Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan *Naïve Bayes* dan Algoritme Genetika

Oman Somantri[1], Mohammad Khambali[2]

***Abstract*— Classification of short stories category based on age of the reader is still difficult. Therefore, a decision support system to classify the short stories category is needed. Naïve Bayes is one of methods suitable for short stories classification. However, Naïve Bayes has flaws in accuracy level, and needs to be optimized. In this paper, Genetic algorithm is proposed to increase the level of accuracy. In this case, genetic algorithm is used for feature selection. The results show an increase in the level of accuracy produced. The accuracy increases from 78,59% to 84,29%. In conclusion, the application of genetic algorithm on Naïve Bayes in classifying the online short stories category can improve the accuracy.**

***Intisari*— Penetapan klasifikasi kategori cerpen yang sesuai dengan keinginan, yaitu adanya kesesuaian jenis cerpen yang dibaca dengan usia pembaca, masih sulit dilakukan. Hal ini menjadikan perlu adanya pendukung keputusan untuk dapat mengklasifikasikan kategori cerpen sehingga jenis cerpen tersebut sesuai dengan kategori pembaca. *Naïve Bayes* merupakan metode yang sesuai untuk digunakan dalam klasifikasi kategori cerpen. *Naïve Bayes* masih mempunyai kekurangan dalam tingkat akurasi yang dihasilkan, sehingga perlu adanya optimasi. Algoritme genetika sebagai algoritme optimasi merupakan metode yang diusulkan untuk meningkatkan tingkat akurasi yang digunakan. Dalam hal ini, algoritme genetika digunakan untuk *feature selection*. Hasil yang diperoleh memperlihatkan adanya sebuah peningkatan akurasi, yaitu dari 78,59% menjadi 84,29%. Dengan demikian, dapat disimpulkan bahwa penerapan algoritme genetika pada *naïve bayes* untuk klasifikasi kategori cerpen *online* dapat meningkatkan akurasi.**

***Kata Kunci*— klasifikasi, kategori cerpen, *Naive Bayes*, algoritme genetika.**

I. PENDAHULUAN

Indonesia merupakan negara yang memiliki berbagai macam budaya dan kultur yang beragam, seperti halnya dalam penciptaan karya sastra, sehingga tidak heran apabila setiap daerah di Indonesia mempunyai karya sastra yang beragam. Cerpen adalah sebuah karya sastra yang sangat disukai oleh semua orang, terutama di Indonesia. Institusi pendidikan pun pada proses pembelajaran yang dilaksanakan di sekolah memberikan materi yang terkait dengan cerita pendek. Hal inilah yang memberikan motivasi kepada setiap orang untuk dapat membuat sebuah cerpen sesuai dengan keinginan dan inspirasi yang muncul dari penulis. Saat ini penulisan cerpen masih didominasi oleh para remaja dan anak-anak, sehingga tidak heran apabila cerpen yang muncul sekarang ini sudah banyak dari lintas generasi baik itu ditulis lewat buku, majalah, *blog* pribadi, maupun media *online* serta situs *website* yang menyediakan cerpen *online* yang menyajikan kumpulan cerpen yang dikirim oleh penulisnya.

Sebuah karya sastra mempunyai dua unsur pokok di dalamnya. Yang pertama adalah unsur intrinsik, yaitu unsur dalam sastra yang ikut memengaruhi terciptanya karya sastra yang terdiri atas tema dan amanat, alur cerita, penokohan atau perwatakan, latar, sudut pandang, dialog, dan gaya bahasa. Sedangkan yang kedua adalah unsur ekstrinsik, yaitu unsur luar sastra yang ikut memengaruhi terciptanya karya sastra, meliputi sosial ekonomi, kebudayaan, sosio politik, keagamaan, dan adat istiadat dalam masyarakat [1]. Cerpen adalah cerita atau kisah pendek dengan jumlah kata kurang dari 10.000 dengan memberikan kesan tunggal dan ceritanya terpusat pada salah satu tokoh [2]. Cerpen adalah cerita fiktif yang belum pasti kebenarannya serta ceritanya relatif pendek dan cerpen bukanlah suatu analisis argumentatif [3]. Cerpen merupakan salah satu bentuk prosa naratif fiktif, pada umumnya berbentuk suatu karangan fiksi seperti fiksi ilmiah, fiksi detektif, fiksi horror, dan lainnya. Penentuan kategori genre sebuah cerpen bagi beberapa kalangan sangatlah berpengaruh, terutama dalam penentuan layak tidaknya sebuah cerpen tersebut dibaca oleh kalangan tertentu, seperti orang dewasa, remaja, dan anak-anak. Cerpen adalah tulisan berbentuk teks yang ditulis oleh penulis sesuai dengan inspirasi penulis, baik itu ditulis secara *online* maupun tidak.

