Jurnal Elektronik Ilmu Komputer Udayana
Volume 10, No 4. May 2022

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Klasifikasi Cerita Pendek Berbahasa Bali Berdasarkan Umur Pembaca dengan Metode *Naive Bayes*

Luh Ristiari[1], AAIN Eka Karyawati[2], I Putu Gede Hendra Suputra[3] , Agus Muliantara[4], I Dewa Made Bayu Atmaja Darmawan[5] ,I Made Widiartha[6]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Bali, Indonesia
[1]luhris21@gmail.com
[2]eka.karyawati@unud.ac.id
[3]hendra.suputra@unud.ac.id
[4]muliantara@unud.ac.id
[5]dewabayu@unud.ac.id
[6]madewidiartha@unud.ac.id

Abstract

Short stories are short stories that tell an event that has happened in a short and clear way. Parents should be able to choose short stories that are suitable for their children because if the stories that parents bring to children are not in accordance with their age, it can affect the development of children. In this study, we will build a system that can classify text. The method used in this research is Nave Bayes with feature selection, namely Genetic Algorithm. This research was conducted to help parents so that their children do not read short stories that are not appropriate for their age so that they do not interfere with their child's development. The data used are children's short stories, youth short stories and adult short stories in Balinese. The best model performance is generated in the training and validation process using new data. The results of testing the Naïve Bayes method without feature selection are 66% accuracy, 66% precision, 67% recall and 66% F1-score. While the Naïve Bayes method uses feature selection, namely 72% accuracy, 72% precision, 78% recall and 73% F1-score.

Keywords: *Naïve Bayes, Genetic Algorithm, Short Stories for Children, Short Stories for Teenagers and Short Stories for Adults*

1. Pendahuluan

Cerpen adalah cerita pendek yang menceritakan suatu kejadian yang pernah terjadi secara singkat dan jelas [7]. Sastra anak-anak berbeda dengan sastra orang dewasa, karena pada sastra anak-anak fokus pada gambaran kehidupan yang bermakna dan mudah dipahami oleh anak-anak. Orang tua harus bisa memilihkan cerpen yang sesuai untuk anaknya karena apabila cerita yang dibawakan orang tua kepada anak tidak sesuai dengan usianya, maka dapat mempengaruhi perkembangan anak. Objek yang belum memiliki label, dapat ditentukan labelnya dengan cara menemukan model yang bisa membedakan setiap label, proses ini disebut dengan klasifikasi. Dalam menyelesaikan proses klasifikasi dapat memanfaatkan peran teknologi sehingga waktu yang diperlukan akan berkurang dan menjadi lebih efesien [4].

Penelitian mengenai klasifikasi yang sudah dilakukan oleh peneliti sebelumnya seperti Seleksi Fitur Klasifikasi Kategori Cerita Pendek Menggunakan *Naïve Bayes* dan Algoritma Genetika yang menghasilkan akurasi 84,29 [9]. Kemudian penelitian mengenai *Text Mining* Untuk Klasifikasi Kategori Cerita Pendek Menggunakan *Naïve Bayes* menghasilkan akurasi 78,59% [8]. Lalu penelitian mengenai Klasifikasi Cerita Bahasa Indonesia Menggunakan Metode *Hybrid* PSO-KNN (*Modified Binary Particle Swarm Optimization* dengan *K-Nearest Neighbor*) menghasilkan akurasi 53% [6]. Selanjutnya penelitian mengenai Klasifikasi Berita Lokal Radar Malang menggunakan Metode *Naive Bayes* dengan Fitur N-Gram menghasilkan akurasi 78,66% [3] dan yang terakhir penelitian mengenai Klasifikasi Teks

Bahasa Bali dengan Metode *Supervised Learning Naïve Bayes Classifier* menghasilkan akurasi 92% [5].

Pada penelitian kali ini akan membangun sebuah sistem yang dapat mengklasifikasikan teks. Metode yang digunakan pada penelitian kali ini adalah *Naïve Bayes* dengan seleksi fitur yaitu Algoritma Genetika. Kategori yang digunakan adalah kategori anak-anak, kategori remaja dan kategori dewasa. Penelitian ini dilakukan untuk membantu orang tua agar anaknya tidak membaca cerpen yang tidak sesuai dengan usianya sehingga tidak mengganggu perkembangan anak.

