Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 3, Maret 2019, hlm. 2173-2179 http://j-ptiik.ub.ac.id

Klasifikasi Dokumen SAMBAT Online Menggunakan Metode Naive Bayes dan Seleksi Fitur Berbasis Algoritme Genetika

Tony Faqih Prayogi[1], Imam Cholissodin[2], Edy Santoso[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]tonyfaqih@gmail.com, [2]imamcs@ub.ac.id, [3]edy144@ub.ac.id

Abstrak

Sistem Aplikasi Masyarakat Bertanya Terpadu (SAMBAT) Online adalah salah satu aplikasi yang menjadi sistem eGov di Kota Malang untuk memberikan tempat bagi masyarakat Kota Malang untuk menyuarakan aspirasinya terhadap permasalahan yang ada untuk kebaikan kota malang itu sendiri. Semua pengaduan yang masuk melalui SAMBAT Online telah dikelompokkan bedasarkan bagian yang ada dan nantinya akan dipilah secara manual dan diteruskan ke bagian Satuan Kerja Perangkat Daerah (SKPD) masing-masing agar dapat segera ditindaklanjuti. Namun karena banyaknya pengaduan yang diterima sehingga cukup lama untuk diproses oleh SKPD masing-masing. Maka dari itu dibuat sebuah sistem untuk klasifikasi dokumen SAMBAT Online. Pada penelitian ini mengimplementasikan metode naïve bayes dan seleksi fitur berbasis algoritme genetika untuk klasifikasi dokumen SAMBAT Online. Proses implementasi itu sendiri terdiri dari proses preprocessing, term weighting, Seleksi Fitur menggunakan algoritme genetika dan proses klasifikasi menggunakan metode naïve bayes. Hasil pengujian yang telah dilakukan, didapatkan akurasi tertinggi sebesar 89.79% pada data uji sebanyak 49 dengan parameter banyak generasi 70, ukuran populasi 20, crossover rate 0.8 dan mutation rate 0.2.

Kata kunci: sistem aplikasi masyarakat bertanya terpadu (SAMBAT), seleksi fitur, algoritme genetika, naïve bayes

Abstract

Integrated Community Asking Application System (SAMBAT) Online is one of application that becomes an eGov system in Malang City to provide a place for the people of Malang City to voice their aspirations towards problems that exist for the good of the city itself. All complaints that enter through SAMBAT Online have been grouped based on the existing parts and later will be sorted manually and forwarded to the respective Regional Work Unit (SKPD) so that they can be immediately followed up. But because of the number of complaints received so long enough to be processed by each SKPD. Therefore a system was created for the classification of SAMBAT Online documents. In this study implemented a naïve bayes method and genetic algorithm-based feature selection for the SAMBAT Online document classification. The implementation process itself consists of preprocessing, term weighting, Feature Selection using genetic algorithms and the classification process using naïve bayes method. The results of the tests that have been done, obtained the highest accuracy of 89.79% in the test of 49 data test with the parameter value of generations 70, population size 20, crossover rate 0.8 and mutation rate 0.2.

Keywords: *integrated community application application system (SAMBAT), feature selection, genetic algorithm, naïve bayes*

1. PENDAHULUAN

Pada era globalisasi, kemajuan teknologi saat ini memberi dampak terhadap kehidupan sehari-hari. Dampak tersebut memberikan perubahan dalam berbagai bidang kehidupan seperti kesehatan, bisnis, kebudayaan, pendidikan bahkan pemerintahan. Pemerintah saat ini telah menerapkan teknologi informasi dalam menjalankan program pemerintah. E-Government atau eGov merupakan salah satu

wadah pemerintah untuk mengakomodir teknologi untuk mendukung administrasi pemerintahan. Sistem eGov sendiri memiliki fungsi untuk memberikan tempat bagi masyarakat umum untuk menyalurkan aspirasinya berupa pertanyaan, kritik, saran dan lain sebaginya kepda pemerintah.

