OPTIMALISASI KLASIFIKASI BERITA MENGGUNAKAN FEATURE INFORMATION GAIN UNTUK ALGORITMA NAIVE BAYES TERHUBUNG RANDOM FOREST

Bobby Suryo Prakoso[1]; Didi Rosiyadi [2]; Dedi Aridarma[3]; Heru Sukma Utama[4]; Fariz Fauzi[5]; Mohammad Arifin Nurul Qhomar [6]

Program Studi Magister Ilmu Komputer
STMIK Nusa Mandiri
www.nusamandiri.ac.id
14002107@nusamandiri.ac.id; didi.rosiyadi@gmail.com; 14002154@nusamandiri.ac.id; 14002126@nusamandiri.ac.id; hsukmautama@gmail.com; 14002102@nusamandiri.ac.id; 14002108@nusamandiri.ac.id

Abstract— *This research is about the classification of news that optimizes with a combination of algorithms. About the dataset used is taken on the online news site. The algorithm used is the Naive Bayes Classifier and Random Forest algorithms by weighting the Information Gain feature selection. The dataset used is 615 datasets with 3 categories or news themes. Get useless models, Delete Useful Attributes, Naive Bayes Classifier-Multinomials, and Random Forest-Feature Selection Information gain. The results of the assessment obtained were an accuracy value of 85.67%, a recall value of 85.67%, and a precision value of 86.23%.*

Keywords*: Remove Useless Attributes, Naive bayes Classifier-Multinomial, Random Forest, Feature Selection Information gain.*

Intisari—Penelitian ini adalah tentang pengklasifikasian berita yang mengoptimalisasi dengan kombinasi antar algoritma. Tentang dataset yang digunakan diambil pada situs pemberitaan online. Algoritma yang digunakan adalah algoritma *Naive Bayes Classifier*, dan *Random Forest* dengan pembobotan seleksi fitur *Information Gain*. Dataset yang digunakan terdapat 615 dataset dengan 3 katagori atau tema berita. Dalam permodelan terdapat 6 model skenario sebagai pembanding untuk menentukan skenario mana yang mendapatkan nilai terbaik, berdasarkan hasil penelitian ini nilai terbaik didapatkan oleh model *Remove Useless Attributes, Naive bayes Classifier-Multinomial,* dan *Random Forest-Feature Selection Information gain.* Hasil evaluasi yang didapatkan adalah nilai *accuracy* 85.67%, nilai *recall* 85.67%, dan nilai *precision* 86.23%.

Kata Kunci*: Remove Useless Attributes, Naive bayes Classifier-Multinomial, Random Forest, Feature Selection Information gain.*

PENDAHULUAN

Tentang klasifikasi terhadap text mining saat ini sudah banyak pengembangan dengan berbagai macam algoritma. Dengan tersebarnya berita saat ini pada situs pemberitaan online, menjadi tantangan tersendiri dalam hal pengkalsifikasi berita. Tantangan yang muncul terhadap pengklasifikasi berita adalah dari dimensi data, ketepatan hasil yang didapatkan, dan kedekatan nilai atas berita yang telah diklasifikasikan.

Sering terjadinya perbedaan perbedaan nilai yang jauh antara ketepatan hasil dengan kedekatan nilai, yang menyebabkan pembacaan data yang bias atas hasil pengklasifikasian, dengan kondisi yang terjadi nilai ketepatan/akurasi tinggi tapi nilai kedekatan jauh, atau nilai ketepatan/ akurasi rendah tetapi nilai kedekatan atau presisi tinggi. Maka dalam penelitian ini maka akan memecahkan permasalahan yang ada tersebut.

Dalam beberapa penelitian yang telah dilakukan, diantaranya ialah milik Betha Nurina Sari, melakukan penelitian tentang penggunaan seleksi fitur *Information Gain*, pada algoritma *Decision Tree, Random Forest, Artificial Neural Network, Support Vector Machine, dan Naive Bayes*(Saifudin, 2018). Hasil atas penelitian tersebut adalah meningkatkan pengklasifikasian atas nilai perhitungan baik dari akurasi, recall, dan presisi atas data yang dihitung, dengan rata-rata kenaikan yaitu sebesar 1.03% sampai dengan 2.5% atas nilai yang didapatkan(Sari, 2016)

