Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 1, No. 10, Oktober 2017, hlm. 1000-1007 http://j-ptiik.ub.ac.id

Klasifikasi Teks Bahasa Indonesia pada Dokumen Pengaduan Sambat Online menggunakan Metode K-Nearest Neighbors (K-NN) dan Chi-Square

Claudio Fresta Suharno[1], M. Ali Fauzi[2], Rizal Setya Perdana[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]claudiofresta@gmail.com, [2]moch.ali.fauzi@ub.ac.id, [3]rizalespe@ub.ac.id

Abstrak

K-Nearest Neighbors (K-NN) merupakan metode klasifikasi yang mudah untuk dipahami. Akan tetapi metode tersebut memiliki beberapa kekurangan, salah satunya adalah metode ini menggunakan semua fitur pada perhitungan klasifikasi. Hal ini dapat mengakibatkan rendahnya nilai akurasi yang dihasilkan yang disebabkan banyaknya fitur tidak penting yang masuk dalam perhitungan klasifikasi. Oleh karena itu, seleksi fitur digunakan sebagai salah satu cara untuk mengatasi kekurangan tersebut. Teknik seleksi fitur mengurangi jumlah fitur yang tidak relevan dalam klasifikasi teks. Metode seleksi fitur yang digunakan adalah menggunakan metode Chi-Square untuk menghitung tingkat dependensi fitur. Proses yang dilakukan adalah mengumpulkan dokumen latih dan dokumen uji, melakukan tahap preprocessing dan seleksi fitur, kemudian dilakukan klasifikasi, dan pada tahap akhir dilakukan pengujian dan analisis terhadap hasil klasifikasi oleh sistem terkait nilai precision, recall, dan F-Measure. Dari 16 hasil pengujian, didapatkan nilai *precision* dan *recall* terbaik didapatkan dengan nilai masing-masing adalah 90% dan 78% pada k = 15 dengan seleksi fitur sebesar 25%. Sedangkan hasil dari F-Measure terbaik didapatkan dengan nilai 78% pada k = 15 dan k = 5 dengan seleksi fitur sebesar 25%. Dari penelitian ini dihasilkan bahwa seleksi fitur dapat meningkatkan nilai F-Measure dalam klasifikasi teks berbahasa Indonesia pada dokumen pengaduan SAMBAT Online dengan menggunakan metode klasifikasi K-Nearest Neighbors.

Kata Kunci: *K-NN, Seleksi Fitur, Chi-Square, Dokumen Pengaduan, Klasifikasi Teks*

Abstract

K-Nearest Neighbors (K-NN) is one classification method that easy to learn. Although, this method has some drawbacks, one of them is this classification could provide a low accuracy casued by a large feature space with irrelevant features among them. Because of that drawback, feature selection is applied to reduce the feature space by reducing number of irrelevant features in text classification. Selection feature method that being used in this experiment is using Chi-Square method. Using Chi-Square method to select important features by measuring dependency level of each feature across classes and documents. The process including in this experiment is collecting training and testing documents, text preprocessing and feature selection, and classification. After classification is being done by the system, we make an observation and analysis towards classification result, including precision, recall, and F-Measure value. From 16 evaluations, the best precision and recall score obtained with 90% precision and 78% recall on k = 15 using 25% feature selection used. While the best F-Measure score obtained with 78% F-Measure on k = 15 and k = 5 using 25% feature selection used. From this experiment, its appear that feature selection take effect in increasing F-Measure value in text classification of SAMBAT Online complaint documents in bahasa using K-Nearest Neighbors classification method.

Keywords: *K-NN, Feature Selection, Chi-Square, Complaint Documents, Text Classification*

1. PENDAHULUAN

Indonesia merupakan negara dengan nilai penetrasi teknologi informasi tertinggi ketiga di Asia. Hal tersebut dapat diakomodir secara positif oleh pemerintah Indonesia untuk

mendukung administrasi pemerintahan, atau bisa disebut juga dengan e-Government, atau eGov (Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi, 2015). Dengan perkembangan pemanfaatan komputer dan internet dalam teknologi informasi, istilah e-Government sering dikonotasikan dengan pemanfaatan internet dalam urusan-urusan pemerintahan berikut pelayanan publiknya kepada masyarakat, termasuk transparansi dan kebijakan regulasinya. Dengan adanya penerapan teknologi informasi dan komunikasi pada administrasi pemerintahan, akan memberikan dampak positif pada aspek perekonomian dan demokratisasi.