Kota Malang sendiri juga menjadi salah satu Kota yang menerapakan sistem eGov ini. Sistem Aplikasi Masyarakat Bertanya Terpadu (SAMBAT) Online adalah salah satu aplikasi yang menjadi sistem eGov di Kota Malang untuk memberikan tempat bagi masyarakat Kota Malang untuk menyuarakan aspirasinya terhadap permasalahan yang ada untuk kebaikan kota malang itu sendiri. Masyarakat dapat mengirimkan pengaduan, pertanyaan, saran atu kritik terhadap kinerja pemerintah Kota Malang melalui situs web secara langsung yaitu pada alamat http://www.sambat.malangkota.go.id atau melalui pesan singkat yang dikirimkan kepada nomor yang ada pada situs web tersebut.

Pengaduan tersebut juga sebuah masukan untuk Kota Malang agar dapat meningkatkan kualitas mereka dan memenuhi kepuasan seluruh masyarakat Kota Malang. Semua pengaduan yang masuk melalui SAMBAT Onine telah dikelompokkan bedasarkan bagian yang ada dan nantinya akan dipilah secara manual dan diteruskan ke bagian Satuan Kerja Perangkat Daerah (SKPD) masing-masing agar dapat segera ditindaklanjuti. Maka dari itu diperlukanlah sistem pengklasifikasian dokumen SAMBAT Online berdasarkan bagian SKPD yang dituju untuk setiap pengaduan yang diberikan oleh masyarakat Kota Malang.

Klasifikasi pada teks merupakan proses pengkategorian suatu dokumen ke dalam suatu kelas yang memiliki kemiripan yang sama. Proses klasifikasi dilakukan untuk memberikan label pada setiap dokumen berdasarkan masing-masing bagian brikorat seperti keuangan, perlengkapan dan lain sebagainya sehingga dokumen-dokumen tersebut dapat diteruskan ke tujuannya. Naïve bayes merupakan salah satu metode klasifikasi yang memiliki hasil akurasi yang cukup tinggi saat diuji dan diimplementasi untuk beberapa permasalahan dalam klasifikasi (Soelistio dan Surendra, 2013). Naïve bayes juga memiliki kelebihan yaitu metode yang sederhana dan proses komputasi yang cepat namun memilik tingkat akurasi yang tinggi. Oleh karena itu metode naïve bayes menjadi metode klasifikasi yang sering digunakan.

Dalam penelitian yang dilakukan oleh Baharsyah (2014), proses klasifikasi dokumen SAMBAT Online ke dalam kategori positif atau negatif menggunakan metode K-Nearest Neighbor (KNN). Dalam penelitian tersebut dihasilkan tingkat akurasi sebesar 81,17647%. Dalam Penelitian lainnya yang dilakukan oleh Saptono, Wiranto dan Suryono (2016), proses klasifikasi keluhan menggunakan metode naïve bayes. Dalam penelitian tersebut menghasilkan tingkat akurasi sebesar 87%. Naïve bayes terbukti menjadi salah satu metode klasifikasi yang memliki tingkat akurasi cukup tinggi.

Dalam pemrosesan teks, dokumen akan dilakukan pre-processing untuk mendapatkan informasi yang diinginkan yaitu term atau kata unik yang dapat dijadikan sebagai fitur untuk proses klasifikasi. Banyaknya dokumen SAMBAT Online yang ada serta term yang didapat, maka fitur yang terdapat dalam sistem juga akan semakin kompleks. Dalam proses klasifikasi teks juga terdapat masalah utama yaitu dimensi dari ruang fitur yang terlalu tinggi (Chen, 2009). Masalah itu sering terjadi pada teks yang memiliki puluhan ribu fitur dan sebagian besar fitur tersebut tidak relevan dengan proses klasifikasi teks atau bahkan dapat mengurangi tingkat akurasi. Untuk itu perlu dilakukan proses seleksi fitur atau pengurangan fitur agar proses dalam sistem semakin efektif dan hasil yang didapat juga lebih maksimal. Penelitian yang dilakukan oleh Bidi dan Elberrichi (2016) menunjukkan bahwa algoritme genetika merupakan salah satu algoritme yang efektif untuk proses seleksi fitur. Pada penelitian tersebut seleksi fitur berbasis algoritme genetika terbukti dapat meningkatkan performa dan akurasi dalam setiap kasus dan setiap metode klasifikasi yang ada..

