan dapat dilakukan secara cepat dan terus-menerus agar kepuasan dari pengguna senantiasa terpenuhi. Pesan yang disampaikan oleh pelanggan merupakan masukkan bagi Unjani yang sangat bernilai dan akan menjadi bahan evaluasi terhadap sistem yang telah berjalan.

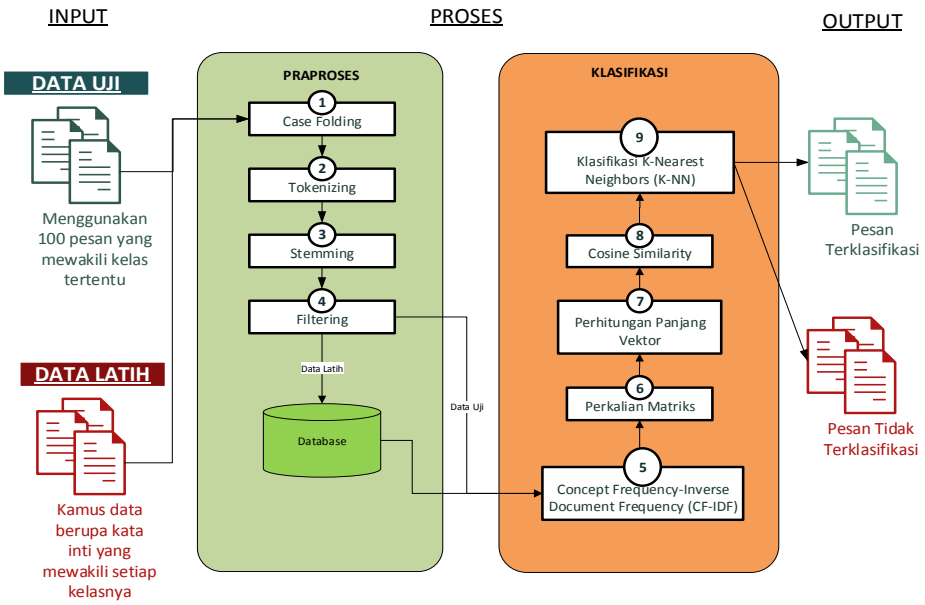
Klasifikasi adalah suatu proses penggolongan atau pengelompokan secara sistematis mengenai sebuah objek ke dalam kelompok atau kelas tertentu berdasarkan ciri-ciri yang sama. Pada beberapa kasus klasifikasi pada dokumen akan sangat penting dan berguna untuk kemudahan pengguna dalam melakukan pencarian dan estimasi waktu yang lebih cepat (Indriani.A, 2014). Klasifikasi yang dilakukan pada pesan ini menggunakan metode K-NN. Algoritma K-NN merupakan algoritma *supervised learning* dimana hasil dari suatu data baru diklasifikasikan berdasarkan kelompok mayoritas dari nilai K buah tetangga tedekat. Algoritma K-NN melakukan klasifikasi berdasarkan kedekatan jarak antar suatu data dengan data yang lainnya (E.Prasetyo, 2012). K-NN merupakan algoritma yang menggunakan seluruh data latih untuk melakukan proses klasifikasi (*complete storage).* Hal ini mengakibatkan proses dalam klasifikasi untuk data latih

ISBN 978-602-99334-5-1

dalam jumlah yang sangat besar akan memakan proses prediksi menjadi sangat lama tetapi menghasilkan akurasi yang baik dalam klasifikasi.

2. METODOLOGI

Pada penelitian ini menggunakan *website* Unjani pada fitur *helpdesk* sebagai objek penelitian yang digunakan. Terdapat 4919 baris data dengan kurun waktu 2013 hingga 2015. Pesan yang telah diuji kebenaran kelasnya akan digunakan sebagai data latih. Untuk lebih jelasnya tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 Data Masukkan

Data yang akan digunakan pada penelitian ini sebanyak 100 pesan yang mewakili kelas tertentu yang akan diuji kebenarannya akan kelas tersebut. Contoh data masukkan dapat dilihat pada Tabel 1.