Pemerintah Kota Malang selaku salah satu penyelenggara pemerintah juga tak luput dalam penerapan eGov. Salah satu penerapan eGov oleh Pemerintah Kota Malang untuk dapat mendukung tujuan demokratisasi adalah dengan disediakannya platform bagi masyarakat Kota Malang untuk dapat menyuarakan aspirasinya terkait permasalahan yang dialami di Kota Malang dalam bentuk situs web SAMBAT Online. Melalui sistem tersebut, pengguna dapat mengirimkan saran, kritik, pertanyaan, atau pengaduan seputar Pemerintah Kota Malang melalui situs web secara langsung atau melalu pesan singkat yang dikirimkan kepada nomor yang telah disediakan.

Dalam penerapan platform SAMBAT Online, setiap pengaduan yang masuk akan dimoderasi sesuai dengan peraturan dan ketentuan yang berlaku. Pengelola sistem berhak untuk tidak menayangkan atau tidak membalas pengaduan yang masuk apabila menyimpang dari prosedur. Sistem pengelolaan dan pengelompokan laporan pada tiap Satuan Kerja Perangkat Daerah (SKPD) masih dilakukan secara manual.

Untuk mempersingkat waktu pekerjaan pengelola sistem tertinggi dalam pemilahan laporan menuju SKPD yang sesuai, dapat menerapkan metode klasifikasi teks. Metode K-Nearest Neighbors (K-NN) merupakan salah satu dari banyak metode klasifikasi teks yang populer yang memberikan hasil yang akurat dan mudah dimengerti (Alhutaish & Omar, 2015).

Akan tetapi metode K-NN memiliki kekurangan, salah satunya adalah pada metode ini menggunakan semua fitur untuk perhitungan besar jarak, yang menyebabkan beratnya komputasi menggunakan metode K-NN[4]. Untuk mengatasi permasalahan luasnya fitur yang digunakan pada metode K-NN, dapat dilakukan tahap seleksi fitur sebelum dokumen diklasifikasikan. Beberapa teknik seleksi fitur yang dapat dilakukan adalah menggunakan metode Document Frequency, Mutual Information, Information Gain, Chi-Square Statistic, dan Term Frequency – Inverse Document Frequency.

Metode yang digunakan pada penelitian ini adalah menggunakan metode Chi-Square Testing. Metode ini merupakan metode pengujian data hipotesis diskrit secara statistik, yang mengevaluasi korelasi antar dua variabel, dan menentukan apakah dua variabel tersebut memiliki hubungan atau tidak (Snedecor & Cohran, 1989). Dalam subjek klasifikasi teks, dua variabel dalam Chi-Square Testing tersebut adalah kata dan kategori (Thabtah, et al., 2009).

Hasil penerapan metode K-NN dalam klasifikasi teks yang dilakukan dalam percobaan penggunaan seleksi fitur dan metode machine learning pada klasifikasi teks otomatis dalam bahasa melayu memberikan hasil yang baik bila dibandingkan dengan metode Naive Bayes dan N-gram pada metode seleksi fitur Chi-Square (Alshalabi, et al., 2013).

Dengan memadukan metode K-NN dengan fitur seleksi Chi-Square, diharapkan pada penelitan ini dapat mengetahui performa sistem klasifikasi teks berbasis K-NN dengan menggunakan metode seleksi fitur sebagai penunjang sistem klasifikasi, serta dapat digunakan sebagai acuan untuk membangun model klasifikasi dokumen laporan otomatis pada platform SAMBAT Online.

2. METODE PENELITIAN

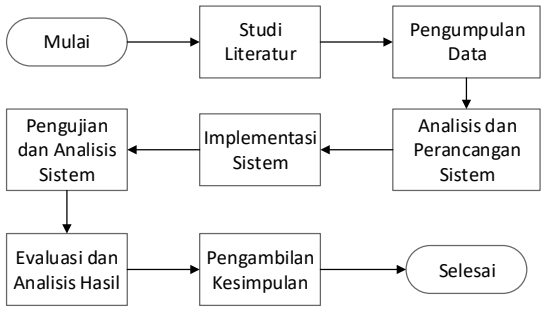
Beberapa tahapan penelitian yang dilakukan meliputi studi literatur, pengumpulan data, analisis dan perancangan sistem, pengambilan implementasi sistem, pengujian dan analisis, evaluasi dan analisis hasil, serta kesimpulan. Alur dari tahapan penelitian tersebut dapat dilihat pada Gambar 1.