2. LANDASAN KEPUSTAKAAN

2.1. SAMBAT *ONLINE*

Terpenuhinya kepuasan masyarakat adalah salah satu tujuan adanya SAMBAT Online di Pemerintah Kota Malang. Demi meningkatkan standar mutu Kota Malang secara konsisten maka diperlukan pengembangan secara berkelanjutan. Pengaduan yang disampaikan oleh masyarakat Kota Malang menjadi masukan yang penting dan perlu ditindaklanjuti agar pengembangan dan perbaikan dapat dilaksanakan dengan segera. Pengaduan masyarakat Kota Malang menjadi wadah untuk melakukan evaluasi terhadap kinerja Kota

Malang

2.2. Pemrosesan Teks

Text mining adalah proses yang berasal wawasan dari teks. Informasi ini biasanya diperoleh melalui menentukan pola dan tren dalam teks melalui metode seperti belajar pola statistik. Hal ini biasanya melibatkan proses penataan teks input, mencari pola dalam data terstruktur, dan akhirnya mengevaluasi dan menafsirkan output.

Pemrosesan teks terdiri dari beberapa tahapan, yaitu :

1. *Cleaning* yang berfungsi untuk pembersihan karakter-karakter yang tidak diperlukan dalam dokumen
2. *Tokenizing* adalah proses memecah aliran teks dalam kata-kata, frase, simbol, atau elemen bermakna lainnya yang disebut tokens. Tujuan dari tokenization adalah eksplorasi kata-kata dalam kalimat.
3. *Filtering* merupakan proses untuk menghilangkan kata-kata tidak memberikan kontribusi pada konteks atau isi dari dokumen
4. *Stemming* merupakan salah satu proses yang digunakan untuk mendapatkan informasi dalam sebuah teks atau dokumen yaitu dengan cara mengubah kata yang ada dalam sebuah dokumen menjadi kata dasar.

2.3. Seleksi Fitur

Seleksi Fitur merupakan salah satu praproses dalam suatu klasifikasi. Proses seleksi fitur bertujuan untuk memilih fitur – fitur yang relevan dalam suatu data. Dengan proses seleksi fitur maka dimensi data akan berkurang dan dapat meningkatkan efektifitas dan efisiensi dalam suatu proses klasifikasi.

2.4. Algoritme Genetika

Algoritme genetika merupakan penggunaan teknik yang terinspirasi dari biologi evolusi seperti seleksi, mutasi, warisan dan rekombinasi untuk memecahkan masalah. Metode yang paling umum digunakan dalam algoritme genetika adalah untuk menciptakan kelompok individu secara acak dari suatu populasi tertentu. Individu yang terbentuk dievaluasi dengan bantuan fungsi evaluasi yang disediakan oleh programmer. Individu kemudian diberikan dengan skor yang secara tidak langsung menyoroti kebugaran untuk situasi tertentu. Dua individu yang terbaik kemudian digunakan untuk membuat satu atau lebih keturunan, setelah mutasi acak dilakukan pada keturunannya. Tergantung pada kebutuhan aplikasi, prosedur berlanjut sampai solusi optimal yang dapat diterima atau sampai sejumlah generasi tertentu.

2.5. Naïve Bayes

Naïve Bayes Classifier berdasarkan teorema Bayes dengan asumsi independence antara prediktor. Sebuah model Naïve Bayes mudah untuk dibangun, tanpa estimasi parameter berulang rumit yang membuatnya sangat berguna untuk dataset yang sangat besar. Meskipun sederhana, classifier Naïve Bayes sering bekerja dengan sangat baik dan secara luas digunakan karena sering melebihi metode klasifikasi yang lebih canggih.