Penelitian berikutnya yang dijadikan refrensi adalah penelitian yang dilakukan oleh Shuo Xu, yang mengklasifikasikan berita menggunakan algortiama *Naive Bayes Classifier* dengan menggabungkan metode *Gausian Event Model* dan *Multinomial Event Model*. Terhadap 20 katagori berita yang digunakan, dengan masukan untuk *Gausian Gausian Event Model* memiliki hasil

yang lebih baik setelah diperbandingkan, dengan hasil akurasi 88% (Xu, 2018)

Selanjutnya penelitian yang dilakukan oleh Novan Dimas Pratama untuk mengklasifikasi ulasan konsumen pada salah satu restoran. Tujuan dari penelitian ini adalah untuk menganalisis pendapat sentimen dari konsumen makanan tradisional serta memberikan rekomendasi lokasi(Hadna, Santosa, & Winarno, 2016) dengan kata kunci yang diinginkan. Perhitungan menggunakan algoritma *Naive Bayes* dan seleksi fitur *Chi Squere*. Hasil akurasi klasifikasi yang didapatkan adalah sebagai berikut, dengan seleksi fitur 25% adalah 81%, dengan seleksi fitur 50% adalah 80% dan dengan seleksi fitur 77% adalah 80%(Pratama, Sari, & Adikara, 2018).

Sebagai bahan refrensi untuk penelitian yang menjadi acuan adalah penelitian yang dilakukan oleh Irfan dan M. Ali Fauzi untuk pengklasifikasian dokumen berbasis teks berita online yang ada pada suatu situs. Serta penggunaan algoritma yaitu dengan menggunakan *Maximum Marginal Relevance-Feature Selection*(MMR-FS) untuk information gain untuk algoritam *Naive Bayes Classifier* dengan mendapatkan akurasi 86%(Irfan & Fauzi, 2018).

Selanjutnya penelitian yang dilakukan Ghulam Asrofi Buntoro pengklasifikasian untuk data *twitter* untuk ujaran kebencian dengan menggunakan algoritma *Naive Bayes Classifier* dan *Support Vector Machine* dengan menggunakan 522 data *tweet* dengan hasil yang didapatkan untuk akurasi adalah akurasi mencapai 66,6%, nilai presisi 67,1%, nilai *recall* 66,7% nilai *TP rate* 66,7% dan nilai *TN rate* 75,8%(Buntoro, 2016).

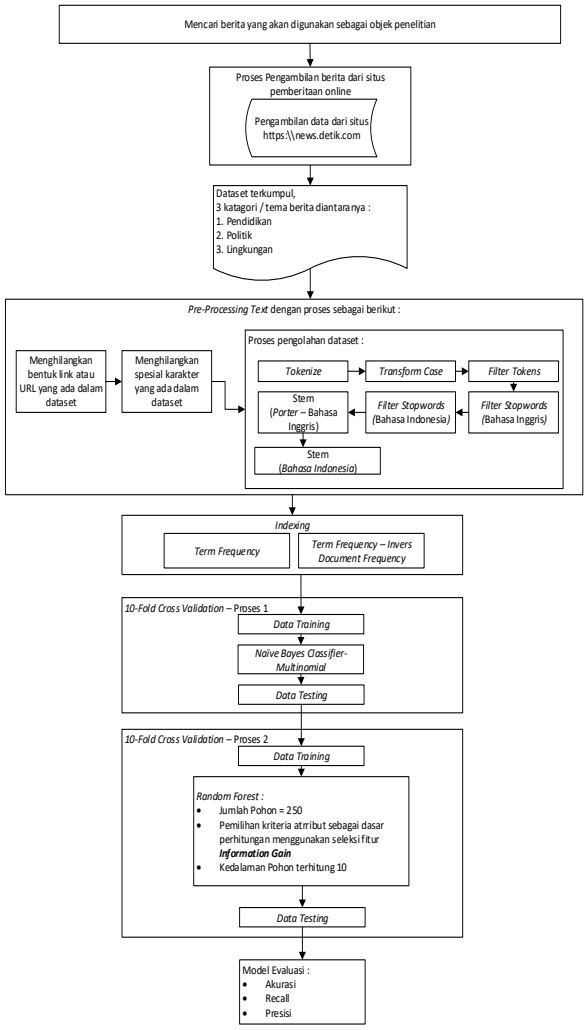
Penelitian tentang pengklasifikasian selanjutnya tentang data berita oleh Yoga Dwitya Pramudita, dengan menggunakan *Naive Bayes Classifier* mendapatkan hasil perhitungan dengan keakuratan sebesar 77%(Pramudita et al., 2018).