2.1. Data Penelitian

Data penelitian yang digunakan merupakan laporan pengaduan dari situs SAMBAT Online yang beralamat di http://sambat.malangkota.go.id dalam bentuk data teks berbahasa Indonesia. Dokumen diambil dari tiga SKPD, antara lain SKPD Dinas Perhubungan (Dishub), Dinas Pekerjaan Umum Perumahan dan Pengawasan Bangunan (DPUPPB), dan Dinas Kebersihan dan

Pertamanan (DKP).

Masing masing data berjumlah sebanyak 100 dokumen untuk SKPD Dishub, 67 dokumen untuk SKPD DPUPPB, dan 37 dokumen untuk SKPD DKP. Koleksi dokumen pada tiap kategori akan dipecah dengan rasio jumlah 80% untuk data latih, dan 20% untuk data uji (Suthaharan, 2015).



Gambar 1. Alur metodologi penelitian

2.2. SAMBAT Online

SAMBAT Online, yang merupakan akronim dari Sistem Aplikasi Masyarakat Bertanya Terpadu Online, merupakan fasilitas bagi masyarakat kota Malang untuk mengirimkan kritik, saran, pertanyaan, atau pengaduan seputar Pemerintah Kota Malang. SAMBAT Online merupakan salah satu jalur yang disediakan oleh Dinas Komunikasi dan Informatika (Diskominfo) untuk memfasilitasi pengaduan melalui jalur online. Pengaduan yang dilaporkan pada SAMBAT Online dapat dikirimkan melalui situs web secara langsung atau melalui pesan singkat.

2.3. Klasifikasi Teks

Klasifikasi teks adalah sebuah pekerjaan untuk menentukan apakah sebuah dokumen adalah milik dari sebuah kategori yang telah ditentukan sebelumnya (Ramya & Pinakas, 2014). Tahapan dalam klasifikasi teks antara lain adalah (Nikhath, et al., 2016) :

1. *Preprocessing.*

Merupakan tahapan untuk merpresentasikan dokumen dalam bentuk fitur vektor, yang berarti harus memisahkan teks menjadi kata terpisah (Ramya & Pinakas, 2014). Dalam tahap ini dilakukan penghapusan *stopwords* pada dokumen, *cleaning*, dan *stemming*.

2. Rekayasa Fitur

Pada tahap ini merupakan tahapan latih yang yang terdiri dari tahapan seleksi fitur, *dictionary construction*, dan *feature weighting*. Tujuan dari rekayasa fitur adalah untuk menghapus semua fitur yang tidak relevan dan selalu muncul pada semua dokumen (Nikhath, et al., 2016).

3. Generasi Model Klasifikasi

Tahap ini merupakan tahap untuk membangun algoritme klasifikasi K-Nearest Neighbor (K-NN) berdasarkan hasil pelatihan oleh dokumen sebelumnya yang akan digunakan untuk mengklasifikasikan dokumen yang tidak diketahui kategorinya.

4. Pengkategorian Dokumen

Merupakan tahapan untuk melakukan klasifikasi dari dokumen baru yang tidak diketahui asal kategori dari dokumen tersebut, dengan catatan bahwa dokumen baru tersebut telah melewati tahap *preprocessing* dan *feature weighting*.

2.4. Seleksi Fitur

Seleksi fitur merupakan bagian dari tahap *preprocessing*, yang merupakan tahap untuk menyeleksi fitur apa yang paling penting. Tahap *preprocessing* dilakukan karena pengurangan fitur yang telah dilakukan pada tahap penghilangan *stopword* dan *stemming* masih dirasa kurang (Nejad, et al., 2014). Seleksi fitur dilakukan dengan menyimpan kata dengan nilai tertinggi sesuai dengan pengukuran derajat penting tidaknya suatu kata yang telah dilakukan sebelumnya (Khan, et al., 2010).

Tujuan dari seleksi fitur adalah untuk meningkatkan performa klasifikasi teks dengan menghilangkan fitur yang dianggap tidak relevan dalam klasifikasi untuk mengurangi dimensi dari himpunan (Alhutaish & Omar, 2015). Teknik seleksi fitur yang digunakan dalam penelitian ini adalah menggunakan teknik *filters* yang menggunakan perhitungan algoritme tersendiri untuk mengevaluasi kemampuan fitur untuk membedakan tiap kelas. Pada penelitian ini menggunakan metode Chi-Square.

