TEXT MINING UNTUK KLASIFIKASI PENGADUAN PADA SISTEM LAPOR MENGGUNAKAN METODE C4.5 BERBASIS *FORWARD SELECTION*

Ali Sofyan[1], Stefanus Santosa[2]
[12]Pasca Sarjana Teknik Informatika Universitas Dian Nuswantoro

ABSTRACT

Report of public complaints on the site becomes a medium of communication between the community and government agencies. The number of incoming documents every day be a source of information to measure the services of government agencies. Classification of documents is very important to do other than to ensure that the intended objectives of the institution, as well as to classify complaints fit the category. C4.5 algorithm is one of the algorithms that can be used for classification. There were some complaints classification research. This study aims to apply the classification of complaints by algritma C4.5 with a selection of features to improve the accuracy of classification. Results of experiments with methods of research division of the number of datasets, cross validation, classification with and without features. The test results obtained by testing the value of the best accuracy with 550 documents with forward selection, with cross valiadtion 9folds with a value of 85.27%. precission 87.8% and 85.3% recall

Keywords : complaint, classification, C4.5, forward selection

1. PENDAHULUAN

Tingkat kemajuan perekonomian Indonesia yang masih tergolong sebagai negara yang sedang membangun[1] setelah didera krisis menuntut campur tangan pemerintah dalam pemulihan dan pergerakan kegiatan perekonomian masyarakat.

Proses perubahan sosial atau pembangunan tersebut perlu dilakukan secara terencana, terkoordinasi, konsisten, dan berkelanjutan[2] melalui peran pemerintah bersama masyarakat dengan memperhatikan kondisi ekonomi, perubahan-perubahan sosio-politik, perkembangan sosial-budaya yang ada, perkembangan ilmu dan teknologi, dan perkembangan dunia internasional atau globalisasi.

Dukungan rakyat terhadap program kerja pemerintah sangat diperlukan sehingga program kerja yang direncanakan dalam jangka panjang ataupun jangka pendek, bisa terealisasi. Selain sebagai perencana program, pemerintah juga berperan dalam pengawasan pelaksanaan program kerja. Keterbatasan pengawasan pemerintah secara langsung dalam pelaksanaan program kerja menjadi kendala, untuk itu peran serta warga masyarakat dalam hal pengawasan sangat diperlukan.

Portal layanan aspirasi dan pengaduan *online* rakyat atau yang disingkat LAPOR, bisa mempermudah pemerintah dalam hal pengawasan dan kontroling. Partisipasi dan interaksi warga masyarakat secara langsung dengan pemerintah lebih mudah dilaksanakan. LAPOR adalah sebuah sarana pengaduan yang di jalankan oleh Kantor Staf Kepresidenan[3].

Dari portal LAPOR banyak sekali rupa laporan dalam bentuk teks. Kemudahan dalam membuat laporan oleh masyarakat, mengakibatkan jumlah laporan bertambah dengan cepatnya. Proses pelaporan pada LAPOR diawali dengan registrasi pelapor, setelah itu pelapor melakukan pelaporan dengan cara mengisi form pelaporan dan memilih klasifikasi laporan. Laporan yang telah dikirim akan menunggu untuk diverifikasi oleh administrator LAPOR. Proses verifikasi membutuhkan waktu yang cukup untuk sampai pada instansi kementrian atau lembaga atau departemen terkait.

Melimpahnya laporan bentuk teks mengalami penambahan yang sangat pesat. Jumlah data laporan

yang bertambah setiap hari membutuhkan juga waktu lama dalam mempersiapkan verifikasi laporan. Persiapan verifikasi laporan dilakukan untuk bisa memilah laporan yang sesuai untuk diteruskan ke instansi terkait. Laporan dalam bentuk teks membutuhkan proses klasifikasi untuk persiapan verifikasi.