Terdapat beberapa metode berbasis *machine learning* dan statistik yang dapat digunakan untuk klasifikasi teks, di antaranya adalah *Decision Tress*, *k-Nearest Neighbors* (kNN), *Neural Networks*, *Naïve Bayes* (NB), *Support Vector Machines* (SVM), dan *Associative Classification*. Penelitian yang dilakukan dengan menggunakan SVM merupakan metode yang sering digunakan oleh para peneliti dalam *text mining* [4]--[6]. Salah satu di antara metode terbaik yang ada, NB, adalah metode yang populer dalam klasifikasi teks, yaitu sebagai salah satu metode komputasi yang efisien dan juga mempunyai *performance predictive* yang baik [7].

Penentuan klasifikasi kategori cerpen dilakukan dengan menggunakan *text mining* sebagai metode yang memungkinkannya. Salah satu di antaranya adalah NB. NB merupakan algoritme yang sering digunakan dalam pengkategorian teks. Konsep dasarnya adalah menggabungkan probabilitas kata-kata dan kategori sebuah dokumen [8]. Di samping kelebihan yang dimilikinya, NB memiliki kelemahan,

[1]*Dosen, Jurusan Teknik Informatika Politeknik Harapan Bersama Tegal, Jln. Mataram No.09 Pesurungan Lor Kota Tegal 52147 INDONESIA (tlp: 0283-352000; fax: 0283-350567; e-mail: oman.somantri@poltektegal.ac.id)*

[2]*Dosen, Jurusan Teknik Listrik Politeknik Negeri Semarang, Jln. Prof. H. Soedarto, S.H., Tembalang, Semarang 50275 INDONESIA (tlp: 024-7473417 ext.126; fax:024-7472396; e-mail: mc.chambali.poltek@gmail.com)*

yaitu sangat sensitif terhadap pemilihan fitur seleksi. Oleh karena itu, sebuah pemilihan fitur yang sesuai dengan model yang diusulkan sangatlah diperlukan. Masalah utama dalam klasifikasi teks adalah dimensi tinggi dari ruang fitur. Hal ini sering terjadi pada teks yang memiliki puluhan ribu fitur. Sebagian besar fitur tersebut tidak relevan dan tidak bermanfaat bagi klasifikasi teks, bahkan dapat mengurangi tingkat akurasi *(accuracy)* [7]. Pada umumnya, atribut dari klasifikasi teks sangat besar dan jika semua atribut tersebut digunakan, maka akan mengurangi kinerja dari *classifier* serta untuk mendapatkan akurasi yang lebih baik, atribut yang ada harus dipilih dengan algoritme yang tepat [9], [10].

Pada beberapa penelitian, sebenarnya peningkatan akurasi pada klasifikasi teks pada NB sudah dilakukan dengan beberapa pendekatan, di antaranya adalah *structure extension, local learning, instance weighting, feature selection,* dan *feature weighting* [11].

Fokus pada makalah ini adalah peningkatan akurasi klasifikasi teks dengan menggunakan pendekatan *feature selection*. *Feature Selection* adalah sebuah cara yang digunakan untuk dapat mengoptimalkan kinerja dari *classifier*. Cara kerjanya berdasar pada pengurangan ruang fitur yang besar, yaitu dengan cara mengeliminasi atribut yang kurang relevan serta dengan menggunakan penggunaan algoritme *feature selection* yang tepat sehingga dapat meningkatkan akurasi [10]. Algoritme genetika (GA) adalah salah satu algoritme optimasi yang dapat digunakan untuk permasalahan *feature selection*, selain *Ant Colony Optimization* (ACO) dan *Particle Swarm Optimization* (PSO). Dengan kemampuannya sebagai algoritme optimasi, maka diharapkan GA menghasilkan sebuah peningkatan akurasi pada klasifikasi kategori cerpen dengan menggunakan NB.