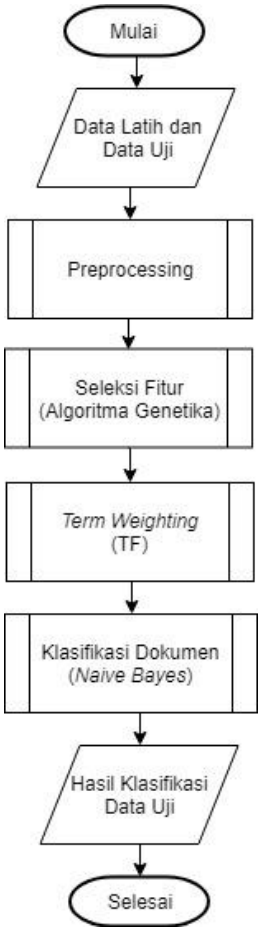
2.5.1. Multinomial Naïve Bayes

Multinomial Naïve Bayes adalah versi khusus dari Naïve Bayes yang dirancang lebih untuk dokumen teks. Sedangkan Naïve Bayes sederhana terhadap model dokumen sebagai kehadiran dan tidak adanya kata-kata tertentu, multinomial Naïve Bayes secara eksplisit model penghitungan kata dan menyesuaikan perhitungan yang mendasari untuk ditangani.

3. METODE

3.1. Perancangan Algoritma

Metode yang diimplementasikan pada penelitian ini adalah metode klasifikasi naïve bayes dan seleksi fitur algoritme genetika.



Gambar 2. *Flowchart* Perancangan Algoritma

Tahap pertama dari implementasi metode adalah proses pembacaan data yaitu 192 banyak data yang terdiri dari 100 data untuk kategori Dishub, 37 data untuk kategori DKP, dan 67 data untuk kategori DPUPPB.

Pada *Preprocessing* yang terdiri dari tahapan cleaning, tokenizing, filtering dan stemming akan dilakukan pada data dokumen yang ada untuk didapatkan informasi yang dibutuhkan.

Seleksi Fitur dilakukan menggunakan algoritme genetika. Pada tahap ini akan diproses seleksi untuk memilih fitur – fitur yang relevan sehingga diimensi data akan berkurang dan meningkatkan efektifitas dan efisiensi dalam proses klasifikasi nantinya. Dari seleksi fitur algoritme genetika tersebut akan dicari individu terbaik dan hasil fitur yang telah terseleksi.

Tahap selanjutnya adalah proses *Term* Weighting. Proses perhitungan dilakukan dengan menghitung berapa kali kata muncul pada semua dokumen pada tiap kategori

Proses Klasifikasi menggunakan naïve bayes merupakan tahapan terakhir. Tahap ini dilakukan dengan menghitung peluang tiap kata yang dapat dihitung dengan persamaan 1

$$P(w_i|c_j) = \frac{count(w_i.c) + 1}{(\sum_{w \in V} count(w.c)) + |V|} \quad (1)$$

Keterangan:

- $count(w_i.c)$ adalah jumlah kata query yang muncul dalam suatu kelas.
- $\sum_{w \in V} count(w.c)$ adalah jumlah seluruh kata yang ada didalam kelas.
- $|V|$ adalah jumlah seluruh kata unik yang ada di semua kelas.

4. HASIL DAN PEMBAHASAN

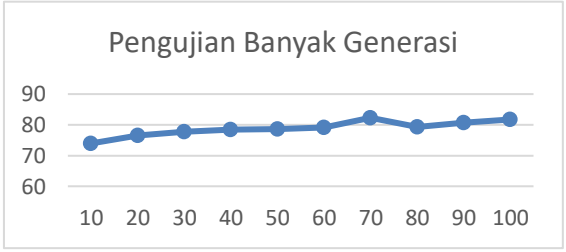
Berikut adalah hasil dari implementasi GA-LVQ2 beserta pembahasan dari pengujian yang telah dilakukan.

