

Klasifikasi *Helpdesk* Menggunakan Metode *K-Nearest Neighbor* dan TF-ABS

Helpdesk Classification Using K-Nearest Neighbor and TF-ABS Methods

Muhammad Azis Suprayogi¹, Riza Adrianti Supono²

^{1,2}Manajemen Sistem Informasi, Program Pasca Sarjana, Universitas Gunadarma
E-mail: ¹azissuprayogi.gunadarma@gmail.com, ²adrianti@staff.gunadarma.ac.id

Abstrak

Helpdesk merupakan aplikasi yang bermanfaat bagi pengguna nya untuk memperoleh informasi mengenai layanan yang ada pada sebuah perusahaan atau instansi pemerintah. Proses disposisi tiket *helpdesk* secara manual dapat memperpanjang masa penyelesaian tiket karena adanya waktu tunggu untuk mendisposisikan tiket menuju unit yang sesuai. Klasifikasi teks *helpdesk* sangat diperlukan untuk mendisposisikan tiket secara tepat dan cepat ke unit yang berwenang menangani tiket. Penelitian ini menggunakan data tiket *helpdesk* Direktorat Jenderal Kekayaan Negara (DJKN) sebagai *dataset* sebanyak 10.537 dokumen. Teks *helpdesk* diklasifikasi ke dalam 8 kategori unit tujuan yaitu Setditjen, Dit.Humas, Dit.PKNSI, Dit.KND, Dit.BMN, Dit.Penilaian, Dit.PNKNL, dan Dit.Lelang. Klasifikasi menggunakan metode *K-Nearest Neighbor* (KNN) dengan jumlah tetangga terdekat (k) yaitu $k=1$ sampai dengan $k=30$ serta metode pembobotan TF-ABS untuk proses seleksi fitur. Jumlah fitur yang digunakan untuk klasifikasi sebanyak 5%, 10%, 15%, 20%, 25% dan 30% dari jumlah seluruh dokumen. Akurasi klasifikasi tertinggi sebesar 90,04% diperoleh pada $k=3$ dan jumlah fitur sebanyak 15%, sedangkan akurasi terendah 84,54% pada $k=19$ dan jumlah fitur sebanyak 30%. Hasil klasifikasi *helpdesk* menggunakan KNN dan TF-ABS menghasilkan akurasi yang cukup baik.

Kata kunci: *Helpdesk*, Klasifikasi, KNN, Pembobotan, TF-ABS

Abstract

Helpdesk is an application that is useful for its users to obtain information about the services that exist in a company or government agency. The manual disposition process of *helpdesk* tickets can extend the ticket settlement period due to waiting time to dispose of tickets to the appropriate unit. *Helpdesk* text classification is very necessary to dispose of tickets accurately and quickly to the unit authorized to handle tickets. This study uses the *helpdesk* ticket data of Direktorat Jenderal Kekayaan Negara (DJKN) as a *dataset* of 10.537 documents. *Helpdesk* texts are classified into 8 categories of unit objectives, namely Setditjen, Dit.Humas, Dit.PKNSI, Dit.KND, Dit.BMN, Dit.Penilaian, Dit.PNKNL, and Dit.Lelang. The classification uses the *K-Nearest Neighbor* (KNN) method with the closest neighbors (k) $k=1$ to $k=30$ and the TF-ABS weighting method for the feature selection. The number of features used for classification is 5%, 10%, 15%, 20%, 25% and 30% of the total number of documents. The highest classification accuracy of 90.04% was obtained at $k=3$ and the number of features 15%, while the lowest accuracy was 84.54% at $k=19$ and the number of features 30%. The *helpdesk* classification results using KNN and TF-ABS can produce quite good accuracy.

Keywords: *classification, helpesk, KNN, term weighting, tf-abs*

1. PENDAHULUAN

Proses disposisi tiket *helpdesk* secara manual oleh petugas *admin* memiliki kelemahan yaitu jumlah tiket yang cukup banyak yang harus didisposisikan dapat menyebabkan waktu respon tiket menjadi semakin lama sehingga mengakibatkan penurunan tingkat kepuasan

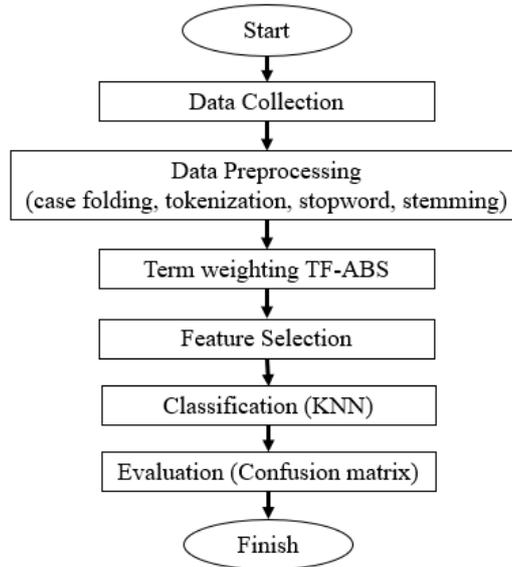
pengguna[1]. Kondisi tersebut membutuhkan metode untuk mengklasifikasikan tiket *helpdesk* secara otomatis yang dapat membantu proses disposisi tiket ke unit tujuan yang sesuai sehingga dapat mengatasi permasalahan pada petugas baru dan mempercepat proses respon tiket. Penelitian ini merupakan penelitian dasar untuk mengklasifikasikan tiket *helpdesk* menggunakan algoritma tertentu dan metode pembobotan kata tertentu sehingga dapat membantu proses disposisi tiket secara otomatis.