Beberapa penelitian menggunakan GA telah dilakukan. GA berbasis *latent semantic features* diusulkan untuk mendapatkan representasi dokumen klasifikasi yang lebih baik [12]. Penelitian lain menerapkan konsep *biological evolution* untuk meningkatkan GA [13]. Penelitian terkait dengan kinerja dari kombinasi DF, IG, MI, dan metode CHI dengan GA juga telah dilakukan. Pada penelitian ini, GA dijadikan sebagai metode *Feature Selection* untuk kategori teks [14], [15]. Sebuah penelitian lain menerapkan metode *hybrid feature selection* berbasis peningkatan GA untuk pengkategorian teks. Pada penelitian ini, teknik pencarian *hybrid* dikombinasikan menjadi metode *filter feature selection* tingkat tinggi *Enhanced* GA (EGA) untuk mengantisipasi dimensi tinggi dari *feature space* dan meningkatkan kinerja kategori teks secara simultan [16].

Dari semua penelitian yang telah dilakukan, model NB masih mempunyai kekurangan dalam tingkat akurasi yang dihasilkan, sehingga perlu dilakukan optimasi. Optimasi dengan model yang tepat dan terbaik untuk klasifikasi teks, khususnya klasifikasi kategori genre cerpen dilakukan dalam penelitian ini, sehingga diperoleh tingkat akurasi yang lebih baik. Pada penelitian ini dilakukan optimasi pada NB dengan menggunakan algoritme optimasi yaitu GA yang akan digunakan sebagai model untuk *feature selection,* sehingga diperoleh nilai akurasi klasifikasi teks kategori cerpen yang lebih baik.

II. NAÏVE BAYES DAN ALGORITME GENETIKA

A. *Naïve Bayes*

NB adalah metode yang digunakan dalam statistika untuk menghitung peluang dari suatu hipótesis. *Naïve Bayes* nmenghitung peluang suatu kelas berdasarkan atribut yang dimiliki dan menentukan kelas yang memiliki probabilitas paling tinggi. NB mengklasifikasikan kelas berdasarkan probabilitas sederhana dengan mengasumsikan bahwa setiap atribut dalam data tersebut bersifat saling terpisah. Metode NB merupakan salah satu metode yang banyak digunakan berdasarkan beberapa sifatnya yang sederhana. Metode ini mengklasifikasikan data berdasarkan probabilitas *P* atribut *x* dari setiap kelas *y* data [17]. Model probabilitas setiap kelas *k* dan jumlah atribut *a* dapat dituliskan seperti persamaan berikut.

$$P = (y_k|x_1, x_2, \ldots . x_a). \quad (1)$$

Perhitungan NB yaitu probabilitas kemunculan dokumen X_a pada kategori kelas Y_k $P(x_a|y_k)$ dikali dengan probabilitas kategori kelas $P(y_k)$. Dari hasil kali tersebut kemudian dilakukan pembagian terhadap probabilitas kemunculan dokumen $P(x_a)$, sehingga diperoleh rumus perhitungan NB seperti dituliskan pada (2).

$$P(y_k|x_a) = \frac{P(y_k)P(x_a|y_k)}{P(x_a)} . \quad (2)$$

Selanjutnya, dilakukan proses pemilihan kelas yang optimal, yaitu dipilih nilai peluang terbesar dari setiap probabilitas kelas yang ada. Didapatkan rumus untuk memilih nilai terbesar seperti pada (3).

$$\mathrm{y}(x_i) = \arg\max \mathrm{P(y)} \prod_{i=1}^{a} P(x_1 |y). \quad (3)$$

Pembobotan suatu atribut kelas dapat meningkatkan pengaruh prediksi. Dengan memperhitungkan bobot atribut terhadap kelas, maka yang menjadi dasar ketepatan klasifikasi bukan hanya probabilitas, melainkan juga bobot setiap atribut kelas.