Tabel 1. Data Masukkan

No	Pesan
1	Apakah ada jalur ekstensi informatika di unjani? terimakasih
2	Mohon informasi jadwal penerimaan mahasiswa baru informatika TA 2016/2017. Terima kasih
3	Saya mau bertanya kalo jurusan informatika di unjani S1 atau D3?
...	...
...	...
100	Maaf, apakah ada program ekstensi kimia di Unjani? Jika ada dimana saya bisa mendapat info resminya dengan lengkap. Terima Kasih

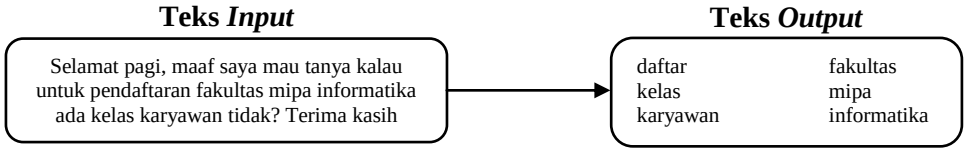
2.2 Tahapan Penelitian

Tahap pertama yang akan dilakukan yaitu ekstraksi dokumen yang dimana didalamnya terdapat empat tahapan yaitu *case folding*, *tokenizing*, *stemming* dan *filtering*. Setelah melalui tahap praproses akan dilakukan pembobotan dengan menggunakan CF-IDF tetapi untuk data latih akan di simpan terlebih dahulu kedalam *database*. Hasil dari proses CF-IDF untuk dilakukan perkalian matriks dari bobot yang dihasilkan sebelumnya lalu menghitung panjang vektor antara data uji dengan data latih. Hasil dari perbandingan tersebut akan digunakan sebagai masukkan untuk metode *Cosine Similarity*. Hasil dari nilai tersebut akan digunakan untuk dihitung kedekatannya

menggunakan K-NN dengan jumlah kedekatan k=1, k=3 dan k=5. Sebagai contoh, misalkan terdapat sebuah dokumen uji (Q) yang akan dilakukan klasifikasi akan kebenaran kelasnya yang berisi "**Selamat pagi, maaf saya mau tanya kalau untuk pendaftaran fakultas mipa informatika ada kelas karyawan tidak? Terima kasih**".

2.2.1 Ekstraksi Dokumen

Pada tahap ekstraksi dokumen terdapat empat tahapan yang dilakukan yaitu *case folding*, *tokenizing*, *stemming* dan *filtering*. Berikut hasil yang didapatkan setelah dilakukan ekstraksi dokumen pada dokumen uji (Q):



2.2.2 Pembobotan CF-IDF

Setelah melalui proses pada ekstraksi dokumen lalu akan dijadikan sebagai masukan dalam pembobotan menggunakan untuk menghitung CF, menghitung DF, menghitung IDF dan meghitung hasil bobot dari suatu dokumen, dimana untuk pemberian bobot terhadap dokumen.

Tabel 2. Concept Frequency dan Document Frequency

No	Concept	Frequency							DF
		Q	D1	D2	D3	...	...	D20	
1	daftar	1	0	0	0	...	...	1	3
2	kelas	1	0	0	0	...	...	0	1
3	karyawan	1	1	1	1	...	...	0	17
4	fakultas	1	0	0	0	...	...	0	1
5	mipa	1	0	0	0	...	...	0	3
6	informatika	1	0	0	0	...	...	0	2

1. Hitung nilai CF dari kata/konsep "daftar":
   a. n_q = 1 dan $\sum_k n_q$ = 6 sehingga CFDQ = 1/6 = **0,167**
   b. n_{d1} = 0 dan $\sum_k n_{d1}$ = 6 sehingga CFD1 = 1/5 = **0.2**
   c. n_{d2} = 0 dan $\sum_k n_{d2}$ = 5 sehingga CFD2 = 1/5 = **0.2**
2. Hitung Inverse Document Frequency (IDF) dari kata/konsep "daftar":
   $\sum D$ = 20 dan df = 3 sehingga IDF=Log(20/3) = **0.824**
3. Hitung Bobot CF-IDF dari kata/konsep "daftar":
   CFQ = 0.167 dan IDF = 0.824 sehingga CFIDF = **0.137**

2.2.3 Menghitung Perkalian Matriks Dokumen

Hasil dari pembobotan CFIDF akan dilakukan perkalian matriks WDQ*WDi.