Penelitian terkait yang pernah dilakukan adalah Klasifikasi *Helpdesk* pada Universitas Jenderal Achmad Yani menggunakan metode CF-IDF atau TF-IDF dan *K-Nearest Neighbor* [2]. Hasil percobaan menunjukkan akurasi terbesar didapatkan dengan menggunakan jumlah kedekatan $k=1$ dengan akurasi sebesar 95%. Penelitian lainnya adalah klasifikasi jenis laporan masyarakat pada dokumen pengaduan Sambat Online menggunakan *K-Nearest Neighbour algorithm* [3]. Algoritma *K-Nearest Neighbour* menghasilkan akurasi yang baik pada klasifikasi data laporan masyarakat ke dalam tiga kategori yaitu pengaduan, permintaan informasi dan aspirasi, berdasarkan evaluasi serta validasi dengan *Confusion Matrix* ditemukan bahwa akurasi tertinggi adalah 82% dengan parameter $k=11$. Penelitian selanjutnya yaitu perbandingan akurasi dan waktu proses algoritma KNN dan SVM dalam analisis sentiment twitter[4] dengan hasil perhitungan akurasi menunjukkan bahwa SVM lebih unggul dengan nilai 89,7% tanpa *K-Fold Cross Validation* dan 88,76% dengan *K-Fold Cross Validation*. Namun untuk perhitungan waktu proses, metode *K-Nearest Neighbor* lebih unggul dengan waktu proses 0,0160 *seconds* tanpa *K-Fold Cross Validation* dan 0,1505 *seconds* dengan *K-Fold Cross Validation*. Penelitian lainnya yaitu Penelitian mengenai perbandingan pembobotan kata pada proses klasifikasi yang telah dilakukan yaitu kategorisasi berita menggunakan metode pembobotan TF-ABS dan TF-CHI dan algoritma *Support Vector Machine (SVM)* pada berita berbahasa inggris [5] dengan hasil bahwa kategorisasi berita tanpa proses stemming menggunakan metode pembobotan TF-ABS menghasilkan akurasi 95,74%, sedangkan menggunakan metode pembobotan TF-CHI menghasilkan akurasi 95,87% sehingga disimpulkan bahwa metode pembobotan TF-ABS dan TF-CHI tidak memberikan beda yang signifikan dalam performansi serta menunjukkan bahwa kedua metode tersebut memiliki performansi yang baik dalam kategorisasi berita. Penelitian lain dilakukan pada klasifikasi topik berita dalam bahasa Indonesia menggunakan *Decision Tree* dan pembobotan kata TF, TF-IDF, TF-RF, TF-ABS, TF-CHI² [5] dengan kesimpulan TF-ABS memiliki akurasi tertinggi sebesar 82,22%.

Berdasarkan hasil beberapa penelitian tersebut, maka pada penelitian ini akan mengklasifikasikan tiket *helpdesk* menggunakan algoritma *K-Nearest Neighbor* yang merupakan salah satu algoritma yang baik untuk klasifikasi teks [6] serta menggunakan metode pembobotan kata TF-ABS. Metode pembobotan kata diperlukan pada klasifikasi teks untuk menentukan tingkat kepentingan suatu kata terhadap suatu dokumen. Dalam penelitian ini dilakukan proses seleksi fitur yaitu dengan cara mereduksi fitur dalam dokumen berdasarkan nilai bobot kata yang dihitung menggunakan metode TF-ABS. Reduksi fitur merupakan cara termudah untuk meningkatkan akurasi klasifikasi dan meningkatkan efisiensi klasifikasi [7]. Penelitian ini menggunakan data tiket *helpdesk* DJKN sebagai *dataset* sebanyak 10.537 tiket *helpdesk*. Klasifikasi dengan algoritma KNN menggunakan *parameter* jumlah *neighbor* (k) yaitu $k=1$ sampai dengan $k=30$. Reduksi fitur dilakukan terhadap kata-kata yang memiliki bobot yang rendah sehingga menyisakan fitur (kata) dengan bobot yang tinggi dengan jumlah fitur 5%, 10%, 15%, 20%, 25%, dan 30% dari jumlah dokumen yang dihitung dengan metode TF-ABS.