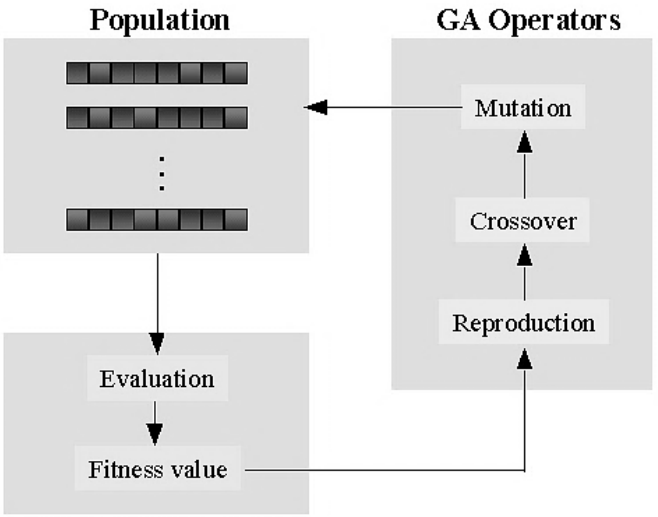
B. *Algoritme Genetika (GA)*

GA adalah sebuah metode pencarian yang disesuaikan dengan proses generik dari organisme biologi yang berdasar pada teori evolusi dan merupakan suatu teknik optimasi yang didasarkan pada prinsip genetika dan seleksi alam [14]. Algoritme genetika terinspirasi dari mekanisme seleksi alam, yaitu individu yang lebih kuat dimungkinkan akan menjadi pemenang dalam lingkungan yang kompetitif dan solusi yang optimal dapat diperoleh dan diwakilkan oleh pemenang akhir dari permainan [18].

Untuk menggunakan GA, solusi permasalahan direpresentasikan sebagai kromosom. Terdapat beberapa aspek penting dalam penggunaan GA, di antaranya adalah

- definisi fungsi *fitness*,
- definisi dan implementasi refresentasi genetik, dan
- definisi dan implementasi operasi genetik.

Adapun gambaran dari alur evolusi GA diperlihatkan seperti pada Gbr. 1 [19].



Gbr. 1 Alur evolusi GA.

III. METODOLOGI

A. Data Set dan Tools

Pada penelitian ini *data set* yang digunakan bersifat *public*. Data diambil dari http://cerpenmu.com/. Data yang diambil berupa teks cerita pendek *online* berdasarkan kategori yang ditentukan, yaitu kategori cerpen anak dan kategori cerpen dongeng. *Data set* diambil dari cerpen yang dibuat dan diunggah antara tahun 2015 sampai 2016 dengan jumlah data sebanyak 121 data cerpen. *Tools* yang digunakan pada adalah Rapid Miner 5.3 untuk analisis data teks.

B. Preprocessing Data

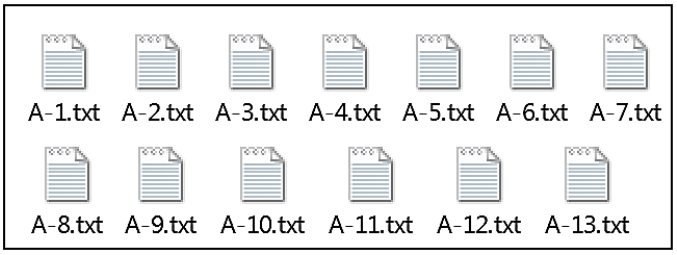
Sebelum *data set* dimasukkan ke dalam model yang disusulkan, terlebih dahulu dilakukan *praprocessing* data. Pada tahapan ini, dilakukan beberapa hal, yaitu konversi data teks menjadi *.txt*, *tokenized*, *transform cases*, *filter tokens*, *filter stopword*, dan *Stem*.

1) Text Converting: Agar teks yang akan dimasukkan kedalam model dapat dibaca oleh model, dilakukan konversi tipe format teks. Tahapan ini adalah dilakukannya sebuah konversi format teks yang diambil dari sumbernya menjadi format file *.txt*. Gbr. 2 menampilkan contoh hasil konversi teks.