Penggalian informasi dari banyaknya laporan pengaduan agar menjadi informasi yang berguna, menjadi objek penelitian *text mining*. Proses klasifikasi teks atau *text clasification* adalah salah satu cabang dalam teknologi *text mining*.*Text Mining* dapat didefinisikan sebagai suatu proses menggali informasi. User berinteraksi dengan sekumpulan dokumen teks menggunakan *tools* analisis untuk memperoleh informasi yang diperlukan yang salah satunya adalah melalui proses klasifikasi[4].

Banyak metode klasifikasi yang bisa digunakan untuk proses teks mining. Salah satunya adalah algoritma C4.5. Algoritma C4.5 atau disebut juga algoritma *decision tree* merupakan metode klasifikasi dan prediksi yang sangat dikenal[5]. Dalam penelitian tentang pengklasifikasian pengaduan masyarakat pada laman Kantor Pertanahan Kota Surabaya dengan metode pohon keputusan didapatkan hasil penelitian bahwa klasifikasi dengan membandingkan 3 (tiga) algoritma pohon keputusan yaitu C4.5, *RandomForest*, dan *RandomTree*, hasil tertinggi akurasinya adalah algoritma C4.5. dengan nilai akurasi yang didapat sebesar 76,39%[6]. Namun ada beberapa kekurangan diantaranya akurasinya yang masih rendah, dan penggunaan *dataset* yang sedikit. Kemudian penelitian tentang klasifikasi pengaduan dengan membandingkan beberapa algoritma, didapatkan hasil akurasi tertinggi SVM dengan nilai 82,61%.[7], [8]

Seleksi fitur dapat mengoptimalkan metode klasifikasi[9]. Seleksi fitur dapat diartikan mengurangi fitur yang besar, misalnya dengan membuang atribut yang kurang relevan[10]. Algoritma seleksi fitur dapat dibedakan menjadi dua tipe, yaitu *filter* dan *wrapper*[11]. Seleksi fitur dengan pendekatan *filter* dilakukan dengan memilih *set* fitur dalam *preprocesing* data itu sendiri tanpa menggunakan algoritma klasifikasi. Sedangkan pada *wrapper* proses set fitur dilakukan dengan menggunakan algoritma klasifikasi. Sehingga hasil pada pendekatan seleksi fitur dengan *wrapper* lebih baik, karena proses pemilihan fitur dilakukan dengan menggunakan algoritma klasifikasi yang telah ditentukan sebelumnya. Contoh dari seleksi fitur tipe *wrapper* adalah *forward selection* dan *backward elimination*[12]. *Forward selection* memasukan variabel signifikan ke dalam model sampai tidak ada lagi variabel independen yang bisa dimasukkan kembali ke dalam model[13]. Pada penelitian di atas tidak dilakukan penambahan seleksi fitur.

Pada penelitian sebelumnya[6]–[8] hasil akurasi tertinggi yang diperoleh yaitu 82,61%. Penggunaan *dataset* yang sedikit, jumlah fitur yang sedikit, dan hasil *rule* klasifikasi yang panjang berpengaruh pada hasil penelitian. Dengan penambahan *dataset* yang lebih banyak, jumlah fitur yang beragam dan signifikan diharapkan dapat meningkatkan akurasi. Namun tidak semua fitur memiliki pengaruh yang relevan dalam menghasilkan informasi yang akurat. Seleksi fitur merupakan pemilihan fitur dalam tahap praproses yang bertujuan untuk mengurangi dimensi data, menghilangkan data yang tidak relevan serta untuk meningkatkan hasil akurasi[14]. Dengan menerapkan seleksi fitur diharapkan akan meningkatkan akurasi dan *rule* klasifikasi menjadi lebih pendek.

Penelitian ini membahas klasifikasi laporan pengaduan menggunakan algoritma C4.5 dengan seleksi fitur menggunakan metode *wrapper*. *Forward selection* salah satu metode seleksi fitur metode *wrapper*. Seleksi fitur dapat mengurangi fitur yang tidak relevan sehingga penambahan *dataset* diharapkan bisa mendapatkan fitur yang relevan dan memotong *rule* yang tidak penting. Fitur yang signifikan terhadap variabel independen dimasukkan satu persatu sampai tidak ada lagi fitur yang tidak signifikan, dengan tujuan mengoptimalkan hasil akurasi dan mengurangi dimensi data.