2. METODE PENELITIAN

Alur penelitian ini terdiri dari 6 tahapan yaitu pengumpulan *dataset (Data Collection)*, persiapan data sebelum proses klasifikasi (*Data Preprocessing*), proses pembobotan kata (*Term Weighting*) menggunakan metode TF-ABS, proses seleksi fitur (*Fitur Selection*), proses klasifikasi (*Classification*) menggunakan algoritma *K-Nearest Neighbor* serta proses evaluasi klasifikasi (*Evaluation*) menggunakan *confusion matrix*. Alur penelitian dapat dilihat pada Gambar 1.



Gambar 1 Alur Penelitian

2.1 Data Collection

Data yang diambil adalah berasal dari tiket *helpdesk* yang dibuat pada tahun 2019 sampai dengan tahun 2020 yang berasal dari jalur *call center*, *email idcc* dan *website helpdesk idcc* seluruhnya berjumlah 10.537 tiket. Atribut yang dibutuhkan untuk klasifikasi teks tiket *helpdesk* adalah uraian isi tiket dan kategori tujuan tiket. Tiket yang menjadi dataset adalah tiket yang berstatus selesai. Adapun kategori tujuan tiket *helpdesk* berjumlah 8 yaitu Sekretariat Direktorat Jenderal (Set.Ditjen), Direktorat Kekayaan Negara Dipisahkan (Dit.KND), Direktorat Barang Milik Negara (Dit.BMN), Direktorat Piutang Negara dan Kekayaan Negara Lain-lain (Dit.PNKNL), Direktorat Pengelolaan Kekayaan Negara dan Sistem Informasi (Dit.PKNSI), Direktorat Penilaian (Dit.Penilaian), Direktorat Lelang (Dit.Lelang), dan Direktorat Hukum dan Hubungan Masyarakat (Dit.Humas). Jumlah dokumen masing-masing kategori adalah 1.462 dokumen Set.Ditjen, 1.226 dokumen Dit.BMN, 50 dokumen Dit.KND, 111 dokumen Dit.PNKNL, 5.145 dokumen Dit.PKNSI, 41 dokumen Dit.Penilaian, 2.279 dokumen Dit.Lelang, dan 223 dokumen Dit.Humas. Tabel 1 memperlihatkan contoh data yang digunakan untuk penelitian.

Tabel 1 Contoh Data Tiket *Helpdesk*

Kategori	Isi Tiket
Set. Ditjen	Bisa minta nomor KPKNL Semarang? apakah ada nomor lain? saya telepon tidak ada respon.
Dit. Bmn	alas hak BMN untuk kementerian/instansi pemerintah itu apa ya? apakah ada peraturan di DJKN terkait hal tersebut?
Dit. KND	MENANYAKAN KEJELASAN TENTANG HIBAH TANAH DIKECAMATAN LIRIK RIAU EX PERTAMINA KE PEMDA.
Dit. PNKNL	PEMANFAATAN BARANG MILIK NEGARA BERUPA SCRAB CABLE DAN BESI TUA YANG ADA DI AREAL KERJA K3 S
Dit. PKNSI	MENANYAKAN TERKAIT SERTIFIKAT BANK BAPINDO KASUSNYA SUDAH 15 TAHUN PROGRESSNYA LAMA DI TANGGAPI DJKN
Dit. Penilaian	Permohonan penilaian untuk PBB apakah bisa melalui DJKN?
Dit. Lelang	Apakah hari ini ada jadwal lelang di KPKNL Semarang?
Dit. Humas	menanyakan nomor direktorat hukum dan humas

2.2 Data Preprocessing

Sebelum melakukan klasifikasi, dokumen teks harus melalui proses *preprocessing* untuk membuat struktur teks dan mengekstrak fitur. Preprocessing secara umum terdiri dari 4 tahap *case folding*, *tokenization*, *stopword* dan *stemming* [8]. Tahap pertama *case folding* yaitu

mengubah karakter teks dokumen menjadi huruf kecil seluruhnya, kemudian tahap *tokenization* yaitu membersihkan tanda baca, spasi yang berulang, mengubah baris baru menjadi spasi, dan memisahkan kata per kata pada setiap dokumen. Tahap *stopword* yaitu menghapus kata yang tidak penting berdasarkan kamus *stopword* bahasa Indonesia. Tahap terakhir yaitu *stemming*, menghilangkan imbuhan kata pada setiap kata dalam dokumen. Proses *preprocessing* dilakukan menggunakan *tools* berbasis *Python* berupa *library Numpy, Pandas, Sastrawi, dan NLTK*.