2) Tokenized: Tokenized erupakan proses untuk memisah-misahkan kata. Hasil dari pemisahan tersebut dinamakan *token*.

3) Transform Cases: Proses ini dilakukan untuk mengubah bentuk kata-kata. Pada proses ini, karakter dijadikan huruf kecil *(lower case)* semua.

4) Filter Tokens: Proses ini mengambil kata-kata yang penting dari token yang sudah dihasilkan berdasarkan jumlah karakter. Parameter yang digunakan ada dalam proses ini adalah *min chars* = 3 dan *max chars* = 25.



Gbr. 2 Contoh format file *.txt* setelah dilakukan konversi *file*.

TABEL I
CONFUSION MATRIX

	Hasil Prediksi	
	Positive	***Negative***
Positive	*True Positive*	*False Positive*
Negtive	*False Negative*	*True Negative*

5) Filter Stopword: Proses ini menghilangkan kata-kata yang sering muncul tetapi tidak memiliki pengaruh apapun dalam ekstraksi klasifikasi teks. Pada proses ini, kata yang termasuk adalah penunjuk waktu, kata Tanya, dan kata sambung.

6) Stem: Stem Merupakan proses pengubahan bentuk kata menjadi kata dasar. Metode ini merupakan proses pengubahan bentuk kata menjadi kata dasar yang menyesuaikan struktur yang digunakan dalam proses *stemming*.

C. Metode yang Diusulkan

Pada penelitian yang dilakukan ini, metode yang diusulkan untuk optimasi peningkatan akurasi pada klasifikasi kategori cerpen adalah penerapan GA pada NB yang digunakan untuk *feature selection*, yaitu pada proses pemilihan atribut yang sesuai pada model sehingga terjadi peningkatan akurasi. Untuk mendapatkan tingkat akurasi yang sesuai, validasi data yang diterapkan adalah *k-Fold Cross Validations*, sehingga validasi hasil eksperimen menjadi lebih baik.

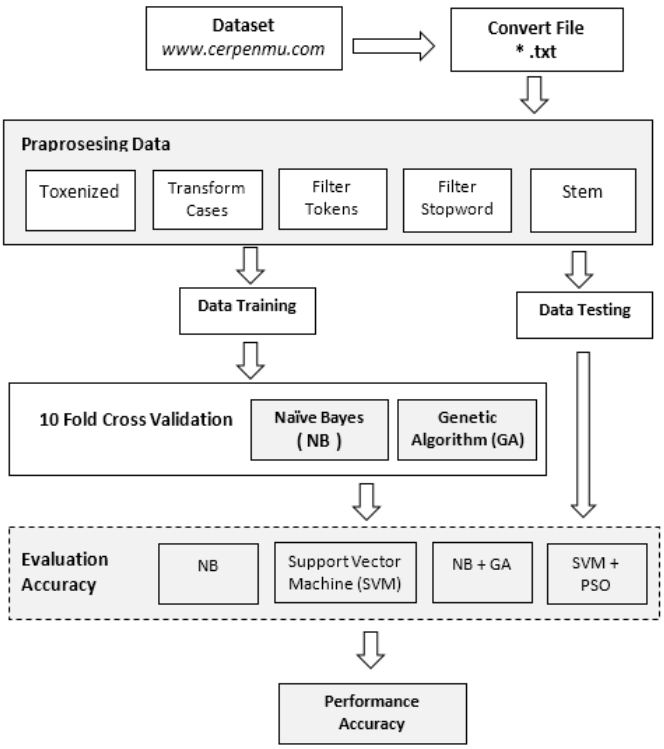
Untuk mengetahui keberhasilan dari model yang diusulkan, evaluasi dilakukan dengan melakukan komparasi antara tingkat akurari model NB klasik, SVM, NB dengan menggunakan GA (NB-GA), dan SVM dengan menggunakan *Particle Swarm Optimization* (SVM-PSO).

Evaluasi dilakukan untuk memperlihatkan perubahan yang terjadi pada model-model secara klasik, dalam hal ini NB dan SVM, dengan model yang telah dioptimalisasi dengan menggunakan *feature selection*.