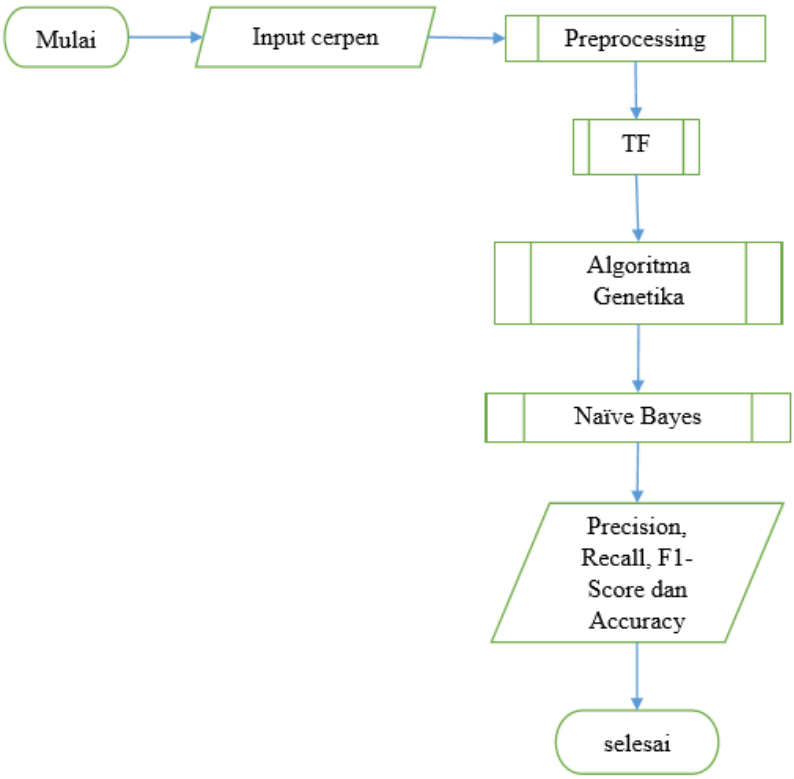
2. Metode Penelitian

2.1 Dataset

Data yang digunakan adalah 90 dokumen cerpen berbahasa Bali. Ada 3 kategori atau kelas yang digunakan yaitu kategori anak-anak, kategori remaja dan kategori dewasa. Data yang digunakan meliputi judul, isi dan label.

2.2 Alur Sistem

Tahap pertama yang dilakukan adalah pengumpulan data berupa cerpen berbahasa Bali dengan 3 kategori yaitu kategori anak-anak, kategori remaja dan kategori dewasa. Setelah data terkumpul maka tahap selanjutnya data dimasukkan ke dalam database lalu dilakukan proses split data. Karena data yang digunakan jumlahnya belum seimbang maka dilakukan proses *under sampling* agar jumlah data menjadi seimbang. Ketika data sudah seimbang, baru dilakukan proses *preprocessing.* Hasil dari *preprocessing* berupa *dictionary* lalu dibuat index sehingga menjadi *vocabolaries.* Setelah itu, *vocabolaries* melalui tahap pembobotan menggunakan *term frequency*. Selanjutnya, setelah tahap pembobotan dilakukan tahapan pemilihan fitur-fitur terbaik menggunakan Algoritma Genetika. Nilai fitness pada Algoritma Genetika didapatkan dari nilai *F1-score* pada proses *Naïve Bayes*.