2.5. Chi-Square

Chi-Square Testing merupakan metode statistika pengujian hipotesis data diskrit yang mengevaluasi korelasi antar dua variabel dan menentukan apakah variabel tersebut tidak berkaitan atau saling terkait (Snedecor & Cohran, 1989). Pada tes keterkaitan, ketika diterapkan pada populasi suatu subjek, menentukan apakah subjek tersebut terkait atau

tidak (Thabtah, et al., 2009). Fungsi dari Chi-Square *testing* dapat dilihat pada Persamaan 1.

$$\chi^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

t : Kata
c : Kelas/Kategori
N : Jumlah dokumen latih
A : Jumlah banyaknya dokumen pada kategori *c* yang memuat *t*
B : Jumlah banyaknya dokumen bukan kategori *c* yang memuat *t*
C : Jumlah banyaknya dokumen pada kategori *c* yang tidak memuat *t*
D : Jumlah banyaknya dokumen bukan kategori *c* yang tidak memuat *t*

Untuk dapat melakukan seleksi fitur yang tidak dipakai berdasarkan nilai Chi-square dari sebuah kata terhadap kategori, diperlukan nilai Chi-square tunggal dari kata. Untuk dapat mengetahui nilai Chi-square tunggal dari suatu kata diperoleh dengan menjumlahkan nilai Chi-square tiap kata antar kategori. Fungsi untuk mendapatkan nilai Chi-square tunggal tiap kata dapat dilihat pada Persamaan 2.

$$X^2(t) = \sum_{c=1}^{k} x^2(t,c) \quad (2)$$

t : Kata
c : Kelas/Kategori

Setelah nilai Chi-Square pada tiap kata diketahui, dilakukan pengurutan kata berdasarkan nilai Chi-Square tertinggi hingga terendah. Hal ini menandakan bahwa semakin besar nilai Chi-Square, semakin dependen suatu fitur, dan semakin penting fitur tersebut untuk digunakan dalam proses klasifikasi.

2.6. *Similarity Measure*

Pada ilmu komputer, *similarity measure* adalah fungsi nilai real untuk mengukur derajat kesamaan antara dua objek (Hugget, 2009). Fungsi perhitungan derajat kesamaan (*similarity function*) yang sering digunakan pada klasifikasi menggunakan metode K-NN antara lain adalah Cosine Similarity, Jaccard Similarity, dan Dice Similarity (Alhutaish & Omar, 2015). Fungsi perhitungan derajat kesamaan yang digunakan pada penelitian ini adalah menggunakan derajat kesamaan Cosine Similarity. Cosine Similarity mengukur kesamaan antar dokumen dengan mencari nilai derajat kosinus antar vektor pada ruang kata. Perhitungan derajat kesamaan antar dokumen *Di* dengan dokumen *Dj* menggunakan fungsi Cosine Similarity dapat dilihat pada Persamaan 3.

$$Sim_{cosine}(D_i, D_j) = \frac{\sum_{k=1}^{m}(W_{ik} \times W_{jk})}{\sqrt{\sum_{k=1}^{m}(W_{ik})^2 \times \sum_{k=1}^{m}(W_{jk})^2}} \quad (3)$$

m : Banyaknya kata unik pada dokumen pada kategori
Di : Dokumen uji
Dj : Dokumen latih
W_{ik} : Bobot nilai dari elemen ke-*k* dari vektor kata *Di*
W_{jk} : Bobot nilai dari elemen ke-*k* dari vektor kata *Dj*

2.7. K-Nearest Neighbor (K-NN)

Metode KNN merupakan algoritme berbasis pembelajaran cepat yang mengategorikan objek berdasarkan ruang fitur terdekat pada himpunan latih (Han, et al., 2001). Himpunan latih dipetakan dalam ruang fitur yang multidimensi. Ruang fitur terbagi dengan basis area sesuai dengan kategori himpunan latih. Titik dari ruang fitur ditetapkan sebagai bagian dari kategori apabila titik tersebut sering dekat dengan kategori tertentu pada data latih dengan *k* terdekat. Pada penelitian ini, menggunakan Cosine Similarity untuk menghitung jarak kedekatan antar titik ruang fitur.

Pada fase pelatihan menggunakan metode K-NN, hanya terdiri dari penyimpanan vektor fitur dan kategori dari set data latih. Pada fase klasifikasi, jarak dari vektor baru, yang merepresentasikan dokumen masuk baru, menuju semua vektor yang telah tersimpan dihitung besar jaraknya dan sampel *k* terdekat akan dipilih. Penentuan kategori dari dokumen tersebut diprediksi berdasarkan titik terdekat yang telah ditentukan pada kategori tertentu.