2.3 Term Weighting TF-ABS

Pembobotan kedua menggunakan metode *Term Frequency – Absolute (TF-ABS)*. *Term Frequency (TF)* ditentukan dengan cara menghitung jumlah *term* yang terdapat pada setiap dokumen, pembobotan ABS merupakan pengukuran kemungkinan suatu *term* t_j yang ada dalam dokumen dengan kategori c_k dibagi dengan kemungkinan *term* t_j yang tidak ada dalam dokumen dengan kategori tersebut dan menggunakan basis log e yang dikenal dengan logit. Perhitungan ABS atau abs-logit dengan *term* t_j dan category c_k dapat dilihat pada persamaan (1).

$$ABS(t_j, c_k) = \left| \ln \left(\frac{(n_{kj}+0.5)(n_{\bar{k}j}+0.5)}{(n_{\bar{k}j}+0.5)(n_{kj}+0.5)} \right) \right| \quad (1)$$

dengan ABS adalah bobot kata, n_{kj} adalah jumlah dokumen pada kategori c_k dan mengandung *term* t_j , $n_{\bar{k}j}$ yaitu jumlah dokumen tidak pada kategori c_k dan mengandung *term* t_j , $n_{k\bar{j}}$ adalah jumlah dokumen pada kategori c_k dan tidak mengandung *term* t_j , $n_{\bar{k}\bar{j}}$ yaitu jumlah dokumen tidak pada kategori c_k dan tidak mengandung *term* t_j , t_j yaitu *term* t_j , c_k yaitu kategori c_k . Persamaan ABS merupakan salah satu metode seleksi kata yang umum digunakan. Persamaan ABS merupakan pengembangan dari metode seleksi *odds ratio*. Sehubungan dengan fokus pada kategorisasi teks adalah untuk kata yang terdistribusi secara berbeda pada kategori c_k dan $c_{\bar{k}}$, tidak penting apakah *term* tersebut lebih lazim pada kategori c_k dan $c_{\bar{k}}$, semakin memadai untuk tujuan ini adalah penggunaan nilai absolut dari *logit* sebagaimana persamaan (1) [9].

2.4 Feature Selection TF-ABS

Seleksi fitur merupakan strategi yang efektif dan efisien untuk menyiapkan data berdimensi tinggi pada permasalahan *data mining* dan *machine learning* dengan tujuan membangun model yang lebih sederhana, membuat data yang lebih bersih, lebih mudah dipahami, dan meningkatkan kinerja *data mining* [7]. Metode reduksi fitur sebagai salah satu dipilih pada penelitian ini karena mudah untuk meningkatkan akurasi dan membuat klasifikasi lebih efisien. Untuk proses fitur reduksi dilakukan dengan dua langkah, pertama memberi peringkat kata sesuai dengan bobot kepentingan suatu fitur berdasarkan skor kata yang dihitung dengan metode TF-ABS. Langkah selanjutnya membuang kata-kata yang memiliki nilai bobot kepentingan yang rendah dan memilih fitur sisanya dengan nilai bobot kepentingan yang tinggi. Jumlah fitur yang dipilih ditentukan sebesar 5%, 10%, 15%, 20%, 25%, dan 30% dari total jumlah dokumen. Selanjutnya fitur hasil seleksi tersebut akan diuji untuk mendapatkan hasil klasifikasi yang terbaik.

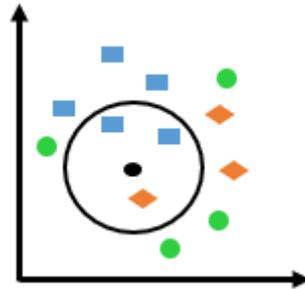
2.5 Classification

Sebelum proses klasifikasi, dokumen dibagi ke dalam data *training* dan data *testing* dengan pembagian 90:10, sehingga jumlah data *training* adalah 9.483, sedangkan jumlah data *testing* adalah 1.054. Proses pembagian data memanfaatkan *library sklearn python train_test_split*. Klasifikasi dilakukan menggunakan algoritma *K-Nearest Neighbor (KNN)* memanfaatkan modul *scikit-learn Python*. Adapun ukuran jarak yang dipilih adalah *euclidean distance*, dan parameter *k-nearest neighbor* (k) $k=1$ sampai dengan $k=30$. KNN digunakan untuk mengklasifikasi dokumen dengan cara menghitung jarak antara dokumen dan semua dokumen dalam pelatihan yang ditentukan menggunakan ukuran variasi atau kesamaan. Kemudian menemukan tetangga k terdekat dan menentukan dokumen ke dalam kategori dengan jumlah tetangga terdekat yang terbanyak [7]. Nilai k dari algoritma KNN adalah faktor yang

menunjukkan jumlah dokumen training terdekat dengan dokumen testing. Klasifikasi menentukan jarak vektor antar dokumen menggunakan persamaan (2).

$$d(x, y) = \sqrt{\sum_{r=1}^n (a_{rx} - a_{ry})^2} \quad (2)$$

dimana $d(x,y)$ adalah jarak antara dua dokumen yang dinamakan jarak *euclidean*. Pengukuran jarak *euclidean* adalah salah satu metode pengukuran yang umum digunakan [11]. a_{rx} adalah bobot term r dalam dokumen x , a_{ry} adalah bobot *term* r dalam dokumen y . Gambar 2 menunjukkan dokumen dalam ruang *euclidean* untuk faktor $k=3$. Jarak *euclidean* yang lebih kecil antara dokumen menunjukkan kesamaan yang lebih tinggi. Jarak 0 berarti dokumen sama.