4.1. Pengujian Banyak Generasi

Pada pengujian banyak generasi ini memiliki tujuan untuk mengetahui berapa banyak generasi yang memiliki hasil paling optimal. Pengujian banyak generasi akan dilakukan sebanyak 5 kali uji coba dengan rentang nilai banyak generasi 10 sampai 100. Pengujian akan dilakukan pada ukuran populasi sebanyak 5, crossover rate 0.5 dan mutation rate 0.5. Hasil pengujian banyak generasi dijabarkan pada Tabel 1 dan Gambar 3.

Tabel 1. Rataan *Fitness* Pengujian Banyak Generasi

PopSize	**Rataan *Fitness***
10	73.95
20	76.56
30	77.80
40	78.43
50	78.64
60	79.05
70	82.29
80	79.26
90	80.62
100	81.76



Gambar 3. Grafik Pengujian Banyak Generasi

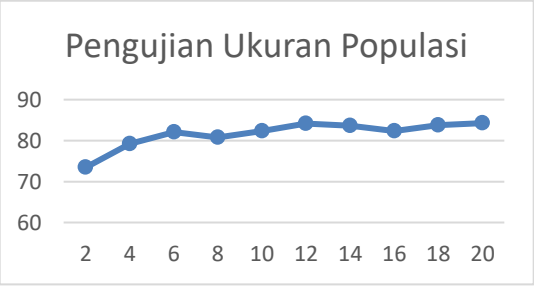
Diliihat darii hasiil pengujian diatas, grafiik menunjukkan hasil fitness yang naiik darii generasi ke 10 hingga generasi ke 70 kemudian turun dari generasi 70 hingga generasi 80 dan kembali naik dari generasi 80 hingga generasi 100. Pada banyak generasi 70 merupakan banyak generasi yang memiliki hasil paling optimal dengan rata-rata fitness 82.29% sehingga banyak generasi 70 akan digunakan pada pengujian selanjutnya.

4.2. Pengujian Ukuran Populasi

Pengujiian ukuran populasi memiliki tujuan untuk mengetahuii banyak generasi yang memiliki hasil optimal. Pengujiian diilakukan sebanyak 5 kali pada setiap ukuran populasi dengan nilai ukuran populasii 2 sampai dengan 20. Pengujian dilakukan dengan banyak generasi 70, crossover rate 0.5 dan mutation rate 0.5. Hasil pengujian ukuran populasi dijabarkan pada Tabel 2 dan Gambar 4.

Tabel 2. Rataan *Fitness* Pengujian Ukuran Populasi

Generasi	Rataan *Fitness*
2	73.43
4	79.26
6	82.08
8	80.83
10	82.39
12	84.16
14	83.64
16	82.39
18	83.74
20	84.26



Gambar 4. Grafik Pengujian Ukuran Populasi

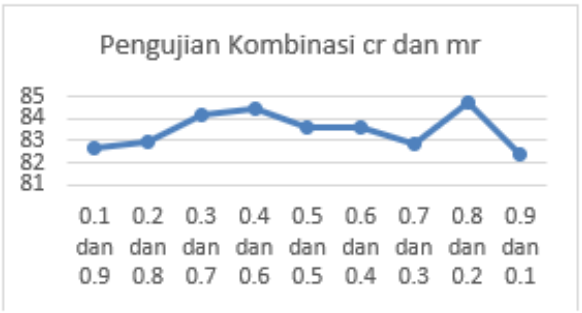
Dari grafik diatas, hasiil pengujiian yang didapatkan tidak stabil, berberapa ukuran populasi mengalami kenaikan, beberapa populasi juga mengalami penurunan. Hal ini terjadi karena pembentukan nilai individu awal yang dilakukan secara acak sehingga nilai fitness yang diapatkan tergantung dengan nilai individu pada inisialisasi awal. Nilai ukuran populasi 20 akan digunakan untuk pengujian selanjutnya karena memiliki rata-rata fitness yang paling tinggi.

4.3. Pengujian Kombinasi *cr* dan *mr*

Pengujian kombinasi cr dan mr akan dilakukan 5 kali dengan nilai 0.1, sampai 0.9. Pengujian kombinasi cr dan mr akan dilakukan dengan nilai banyak generasi 70 dan ukuran populasi. Hasil pengujian kombinasi *crossover rate* dan *mutation rate* dapat dilihat pada Tabel 3 dan Gambar 5.