Gbr. 3 memperlihatkan alur metode penelitian yang diusulkan pada proses klasifikasi kategori cerita pendek.

D. Evaluasi dan Validasi

Untuk mengetahui dan mendapatkan hasil model yang diusulkan sesuai dengn yang diharapkan, maka tahapan evaluasi dilakukan. Evaluasi model pada tahapan ini menggunakan evaluasi *matrix confusion*, seperti ditunjukkan pada Tabel I.



Gbr. 3 Metode yang diusulkan pada penelitian.

Untuk menghitung tingkat akurasi, digunakan persamaan sebagai berikut.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

dengan

- *True Positive* (TP) menunjukkan dokumen yang termasuk dalam hasil pengelompokan oleh sistem memang merupakan anggota kelas,
- *False Positive* (FP) menunjukkan dokumen yang termasuk dalam hasil pengelompokan oleh sistem ternyata seharusnya bukan merupakan anggota kelas,
- *False Negative* (FN) menunjukkan dokumen yang tidak termasuk dalam hasil pengelompokan oleh sistem ternyata seharusnya merupakan anggota kelas, dan
- *True Negative* (TN) menunjukkan bahwa dokumen yang tidak termasuk dalam hasil pengelompokan oleh sistem ternyata seharusnya bukan merupakan anggota kelas.

IV. HASIL DAN PEMBAHASAN

Penelitian dilakukan menggunakan komputer dengan spesifikasi CPU Intel Core i5 2,67 GHz, memori RAM 4 GB, sistem operasi Windows 7 Professional SP1 32-bit.

A. *Hasil Eksperimen Naïve Bayes*

Hasil eksperimen terhadap *data set* yang sudah diperoleh menggunakan model NB disajikan pada Tabel II.

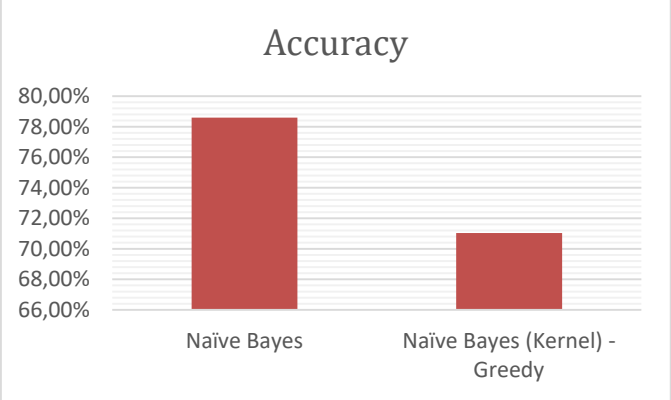
Pada Tabel II diperlihatkan bahwa tingkat akurasi dari hasil eksperimen menunjukkan NB menghasilkan tingkat akurasi sebesar 78,59%.

TABEL II
HASIL EKSPERIMEN MODEL NB

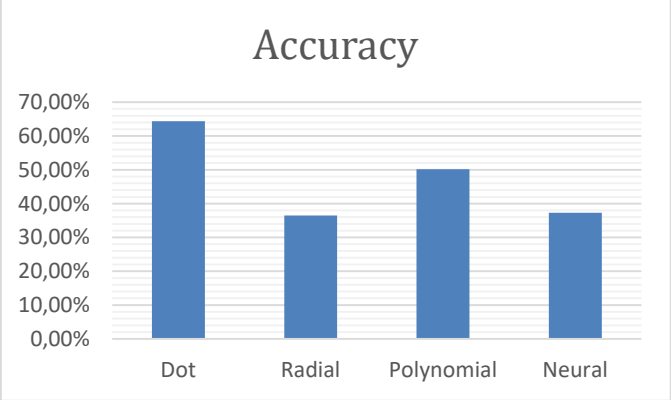
No	Hasil Eksperimen	
	Model	Akurasi
1	*Naïve Bayes*	78,59%
2	*Naïve Bayes (Kernel) - Greedy*	71,03%

TABEL III
HASIL EKSPERIMEN MODEL SVM

No	Hasil Eksperimen SVM	
	Tipe *Kernel*	Akurasi
1	*Dot*	64,36%
2	*Radial*	36,47%
3	*Polynomial*	50,19%
4	*Neural*	37,31%



Gbr. 4 Hasil eksperimen NB.



