

Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan SAMBAT Online Menggunakan Metode *Naïve Bayes* dan Kombinasi Seleksi Fitur

Hilmy Khairi Idris¹, Mochammad Ali Fauzi², Indriati³

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: ¹hilmykhairii@student.ub.ac.id, ²moch.ali.fauzi@ub.ac.id, ³indriati.tif@ub.ac.id

Abstrak

SAMBAT *Online* merupakan bentuk realisasi *e-Government* yang ada di Kota Malang. SAMBAT *Online* atau Sistem Aplikasi Masyarakat Bertanya Terpadu *Online* suatu platform berupa situs web yang disediakan oleh pemerintah Kota Malang untuk menerima laporan pengaduan, kritik, saran, maupun pertanyaan kepada pemerintah. Setiap laporan yang masuk akan dikelompokkan secara manual oleh pengelola sistem SAMBAT *Online*. Pengelompokan yang dilakukan berdasarkan Satuan Kerja Perangkat Daerah (SKPD) yang dituju secara manual. Oleh karena ini dibangunlah sistem klasifikasi untuk mengefesienkan waktu pada proses pengelompokan laporan ke SKPD yang dituju menggunakan metode *Naïve Bayes* dan Kombinasi Seleksi Fitur antara *Chi-Square* dan *Information Gain*. Pada pengujian yang dilakukan, sistem berhasil memberikan hasil akurasi yang lebih baik apabila menggunakan seleksi fitur dibanding tanpa menggunakan seleksi fitur dengan nilai akurasi sebesar 83,33%. Selanjutnya pada saat dilakukan kombinasi seleksi fitur, hasil akurasi yang didapatkan sama dengan hasil tanpa dilakukan kombinasi yaitu 83,33%. Sehingga kombinasi seleksi fitur belum bisa memberikan hasil yang lebih baik.

Kata kunci: SAMBAT Online, Klasifikasi Teks, *Chi-Square*, *Information Gain*, *Naïve Bayes*

Abstract

SAMBAT *Online* is a form of *e-Government* realization in Malang City. SAMBAT *Online* or The Integrated Online Community Application System is a platform of a website provided by the Malang City Government to receive complaints, criticisms, suggestions, or questions to the government. Each incoming report will be grouped manually by the SAMBAT *Online* system manager. Grouping is based on The Regional Work Unit (SKPD) which is handled manually. Therefore, a classification system was built to save time in the process of grouping reports to SKPD using the *Naïve Bayes* method and the Combination of Feature Selection between *Chi-Square* and *Information Gain*. In the tests conducted, the system succeeded in providing better accuracy results when using feature selection than without using feature selection with an accuracy value of 83.33%. Furthermore, when a feature selection combination is performed, the results of the accuracy obtained are the same as the results without a combination of 83.33%. So, the combination of selection has not been able to provide better results.

Keywords: SAMBAT Online, Text Classification, *Chi-Square*, *Information Gain*, *Naïve Bayes*

1. PENDAHULUAN

Dalam hal mewujudkan pelayanan yang optimal memerlukan sebuah sistem yang baik dan terintegrasi mulai dari tingkat pusat sampai pada tingkat daerah. *E-Government* merupakan sebuah sistem terintegrasi dengan menggunakan media teknologi informasi dalam pelaksanaannya. *E-Government* atau *eGov* merupakan sebuah proses pemanfaatan teknologi informasi sebagai alat bantu

pemerintah dalam menjalankan sistem pemerintahan secara efisien yang dapat meningkatkan hubungan pemerintah dengan pihak lain (Somantri, *et al.*, 2017).

SAMBAT *Online* merupakan bentuk realisasi *e-Government* yang ada di Kota Malang. SAMBAT *Online* atau Sistem Aplikasi Masyarakat Bertanya Terpadu *Online* suatu platform berupa situs web yang disediakan oleh pemerintah Kota Malang untuk menerima laporan pengaduan, kritik, saran, maupun

pertanyaan kepada pemerintah. Masyarakat Kota Malang dapat memberikan laporannya melalui pesan singkat atau mengirimkan langsung melalui situs web yang sudah disediakan.