Tabel 3. Perkalian Matriks

	D1	D2	D3	...	...	D13	...	...	D19	D20
	0	0	0	...	...	0	...	...	0	0
	0	0	0	...	...	0	...	...	0	0
	0	0	0	...	...	0	...	...	0	0
Total	0	0	0	...	...	0	...	...	0.06	0.03

2.2.4 Menghitung Panjang Vektor

Kemudian menghitung panjang setiap dokumen termasuk dokumen uji (Q). Dengan cara kuadratkan hasil dari bobot CFIDF pada setiap konsep dalam setiap dokumen lalu jumlahkan nilai kuadrat dan terakhir akarkan.

ISBN 978-602-99334-5-1

Tabel 4. Perhitungan Panjang Vektor

	Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
	0.019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.002	0
	0.047	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	0	0	0	0	0
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Total	0.17	0.06	0.079	0.079	0.052	0.079	0.064	0.108	0.108	0.06	0.108	0.1	0.063	0.08	0.107	0.107	0.176	0.148	0.148	0.076	0.038
Akar	0.411	0.25	0.281	0.281	0.227	0.281	0.252	0.329	0.329	0.245	0.329	0.316	0.252	0.28	0.327	0.327	0.419	0.385	0.385	0.275	0.196

2.2.5 Cosine Similarity

Metode *euclidean distance* dapat digunakan untuk menentukan jarak antara data latih dengan data uji. Selain itu dapat digunakan pula metode pengukuran kesamaan (*similarity*) untuk mengukur kemiripan antara data latih dengan data uji. Salah satu metode yang dapat digunakan adalah *cosine similarity*. Rumus *cosine similarity* yang dihitung menggunakan persamaan dibawah ini.

$$sim(q,d) = \frac{\sum_t w_{t,d} \cdot w_{t,q}}{\sqrt{\sum_t w_{t,d}^2} \cdot \sqrt{\sum_t w_{t,q}^2}} \qquad (1)$$

Dimana, sim(q,d) adalah kemiripan antara data latih (q) dengan data uji (d), $w_{t,d}$ nilai bobot konsep pada data latih dan $w_{t,q}$ adalah nilai bobot konsep pada data uji.

Setelah mendapatkan hasil perkalian dokumen uji (Q) dengan 20 dokumen lainnya dan menghasilkan nilai panjang total dokumen dilanjutkan mengakarkan jumlah nilai total tersebut. Dengan perhitungan *similarity* dokumen D1 terhadap dokumen uji Q atau sebaliknya dan seterusnya terhadap dokumen yang akan dibandingkan dengan nilai Q sehingga akan didapat nilai kemeripan terdekat Q.

1. Cos(Q,D13) = 0,037/(0,41 x 0,28)
   = 0,037/0,1148
   = **0,323**

2. Cos(Q,D19) = 0,006/(0,41 x 0,28)
   = 0,006/0,1148
   = **0,053**

2.2.6 Klasifikasi K-NN

K-*Nearest Neighbor* merupakan salah satu metode yang digunakan dalam pengklasifikasian. Prinsip kerja K-NN adalah mencari jarak terdekat antara data yang akan dievaluasi dengan K tetangga (*neighbor*) terdekatnya dalam data pelatihan. Berikut rumus pencarian jarak menggunakan rumus *Euclidian*:

$$d_i = \sqrt{\sum_{i-1}^{p} (x_{2i} - x_{1i})^2} \qquad (2)$$

Dimana, x1 adalah sampel data, x2 data uji, p adalah dimensi data dan d adalah jarak.

Hasil tabel dibawah ini adalah hasil dari perhitungan *cosine similarity*, yang akan digunakan sebagai masukkan terhadap metode K-NN.

Tabel 5. Hasil Cosine Similarity

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
0	0	0	0	0	0	0	0	0	0	0	0	0.327	0	0	0	0	0	0.053	0.034

Lalu mengurutkan dokumen berdasarkan nilai yang paling dekat kemiripannya.