Penelitian ini ditujukan untuk memperoleh model klasifikasi pengaduan pelaporan rakyat dengan akurasi yang lebih tinggi daripada penelitian sebelumnya. Selain itu juga diharapkan dapat memberikan sumbang saran kepada pengelola portal LAPOR bagi peningkatan sistem LAPOR agar layanan aspirasi dan pengaduan *online* rakyat bisa lebih mudah, cepat, dan tepat dalam merespon laporan rakyat.

2. TINJAUAN PUSTAKA

2.1. Penelitian yang Relevan

Penelitian terkait tentang klasifikasi pengaduan pernah dilakukan oleh Yulia Sulistyaningsih, dkk. Dalam penelitian yang berjudul pengklasifikasian pengaduan masyarakat pada laman Kantor Pertanahan Kota Surabaya dengan metode pohon keputusan, 2013[6]. Berdasarkan hasil uji coba perbandingan beberapa metode pohon keputusan (C4.5, Random Forest, dan RandomTree) didapatkan bahwa metode C4.5 mendapatkan hasil yang terbaik, dan juga penggunaan stemming mampu meningkatkan akurasi. Nilai akurasi tertinggi yang diperoleh dalam penelitian ini sebesar 76,39%.

Masayu Leylia Khodra, Ahmad Fauzan Pengkategorian otomatis multilabel pada pemerintah Kota Bandung[7]. Dari penelitian klasifikasi pengaduan, dengan membandingkan beberapa algoritma diantaranya SVM, ANN, Naive Bayes, J48(C4.5). Didapatkan bahwa nilai akurasi terbaik dengan *singelabel* adalah algoritma SVM dengan hasil 65,65%.

Maya Kurniawati, Imam Cholissodin, Indriati [8]. Penelitian ini adalah mengklasifikasi laporan pengaduan pada layanan kampus. penelitian menggunakan algoritma SVM. Dengan melakukan pengujian pada jumlah data serta pemakaian *stemming* pada pengelolaan awal data. Hasil klasifikasi tertinggi diperoleh sebesar 82,61%.

2.2. Landasan Teori

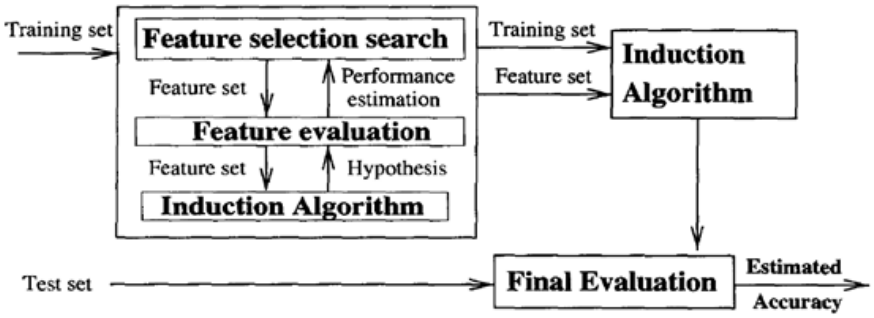
Text mining adalah salah satu teknik yang dapatdigunakan untuk melakukan klasifikasi teks dimana, t*extmining* merupakan bagian dari *data mining* yang berusaha menemukan informasi yang dibutuhkan dari sekumpulan data teks yang berjumlah besar[15].Selain klasifikasi, *text mining* juga digunakan untuk menangani masalah *clustering*, *information extraction*, dan *information retrival*[16].

Klasifikasi atau kategorisasi teks adalah proses penempatan suatu dokumen ke suatu kategori atau kelas sesuai dengan karakteristik dari dokumen tersebut. Dalam *text mining*, klasifikasi mengacu kepada aktifitas menganalisis atau mempelajari himpunan dokumen teks untuk memperoleh suatu model atau fungsi yang dapat digunakan untuk mengelompokkan dokumen teks lain yang belum diketahui kelasnya ke dalam satu atau lebih kelas [17]. Dokumen yang digunakan untuk pembelajaran dinamakan contoh (sample atau training data set) yang dideksripsikan oleh himpunan atribut atau variabel. Salah satu atribut mendeskripsikan kelas yang diikuti oleh suatu contoh, hingga disebut atribut kelas. Atribut lain dinamakan atribut independen atau *predictor*.