Pada penelitian ini, sistem klasifikasi K-NN menggunakan aturan *sum*, dengan menjumlahkan nilai Cosine Similarity dokumen pada tiap kategori pada subhimpunan *k*, dengan dipilih kategori dengan nilai sum Cosine Similarity tertinggi. Untuk fungsi penilaian skor dari K-NN dengan aturan *sum* dapat dilihat pada Persamaan 4.

$$Score(D, C_i) = \sum_{D_j \in KNN_D} sim(D_j|D) \times \delta(D_j, C_i) \quad (4)$$

D_j : Dokumen latih ke-j
C_i : Kategori ke-i
D : Dokumen uji
$sim(D_j|D)$: Nilai derajat kesamaan antara dokumen uji dengan dokumen latih
$\delta(D_j, C_i)$: Bernilai 1 jika kategori D_j adalah berasal dari kategori C_i dan bernilai 0 jika kategori D_j adalah bukan dari kategori C_i

2.8. Alur Kerja Sistem

Pada penelitian klasifikasi teks bahasa Indonesia pada dokumen pengaduan SAMBAT Online menggunakan metode K-Nearest Neighbors (K-NN) dan Chi-Square, sistem dapat memproses data latih dan data uji sesuai dengan tahapan klasifikasi teks. Alir kerja sistem secara umum dapat dilihat pada Gambar 2.

Tahap dari alur kerja sistem terdiri dari beberapa bagian utama:

1. *Preprocessing*

Pada tahap ini, dilakukan proses *filtering*, *stemming*, dan tokenisasi dari dokumen latih dan dokumen uji. Keluaran dari tahap ini adalah terbentuknya representasi fitur dalam bentuk *bag of words.*

2. Seleksi Fitur

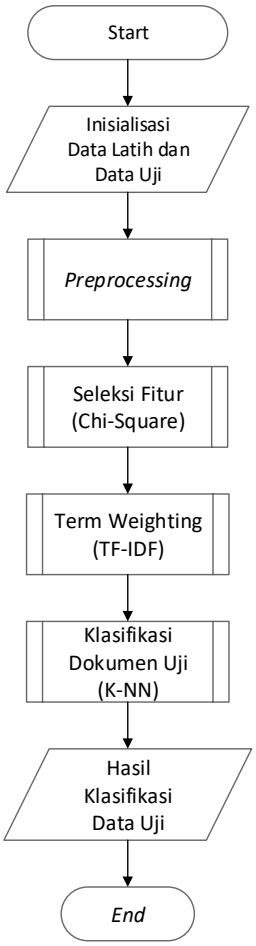
Pada tahap ini, dilakukan proses perhitungan nilai Chi-Square dari tiap fitur yang ada pada *bag of words.* Dengan didapatkannya nilai Chi-Square dari tiap kata pada tiap kategori, dapat dilakukan eliminasi fitur-fitur berdasarkan rasio jumlah yang diinginkan. Dengan pilihan rasio jumlah penggunaan fitur sebesar 25%, 50%, 75%, dan 100%.

3. *Term Weighting*

Pada tahap ini, dilakukan perhitungan bobot fitur yang telah terseleksi dari tahap sebelumnya. Nilai bobot yang dihitung antara lain adalah *Term Frequency* (TF), *Document Frequency* (DF), *Inverse Document Frequency* (IDF), nilai TF-IDF, dan nilai kuadrat dari TF-IDF.

4. Klasifikasi Dokumen Uji

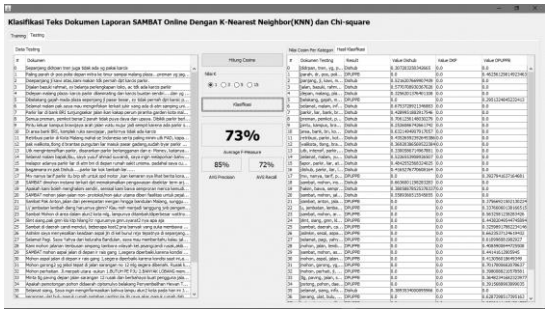
Pada tahap ini, dilakukan perhitungan nilai vektor dan derajat kemiripan dari tiap dokumen, serta melakukan proses klasifikasi dari dokumen uji terhadap dokumen latih pada semua kategori menggunakan metode K-Nearest Neighbor (K-NN).