Setiap laporan yang masuk akan dikelompokkan secara manual oleh pengelola sistem SAMBAT *Online*. Pengelompokan yang dilakukan berdasarkan Satuan Kerja Perangkat Daerah (SKPD) yang dituju. Klasifikasi teks dapat dilakukan untuk mengefesienkan waktu pada proses pengelompokan laporan ke SKPD yang dituju oleh pengelola sistem SAMBAT *Online*.

Pada penelitian sebelumnya dalam melakukan klasifikasi pengaduan SAMBAT *online* yang dilakukan oleh Claudio Fresta Suharno dengan menggunakan *KNN* menghasilkan akurasi sebesar 78%. Selain metode tersebut masih banyak metode klasifikasi yang dapat digunakan salah satunya adalah *Naïve Bayes*. Metode *Naïve Bayes* merupakan metode klasifikasi sederhana tetapi memiliki akurasi dan performansi yang tinggi dalam pengklasifikasian teks (Nugroho, *et al.*, 2016). Pada penelitian yang dilakukan oleh Didik Garbian Nugroho yang menggunakan metode *Naïve Bayes* dalam melakukan pengklasifikasian analisis sentimen pada jasa ojek *online* menghasilkan akurasi sebesar 80%. Oleh karena itu, metode *Naïve Bayes* diharapkan dapat menghasilkan akurasi lebih baik dari penelitian sebelumnya dalam melakukan klasifikasi teks Bahasa Indonesia pada pengaduan SAMBAT *Online*.

Dalam melakukan pengklasifikasian teks, setiap dokumen dapat menjadi milik banyak kategori. Untuk itu dilakukan tahapan seleksi fitur yang dapat meningkatkan skalabilitas, efisiensi, dan akurasinya (Zheng, *et al.*, 2014).

Pada penelitian ini menggunakan kombinasi seleksi fitur. Seleksi fitur yang digunakan adalah terdiri dari dua seleksi fitur. Pertama adalah *Information Gain* yang merupakan algoritma seleksi fitur yang efisien dalam mengukur jumlah bit informasi yang diperoleh pada proses klasifikasi untuk mengetahui keberadaan suatu fitur pada sebuah dokumen kemudian memilih subset optimal (Putra, *et al.*, 2016). Seleksi fitur kedua adalah *Chi-Square* yang merupakan salah satu metode seleksi fitur yang dapat membuang banyak jumlah fitur tanpa mempengaruhi tingkat akurasinya (Sun, *et al.*, 2009).

Dasar pemikiran kombinasi seleksi fitur ini berasal dari bidang *ensemble learning*, dimana kombinasi penggolongan untuk mendapatkan

hasil yang lebih stabil, dan hasilnya kombinasi penggolongan memiliki kinerja lebih baik. Oleh karena itu, orang bisa berpikir juga pada teknik seleksi fitur dengan menggabungkan dapat menghasilkan lebih stabil dibanding tunggal. (Saeys, *et al.*, 2008).

Perpaduan metode *Naïve Bayes* dengan kombinasi seleksi fitur, diharapkan dapat mengetahui hasil dari sistem klasifikasi teks dengan metode *Naïve Bayes* yang dipadukan dengan metode untuk menunjang sistem klasifikasi teks yaitu metode kombinasi seleksi fitur. Selain itu dapat digunakan sebagai acuan pada implementasi pada klasifikasi dokumen laporan SAMBAT *Online* dengan otomatis serta dapat menghasilkan tingkat akurasi yang lebih baik dari penelitian sebelumnya.

2. KAJIAN PUSTAKA

2.1 SAMBAT *Online*

SAMBAT *Online* atau Sistem Aplikasi Masyarakat Bertanya Terpadu *Online* merupakan fasilitas yang disediakan oleh Dinas Komunikasi dan Informatika (Diskominfo) untuk masyarakat kota Malang yang ingin menyampaikan pengaduan, kritik, saran, maupun pertanyaan terhadap permasalahan yang ada di wilayah kota Malang terkait apapun yang ingin disampaikan atau ditanyakan ke pemerintah kota Malang.