Algoritma C4.5 diperkenalkan oleh Quinlan sebagai versi perbaikan dari ID3. Dalam ID3, induksi *decision tree* hanya bisa dilakukan pada fitur bertipe kategorikal (nominal atau ordinal), sedangkan tipe numerik (interval atau rasio) tidak dapat digunakan. Perbaikan yang membedakan algoritma C4.5 dari lD3 adalah dapat menangani fitur dengan tipe numerik, melakukan pemotongan (pruning) *decision tree*. dan penurunan (deriving) *rule set*[18]. Algoritma C4.5 juga menggunakan kriteria *gain* dalam menentukan fitur yang menjadi pemecah *node* pada pohon yang diinduksi.

Feature Selection atau *Feature Reduction* adalah suatu kegiatan yang umumnya bisa dilakukan secara preprocessing dan bertujuan untuk memilih *feature* yang berpengaruh dan mengesampingkan *feature* yang tidak berpengaruh dalam suatu kegiatan pemodelan atau penganalisisan data. Ada banyak alternatif yang bisa digunakan dan harus dicoba-coba untuk mencari yang cocok. Secara garis besar ada dua kelompok besar dalam pelaksanaan *feature selection*: *Ranking Selection* dan *Subset Selection*.

Wrapper Feature Selection merupakan suatu pendekatan untuk mengatasi permasalahan dalam algoritma pembelajaran (classifier) yang melibatkan atribut-atribut untuk fokus di dalam algoritma tersebut[19]. Pada algoritma pembelajaran, untuk mendapatkan akurasi yang tinggi saat membangkitkan *classifier* dibutuhkan atribut-atribut yang relevan.



Gambar 1. Tahapan *Wrapper Feature Selection*

Dalam penelitian ini evaluasi dan validasi menggunakan akurasi *precision*, dan *recall*, pengukuran menggunakan tabel klasifikasi prediktif yaitu *confusion matrix*. Akurasi dalam klasifikasi adalah persentase ketepatan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. Presisi adalah ukuran dari akurasi dari suatu kelas tertentu yang telah diprediksi. *Recall* merupakan persentase dari data dengan nilai positif yang nilai prediksinya juga positif [20].

3. METODE PENELITIAN

Metode penelitian diawali dengan proses pengumpulan data. Data yang digunakan pada penelitian ini dikumpulkan dari laporan pengaduan rakyat yang ada di *website* https://lapor.ukp.go.id/beranda/laporan-aspirasi-pengaduan-*online* -masyarakat-indonesia.html. Penelitian ini menggunakan data sebanyak sebelas kategori. Adapun kategori yang dipilih adalah sebagai berikut : Administrasi Kependudukan, Reformasi Birokrasi dan Tata Kelola, Beras Miskin, Bidang Polhukam, Energi dan Sumber Daya Alam, Infrastruktur, Kesehatan, Lingkungan Hidup dan Penanggulangan Bencana, Pendidikan, Perhubungan, dan Program Keluarga Harapan.

Berikutnya dilakukan praprosessing data dengan tahapan tokenisasi, *filtering, steming, tagging,* dan *analyzing*.

Eksperimen dilakukan dengan melakukan uji coba klasifikasi pada *dataset* yang diambil. Pengujian pertama dilakukan pada klasifikasi dengan menggunakan metode C4.5 dengan jumlah data 1100 dokumen, 550 dokumen dan 275 dokumen. kemudian pengujian kedua dilakukan pada klasifikasi dengan C4.5 dengan *forward selection*, dengan jumlah dokumen 1100 dokumen, 550 dokumen dan 275 dokumen.