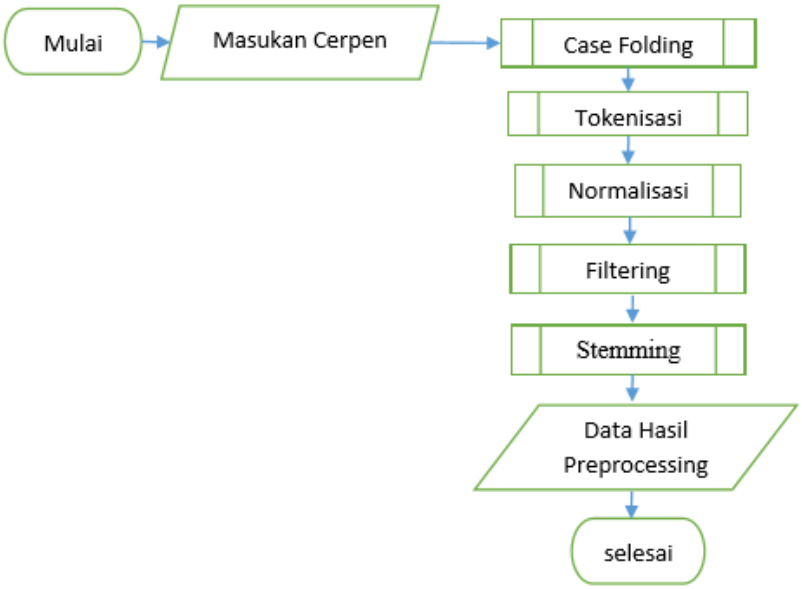


Gambar 1. Alur Penelitian

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Dalam *Naïve Bayes* tidak ada penentuan *hyper parameter* maka digunakan *K-Fold Cross Validation* untuk proses pengujian dan validasi. Pada proses *Naïve Bayes* menggunakan *K-Fold Cross Validation* menghasilkan *output* berupa *term* probabilitas yang ditetapkan sebagai model terbaik. Selanjutnya dilakukan pengujian model terbaik menggunakan *new* data, *output* dari pengujian model terbaik berupa hasil klasifikasi cerpen. Hasil akhir menghasilkan nilai rata-rata hasil evaluasi klasifikasi cerita pendek berbahasa Bali, yaitu rata-rata nilai *Precision, Recall, F1-Score* dan Akurasi. Alur penelitian bisa dilihat pada Gambar 1.

2.3 *Preprocessing Data*



Gambar 2. Alur Preprocessing

Dalam proses ini dokumen melalui tahap *case folding* yaitu merubah semua karakter menjadi huruf kecil. Lalu dilakukan proses tokenisasi yaitu memisahkan dokumen menjadi beberapa token. Setelah itu dilakukan proses normalisasi yaitu mengubah huruf é menjadi e. Selanjutnya dilakukan proses *filtering* yaitu menghilangkan token yang tidak penting. Setelah itu dilakukan proses *stemming* yaitu mengubah semua kata ke dalam bentuk kata dasar [1]. Hasil dari *preprocessing* berupa *dictionary* lalu dilakukan proses *indexing* sehingga menjadi *vocabolaries.* Alur proses *preprocessing* bisa dilihat pada Gambar 2.

2.4 *Term Frequency*

Setelah data melalui proses *preprocessing*, maka selanjutnya data yang berupa *vocabolaries* akan diberi bobot menggunakan Persamaan :

$$tf_{t,d} = \log f_{t,d} + 1 \quad (1)$$

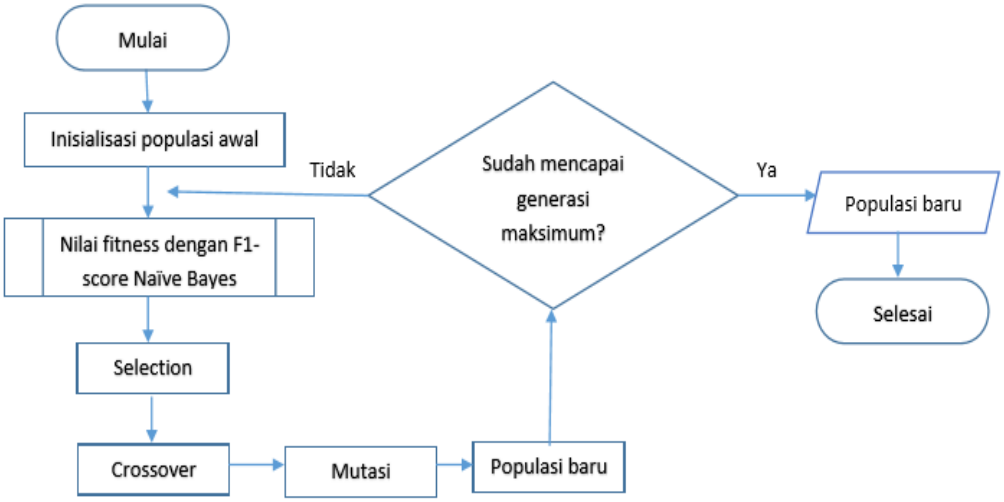
Keterangan:

$tf_{t,d}$ = *term frequency*

$f_{t,d}$ = jumlah kemunculan *term* t di dalam dokumen d

2.5 Algoritma Genetika

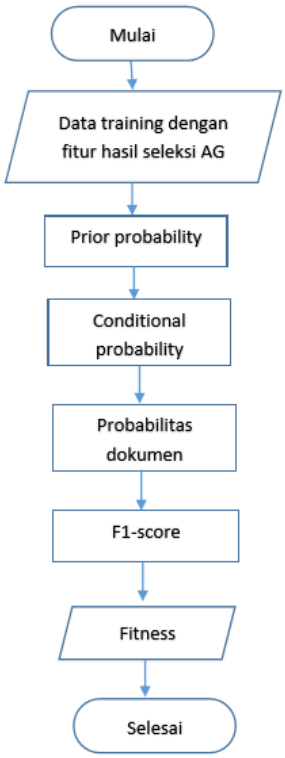
Tahap pertama yang dilakukan adalah inisialisasi kromosom awal yaitu dibangkitkan bilangan acak yaitu 1 dan 0 sebanyak fitur pada proses *TF*. Setelah itu mencari nilai *fitness* yang didapatkan dari hasil evaluasi *F1-score* pada *Naïve Bayes*. Setelah itu, dilakukan tahapan seleksi. Pada tahapan seleksi ini menggunakan roulette wheel, langkah pertama kita mencari nilai *fitness* total, lalu menentukan peluang relatif setiap kromosom dan menentukan peluang komulatif setiap kromosom. Setelah itu kita membangkitkan bilangan acak sejumlah kromosom yang ada. Jika bilangan acak lebih besar dari *PK (peluang komulatif)* dan kurang dari *P(peluang kromosom)* maka kromosom tersebut terpilih untuk diseleksi. Setelah kromosom diseleksi, maka selanjutnya dilakukan inisialisasi parameter Pc, dimana nilai awal yang ditetapkan adalah 0.5. Apabila bilangan acak kurang dari probabilitas *crossover* maka dilakukan proses crossover. Setelah itu, dilakukan inisialisasi parameter Pc, dimana nilai awal yang ditetapkan adalah 0.5. Probabilitas mutasi mempengaruhi jumlah gen yang dimutasi. Proses mutasi menghasilkan kromosom baru. Ketika sudah mencapai generasi maksimum proses Algoritma Genetika berhenti dan menghasilkan populasi baru berupa fitur-fitur pilihan yang digunakan pada proses klasifikasi. Alur proses Algoritma Genetika bisa dilihat pada Gambar 3.



Gambar 3. Alur Proses Algoritma Genetika

Dalam menentukan nilai fitness dilakukan beberapa tahapan yaitu menginputkan *term unique* yang bernilai 1 sesuai bilangan acak yang dibangkitkan pada Algoritma Genetika. Kemudian menentukan *prior probability*, lalu menentukan *conditional probability,* lalu menentukan probabilitas dokumen. Setelah itu menentukan nilai *fitness* atau *F1-score* kategori anak, menentukan nilai *fitness* atau *F1-score* kategori remaja dan menentukan nilai *fitness* atau *F1-score* kategori dewasa. Setelah itu mencari rata-rata nilai *fitness* atau *F1-score* semua kategori. Sehingga didapatkan satu nilai fitness atau F1-score. Alur dalam menentukan nilai *fitness* bisa dilihat pada Gambar 4.