Tabel 6. Hasil Urut Cosine Similarity

D13	D19	D20	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D14	D15	D16	D17	D18
0.327	0.053	0.034	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Langkah yang terakhir yaitu menentukan nilai ketetanggaan menggunakan (K-NN) K-Nearest Neighbors dengan menggunakan k=1, k=3 dan K=5.

Tabel 7. Hasil Menentukan Nilai Ketetanggan K=1

D13
0.327

Dokumen Uji (Q) dengan menggunakan kedekatan K=1 dikenali sebagai MIPA karena D13 merupakan anggota dari kelas MIPA Informatika, K=3 termasuk kedalam Universitas karena D19 dan D20 anggota dari kelas Universitas dan jika K=5 kelas tidak terklasifikasi.

3. HASIL DAN PEMBAHASAN

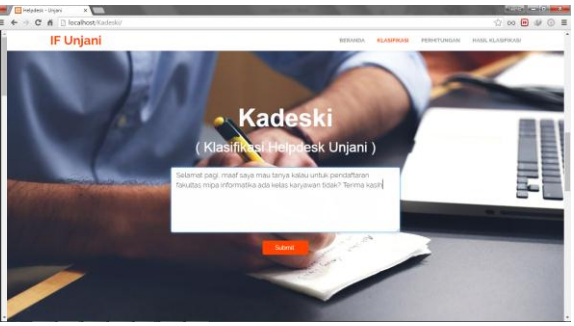
3.1 Uji Coba

Tahapan uji coba pada perangkat lunak klasifikasi pesan *helpdesk*, seperti yang terlihat pada Gambar 1 yaitu yang terdiri dari 9 proses utama. Tahapan uji coba tersebut akan dijelaskan secara singkat berikut ini:

1. Input Data Uji

Tahap ini yaitu menguji data baru untuk mengetahui akan kebenaran pesan tersebut termasuk kedalam kelas mana. Implementasi perangkah lunak menggunakan bahasa pemrograman PHP dan database MySQL. Antarmuka input data uji dapat dilihat pada Gambar 3.

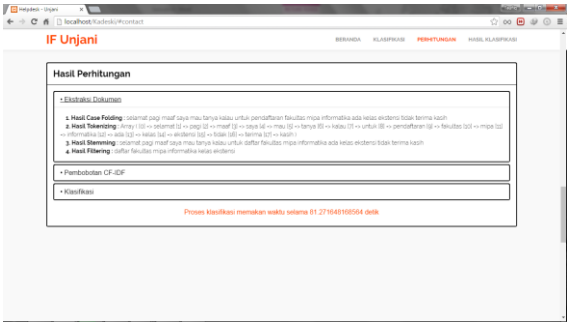
Gambar 2. Tampilan Utama Sistem



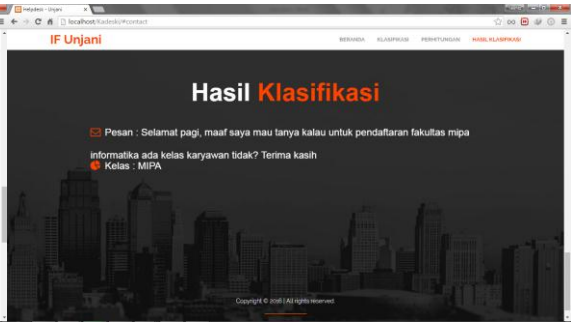
Gambar 3. Tampilan Input Data Uji

2. Perhitungan

Pada tahap ini sistem melakukan perhitungan kesulurhan dalam melakukan klasifikasi dari mulai ekstraksi dokumen hingga klasifikasi menggunakan K-NN. Tahapan perhitungan dapat dilihat pada Gambar 4.



Gambar 4. Hasil Perhitungan



Gambar 5. Hasil KlasifikasiHasil

3.2 Klasifikasi

Pada tahap akhir ini akan terlihat hasil dari kelas pesan yang diujikan. Antarmuka hasil klasifikasi dapat dilihat pada Gambar 5.