Tabel 1. Jumlah Dokumen Latih

Klasifikasi C4.5	Klasifikasi C4.5 + *forward selection*
1100 dokumen (100% *dataset*)	1100 dokumen (100% *dataset*)
550 dokumen (50% *dataset*)	550 dokumen (50% *dataset*)
275 dokumen (25% *dataset*)	275 dokumen (25% *dataset*)

Pada penelitian klasifikasi untuk evaluasi hasil pengujian dilakukan dengan menghitung nilai dari *accuracy, precision* dan *recall*[13]. *Accuracy* adalah tingkat dari dokumen yang benar diidentifikasi, sedangkan *recall* adalah tingkat keberhasilan sistem dalam menemukan kembali informasi dan *precision* adalah perbandingan jumlah data yang sesuai dengan data yang diminta. Persamaan dari ketiganya bisa

digambarkan sebagai berikut.

Tabel 2. *Confusion* Matrix

		Predicted Class	
		Class=Yes	*Class=No*
Actual Class	*Class=Yes*	*TP(true positif)*	*FN(false negatif)*
	Class=No	*FP(false positif)*	*TN(true negatif)*

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.7)$$

$$\text{Recall} = \frac{TP}{TP+FP} \quad (3.8)$$

$$\text{Precision} = \frac{TP}{TP+FN} \quad (3.9)$$

Selanjutnya setelah *confusion matrix* dilakukan uji *cross validation* untuk mendapatkan akurasi terbaik, uji *cross validation* dilakukan sampai 10-fold untuk hasil terbaik.

4. HASIL DAN PEMBAHASAN

Pada tahap tokenisasi dilakukan pemecahan dari kalimat kalimat yang menyusun paragraf laporan pengaduan ke dalam satuan kata penyusunnya. Proses tokenisasi dengan bantuan WEKA yaitu dengan WordTokenizer. Tanda baca yang tidak penting atau tidak digunakan dihilangkan, adapun tanda baca yang dihilangkan adalah .,;:'"()?!1234567890/#-+*&%={}. Frekuensi kata yang muncul sebanyak 3 kali. Hasil tokenisasi dari *dataset* dengan 1100 laporan dokumen, menghasilkan 1316 fitur.

Tahap *filtering* adalah mengambil kata – kata yang penting dari hasil token dengan menggunakan algoritma, yaitu dengan membuang kata-kata yang tidak penting atau disebut dengan *stop list*. Contoh kata yang tidak penting dalam bahasa Indonesia adalah ‘yang’, ‘di’, ‘ke’. sehingga kata tersebut tidak menjadi fitur.

Tahap *stopwords* adalah tahap lanjutan untuk membuang kata yang tidak memiliki hubungan dengan dokumen. Hal ini juga dapat mengurangi dimensi data dari dokumen. Perbedaan dengan tahap sebelumnya, yaitu pada kata yang dibuang.

Tahapan *term weighting* mengunakan pembobotan dengan TF/IDF. Pertama ditentukan nilai TF dari dokumen, TF diperoleh dengan mengubah kata pada suatu dokumen, jika ada dalam dokumen diberi nilai 1 (satu) jika tidak ada maka diberi nilai 0 (nol). Hasil TF dengan menggunakan WEKA didapat gambaran seperti yang tercantum pada tabel di bawah ini. Setelah ditentukan TF kemudian ditentukan nilai *Inver Document Frekuensi* (1). IDF untuk mengurangi bobot term yang tersebar di seluruh dokumen. Bobot suatu term dicari dengan mengalikan nilai TF dengan IDF (2).

$$idf_j = \log\left(\frac{D}{df_j}\right) \quad (1)$$

$$W_{ij} = tf_{ij} * idf_j$$
$$W_{ij} = tf_{ij} * \log\left(D/df_j\right) \quad (2)$$

Tabel 3. Term Frekuensi (TF)

No.	1: ada Numeric	2: adalah Numeric	3: adanya Numeric	4: agar Numeric	5: akan Numeric	6: akhirnya Numeric	7: aktivasi Numeric	8: alamat Numeric	9: alasan Numeric	10: alat Numeric
1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
4	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
8	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
14	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
19	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
21	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tabel 4. Pembobotan Term (TF-IDF)