2.2 Klasifikasi Teks

Klasifikasi teks adalah salah satu pengaplikasian Text Mining dalam pengelompokan data (Li, *et al.*, 2016). Metode klasifikasi sangat banyak yang dapat digunakan yaitu *Secision/class classification trees*, *Bayesian classifiers/Naïve Bayes classifiers*, *Neural Network*, Analisis Statistik, Algoritme Genetika, *Rough sets*, *Memory based reasoning*, dan *Support Vector Machine (SVM)*.

2.3 Seleksi Fitur

Pemilihan fitur telah diterapkan untuk kategorisasi teks untuk meningkatkan skalabilitas, efisiensi, dan akurasi (Zheng, *et al.*, 2014). Seleksi fitur adalah salah satu bagian dari tahapan preprocessing.

Tujuan dari seleksi fitur adalah untuk meningkatkan performa klasifikasi teks dengan menghilangkan fitur yang dianggap tidak

relevan dalam klasifikasi untuk mengurangi dimensi dari himpunan (Suharno, *et al.*, 2017). Sejumlah metrik seleksi fitur telah dieksplorasi, beberapa yang terkemuka di antaranya adalah *Informasi Gain* (IG), *Chi-square* (CHI), Koefisien Korelasi (CC), dan *Odds Ratio* (OR). Untuk penelitian ini menggunakan kombinasi seleksi fitur dari *Chi-square* dan *Information Gain*.

2.4 Chi-Square

Untuk Chi-Square Testing adalah metode statistika pengujian hipotesis data diskrit yang menentukan apakah sebuah variabel tersebut saling berkaitan atau tidak dan mengevaluasi antara dua variable seberapa besar korelasinya. Persamaan Chi-Square sebagai berikut (Suharno, *et al.*, 2017).

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

Keterangan :

- t : Kata
- c : Kelas atau kategori
- N : Jumlah dokumen
- A : Banyaknya dokumen pada kategori c yang memuat t
- B : Banyaknya dokumen bukan kategori c yang memuat t
- C : Banyaknya dokumen pada kategori c yang tidak memuat t
- D : Banyaknya dokumen bukan kategori c yang tidak memuat t

2.5 Information Gain

Information Gain merupakan seleksi fitur dimana ukuran jumlah bit informasi yang diperoleh untuk prediksi kategori dengan mengetahui ada atau tidaknya istilah dalam dokumen (Zheng, *et al.*, 2014). Persamaan IG dapat dilihat sebagai berikut :

$$IG(t) = -(\sum_{i=1}^{|c|} P(ci) \log P(ci) + P(t) \sum_{i=1}^{|c|} P(ci|t) \log P(ci|t) + P(\bar{t}) \sum_{i=1}^{|c|} P(ci|\bar{t}) \log P(ci|\bar{t})) \quad (2)$$

Keterangan :

- $P(c_i)$: Probabilitas kelas c_i
- $P(t)$: Probabilitas kata
- $P(c_i|t)$: Probabilitas kondisional dari

nilai dokumen c_i yang diberikan t
 $P(c_i|t)$: Probabilitas kondisional dari nilai dokumen c_i yang diberikan bukan t

2.6 Naïve Bayes

Metode Naïve Bayes melalui tahap training dan klasifikasi pada proses klasifikasinya. Pada tahap pelatihan dilakukan proses analisis pada sampel dokumen berupa pemilihan vocabulary, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. (Hamzah, 2012). Kemudian dari sampel dokumen dicari probabilitas prior nya pada setiap kategori. Untuk tahapan klasifikasinya dari satu dokumen ditentukan nilai kategorinya berdasarkan term yang muncul dalam dokumen yang diklasifikasi. Persamaan *Naïve Bayes* bisa dilihat sebagai berikut :

$$P(C_j|W_i) = \frac{P(C_j) \times P(W_i|C_j)}{P(W_i)} \quad (3)$$

Keterangan :