Gambar 2 Jarak *Euclidean* untuk $k=3$

2.6 Evaluation

Evaluasi klasifikasi menggunakan pengukuran nilai akurasi dan *score-f1* dalam satuan persen berdasarkan rumusan menggunakan *Confusion Matrix*. *Confusion matrix* merupakan alat yang berfungsi menganalisis seberapa baik suatu *classifier* dalam melakukan klasifikasi [12].

Tabel 2 *Confusion Matrix*

Terprediksi	Aktual	
	Positif	Negatif
Positif	<i>True Positive (TP)</i>	<i>False Positif (FP)</i>
Negatif	<i>False Negative (FN)</i>	<i>True Negative (TN)</i>

Dimana *True Positive (TP)* adalah data yang diklasifikasikan berkaitan dengan kategori yang benar, *False Positive (FP)* adalah data yang diklasifikasikan berkaitan dengan kategori yang salah, *True Negative (TN)* adalah data yang diklasifikasikan tidak berkaitan dengan kategori yang benar, *False Negative (FN)* adalah data yang diklasifikasikan tidak berkaitan dengan kategori yang salah[13]. Pengukuran untuk menghitung performansi klasifikasi di antaranya yaitu akurasi. Akurasi adalah jumlah proporsi prediksi yang benar[14], didefinisikan dalam persamaan (3)

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (3)$$

3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk memperoleh hasil akurasi terbaik dengan menggunakan algoritma KNN dan metode pembobotan TF-ABS dalam mengklasifikasi tiket *helpdesk* sesuai dengan unit tujuan yang berwenang berdasarkan dataset tiket *helpdesk*.

3.1 Hasil Preprocessing

Tahapan *preprocessing* yang dilakukan, yaitu *case folding*, *tokenization*, *stopword*, dan *stemming*. Contoh dokumen hasil *preprocessing* pada teks *helpdesk* dapat dilihat pada tabel 4.

Tabel 3 Contoh Teks Hasil *Preprocessing*

Tahapan	Hasil
Raw Data	alas hak BMN untuk kementerian/instansi pemerintah itu apa ya? apakah ada peraturan di DJKN terkait hal tersebut?
Case Folding	alas hak bmn untuk kementerian/instansi pemerintah itu apa ya? apakah ada peraturan di djkn terkait hal tersebut?
Tokenization	alas, hak, bmn, untuk, kementerian, instansi, pemerintah, itu, apa, ya, apakah, ada, peraturan, di, djkn, terkait, hal, tersebut
Stopword	alas, hak, bmn, kementerian, instansi, pemerintah, peraturan, djkn
Stemming	alas, hak, bmn, menteri, instansi, perintah, atur, djkn

3.2 Pembobotan TF-ABS

Proses pembobotan dilakukan dengan cara menghitung nilai *Term Frequency (TF)*, menghitung nilai ABS menggunakan persamaan (1) kemudian mengalikan nilai TF dan ABS untuk setiap kata pada dokumen. Contoh hasil perhitungan TF-ABS dapat dilihat pada tabel 4.

Tabel 4 Contoh Perhitungan Nilai TF dan ABS

Term	TF	ABS	TF x ABS
alas	1	0.13619414503589636	0.13619414503589636
hak	1	0.5299737297619499	0.5299737297619499
bmn	1	2.105473169580813	2.105473169580813
menteri	1	0.9153215381339113	0.9153215381339113
instansi	1	0.7498229855449962	0.7498229855449962
perintah	1	0.42115822198769404	0.42115822198769404
atur	1	0.6024313004418819	0.6024313004418819
djkn	1	0.5765265535425604	0.5765265535425604

3.3 Seleksi Fitur

Tahap awal pada proses ini yaitu memberi peringkat kata berdasarkan bobot kepentingan suatu kata yaitu nilai TF dikali ABS atau TF-ABS. Kata-kata yang sudah dihitung nilai TF-ABS nya kemudian diurutkan seluruhnya dari nilai terbesar hingga nilai yang paling kecil. Contoh 20 kata dengan nilai TF-ABS tertinggi dapat dilihat di tabel 5.