No.	administrasi Numeric	ajukan Numeric	akta Numeric	akte Numeric	alamat Numeric	alasan Numeric	ambil Numeric	ambon Numeric	antrean Numeric	antri Numeric	antrian Numeric	aparat Numeric
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.890...
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.890...
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.331...	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.331...	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	2.370...	3.131...	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	3.331155	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	3.258...	0.0	0.0	0.0	0.0	0.0	3.331...	0.0	0.0
19	0.0	0.0	0.0	3.258...	0.0	0.0	0.0	4.092...	0.0	0.0	0.0	0.0
20	3.331155	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	3.893...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	3.893...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	3.331155	3.893...	0.0	3.258...	0.0	0.0	0.0	0.0	0.0	0.0	2.890...	0.0
25	0.0	0.0	0.0	3.258...	0.0	0.0	0.0	4.092...	0.0	0.0	0.0	0.0

Eksperimen ini menggunakan data sebanyak 1100 *record*. Berdasarkan eksperimen didapatkan nilai dari *confusion matrik* sebagai berikut.

Tabel 5. Hasil *Confusion Matrix*

Confusion Matrix		Nilai Sebenarnya										
		a	b	c	d	e	f	g	h	i	j	k
Nilai Prediksi	a	72	2	0	13	0	0	4	0	2	7	0
	b	7	70	0	12	0	1	0	0	2	7	1
	c	0	0	97	0	0	0	0	0	0	0	3
	d	6	5	0	60	2	3	1	3	2	17	1
	e	1	0	0	4	90	1	0	0	0	4	0
	f	0	0	0	2	2	91	0	0	0	5	0
	g	1	0	0	4	0	0	88	6	0	1	0
	h	1	1	0	10	2	0	8	70	0	8	0
	i	4	3	1	8	1	0	0	0	77	6	0
	j	1	1	0	15	4	7	1	5	4	62	0
	k	1	0	2	5	0	0	0	1	1	0	90

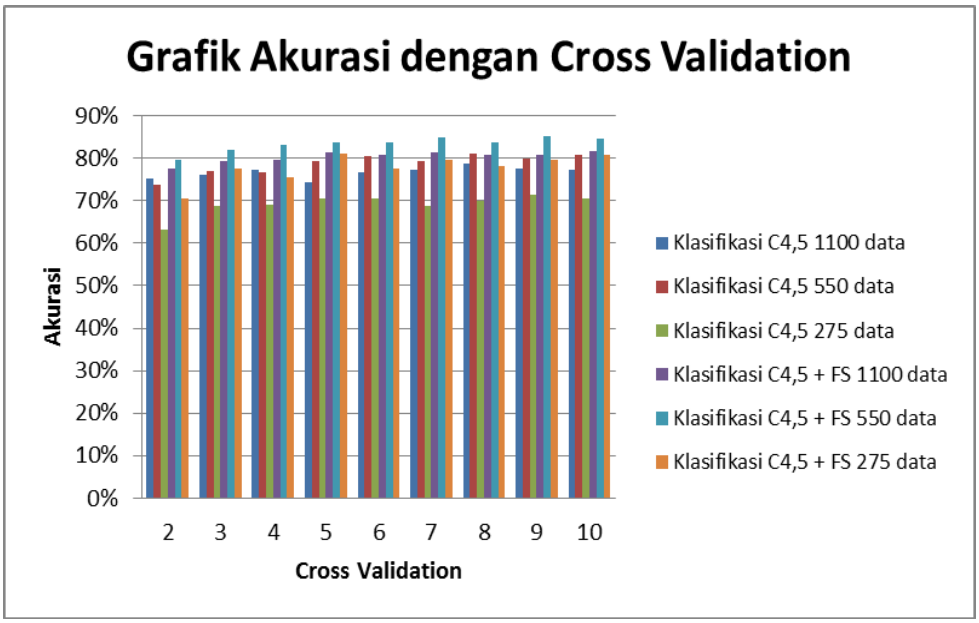
Keterangan:

a = Administrasi Kependudukan,
b = Reformasi Birokrasi dan Tata Kelola
c = Beras Miskin
d = Bidang Polhukam
e = Energi dan Sumber Daya Alam
f = Infrastruktur
g = Kesehatan
h = Lingkungan Hidup dan Penanggulangan Bencana
i = Pendidikan
j = Perhubungan
k = Program Keluarga Harapan