Tabel 3. Rataan *Fitness* Pengujian Kombinasi *cr mr*

cr	mr	Rataan *Fitness*
0.1	**0.9**	82.70
0.2	**0.8**	82.91
0.3	**0.7**	84.16
0.4	**0.6**	84.47
0.5	**0.5**	83.64
0.6	**0.4**	83.64
0.7	**0.3**	82.80
0.8	**0.2**	84.68
0.9	**0.1**	82.39



Gambar 5. Grafik Pengujian Kombinasi *cr mr*

Dari gambar grafik hasil pengujian diatas menunjukkan hasil yang tidak stabil. Pada kombinasi 1 sampai kombinasi 4 memiliki hasil yang naik namun nilai fitness turun pada kombinasi 4 ke kombinasi 7, kemudian naik pada kombinasi 8 dan kembali turun pada kombinasi 9. Kombinasi 7 yaitu dengan nilai cr 0.7 dan mr 0.3 memiliki hasil rata-rata fitness yang paling tinggi dari pada nilai kombinasi lainnya dengan nilai 84.68. Oleh karena itu nilai kombinasi cr 0.8 dan mr 0.2 akan digunakan pada pengujian yang selanjutnya.

4.4. Pengujian Hasil Akurasi

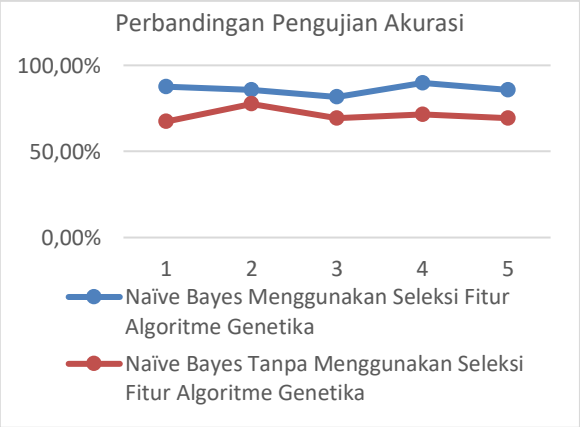
Pada pengujian hasil akurasi dilakukan pengujian terhadap fitur-fitur yang telah terseleksi berdasarkan hasil individu terbaik dengan data uji yang telah disiapkan. Pengujian dilakukan dengan cara membandingkan kelas data uji dengan kelas hasil klasifikasi sistem. Pengujian hasil akurasi ini menggukan nilai parameter Algoritme Genetika yang terbaik hasil

pengujian sebelumnya yaitu banyak banyak generasi 70, ukuran populasi sebesar 20, nilai *crossover rate* 0.8 dan *mutation rate* 0.2.

Hasil perbandingan pengujian akurasi Naïve bayes dan seleksi fitur berbasis algoritme genetika memilih hasil akurasi tertinggi berdasarkan individu yang ada. Tabel 4 akan menunjukkan hasil dari perbandingan pengujian akurasi dengan melakukan percobaan sebanyak 5 kali.

Tabel 4. Hasil Perbandingan Pengujian Akurasi

Percobaan ke	Naïve Bayes-AG	Naïve Bayes
1	87.75%	67.34%
2	85.71%	77.55%
3	81.63%	69.38%
4	89.79%	71.42%
5	85.71%	69.38%



Gambar 6. Grafik Pengujian Akurasi

Dari hasil grafik diatas dapat dilihat bahwa hasil perbandingan pengujian akurasi antara naïve bayes mengguanak seleksi fitur algoritme genetika dengan yang tanpa menggunakan seleksi fitur algoritme gentika masih tidak stabil. Hal ini diikarenakan pemberiian nilaii individu awal sangaat berpengaaruh padaa hasiil pengujian. Dari 5 kalii percobaan penguijian akurasii Naïve Bayes Menggunakan Seleksi Fitur Algoritme Genetika didapatkan rata-rata akurasi sebesar 86.12% dan akurasi tertinggi yang dihasilkan adalah 89.79%. Sedangkan untuk hasil pengujian akurasi Naïve Bayes Tanpa Menggunakan Seleksi Fitur Algoritme Genetika didapatkan rata-rata akurasi sebesar 71.01% dan akurasi tertinggi yang dihasilkan adalah 77.55%. Ini artinya implementasi Naïve Bayes Menggunakan Seleksi Fitur Algoritme Genetika memiliki hasil yang lebih baik.