Dalam penelitian ini akan melakukan pengklasifikasian teks dengan menggunakan model pengujian yang dikembangan terhadap refrensi sebelumnya, yaitu akan melakukan pengujian ganda dengan algoritma yang berbeda. Algoritma yang diujikan pada tahap pertama menggunakan algoritma *Naive Bayes Classifiers*, lalu mengujikan pada algoritma *Random Forests* dengan seleksi kriteria atribut *Information Gain*. Bertujuan untuk menemukan permodelan yang sesuai atas klasifikasi berita dengan nilai akurasi dan presisi yang tinggi dengan tidak mengalami perbedaan nilai yang tidak terlalu jauh antara keduanya,

Atas dasar kajian pustaka yang telah disebutkan tersebut maka penelitian ini memiliki tujuan untuk mengkombinasikan penggunaan algoritma untuk mengoptimalisasikan hasil perhitungan klasifikasi, dengan menggunakan metode algoritma *Naive Bayes Classifier* terhubung dengan *Random Forest* dengan menggunakan seleksi fitur untuk pembobotan yaitu dengan algoritma *Information gain*. Seleksi fitur *Information gain* sendiri sudah tersedia pada parameter yang disediakan untuk *Random Forest* pada *tools Rapidminer.*

BAHAN DAN METODE

Atas usulan serta tujuan penelitian ini, rancangan atas penelitan dapat digambarkan dalam bentuk kerangka berfikir sebagai berikut :



Sumber: (Prakoso, 2019)

Gambar1. Kerangka Berfikir Penelitian

Berdasarkan rancangan penelitian yang digambarkan dengan kerangka berfikir bahwa teknik pengumpulan data bersumber dari situs pemberitaan *online* yaitu http:\\news.detik.com, dengan mengambil katagori / tema berita Pendidikan, Politik, dan Lingkungan .

Pengambilan data menggunakan teknik dari google, dengan menggunakan google spreadsheet untuk formula yang digunakan adalah formula *scrapping* data web.

Tabel 1. Formula Google Spreadsheet Pengambilan Konten

No	Formula Google Spreadsheet	Fungsi
1	=IMPORTXML("link hasil pencarian","//a/@href")	Mengekstrak link pencarian
2	=IMPORTXML("link berita yang digunakan"//*[@id='detikdetailtext']"")	Mengekstrak konten berita yang digunakan sebagai data

Sumber: (Prakoso, 2019)

Setelah melakukan proses tersebut data yang didapatkan mencapai 650 data per katagori berita, dengan total keseluruhan data mencapai 1950 konten berita yang akan digunakan untuk testing data.

Setelah data didapatkan, maka proes yang dilakukan adalah akan memproses data tersebut dengan memberi label sesuai dengan hasil pencarian data berdasarkan katagori atau tema berita dalam file excel. Maka data tersebut siap untuk diproses ketahap selanjutnya dengan menggunakan tools Rapidminer untuk menentukan klasifikasi berdasarkan algoritma yang akan digunakan.

Penggunaan algoritma dalam penelitian ini menggunakan algoritma Naive Bayes Classiffier, Naive Bayes adalah metode algoritma yang bekerja atas bagaimana menghitung frekuensi atas setiap term yang ada dalam dokumen(Fanissa, Fauzi, & Adinugroho, 2018). Dokumen dengan urutan kejadian yang muncul atas kata terhadap dokumen akan diabaikan, menyebabkan pengolahan kata menggunakan distiribusi yang multinomial(Feng, Li, Yuan, Zeng, & Sun, 2018).

Berikut persamaan rumus atas Naive Bayes Classifier(Budiman et al., 2018) :

$$P(c|d) = P(c)\prod_{i=1}^{n} P(w_i|c) \quad \text{.....(1)}$$

d : besaran dokumen

n : jumlah semua kata yang ada pada dokumen

Selanjutnya nilai atas variabel $P(c)$ diperoleh dengan rumus berikut :

$$P(c) = \frac{N_c}{N} \quad \text{.....(2)}$$

$P(c)$: peluang kelas c

N : jumlah seluruh dokumen

Selanjutnya untuk menghitung peluang kata ke- i pada kelas c menggunakan rumus berikut :

$$P(w_i|c) = \frac{count\ (w_i,c)+1}{count\ (c)+|V|} \quad \text{.....(3)}$$

$P(w_i|c)$: Peluang kata ke- i pada kelas c

$count\ (w_i, c)$: Jumlah kata ke- i pada kelas c

$count\ (c)$: Jumlah semua kata pada kelas c

$|V|$: Jumlah kata unik terhadap semua Kelas

Selanjutnya penggunaan algoritma *random forest* untuk proses ke 2 pada validasi data untuk klasifikasi teks. Metode random forest yaitu penerapan metode *bootstrap aggregating* (*bagging*) dan *random feature selection*. Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (forest), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas n amatan dan p peubah penjelas(Dewi, Syafitri, Mulyadi, Statistika, & Statistika, 2011).