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 4. Alur Proses Menentukan Nilai *Fitness* dengan *F1-score Naïve Bayes*

2.6 Naïve Bayes

Langkah pertama kita menginputkan fitur berdasarkan inisialisasi bilangan acak yang bernilai 1 pada Algoritma Genetika untuk proses *Naïve Bayes* dengan seleksi *fitur.* Lalu untuk proses *Naïve Bayes* tanpa seleksi *fitur* inputannya berupa *term* hasil proses *TF.* Kemudian menentukan *prior probability* menggunakan Persamaan :

$$p(c) = \frac{n_c}{n} \tag{2}$$

Lalu menentukan *conditional probability* menggunakan Persamaan :

$$p(t_k|c) = \frac{n_k+1}{|v|+n'} \tag{3}$$

Dan yang terakhir menentukan probabilitas dokumen menggunakan Persamaan :

$$p(c|d) \propto p(c) \prod_{1 \le k \le n_d} p(t_k|c) \tag{4}$$

Hasil dari proses *Naïve Bayes* menggunakan K-Fold Cross Validatin berupa *term* probabilitas atau model terbaik. Lalu hasil dari pengujian model terbaik menggunakan *new data* berupa hasil klasifikasi cerpen.

2.7 Evaluasi

Confusion matrix menampilkan prediksi klasifikasi dan klasifikasi yang aktual. *Precision* yaitu mengukur peforma dokumen yang bersifat relevan dan bernilai positif diantara seluruh dokumen yang bersifat relevan. *Recall* yaitu mengukur performa dokumen yang bersifat relevan dan bernilai positif diantara seluruh dokumen yang bernilai benar. *F1-score* yaitu mengukur rata-rata *harmonic* dari *precision* dan

recall. Akurasi yaitu mengukur peforma dokumen yang bernilai positif diantara seluruh dokumen yang ada [2]. Tabel *confusion matrix* bisa dilihat pada Tabel 1.

Tabel 1. *Confusion Matrik*

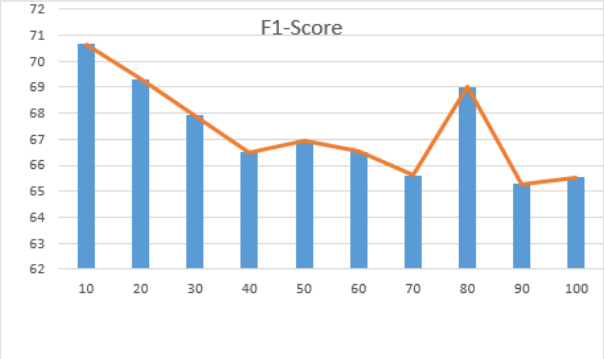
Prediksi / Aktual	Anak	Remaja	Dewasa
Anak	TA	FAR	FAD
Remaja	FRA	TR	FRD
Dewasa	FDA	FDR	TD

3. Hasil dan Diskusi

Pada pengujian metode *Naïve Bayes* tanpa menggunakan seleksi fitur, untuk mendapatkan rata-rata hasil evaluasi digunakan *K-Fold Cross Validation* untuk proses validasi dan pelatihan, dengan k = 3. Didapatkan hasil evaluasi yaitu akurasi 65.278%, *precision* 65.278% *recall* 69.048% dan *F1-Score* 65.021%. Sedangkan pada pengujian metode *Naïve Bayes* menggunakan Algoritma Genetika, akan dilakukan perubahan pada parameter jumlah iterasi, jumlah kromosom, probabilitas *crossover* dan probabilitas mutasi. Dengan menetapkan kombinasi parameter awal yaitu jumlah iterasi 10, jumlah kromosom 2, probabilitas *crossover* 0.5 dan probabilitas mutasi 0.5.