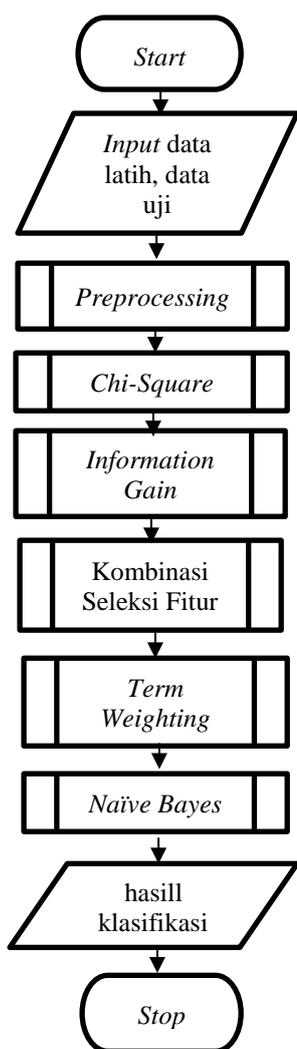
- $P(C_j|W_i)$: Perhitungan *Posterior*, dimana peluang kemunculan kategori j dengan syarat kemunculan kata i.
- $P(C_j)$: Perhitungan *Prior*, dimana peluang kemunculan setiap dokumen pada kategori j.
- $P(W_i|C_j)$: Perhitungan *Likelihood* atau *Conditional Probability*, dimana peluang setiap kata i dengan syarat kategori j.
- $P(W_i)$: Perhitungan *Evidence*, dimana peluang kemunculan setiap kata.
- i : indeks untuk menyimpan kata dari kata 1 sampai dengan kata ke-n.
- j : indeks untuk menyimpan kategori dari kategori 1 sampai dengan kategori ke-n.

3. METODOLOGI

3.1 Perancangan Algoritme

Proses implementasi algoritma yang digunakan diawali dengan tahapan *pre-processing* pada data latih maupun data uji.

Kemudian melakukan seleksi fitur dimana pada penelitian ini melakukan kombinasi seleksi fitur yaitu menggunakan *Information Gain* dan *Chi-Square*. Selanjutnya melakukan proses klasifikasi dimana pada penelitian ini menggunakan *Naïve Bayes*. Berikut tahapan-tahapan yang harus dilakukan pada penelitian ini yang dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alir Klasifikasi

3.3 Data Penelitian

Pada penelitian kali ini, data yang digunakan adalah data teks Bahasa Indonesia pada dokumen pengaduan SAMBAT Online. Total dokumen yang diambil adalah 204 dokumen dimana 80% dokumen akan dijadikan data latih dan 20% dokumen akan dijadikan data uji. Dari 204 dokumen terdiri dari tiga SKPD dengan pembagian 37 dokumen untuk SKPD Dinas Kebersihan dan Pertamanan (DKP), 67 dokumen untuk SKPD Dinas Pekerjaan Umum Perumahan dan Pengawasan Bangunan

(DPUPPB), dan 100 dokumen untuk SKPD Dinas Perhubungan (Dishub). Dinas Kebersihan dan Pertamanan (DKP).

4. HASIL DAN PEMBAHASAN

4.1 Pengujian Tanpa Seleksi Fitur

Pada hasil pengujian Tabel 1 dalam melakukan klasifikasi tanpa menggunakan seleksi fitur didapatkan hasil *accuracy* sebesar 80,95%, *precision* sebesar 73,45%, *recall* sebesar 76,14%, dan *f-measure* sebesar 73,23%. Dari hasil tersebut akan dibandingkan dengan hasil dari pengujian klasifikasi menggunakan seleksi fitur yaitu *Chi-square* dan *Information Gain*.

Tabel 1. Hasil Pengujian Tanpa Seleksi Fitur

<i>Accuracy</i>	80,95%
<i>Precision</i>	73,45%
<i>Recall</i>	76,14%
<i>F-Measure</i>	73,23%

Setelah dilakukan pengujian tanpa menggunakan seleksi fitur, dihasilkan tingkat akurasi sebesar 80,95%. Dimana sistem mampu mengklasifikasikan 34 dokumen uji sesuai dengan kelasnya. Sehingga dari hasil pengujian tanpa menggunakan seleksi fitur, dapat dikatakan metode *Naïve Bayes* mampu memberikan hasil yang baik dalam melakukan klasifikasi dokumen pengaduan SAMBAT online.