Gambar 2. Alur Kerja Sistem

2.9. Implementasi

Implementasi sistem berbasis Java yang berjalan pada perangkat *Desktop*. Fitur yang ada pada sistem meliputi *preprocessing*, hitung Chi-Square, seleksi fitur, *term weighting*, hitung vektor dokumen, hitung Cosine Similarity, dan klasifikasi dokumen. Sistem diterapkan berdasarkan alur implementasi yang telah disusun sebelumnya. Untuk hasil dari implementasi sistem dapat dilihat pada Gambar 3. Contoh data keluaran sebagai hasil klasifikasi oleh sistem dapat dilihat pada Tabel 1.



Gambar 3 Implementasi Sistem

Tabel 1 Contoh Data Keluaran Hasil Klasifikasi Sistem

Dokumen Uji	Hasil	Nilai Dishub	Nilai DPUPPB	Nilai DKP
D_1	Dishub	0,307	0	0
D_2	DKP	0	0	0,462
...	...	...	...	...
D_n	DPUPPB	0	0,464	0

3 HASIL DAN PEMBAHASAN

Untuk mendapatkan hasil dari penelitian, dilakukan pengujian terhadap hasil keluaran sistem menggunakan pengujian *precision*, *recall*, dan F-Measure. Ketiga nilai tersebut digunakan sebagai tolak ukur analisis hasil dari penelitian ini. Hasil dari nilai *precision*, *recall*, dan F-Measure pada tiap skenario pengujian dari penelitian dapat dilihat pada Tabel 2.

Tabel 2 Hasil Skenario Pengujian Kombinasi Nilai *k* dan Rasio Jumlah Fitur

k	Rasio Fitur	*Precision*	*Recall*	F-Measure
1	25%	85%	72%	73%
1	50%	71%	72%	72%
1	75%	71%	72%	70%
1	100%	75%	76%	75%
3	25%	87%	74%	74%
3	50%	74%	72%	72%
3	75%	70%	67%	68%
3	100%	72%	71%	71%
5	25%	83%	**78%**	**78%**
5	50%	79%	73%	75%
5	75%	77%	71%	72%
5	100%	69%	67%	67%
15	25%	**90%**	**78%**	**78%**
15	50%	76%	69%	71%
15	75%	86%	70%	70%
15	100%	76%	70%	70%
20	25%	87%	75%	76%
20	50%	78%	71%	73%
20	75%	76%	70%	69%
20	100%	74%	68%	68%

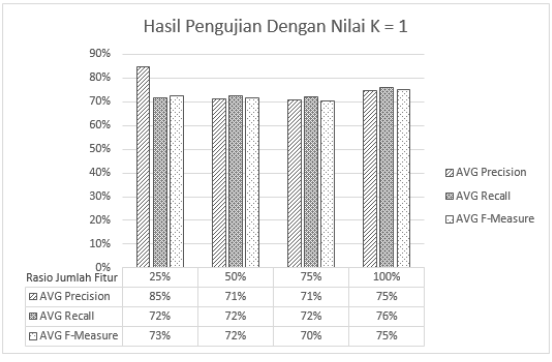
Dari Tabel 2, didapatkan nilai puncak F-Measure adalah sebesar 78% pada nilai rasio penggunaan jumlah fitur sebanyak 25% dengan nilai ketetanggaan *k* sebesar 15 dan 5. Sedangkan untuk *precision* tertinggi adalah sebesar 90% berada pada nilai rasio penggunaan jumlah fitur sebanyak 25% pada nilai ketetanggaan *k* sebesar 15. Dan untuk *recall* tertinggi adalah sebesar 78% berada pada nilai rasio penggunaan fitur sebesar 25% pada nilai ketetanggaan *k* sebesar 15.

3.1. Pengaruh Nilai *k*

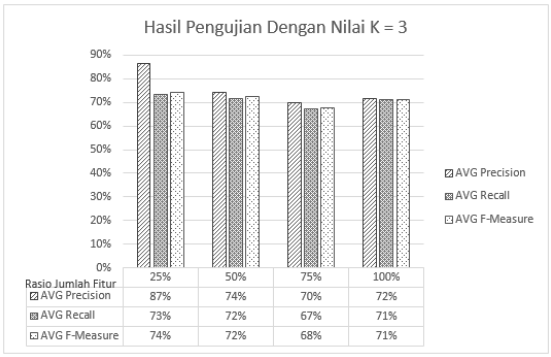
Dari pengujian, dapat diamati pola dari hasil F-Measure akan cenderung semakin menurun seiring dengan bertambahnya rasio jumlah fitur yang digunakan pada semua nilai *k* yang digunakan. Hal ini disebabkan karena model klasifikasi K-NN yang digunakan menggunakan teknik *sum* yang memperhitungkan besar nilai Cosine Similarity pada tiap kategori yang ada pada subhimpunan *k* untuk melakukan klasifikasi. Dengan teknik ini, maka meskipun nilai *k* bernilai besar, hasil klasifikasi tetap menunjukkan pola yang sama dengan nilai *k* yang lain, dikarenakan mempertimbangkan nilai Cosine Similarity dari tetangga data uji, tidak hanya menggunakan kelas terbanyak dari tetangga yang terdekat dari data uji.