Gbr. 5 Hasil eksperimen SVM.

Gbr. 4 memperlihatkan grafik tingkat akurasi yang dihasilkan dari eksperimen dengan mengggunakan model NB.

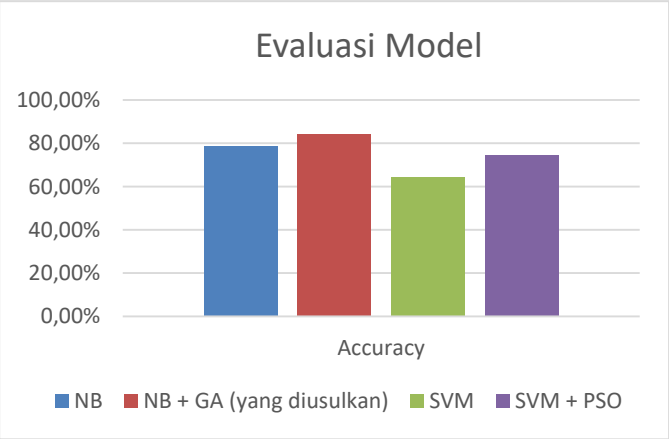
B. *Hasil Eksperimen Support Vector Machine*

Eksperimen berikutnya adalah dengan memasukkan *data set* pada model SVM. Hasil eksperimen yang dilakukan diperlihatkan pada Tabel III.

Dari Tabel III terlihat bahwa tingkat akurasi yang paling tinggi pada SVM adalah ketika digunakan tipe *kernel dot*, yaitu sebesar 64,36%.

TABEL IV
HASIL EKSPERIMEN OPTIMASI NB

No	Hasil Eksperimen Optimasi	
	Model	Akurasi
1	NB	78,59%
2	NB-GA (yang diusulkan)	84,29%
3	SVM	64,36%
4	SVM-PSO	74,49%



Gbr. 6 Hasil eksperimen optimasi model.

Pada Gbr. 5 diperlihatkan grafik tingkat akurasi hasil eksperimen menggunakan SVM dengan berbagai ipe *kernel*.

C. Hasil Optimasi Naïve Bayes dengan Algoritme Genetika

Setelah diperoleh hasil eksperimen yang dilakukan ke dalam model NB dan SVM, langkah selanjutnya adalah melakukan optimasi *feature selection* dengan menerapkan GA, yaitu dengan menerapkan beberapa parameter sebagai berikut.

- *min number of atrribut* = 1.
- *population size* = 5.
- *max number of generation* = 30.
- *selection scheme = roulette wheel.*
- *p mutation* = -1,0.
- *p crossover* = 0,5.
- *crossover type = uniform.*

Setelah dilakukan eksperimen dengan parameter yang telah ditetapkan pada GA serta dilakukan komparasi dengan model yang lain, yaitu SVM dan PSO, maka didapatkan hasil seperti pada Tabel IV.

Berdasarkan hasil yang diperoleh, akurasi yang dihasilkan oleh NB-GA lebih baik dibandingkan dengan yang lainnya. Tingkat akurasi mengalami peningkatan dari 78,59% menjadi 84,29%. Berdasarkan hasil tersebut, tmpak bahwa GA dapat meningkatkan akurasi NB.

Pada Gbr. 6 diperlihatkan grafik tingkat akurasi hasil penelitian terhadap berbagai macam model yang digunakan sehingga didapatkan model terbaik.

V. KESIMPULAN

Berdasarkan hasil eksperimen yang dilakukan, terlihat bahwa GA yang merupakan salah satu algoritme optimasi dapat meningkatkan akurasi klasifikasi kategori cerpen *online* dengan *feature selection*. Peningkatan akurasi yang dihasilkan cukup signifikan, yaitu 5,7%, dari 78,59% menjadi 84,29%, Peningkatan akurasi yang dihasilkan tentunya belum sempurna, sehingga perlu adanya upaya-upaya lain dalam meningkatkan akurasi yang dihasilkan.