4.2 Pengujian Menggunakan *Chi-Square*

Seleksi fitur pertama yang digunakan adalah *Chi-square*. Jumlah fitur yang digunakan pada pengujian ini akan dilakukan variasi yaitu, 25%, 50%, 75%, dan 100%. Untuk hasil dari pengujian dengan menggunakan seleksi fitur *Chi-square* dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pengujian Menggunakan Seleksi Fitur *Chi-Square*

	Jumlah Fitur			
	25%	50%	75%	100%
<i>Accuracy</i>	83,33%	83,33%	80,95%	80,95%
<i>Precision</i>	75,12%	75,12%	70,95%	73,45%
<i>Recall</i>	81,14%	81,14%	76,70%	76,14%
<i>FMeasure</i>	75,41%	75,41%	70,05%	73,23%

Pada Tabel 2 dapat dilihat hasil pengujian

dari skenario pengujian menggunakan seleksi fitur dengan algoritme *Chi-square*. Pada pengujian ini menggunakan jumlah fitur sebesar 25%, 50%, 75%, dan 100% dari keseluruhan fitur yang ada. Hasil pengujian sistem tersebut dapat dilihat variasi *threshold* fitur yang optimal adalah 25% dan 50%. Dimana *threshold* tersebut dihasilkan nilai *accuracy* yang didapatkan adalah 83,33%, *precision* sebesar 75,12%, *recall* sebesar 81,14%, dan *f-measure* sebesar 75,41%. Dengan menggunakan fitur sebesar 25% dan 50%, *Chi-Square* mampu memberikan hasil urutan kata yang relevan di 25% dan 50% kata tersebut. Sebaliknya dengan menggunakan 75% dan 100% fitur membuat sistem menggunakan lebih banyak kata dan keseluruhan kata dimana terdapat kata yang tidak terlalu relevan untuk digunakan. Sehingga penambahan jumlah fitur pada hasil *Chi-Square* membuat kata yang tidak relevan memungkinkan untuk ikut digunakan dan membuat hasil sistem tidak optimal.

4.3 Pengujian Menggunakan *Information Gain*

Metode seleksi fitur lainnya yang digunakan adalah *Information Gain*. Jumlah fitur yang digunakan pada pengujian kali ini disamakan dengan variasi jumlah fitur yang digunakan pada pengujian menggunakan *Chi-square* yaitu 25%, 50%, 75%, dan 100%. Untuk hasil dari pengujian dengan menggunakan seleksi fitur *Information Gain* dapat dilihat pada Tabel 3.

Tabel 3. Hasil Pengujian Menggunakan Seleksi Fitur *Information Gain*

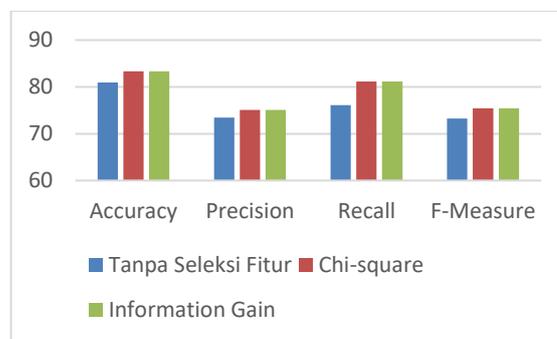
	Jumlah Fitur			
	25%	50%	75%	100%
<i>Accuracy</i>	78,57%	83,33%	83,33%	80,95%
<i>Precision</i>	68,57%	75,12%	75,12%	73,45%
<i>Recall</i>	74,94%	81,14%	81,14%	76,14%
<i>FMeasure</i>	68,03%	75,41%	75,41%	73,23%