Tentang *Information Gain* merupakan algoritma yang berfungsi sebagai penentu batas yang akan digunakan atas atribut yang tersedia, bisa hanya dalam 1 atribut atau lebih dari 1atribut yang digunakan, yang melambangkan refleksi atas kualitas suatu atribut yang akan digunakan(Budiman et al., 2018).

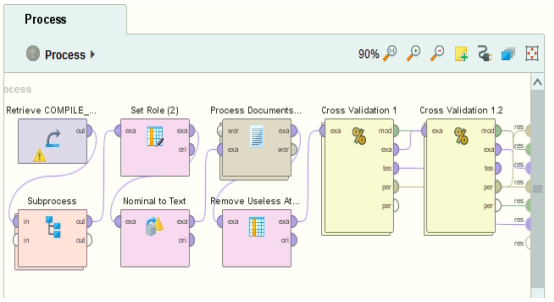
Serta dalam perhitungannya dapat dirumuskan dalam rumus berikut ini :

$$gain(y, A) = entropy(y)\sum_{0} \in nilai\ (A)\frac{yc}{y} entropy(yc) \quad \text{.....(4)}$$

HASIL DAN PEMBAHASAN

Berdasarkan penjelasan pada bahan dan metode maka terdapat lanjutan beberapa jabaran pada bab hasil dan pembahasan diantaranya adalah proses pengolahan data. Proses pengolahan data menggunakan tools Rapidminer yang merupakan tools Machine learning, didalamnya terdapat untuk pengolahan data mining dan teks mining.

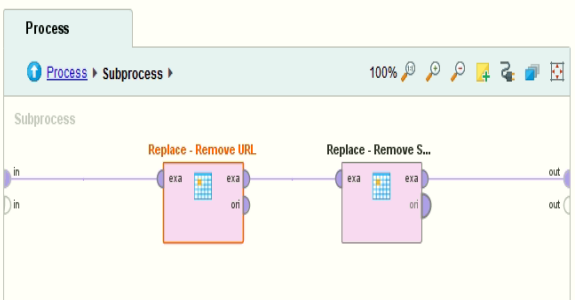
Atas penjelasan tersebut gambaran awal untuk model yang digunakan dalam rapid miner adalah sebagai berikut :



Sumber: (Prakoso, 2019)

Gambar2. Model Perhitungan Rapidminer

Dengan penjelasan per-proses diantaranya untuk yang pertama adalah proses yang ada dalam subproses dengan terdapat 2 operator *replace*. Operator *replace* yang pertama digunakan untuk menghilangkan bentuk link dan URL yang ada dalam komponen data dengan menggunakan Regular Expresion sebagai berikut *http\S+/\S+co\S+*, berikut merupakan operator yang digunakan :

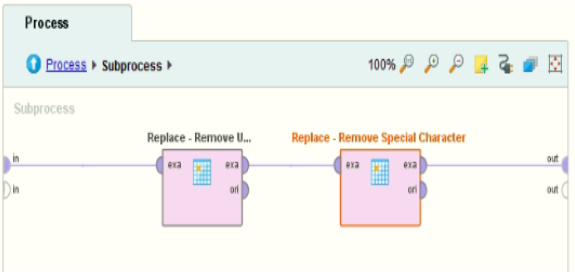


Sumber: (Prakoso, 2019)

Gambar3. Operator *Replace* untuk menghilangkan link atau URL

Setelah itu adalah operator *replace* untuk menghilangkan segala bentuk macam spesial karakter yang terdapat pada data, dengan menggunakan Regular Expression sebagai berikut : [-!"#$%&'()*+,./:;<=>?@\[\\\]_`{|}~].