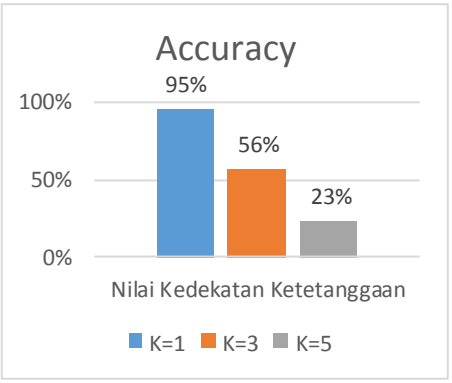
ISBN 978-602-99334-5-1

3.3 Evaluasi

Pada tahap pengujian ini melakukan pengujian menggunakan data uji baru yang belum pernah dilakukan pengujian sebelumnya. Data uji yang digunakan sebanyak 100 pesan masukkan yang diambil secara acak yang akan diuji kebenarannya akan kelas tersebut. Hasil pengujian dengan menggunakan K=1 menghasilkan 95 pesan benar dan terdapat 5 yang salah, hasil pengujian dapat dilihat pada Tabel 8 dan hasil uji akurasi Sistem dapat dilihat pada Gambar 2.

Table 8. Pengujian Data K=1

No	Data Uji	Kebenaran Kelas
1	Data Uji ke-1	Ya
2	Data Uji ke-2	Ya
3	Data Uji ke-3	Tidak
4	Data Uji ke-4	Ya
5	Data Uji ke-5	Ya
6	Data Uji ke-6	Tidak
7	Data Uji ke-7	Ya
8	Data Uji ke-8	Ya
...	...	...
...	...	...
100	Data Uji ke-100	Ya



Gambar 6. Akurasi Sistem Klasifikasi

4. KESIMPULAN

Berdasarkan hasil pengujian sistem klasifikasi *helpdesk* Unjani dengan menggunakan data uji yang diambil secara acak sebanyak 100 pesan dapat ditarik kesimpulan sebagai berikut:

1. Untuk menghasilkan akurasi yang terbaik menggunakan K=1 dengan nilai akurasi sebesar 95% karena dokumen pada setiap kelas jumlahnya terbatas yaitu paling sedikit terdapat 1 dokumen pada 1 kelas dan paling banyak terdapat 6 dokumen.
2. Diperlukan kamus kata yang besar untuk menambah nilai akurasi pada ekstraksi dokumen karena pesan yang diterima tidak semua menggunakan Bahasa Indonesia yang baku.

DAFTAR PUSTAKA

A. Indriani, "Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier," *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, no. ISSN: 1907 - 5022, pp. G5-G10, 2014.

E. Prasetyo, "Fuzzy K-Nearest Neighbor in Every Class Untuk Klasifikasi Data," *Seminar Nasional Teknik Informatika (SANTIKA)*, no. ISSN 2252-3081, pp. 57-60, 2012.

F. R. Hariri, E. Utami and A. Ambrowati, "Learning Vector Quantization untuk Klasifikasi Abstrak Tesis," *Citec Journal*, vol. 2, no. ISSN: 2354-5771, pp. 128-143, 2015.

Jumadi and E. Winarko, "Penggunaan KNN (K-Nearest Neighbors) Untuk Klasifikasi Teks Berita Yang Tak-Terkelompokkan Pada Saat Pengklasteran Oleh STC (Suffix Three Clustreing)," *International Journal of Nusantara Islam*, vol. Volume IX No. 1, no. ISSN 1979-8911, pp. 50-81, 2015.

S. Nursalim and H. Himawan, "Klasifikasi Bidang Kerja Kelulusan Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Teknologi Informasi*, vol. 10 Nomor 1, no. ISSN 1414-9999, pp. 31-43, 2014.

Z. Wei, H. Zhang, Z. Zhang, W. Li and D. Miao, "A Naive Bayesian Multi-label Classification Algorithm A Naive Bayesian Multi-label Classification Algorithm," *International Journal of Advanced Intelligence*, vol. 3 Number 2, pp. 173-188, 2011.