Pada Tabel 3 dapat dilihat hasil pengujian dari skenario pengujian menggunakan seleksi fitur dengan algoritme *Information Gain*. Pada pengujian ini menggunakan variasi jumlah fitur sebesar 25%, 50%, 75%, dan 100% dari keseluruhan fitur yang ada. Hasil pengujian sistem tersebut dihasilkan nilai *threshold* yang paling optimal adalah 50% dan 75% dengan nilai *accuracy* yang didapatkan adalah 83,33%, *precision* sebesar 75,12%, *recall* sebesar 81,14%, dan *f-measure* sebesar 75,41%. Dengan menggunakan fitur sebesar 25%, *Information Gain* belum memberikan hasil urutan kata yang

relevan di 25% kata tersebut. Sebaliknya dengan menggunakan 100% fitur membuat sistem menggunakan keseluruhan kata dimana terdapat kata yang tidak terlalu relevan untuk digunakan. Sehingga bisa dikatakan jika terlalu sedikit fitur hasil *Information Gain* digunakan belum tentu mendapatkan kata yang paling relevan, begitupun dengan terlalu banyak kata digunakan akan memberikan hasil yang tidak optimal.

4.4 Pengaruh Seleksi Fitur

Seleksi fitur berfungsi untuk mengetahui tingkat relevan suatu kata. Dari skenario pengujian menggunakan seleksi fitur *Chi-square* ataupun *Information Gain* didapatkan hasil urutan setiap kata/fitur dari yang paling relevan hingga paling tidak relevan.



Gambar 2. Pengujian Pengaruh Seleksi Fitur

Grafik pada Gambar 2 membandingkan hasil pengujian dari pengujian tanpa menggunakan seleksi fitur dan menggunakan seleksi fitur. Dapat dilihat dari grafik tersebut hasil pengujian dari sisi *accuracy*, *precision*, *recall*, maupun *f-measure*, sistem yang melalui proses seleksi fitur baik menggunakan *Chi-square* maupun *Information Gain* hasilnya lebih baik dibanding tanpa menggunakan seleksi fitur. Dengan menggunakan seleksi fitur, didapatkan hasil akurasi sebesar 83,33% dengan ekstraksi fitur masing-masing sebanyak 50%. Seleksi fitur *Chi-Square* maupun *Information Gain* berhasil mengekstraksi fitur yang relevan untuk digunakan sehingga mendapatkan hasil yang lebih baik. Dengan tidak menggunakan seleksi fitur, membuat sistem menggunakan semua fitur yang ada. Hasil yang kurang akurat memungkinkan terjadi karena terdapat beberapa fitur yang muncul di semua kelas (tidak merepresentasikan satu kelas spesifik) sehingga bisa membuat nilai *posterior* atau kelas yang terpilih bukan kelas sebenarnya.

4.5 Pengujian Kombinasi Seleksi Fitur

Pada pengujian kombinasi seleksi fitur yang pertama, digunakan operasi AND untuk mengkombinasikan fitur hasil *Chi-square* dan *Information Gain*. Hasil kombinasi merupakan fitur yang terdapat di dalam kumpulan fitur *Chi-square* dan *Information Gain*. Fitur yang hanya terdapat di salah satu kumpulan fitur saja tidak akan digunakan. Untuk hasil dari pengujian dengan menggunakan operasi AND dalam melakukan kombinasi seleksi fitur dapat dilihat pada Tabel 4.

Tabel 4. Hasil Pengujian Kombinasi Seleksi Fitur Menggunakan Operasi AND

<i>Accuracy</i>	83,33%
<i>Precision</i>	75,12%
<i>Recall</i>	81,14%
<i>F-Measure</i>	75,41%

Tabel 4 merupakan hasil pengujian kombinasi seleksi fitur menggunakan operasi AND. Pada pengujian kali ini mengekstraksi fitur sebanyak 50% dari fitur *Chi-square* dan *Information Gain*. Hasil pengujian tersebut dihasilkan nilai *accuracy* yang didapatkan adalah 83,33%, *precision* sebesar 75,12%, *recall* sebesar 81,14%, dan *f-measure* sebesar 75,41%.