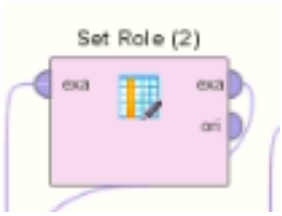
Berikut meripakan bentuk operator yang digunakan untuk menghilangkan karakter spesial pada tools Rapidminer :



Sumber (Prakoso, 2019)

Gambar3. Operator *Replace* untuk menghilangkan karakter spesial

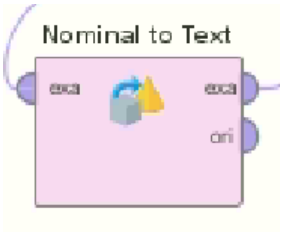
Selanjutnya merupakan operator yang digunakan adalah *Set Role*, operator ini digunakan adalah untuk menentukan *field* atau bagian mana yang akan digunakan sebagai label. Pada pengaturan operator *Set Role* penelitian ini yang akan digunakan sebagai label adalah *field* TEMA_BERITA, yang mempunyai 3 nilai yaitu PENDIDIKAN, POLITIK, dan BUDAYA. Berikut merupakan bentuk operator yang digunakan :



Sumber: (Prakoso, 2019)

Gambar4. Operator *Set Role* untuk menentukan *Label*

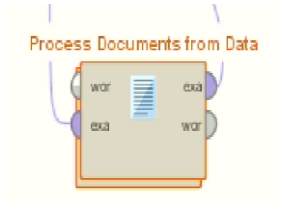
Setelah melalui operator *Set Role* maka operator yang selanjutnya adalah operator *Nominal to Text,* operator tersebut digunakan untuk mengubah jenis atribut nominal yang dipilih untuk teks. Serta memetakan semua nilai atribut kedalam nilai *string* yang sesuai. Berikut merupakan jenis operator yang digunakan :



Sumber: (Prakoso, 2019)

Gambar5. Operator *Set Role* untuk menentukan *Label*

Jika sudah melewati proses tersebut maka proses yang ada selanjutnya adalah proses dengan operator *Process Documents from Data*, yang dimana proeses tersebut tergambarkan sebagi berikut :

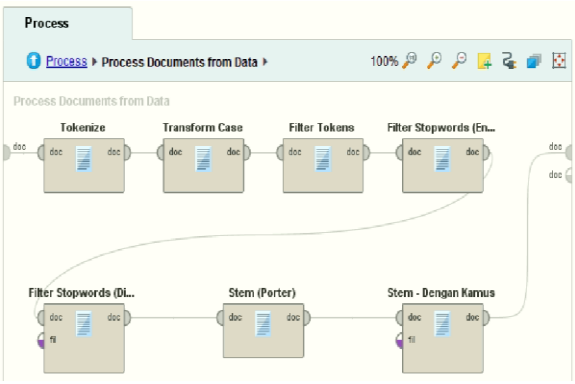


Sumber: (Prakoso, 2019)

Gambar6. Operator *Process Documents from Data*

Dalam operator *Process Documents from Data*, terdapat beberapa proses yang digunakan untuk membersihkan data agar menjadi *vector*

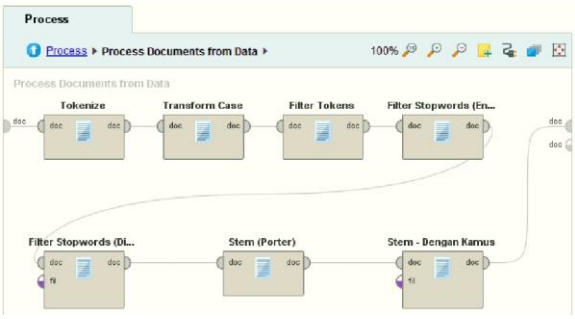
yang dapat digunakan untuk perhitungan algoritma diantaranya Tokenize, Transform Case, Filter Tokens, Filter Stopwords – Bahasa Inggris, Filter Stopword – Bahasa Indonesia, Stem – Porter(Bahasa inggris), dan Stem – Dictionary (Bahasa Indonesia), selanjutnya adalah gambar operator yang digunakan :



Sumber: (Prakoso, 2019)

Gambar7. Operator-operator yang digunakan pada *Process Documents from Data*

Setelah dokumen menjadi vector yang dapat dihitung karena nilai text tersebut berubah jadi nominalmaka selanjutnya adalah proses validasi menggunakan Cross Validation, dengan ketentuan k-fold sebesar K-10, dalam penerapan ini terdapat 2 proses yang digunakan untuk proses validasi diantaranya adalah proses validasi menggunakan algoritma *Naive Bayes Classifier – Multinomial*, berikut untuk gambar operator yang digunakan :



Sumber: (Prakoso, 2019)