Tabel 5 Contoh 20 kata bobot TF-ABS tertinggi

Term	Bobot
biad	100.6757
where	96.571777
lelang	75.530473
lpk	65.774021
tiket	65.507623
siswa	57.847644
investasi	57.142387
paksa	55.421126
nip	49.988319
hutang	42.248579
dukcapil	40.049191
kpp	39.735049
nilai	38.963798
tanggal	38.871237
pratama	38.610952
kpkp	37.960524
wkn	37.906818
set	37.094168
ip	36.814411
ujl	36.097306

Langkah selanjutnya mereduksi fitur (kata) dengan bobot kepentingan yang rendah dan

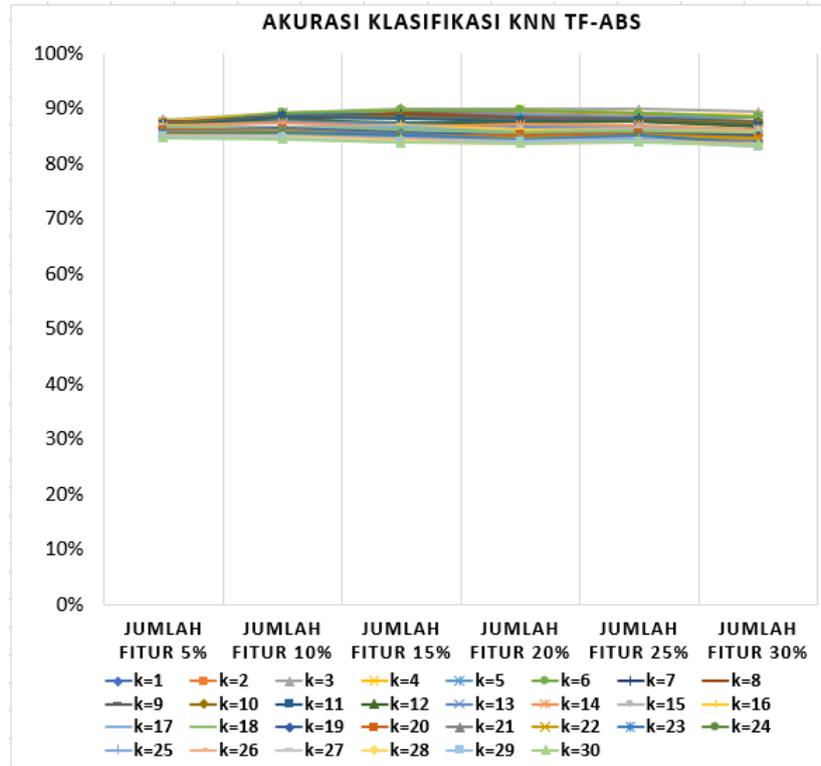
memilih fitur sisanya dengan nilai bobot kepentingan yang tinggi. Jumlah fitur yang dipilih ditentukan sebanyak 5% (526 fitur), 10% (1.053 fitur), 15% (1.580 fitur), 20% (2.107 fitur), 25% (2.634 fitur), dan 30% (3.161 fitur) dari total jumlah dokumen. Selanjutnya fitur hasil seleksi tersebut akan diuji untuk mendapatkan hasil klasifikasi yang terbaik.

3.4 Klasifikasi dan pengujian

Dataset terlebih dahulu dibagi menjadi data *training* dan data *testing* dengan perbandingan 90:10 yaitu 9.483 data training dan 1.054 data testing. Data training diproses ke dalam model klasifikasi KNN kemudian melakukan prediksi terhadap 1.054 data testing. Data diklasifikasi menggunakan variasi nilai $k=1$ sampai dengan $k=30$ serta menggunakan variasi jumlah fitur yang dihasilkan dari proses seleksi fitur untuk dicari hasil yang memiliki akurasi terbaik. Perhitungan akurasi menggunakan *confusion matrix*. Nilai akurasi dihitung secara keseluruhan pada 8 kategori tiket *helpdesk* yaitu Set.Ditjen, Dit. Bmn, Dit.KND, Dit.PNKNL, Dit.PKNSI, Dit.Penilaian, Dit.Lelang, Dit.Humas. Nilai akurasi klasifikasi pada setiap jumlah fitur dan setiap nilai k dapat dilihat pada tabel 6.

Tabel 6 Akurasi klasifikasi TF-ABS (%)