Eksperimen klasifikasi dilakukan dengan 6 kondisi fitur. Pertama klasifikasi C4.5 dengan 1100 *dataset*, 550 *dataset*, dan 275 *dataset,* selanjutnya klasifikasi C4.5 dengan seleksi fitur dengan 1100 *dataset*, 550 *dataset* dan 275 *dataset*. Setelah dilakukan percobaan dari semua klasifikasi tersebut didapatkan hasil nilai tertinggi pada klasifikasi 550 *dataset* dengan *forward selection*. Perbandingan hasil penelitian bisa dilihat pada tabel dan grafik berikut ini.

Tabel 6. Perbandingan Akurasi dengan *Cross Validation*

Cross Validation	Klasifikasi C4.5			Klasifikasi C4.5 + *Forward selection*		
	1100 data	550 data	275 data	1100 data	550 data	275 data
2	75,18 %	73,63 %	63,27 %	77,54 %	79,63 %	70,5 %
3	76 %	76,9 %	68,72 %	79,36 %	82 %	77,45 %
4	77,18 %	76,72 %	69,09 %	79,63 %	83,27 %	75,63 %
5	74,27 %	79,27 %	70,54 %	81,27 %	83,82 %	81,09 %
6	76, 81 %	80,54 %	70,54 %	80,81 %	83,82 %	77,45 %
7	77,36 %	79,27 %	68,72 %	81,45 %	84,9 %	79,64 %
8	78,81 %	81,09 %	69,81 %	80,72 %	83,82 %	78,2 %
9	77,72 %	80 %	71,27 %	80,90 %	85,27 %	79,63 %
10	77,18 %	80,72 %	70,54 %	81,54 %	84,75 %	80,73 %



Gambar 2. Grafik Perbandingan Akurasi Penelitian *Cross Validation*

Dari hasil penelitian di atas dapat diketahui bahwa algoritma C4.5 dengan menggunakan *forward selection* bisa meningkatkan akurasi klasifikasi. Hal ini bisa dilihat dari nilai akurasi hasil klasifikasi meningkat dari klasifikasi tanpa seleksi fitur dengan menggunakan seleksi fitur. Sedangkan hasil akurasi dengan melihat jumlah *dataset* yang diujikan, hasil akurasi tidak tergantung dari jumlah *dataset*. Karena semakin banyak jumlah data tidak menjamin semakin baik fitur yang relevan. Hasil akurasi tertinggi yang diperoleh sebelum dilakukan seleksi fitur adalah tertinggi 81,09% dengan 550 *dataset* pengujian. Sedangkan dengan menggunakan *forward selection* meningkat menjadi nilai tertinggi akurasi adalah sebesar 85,27% pada 1100 *dataset*.

Hasil seleksi fitur *forward selection* menghasilkan jumlah fitur yang berbeda pada tiap data, pada 1100 *dataset* dari hasil seleksi didapat 44 fitur, 550 *dataset* dihasilkan sebanyak 28 fitur, sedangkan pada 275 *dataset* didapat 24 fitur. Semakin banyak data semakin banyak fitur yang dihasilkan, tetapi akurasi yang dihasilkan tidak mengalami kenaikan. Hal ini bisa dilihat pada pengujian 275 data nilai akurasi tertinggi 81,09% naik pada 550 data naik menjadi 85,27% dan 1100 data menjadi 81,54%. Karena semakin banyak data jumlah persebaran fitur yang sama pada setiap kategori semakin merata hal ini bisa dilihat pada tabel *confusion matrix*.

Pada pengujian data proses seleksi fitur memakan waktu yang lama, semakin banyak data semakin lama waktu yang diproses. Tetapi metode *forward selection* mampu menghasilkan fitur yang signifikan sehingga mengurangi dimensi data serta mempercepat proses klasifikasi data.