3.2. Pengaruh Seleksi Fitur

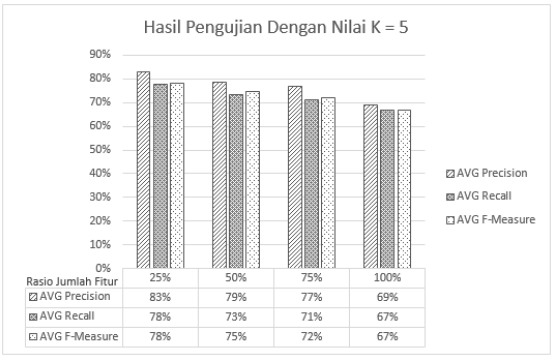
Dari hasil pengujian dapat diamati pola dari hasil F-Measure akan cenderung semakin menurun seiring dengan bertambahnya rasio jumlah fitur yang digunakan. Hasil tersebut dapat diamati dari grafik pada Gambar 4, Gambar 5, Gambar 6, Gambar 7, dan Gambar 8.



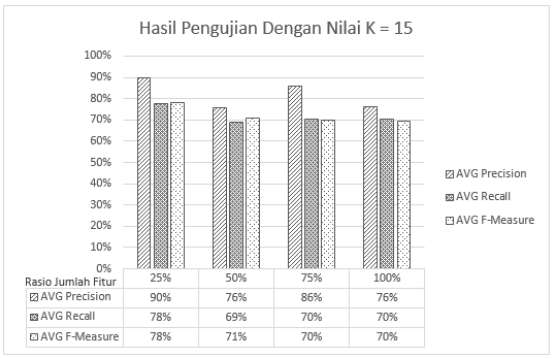
Gambar 4 Grafik Hasil Pengujian Nilai K = 1



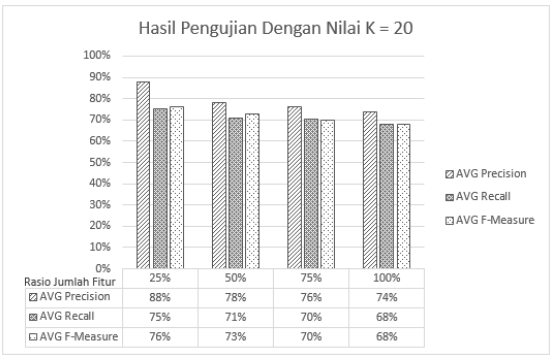
Gambar 5 Grafik Hasil Pengujian Nilai K = 3



Gambar 6 Hasil Pengujian Nilai K = 5



Gambar 7 Hasil Pengujian Nilai K = 15



Gambar 8 Hasil Pengujian Nilai K = 20

Dari fenomena ini dapat diketahui bahwa kumpulan koleksi fitur memiliki banyak fitur yang tidak relevan untuk digunakan dalam perhitungan sistem klasifikasi. Hal ini dapat dibuktikan dengan perbandingan 10 fitur terbaik yang digunakan dengan 10 fitur terendah yang digunakan. Perbandingan fitur ini dapat dilihat pada Tabel 3.