Terdapat beberapa saran untuk penelitian selanjutnya, yaitu sebagai berikut. Perlu dilakukan eksperimen dengan menggunakan algoritme optimasi lain seperti ACO, PSO, dan yang lainnya. Pemilihan parameter untuk optimasi juga sangat berpengaruh terhadap tingkat akurasi yang dihasilkan, sehingga perlu adanya penetapan parameter yang sesuai dengan model yang diusulkan. Selain itu, terdapat optimasi peningkatan akurasi kembali dengan menggunakan metode lain yang sesuai, sehingga dimungkinkan untuk diterapkannya konsep model *hybrid* lain.

REFERENSI

[1] Suroto, *Apresiasi Sastra Indonesia*, Jakarta: Erlangga, 1989.

[2] Pembinaan, Tim Penyusun Kamus Pusat. "Kamus Besar Bahasa Indonesia Online." Balai Pustaka. Depdikbud (2011).

[3] Sumardjo, Jacob dan Saini K.M, *Apresiasi Kesusastraan*, Jakarta: PT.Gramedia, 1988.

[4] Somantri, O. and Wiyono, S., "Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM)", *Scientific Journal of Informatics*, 3(1), pp.34-45,2016.

[5] Fikri, R., Arnia, F. and Muharar, R., "Pengenalan Karakter Tulisan Tangan Jawi Menggunakan Metode New Relative Context dan SVM", *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, *5*(3), 2016..

[6] Putranto, H.A., Setyawati, O. and Wijono, W., "Pengaruh Phrase Detection dengan POS-Tagger terhadap Akurasi Klasifikasi Sentimen menggunakan SVM", *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, *5*(4), 2016.

[7] Chen, J., Huang, H., Tian, S., & Qu, Y., "Feature selection for text classification with Naïve Bayes", *Expert Systems with Applications*, 36, 5432–5435, 2009.

[8] Zhang, & Gao, F. "An Improvement to NB for Text Classification", *Procedia Engineering*, 15, 2160–2164, 2011.

[9] Wang, S., Li, D., Zhao, L., & Zhang, J., "Sample cutting method for imbalanced text sentiment classification based on BRC", *Knowledge-Based Systems*, 37, 451–461, 2013.

[10] Xu, T., Peng, Q., & Cheng, Y., "Identifying the semantic orientation of terms using S-HAL for sentiment analysis". *Knowledge-Based Systems*, 35, 279–289. doi:10.1016/j.knosys.2012.04.011, 2012.

[11] Zhang, L., Jiang, L., & Li, C., "A new feature selection approach to naive bayes text classifiers", *International Journal of Pattern Recognition614 and Artificial Intelligence*, 30(2), 1650003:1–1650003:17, 2016.

[12] Uysal, A. K., & Gunal, S., "Text classification using genetic algorithm oriented latent semantic features", *Expert Systems with Applications*, 41(13), 5938–5947, 2014.

[13] Tsai, C-F., Chen, Z-Y., & Ke, S-W., "Evolutionary instance selection for text classification", *Journal of Systems and Software*, 90, 104–113, 2014.

[14] Lei, S. "A feature selection method based on information gain and genetic algorithm", *Proceedings of international conference on computer science and electronics engineering (ICCSEE)*, (pp. 355–358), 2012.

[15] Fang, Y., Chen, K., & Luo, C., "The algorithm research of genetic algorithm combining with text feature selection method", *Journal of Computational Science and Engineering*, 1(1), 9–13, 2012.

[16] Ghareb, Abdullah Saeed, Azuraliza Abu Bakar, and Abdul Razak Hamdan. "Hybrid feature selection based on enhanced genetic algorithm for text categorization." *Expert Systems with Applications* 49: 31-47, 2016.

[17] McCallum, A. & Nigam, K., "A comparison of event models for naive Bayes text classification", *AAAI-98 workshop on learning for text categorization*, 1998.

[18] Haupt, Randy L., and Sue Ellen Haupt. "The binary genetic algorithm." *Practical Genetic Algorithms, Second Edition*: 27-50, 2004.

[19] J.H. Holland, "Adaptation in natural and artificial systems", Ann Arbor : *The University of Michigan Press*, 1975.