3.1 Pengujian Jumlah Iterasi

Pada pengujian ini akan dilakukan perubahan parameter jumlah iterasi yaitu 10, 20, 30, 40, 50, 60, 70, 80, 90, dan 100. Dapat dilihat bahwa hasil pengujian menunjukkan semakin kecil jumlah iterasi maka ukuran evaluasi cenderung meningkat, dan mencapai ukuran evaluasi terbaik pada iterasi ke-20. Dibandingkan iterasi ke-20, ukuran evaluasi iterasi ke-10 memang lebih besar namun performa ukuran evaluasi menggunakan *new* data terjadi overfitting. Rata-rata hasil evaluasi pada iterasi ke-20 yaitu akurasi 69.444%, *precision* 69.444% *recall* 75.026% dan *F1-Score* 69.311%. Pengaruh Jumlah Iterasi Terhadap *F1-Score* dapat dilihat pada Gambar 5.

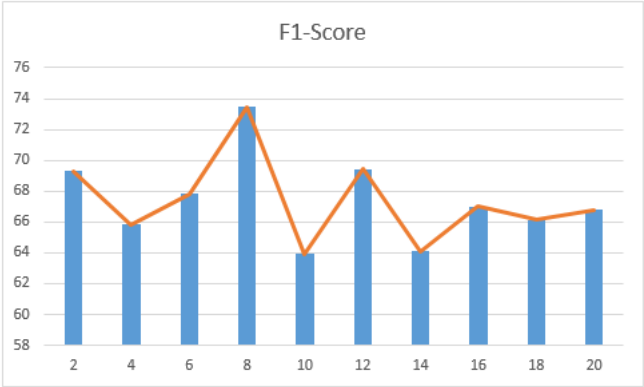


Gambar 5. Pengaruh Jumlah Iterasi Terhadap *F1-Score*

3.2 Pengujian Jumlah Kromosom

Pada pengujian ini, akan digunakan iterasi ke-20 dengan melakukan perubahan parameter jumlah kromosom yaitu 2, 4, 6, 8, 10, 12, 14, 16, 18, dan 20. Dapat dilihat bahwa hasil pengujian menunjukkan semakin besar jumlah kromosom semakin baik, tetapi setelah jumlah kromosom 8 ukuran evaluasi mengalami penurunan. Rata-rata hasil evaluasi pada jumlah kromosom 8 yaitu akurasi 73.611, *precision* 73.611% *recall* 79.349% dan *F1-Score* 73.465%. Pengaruh Jumlah Kromosom Terhadap *F1-Score* dapat dilihat pada Gambar 6.

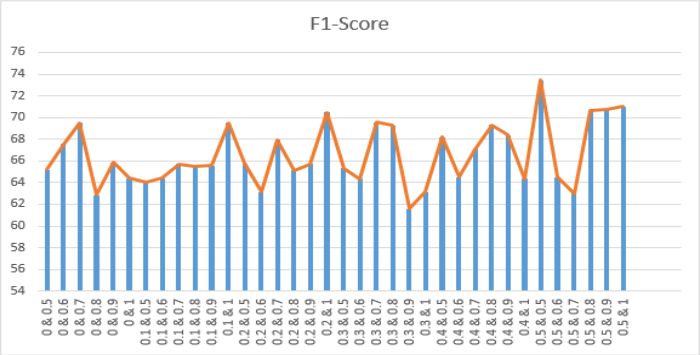
p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 6. Pengaruh Jumlah Kromosom Terhadap *F1-Score*

3.3 Pengujian Pc dan Pm

Pada pengujian ini akan digunakan iterasi ke-20 dan jumlah kromosom 8 dengan melakukan perubahan parameter Pc 0.5, 0.6, 0.7, 0.8, 0.9, 1 dan Pm yaitu 0.1, 0.2, 0.3, 0.4, 0.5. Dapat dilihat bahwa hasil evaluasi pada pengujian ini meningkat walaupun hasilnya tidak stabil. Rata-rata hasil evaluasi pada Pc 0.5 dan Pm 0.5 yaitu akurasi 73.611, *precision* 73.611% *recall* 79.349% dan *F1-Score* 73.465%. Pengaruh Pm dan Pc Terhadap *F1-score* dapat dilihat pada Gambar 7.