5. PENUTUP

5.1. Kesimpulan

Dari hasil yang didapatkan pada penelitian klasifikasi dokumen sambat online menggunakan metode naïve bayes dan seleksi fitur berbasis algoritme genetika dapat diambil kesimpulan :

1. Klasifikasi dokumen sambat online menggunakan metode naïve bayes dan seleksi fitur berbasis algoritme genetika dapat diimplementasikan sesuai dengan perancangan yang telah dibuat.
2. Dalam Implementasi metode naïve bayes dan seleksi fitur berbasi algoritme genetika untuk klasifikasi dokumen SAMBAT *online* menghasilkan rata-rata akurasi sebesar 86.12% dengan nilai akurasi tertinggi sebesar 89.79% dalam 5 kali percobaan menggunakan data uji sebanyak 49 data dan dengan nilai parameter terbaik yaitu banyak generasi 70, ukuran populasi sebesar 20, nilai *crossover rate* 0.8 dan *mutation rate* 0.2.

5.2. Saran

Penelitian yang dilakukan ini, bagi penulis masih memiliki beberapa kekurangan. Oleh karena untuk penelitian selanjutnya, penulis akan memberikan beberapa saran seperti berikut:

1. Pada penelitian berikutnya dapat menambahkan jumlah data yang digunakan agar akurasi yang didapatkan lebih baik.
2. Untuk penelitian lainnya bisa menggunakan metode klasifikasi yang lainnya seperti K-Nearest Neighbor, Support Vector Machine atau metode klasifikasi lainnya untuk dijadikan perbandingan dengan penelitian ini.
3. Pada penelitian berikutnya dapat menambahkan metode seleksi fitur lainnya seperti Information Gain, Forward Selection atau algoritme seleksi fitur lainnya.

6. DAFTAR PUSTAKA

Bidi, Noria & Elberrichi, Zakaria. 2016. Feature Selection For Text Classificationusing Genetic Algorithms. 8th International Conference on Modelling, Identification

and Control (ICMIC-2016). Algiers, Algeria, 15-17 November 2016.

Baharsyah, Ikram. 2014. Klasifikasi Deep Sentiment Analysis SAMBAT Online Universitas Brawijaya Menggunakan Metode K-Nearest Neighbor. S1. Universitas Brawijaya Malang.

Pratama, Enda Esyudha & Trilaksono, Bambang Riyanto, 2015. Klasifikasi Topik Pengaduan Pelanggan Berdasarkan Tweet Dengan Menggunakan Penggabungan Feature Hasil Ekstraksi Pada Metode Support Vector Machine (SVM). Jurnal Edukasi dan Penelitian Informatika. Volume 1, No. 2, 2015.

Saptono, Restu., Wiranto & Suryono, Wachid Daga. 2016. Sistem Klasifikasi Pengaduan Pelanggan Di UPT TIK UNS Menggunakan Algoritme Navie Bayesian Classifier. Seminar Nasional Teknologi Informasi dan Komunikasi 2016. Yogyakarta, 18-19 Maret 2016.

Sarwaswati, Ni Wayan Sumartini. 2011. Text Mining Dengan Metode Naïve Bayes Dan Support Vector Machines Untuk Sentiment Analysis. S2. Universitas Udayana.

Soelistio, Yustinus Eko & Surendra, Martinus Raditia Sigit. 2013. Simple Text Mining For Sentiment Analysis Of Political Figure Using Naïve Bayes Classifier Method. The Proceedings of The 7th ICTS. Bali, Indonesia, 15-16 Mei 2013.