Pada pengujian kombinasi seleksi fitur yang kedua, digunakan operasi OR untuk mengkombinasikan fitur hasil *Chi-square* dan *Information Gain*. Hasil kombinasi merupakan fitur yang terdapat di dalam kumpulan fitur *Chi-square* dan *Information Gain*. Selain itu, fitur yang hanya terdapat di salah satu kumpulan fitur saja akan tetap digunakan. Hasil yang diberikan berupa hasil *accuracy*, *precision*, *recall*, dan *f-measure*. Untuk hasil dari pengujian dengan menggunakan operasi OR dalam melakukan kombinasi seleksi fitur dapat dilihat pada Tabel 5.

Tabel 5. Hasil Pengujian Kombinasi Seleksi Fitur Menggunakan Operasi OR

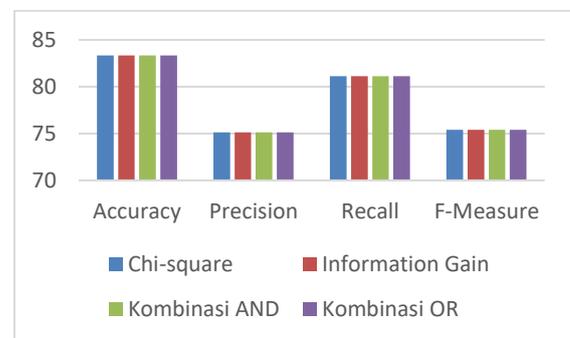
<i>Accuracy</i>	83,33%
<i>Precision</i>	75,12%
<i>Recall</i>	81,14%
<i>F-Measure</i>	75,41%

Tabel 5 merupakan hasil pengujian kombinasi seleksi fitur menggunakan operasi OR. Pada pengujian kali ini mengekstraksi fitur

sebanyak 75% dari fitur *Chi-square* dan *Information Gain*. Hasil pengujian tersebut dihasilkan nilai *accuracy* yang didapatkan adalah 83,33%, *precision* sebesar 75,12%, *recall* sebesar 81,14%, dan *f-measure* sebesar 75,4%. Hasil pengujian tersebut dapat dilihat hasil sistem menggunakan operasi AND atau OR menghasilkan hasil yang sama atau bisa dikatakan kedua operasi tersebut sama baiknya.

4.6 Pengaruh Kombinasi Seleksi Fitur

Perbandingan melakukan kombinasi seleksi fitur dan tanpa melakukan kombinasi, dapat dilihat pada Gambar 3.



Gambar 3. Pengujian Pengaruh Kombinasi Seleksi Fitur

Dari grafik di atas dapat dilihat perbandingan hasil pengujian antara melakukan kombinasi seleksi fitur dan tidak melakukan kombinasi seleksi fitur. Dapat dilihat hasil *accuracy*, *precision*, *recall*, dan *f-measure* yang dihasilkan *Chi-square*, *information Gain*, dan kombinasi menggunakan operasi AND maupun OR sama persis. Fitur dengan hanya menggunakan *Chi-square* atau *Information Gain* akan lebih banyak dikarenakan tidak ada fitur yang dibuang. Sedangkan menggunakan operasi OR menghasilkan lebih banyak lagi fitur karena kedua hasil fitur saling melengkapi. Tetapi pada saat menggunakan operasi AND juga didapatkan hasil yang sama baiknya walaupun hanya menggunakan fitur yang terdapat di kedua hasil seleksi fitur. Sehingga dari hasil pengujian dapat disimpulkan bahwa pengaruh kombinasi seleksi fitur tidak berpengaruh dengan hasil sistem yang diberikan. Hal ini dikarenakan dari sisi *Chi-square* dan *Information Gain* berhasil mengekstraksi fitur yang hampir sama sehingga walaupun dilakukan kombinasi, hasil fitur yang dihasilkan relevan untuk digunakan.