1. KESIMPULAN

Hasil penelitian klasifikasi laporan pengaduan dengan menggunakan algoritma C4.5 dengan menggunakan seleksi fitur *forward selection* menghasilkan akurasi yang lebih baik, yakni sebesar 85,27%. Dibanding penelitian sebelumnya hasil akurasi tertinggi yang dicapai sebesar 82,61%. Penggunaan *dataset* juga tidak berpengaruh pada kenaikan akurasi pada klasifikasi dengan seleksi fitur, hal ini bisa dilihat pada pengujian 275 data, nilai akurasi tertinggi 81,09%. Sedangkan pada pengujian 550 data nilai akurasi tertinggi 85,27%. Tetapi pada pengujian tanpa seleksi fitur, jumlah tidak berpengaruh pada hasil akurasi, hal ini bisa dilihat dari hasil pengujian 1100 data, nilai tertinggi 78,81%, kemudian naik pada pengujian 550 data sebesar 81,09%, tetapi turun kembali pada pengujian 275 data menjadi 71,27%.

PERNYATAAN ORIGINALITAS

"Saya menyatakan dan bertanggung jawab dengan sebenarnya bahwa Artikel ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya" [Ali Sofyan – P31.2013.01348]

DAFTAR PUSTAKA

[1] S. Drs. T. Gilarso, *Pengantar Ilmu Ekonomi Makro*. Yogyakarta: Kanisius, 2008.

[2] K. Maryati, *Sosiologi*. Jakarta: Erlangga, 2007.

[3] Situs Lapor, "https://www.lapor.go.id/lapor/tentang_lapor/tentang-layanan-aspirasi-dan-pengaduan-online-rakyat.html." .

[4] T. Candra, "Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia," Institut Teknologi Telkom, 2009.

[5] E. T. L. Kusrini, *Algoritma Data Mining*. Yogyakarta: ANDI, 2009.

[6] Y. Sulistyaningsih, A. Djunaidy, and R. P. Kusumawardani, "Pengklasifikasian Pengaduan Masyarakat pada Laman Kantor Pertanahan Kota Surabaya I dengan Metode," pp. 1–6.

[7] A. Fauzan, "Automatic Multilabel Categorization using Learning to Rank Framework for Complaint Text on Bandung Government."

[8] M. K. Maya Kurniawati1, Imam Cholissodin, S.Si., M.Kom, Indriati, ST., "Klasifikasi Dokumen E-Complaint Kampus Menggunakan Directed Acyclic," 2014.

[9] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8696–8702, 2011.

[10] J. Koncz, P., & Paralic, "An approach to feature selection for sentiment analysis," *IEEE Int. Conf. Intell. Eng. Syst.*, pp. 357–362, 2011.

[11] S. Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, "An Improved Particle Swarm Optimization for Feature Selection," *J. Bionic Eng.*, vol. 8(2), pp. 191–200, 2011.

[12] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making (Google eBook)*, no. 2004. 2011.
[13] T. Deepa and L. Ladha, "Feature Selection Methods And Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 5, pp. 1787–1797, 2011.
[14] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data : A Fast Correlation-Based Filter Solution," 2003.
[15] S. Ronen, Feldman; James, *No TitleAdvanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007.
[16] U. W.Berry, Michael(University of Tennessee and U. Kogan, Jacob(University of Maryland Baltimore Country, *Text Mining Application and Theory*. 2010.
[17] F. Sebastiani, "Machine Learning in Automated Text Categorization," 2001.
[18] Eko Prasetyo, *Data Mining Mengolah Data Menjadi Informasi menggunakan Matlab*. Yogyakarta: ANDI, 2014.
[19] R. Kohavi and H. John, "Artificial Intelligence Wrappers for feature subset selection," vol. 97, no. 97, pp. 273–324, 2011.
[20] M. Han, J., & Kamber, *Data Mining Concepts And Techniques 2nd Edition*. San Fransisco: Elsevier B.V., 2006.