Gambar 7. Pengaruh Pm dan Pc Terhadap *F1-score*

4. Kesimpulan

1. Pada pengujian seleksi fitur Algoritma Genetika, pengaruh perubahan parameter jumlah iterasi yaitu semakin kecil jumlah iterasi maka ukuran evaluasi cenderung meningkat, dan mencapai ukuran evaluasi terbaik pada iterasi ke-20. Dibandingkan iterasi ke-20, ukuran evaluasi iterasi ke-10 memang lebih besar namun performa ukuran evaluasi menggunakan *new* data terjadi overfitting. Pengaruh perubahan parameter jumlah kromosom yaitu semakin besar jumlah kromosom semakin baik, tetapi setelah jumlah kromosom 8 ukuran evaluasi mengalami penurunan. Ukuran evaluasi terbaik diperoleh ketika Pc = 0.5 dan Pm = 0.5. Sehingga menghasilkan kombinasi parameter terbaik yaitu jumlah iterasi 20, jumlah kromosom 8, Pc = 0.5 dan Pm = 0.5. Kombinasi parameter tersebut menyeleksi 3185 fitur dengan nilai fitness tertinggi.
2. Pada penelitian ini dilihat performa model terbaik yang dihasilkan pada proses pelatihan dan validasi. Hasil dari pengujian metode *Naïve Bayes* tanpa seleksi fitur yaitu akurasi 66%, *precision*

66%, *recall* 67% dan *F1-score* 66%. Sedangkan pada metode *Naïve Bayes* menggunakan seleksi fitur yaitu akurasi 72%, *precision* 72%, *recall* 78% *dan F1-score* 73%.

Daftar Pustaka

[1] Abimanyu, C.G., Sanjaya ER, N.A dan Karyawati, A.A.I.N.E. 2020. Balinese Automatic Text Summarization Using Genetic Algorithm. *JITK*. 6(1).p. 13-20.

[2] Arini, Wardhani, L.K., dan Octaviano, Dimas. 2020. Perbandingan Seleksi Fitur Term Frequency & Tri-Gram Character Menggunakan Algoritma Naïve Bayes Classifier (Nbc) Pada Tweet Hastag #2019gantipresiden. *KILAT.* 9(1).p.103-114.

[3] Chandra, D.N., Indrawan, Gede., dan Sukajaya, I.N. 2019. Klasifikasi Berita Lokal Radar Malang menggunakan Metode Naive Bayes dengan Fitur N-Gram. *Jurnal Ilmu Komputer Indonesia (JIKI)*. 4(2). p.10-20.

[4] Khadijah. 2016. *Pengembangan Kognitif Anak Usia Dini*. Perdana Publishing. Medan.

[5] Putra, I.B.G.W., Sudarma, Made., dan Kumara, I.N.S. 2016. Klasifikasi Teks Bahasa Bali dengan Metode Supervised Learning Naïve Bayes Classifier. *Teknologi Elektro*. 15(2). p.81-86.

[6] Rahayu, Anita. dan Rochmawati, Naim. 2019. Klasifikasi Cerita Bahasa Indonesia Menggunakan Metode Hybrid PSO-KNN (Modified Binary Particle Swarm Optimization dengan K-Nearest Neighbor). *JINACS*. 1(1). p.64-69.

[7] Ruswati, S.O. 2020. *Bahasa Indonesia PAKET B Setara SMP/Mts Kelas IX.* Direktorat Pendidikan Masyarakat dan Pendidikan Khusus-Direktorat Jendral Pendidikan Anak Usia Dini, Pendidikan Dasar, dan Pendidikan Menengah-Kementrian Pendidikan dan Kebudayaan. Jakarta.

[8] Somantri, Oman. 2017. Text Mining untuk Klasifikasi Kategori Cerita Pendek menggunakan Naïve Bayes (NB). *Jurnal Telematika*. 12(1). p.7-11 (1).

[9] Somantri, Oman. dan Khambali, Mohammad. 2017. Seleksi Fitur Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritma Genetika. *JNTETI*. 6(3). p.301-306 (2).