Tabel 3 Perbandingan 10 Fitur Tertinggi dan 10 Fitur Terendah

No.	10 Fitur Tertinggi		10 Fitur Terendah	
	Term	Nilai	Term	Nilai
1	parkir	78,913	kenan	3,314
2	karcis	37,406	dar	3,314
3	jalan	33,705	alias	3,314
4	sampah	32.219	masak	3,314
5	sambat	27,548	nomor	3,314
6	liar	24,764	marak	3,314
7	tukang	24,325	marah	3,314
8	akses	24,200	hotline	3,314
9	dkp	22,779	pikir	3,314
10	tertib	22,576	henti	3,314

Dari Tabel 3 dapat diamati bahwa terbukti semakin semakin tinggi nilai kata, semakin relevan untuk digunakan. Hal ini dapat dibuktikan dengan adanya informasi yang bisa diperoleh dari 10 fitur tertinggi sebagai informasi pengaduan. Hal ini dibuktikan dengan adanya kata ‘parkir’, ‘karcis’, ‘jalan’, ‘sampah’, ‘tukang’, ‘akses’, ‘dkp’ yang bisa jadi merupakan topik dari suatu dokumen pengaduan bila dibandingkan dengan kata ‘nomor’, dan ‘hotline’ pada kelompok 10 fitur terendah.

4 KESIMPULAN

Dari penelitian yang telah dilakukan, dapat diambil kesimpulan dari hasil yang telah didapatkan. Tingkat akurasi *precision*, *recall*, dan F-Measure pada penggunaan metode K-Nearest Neighbor (K-NN) dan Chi-Square pada klasifikasi teks bahasa Indonesia pada dokumen pengaduan berdasarkan bidang SKPD yang sesuai adalah, nilai F-Measure yang dihasilkan cenderung memiliki pola peningkatan yang sama, tidak terpengaruh dengan nilai *k* yang digunakan.

Meningkatnya nilai F-Measure bertolak belakang dengan nilai rasio jumlah fitur yang digunakan. Semakin banyak fitur yang digunakan, nilai F-Measure yang dihasilkan akan cenderung semakin rendah, dan sebaliknya, semakin sedikit fitur yang digunakan, nilai F-Measure yang dihasilkan akan cenderung semakin tinggi. Dari penelitian ini menunjukkan bahwa hasil yang didapatkan dengan menggunakan seleksi fitur lebih baik daripada tanpa adanya proses seleksi fitur. *Precision* dan *recall* terbaik didapatkan pada *k* = 15 dengan seleksi fitur sebesar 25%.

Sedangkan hasil dari F-Measure terbaik didapatkan dengan nilai 78% pada *k* = 15 dan *k* = 5 dengan seleksi fitur sebesar 25%.

5 DAFTAR PUSTAKA

Alhutaish, R. & Omar, N., 2015. Arabic Text Classification using K-Nearest Neighbour Algorithm. The International Arab Journal of Information Technology, 12(2), pp. 190-195.

Alshalabi, H., Tiun, S., Omar, N. & Albared, M., 2013. Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization. International Conference on Electrical Engineering and Informatics, Volume 8C, pp. 734-739.

Han, E.-H., Karypis, G. & Kumar, V., 2001. Text Categorization Using Weighted Adjusted k-Nearest Neighbor. Berlin, Springer.

Hugget, M., 2009. Similarity and Ranking Operation. Dalam: L. Liu & M. T. Ozsu, penyunt. Encyclopedia of Database Systems. New York: Springer, p. 2647.

Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi, 2015. Kementrian Pendayagunaan Aparatur Negara dan Reformasi Birokrasi - Naskah Akademik RUU Egov. [Online] Available at: http://www.menpan.go.id/download/file/4977-naskah-akademik-ruu-egov [Diakses 2 2017].

Khan, A., Baharudin, B., Lee, L. H. & Khan, K., 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, 1(1), pp. 4-20.

Nejad, A. M. B., Hashemi, B. S. M., Sayahi, C. A. & Kiaeimehr, D. B., 2014. Feature Selection Techniques for Text Classification. International journal of Computer Science & Network Solutions , 2(1), pp. 90-94.

Nikhath, A. K., Subrahmanyam, K. & Vasavi, R., 2016. Building a K-Nearest Neighbor Classifier for Text Categorization. (IJCSIT) International Journal of Computer Science and Information Technologies, 7(I), pp. 254-256.

Ramya, M. & Pinakas, J. A., 2014. Different Type of Feature Selection for Text Classification. International Journal of Computer Trends and Technology (IJCTT), 10(2), pp. 102-107.

Salton, G., 1983. An Introduction to Modern Information Retrieval. s.l.:Mc Graw Hill.

Snedecor, W. & Cochran, W., 1989. Statistical Methods. 8th penyunt. s.l.:Iowa State University Press.

Suthaharan, S., 2015. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. s.l.:Springer.

Thabtah, F., Eljinini, M., Zamzeer, M. & Hadi, W., 2009. Naïve Bayesian Based on Chi Square to Categorize Arabic Data. Communications of the IBIMA, Volume 10, pp. 158-163.