Nilai k	Jumlah fitur					
	5%	10%	15%	20%	25%	30%
k=1	86.81	88.99	89.56	89.56	89.09	88.52
k=2	87.57	88.99	89.66	89.56	88.99	88.61
k=3	87.95	89.09	90.04	89.94	89.94	89.37
k=4	87.58	89.28	89.66	89.75	89.18	88.71
k=5	87.00	89.09	88.99	88.99	88.33	88.52
k=6	87.19	89.28	89.75	89.75	89.09	88.52
k=7	87.19	88.71	88.99	88.05	88.24	87.48
k=8	87.67	87.57	89.18	88.33	88.24	87.76
k=9	87.38	88.71	88.80	88.14	88.14	87.76
k=10	87.19	88.52	88.43	87.76	87.76	87.29
k=11	86.91	88.43	88.05	87.67	87.67	86.81
k=12	86.81	87.96	87.48	87.48	87.57	86.62
k=13	86.62	87.57	87.10	86.62	86.72	85.96
k=14	86.81	87.29	86.62	87.10	86.91	86.24
k=15	86.62	86.62	86.53	86.24	86.34	85.96
k=16	86.91	86.24	87.00	86.05	85.86	85.67
k=17	86.72	86.43	86.62	85.67	85.96	85.20
k=18	86.43	86.34	86.24	85.58	85.77	85.39
k=19	85.77	86.34	85.58	85.10	85.48	85.10
k=20	86.05	86.24	85.10	85.20	85.39	84.54
k=21	85.77	86.15	85.39	84.72	85.01	84.06
k=22	85.29	85.96	84.91	84.72	85.01	84.54
k=23	85.39	85.67	85.10	84.25	85.01	83.59
k=24	85.30	85.29	84.63	84.06	84.63	83.30
k=25	85.10	85.10	84.63	84.16	84.54	83.21
k=26	85.01	84.91	84.44	83.97	84.16	83.49
k=27	84.91	84.72	83.97	83.78	84.16	83.11
k=28	84.72	84.91	84.16	83.68	84.07	83.11
k=29	84.91	84.63	83.87	83.87	84.06	83.02
k=30	84.72	84.44	83.78	83.68	83.87	83.40



Gambar 3 Akurasi Klasifikasi KNN TF-ABS

Tabel 6 dan Gambar 3 menunjukkan bahwa akurasi meningkat mulai penggunaan jumlah fitur 5% sampai dengan 15%, pada titik ini diperoleh akurasi tertinggi, kemudian akurasi mengalami penurunan setelah jumlah fitur diubah menjadi 20%, 25% dan 30%. Penggunaan jumlah fitur yang terlalu sedikit mengakibatkan adanya fitur-fitur yang memiliki nilai pembobotan kata yang cukup tinggi yang terabaikan, sebaliknya penggunaan jumlah fitur yang terlalu banyak mengakibatkan adanya fitur-fitur yang memiliki nilai pembobotan kata yang rendah diperhitungkan sehingga mengakibatkan akurasi rendah. Akurasi juga mengalami perubahan dengan nilai k yang berbeda-beda. Akurasi meningkat pada k=1 hingga k=3, pada titik ini diperoleh akurasi tertinggi, kemudian akurasi menurun setelah nilai k diubah menjadi k=5 sampai dengan k=30. Penggunaan nilai k yang semakin besar dapat memperbesar jumlah kesalahan prediksi disebabkan karena sebaran data yang tidak merata pada *data testing* sehingga mengakibatkan akurasi menurun. Akurasi tertinggi klasifikasi diperoleh pada jumlah fitur 15% dan k=3 sebesar 90,04% yang berarti bahwa penggunaan jumlah fitur dan k tersebut adalah yang paling sesuai dengan model klasifikasi. Model pengujian *confusion matrix* data *testing* 1.054 dokumen yang dihasilkan klasifikasi menggunakan jumlah fitur 15% dan k=3 dapat dilihat pada tabel 7.

Tabel 7 Confusion Matrix Algoritma KNN TF-ABS

	Actual							
Setditjen	130	7	27	0	7	0	2	0
Dit.Humas	3	5	3	0	0	0	0	0
Dit.PKNSI	6	5	473	0	9	3	1	5
Dit.KND	0	0	0	4	0	0	0	0
Dit.BMN	4	1	1	2	112	0	1	1
Dit.Penilaian	1	0	1	0	0	3	0	0
Dit.PNKNL	0	0	1	2	1	0	8	0
Dit.Lelang	1	0	10	0	0	0	0	214

Berdasarkan tabel 7 terdapat rincian jumlah *True Positive (TP)* Setditjen 130, *False Negative (FN)* Setditjen 15, *False Positive (FP)* Setditjen 43, *True Negative (TN)* Setditjen 866. Jumlah *True Positive (TP)* Dit.Humas 5, *False Negative (FN)* Dit.Humas 13, *False Positive (FP)* Dit.Humas 6, *True Negative (TN)* Dit.Humas 1030. Jumlah *True Positive (TP)* Dit.PKNSI 473, *False Negative (FN)* Dit.PKNSI 43, *False Positive (FP)* Dit.PKNSI 29, *True Negative (TN)* Dit.PKNSI 509. Jumlah *True Positive (TP)* Dit.KND 4, *False Negative (FN)* Dit.KND 4, *False Positive (FP)* Dit.KND 0, *True Negative (TN)* Dit.KND 1046. Jumlah *True Positive (TP)* Dit.BMN 112, *False Negative (FN)* Dit.BMN 17, *False Positive (FP)* Dit.BMN 10, *True Negative (TN)* Dit.BMN 915. Jumlah *True Positive (TP)* Dit.Penilaian 3, *False Negative (FN)* Dit.Penilaian 3, *False Positive (FP)* Dit.Penilaian 2, *True Negative (TN)* Dit.Penilaian 1046. Jumlah *True Positive (TP)* Dit.PNKNL 8, *False Negative (FN)* Dit.PNKNL 4, *False Positive (FP)* Dit.PNKNL 4, *True Negative (TN)* Dit.PNKNL 1038. Jumlah *True Positive (TP)* Dit.Lelang 214, *False Negative (FN)* Dit.Lelang 6, *False Positive (FP)* Dit.Lelang 11, *True Negative (TN)* Dit.Lelang 823. Perhitungan akurasi *confusion matrix* adalah sebagai berikut:

$$\begin{aligned} \text{Akurasi} &= (130+5+473+4+112+3+8+214)/1054 \times 100\% \\ &= 90,04\% \end{aligned}$$

4. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, klasifikasi teks *helpdesk* dapat dilakukan menggunakan algoritma KNN dan metode pembobotan TF-ABS dengan hasil akurasi yang cukup baik. Percobaan menggunakan variasi jumlah fitur dan nilai k diperoleh akurasi tertinggi sebesar 90,04% menggunakan jumlah fitur sebanyak 15% dan nilai k=3, sedangkan akurasi terendah sebesar 84,54% diperoleh menggunakan jumlah fitur sebanyak 30% dan nilai k=19. Metode klasifikasi ini dapat digunakan untuk membantu proses disposisi tiket *helpdesk* ke unit tujuan dengan cepat.

Saran untuk penelitian selanjutnya menggunakan metode yang dapat menangani data yang tidak seimbang pada dataset sehingga dapat memperkecil perbedaan jumlah data training pada setiap kategori dengan harapan untuk meningkatkan akurasi klasifikasi.

DAFTAR PUSTAKA

- [1] M. Altintas and A. C. Tantug, "Machine Learning Based Ticket Classification In Issue Tracking Systems," vol. 21, no. 3, pp. 33–44, 2014.
- [2] T. A. Herawan, Y. H. Chrisnanto, and A. I. Hadiana, "Klasifikasi Helpdesk Universitas Jenderal Achmad Yani Menggunakan Concept Frequency-Inverse Document Frequency (CF-IDF) dan K-Nearest Neighbor," *Pros. SNST*, vol. 7, pp. 108–113, 2016.
- [3] C. F. Suharno, M. A. Fauzi, and R. S. Perdana, "Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors Dan Chi-square," *Syst. Inf. Syst. Informatics J.*, vol. 3, no. 1, pp. 25–32, 2017, doi: 10.29080/systemic.v3i1.191.
- [4] M. R. A. Nasution and M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *J. Inform.*, vol. 6, no. 2, pp. 226–235, 2019, doi: 10.31311/ji.v6i2.5129.
- [5] M. A. Kurniawan, Y. Sibaroni, and K. L. Muslim, "Kategorisasi Berita Menggunakan Metode Pembobotan TF.ABS dan TF.CHI," *Indones. J. Comput.*, vol. 3, no. 2, p. 83, 2018, doi: 10.21108/indoic.2018.3.2.236.
- [6] V. C. Gandhi and J. A. Prajapati, "Review on Comparison between Text Classification Algorithms," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 1, no. 3, pp. 1–4, 2012.
- [7] A. H. Aliwy and E. H. A. Ameer, "Comparative study of five text classification algorithms with their improvements," *Int. J. Appl. Eng. Res.*, vol. 12, no. 14, pp. 4309–4319, 2017, doi: 10.113/J.0973-4562.

- [8] M. A. Rosid, A. S. Fitriani, I. Ratna, and I. Astutik, "Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi," 2020, doi: 10.1088/1757-899X/874/1/012017.
- [9] L. A. Matsunaga and N. F. F. Ebecken, "Two Novel Weighting for Text Categorization," in *Data Mining IX - Data Mining, Protection, Detection and other Security Technologies, IX.*, A. Zanasi, D. Almorza Gomar, N. F. . Ebecken, and C. . Brebbia, Eds. Rio de Janeiro, Brazil: WITPRESS, 2008, pp. 105–114.
- [10] J. Li *et al.*, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017, doi: 10.1145/3136625.
- [11] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," *Int. Conf. Electr. Electron. Optim. Tech. ICEEOT 2016*, no. March 2016, pp. 61–66, 2016, doi: 10.1109/ICEEOT.2016.7754750.
- [12] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques - third edition*. 2012.
- [13] D. Yuliana and C. Supriyanto, "Klasifikasi Teks Pengaduan Masyarakat Dengan Menggunakan Algoritma Neural Network," *UPI YPTK J. KomTekInfo*, vol. 5, no. 3, pp. 92–116, 2019.
- [14] L. A. Andika, P. A. N. Azizah, and R. Respatiwan, "Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier," *Indones. J. Appl. Stat.*, vol. 2, no. 1, p. 34, 2019, doi: 10.13057/ijas.v2i1.29998.