5. PENUTUP

Dari penelitian yang dilakukan dapat ditarik kesimpulan bahwa pada saat melakukan klasifikasi menggunakan metode *Naïve Bayes* tanpa melalui proses seleksi fitur, didapatkan hasil akurasi sebesar 80,95%. Sedangkan pada saat dilakukan klasifikasi menggunakan metode *Naïve Bayes* dan sebelumnya dilakukan proses ekstraksi fitur sebanyak 50% menggunakan *Chi-Square*, didapatkan hasil akurasi sebesar 83,33%. Hasil yang sama didapatkan pada saat menggunakan *Information Gain* dalam melakukan ekstraksi fitur sebanyak 50%. Dapat disimpulkan bahwa seleksi fitur berpengaruh terhadap hasil klasifikasi menggunakan metode *Naïve Bayes* untuk mendapatkan yang lebih baik.

Selain itu, pada saat melakukan klasifikasi menggunakan metode *Naïve Bayes* dengan dilakukannya kombinasi seleksi fitur sebelumnya, dimana fitur yang diekstraksi sebesar masing-masing 50% didapatkan hasil akurasi sebesar 83,33% dengan menggunakan operasi AND dalam pengkombinasianya. Sedangkan menggunakan operasi OR dalam pengkombinasianya didapatkan hasil akurasi sebesar 83,33%. Sehingga dapat disimpulkan bahwa kombinasi dalam seleksi fitur tidak berpengaruh untuk menghasilkan hasil yang lebih baik dikarenakan hasil yang diberikan sama dengan hasil apabila tidak dilakukan kombinasi.

DAFTAR PUSTAKA

- Hamzah, A., 2012. Klasifikasi Teks Dengan *Naïve Bayes Classifier (NBC)* Untuk Pengelompokan Teks Berita dan Abstrak Akademis, [online] Tersedia di: <http://repository.akprind.ac.id/sites/files/conference-proceedings/2012/hamzah_15430.pdf>
- Li, Z., Shang, W. & Yan, M., 2016. *News Text Classification Model Based on Topic Model*, [online] Tersedia di : <<https://ieeexplore.ieee.org/>> [Diakses 20 Agustus 2018].
- Nugroho, D. G., Chrisnanto, Y. H. & Wahana, A. 2016. Analisis Sentimen Pada Jasa Ojek *Online* Menggunakan Metode *Naïve Bayes*, [online] Tersedia di : <https://publikasiilmiah.unwahas.ac.id/index.php/PROSIDING_SNST_FT/article/view/1526> [Diakses 16 Juli 2018]
- Putra, I. B. G. W., Sudarma, M. & Kumara, I. N. S., 2016. Klasifikasi Teks Bahasa Bali Dengan Metode *Supervised Learning Naive Bayes Classifier*, [online] Tersedia di : <<https://text-id.123dok.com/>> [Diakses 17 September 2018]
- Saeyes, Y., Abeel, T. & Peer, Y. Vd. 2008. *Robust Feature Selection Using Ensemble Feature Selection Techniques*. *Indonesian Journal of Electrical Engineering and Computer Science*, [e-journal] pp. 313-325. Tersedia melalui: <<http://www.iaescore.com/>> [Diakses 17 September 2018].
- Somantri, O. & Hasta, I. D., 2017. Implementasi *e-Government* Pada Kelurahan Pesurungan Lor Kota Tegal Berbasis *Service Oriented Architecture*, [online] Tersedia di: <<http://ejournal.poltektegal.ac.id/>> [Diakses 4 April 2018].
- Suharno, C. F., Fauzi, M. A. & Perdana, R. S., 2017. Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan SAMBAT Online Menggunakan Metode *K-Nearest Neighbor* Dan *Chi-Square*. Malang: Universitas Brawijaya.
- Sun, Changqiu, Xiaolong Wang, dan Jun Xu. 2009. *Study on Feature Selection in Finance Text Categorization*. *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*. San Antonio.
- Zheng, Z., Wu, X. & Srihari, R., 2014. *Feature Selection for Text Categorization on Imbalance Data*. *ACM SIGKDD Exploration Newsletter – Special Issue on learning from imbalance datasets*, [online] Tersedia di: <<https://dl.acm.org/citation.cfm?id=1007741>> [Diakses 29 Juni 2018].