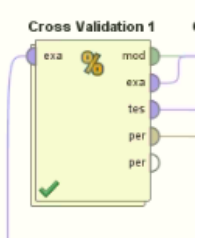
Gambar8. Operator-operator didalam *Process Documents from Data*

Setelah proses tersebut adalah proses menghilangkan atribut yang tidak terpakai dengan menggunakan operator *Remove Useless Attributes,* dengan ketentuan menghapus *Exemple Set* yang tidak berguna, sedangkan untuk *thresholds* atau ambang batas ditentukan pada parameter, pada penelitian ini menggunakan ketentuan *thresholds* nominal di atas 1.00 tidak terpakai, dan nominal dibawah 0.00 tidak terpakai. Berikut untuk gambar operator yang digunakan :

Sumber: (Prakoso, 2019)

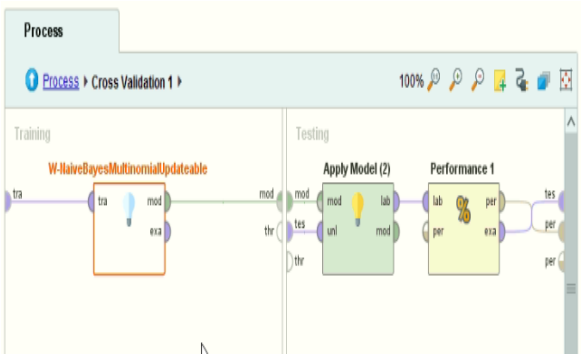
Gambar9. Operator *Remove Useles Attributes*

Proses berlanjut pada proses validasi dengan menggunakan *Cross Validation,* dengan ketentuan *K-Fold* sebesar 10-*fold cross validation*, yang didalam menggunakan algoritma *Naive Bayes Classifier – Multinomial* penggunaan algoritma tersebut merupakan algoritma yang sesuai dengan label lebih dari 2. Berikut merupakan gambar operator dan operator didalamnya :



Sumber: (Prakoso, 2019)

Gambar10. Operator *Cross Validation 1*

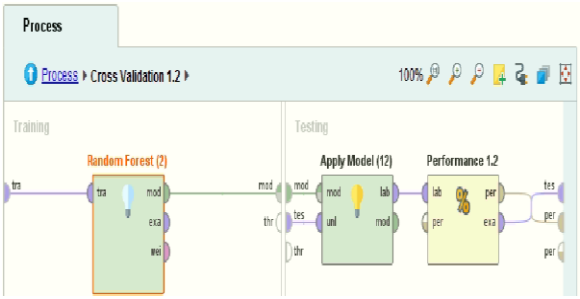


Sumber: (Prakoso, 2019)

Gambar11. Operator yang digunakan didalam *Cross Validation 1*

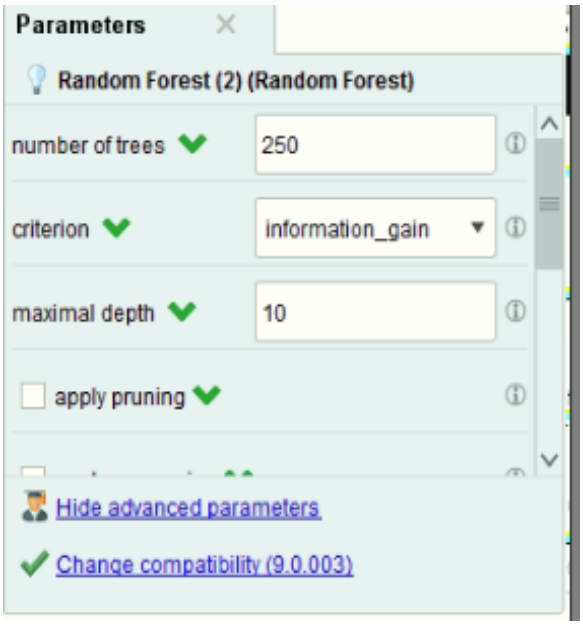
Selanjutnya proses validasi pada proses *cross valdation 1* maka berlanjut pada *cross valdation 1.2* yang dimana didalamnya menggunakan algoritma *Random Forest* yang didalamnya menggunakan jumlah pohon keputusan sebesar 250. Selanjutnya pemilihan kriteria berdasarkan atribut, pada penelitian ini pemilihan kriteria berdasarkan attribut *Information Gain.* Terakhir kedalaman pohon menggunakan maksimal 10, penggunaan parameter ini dimaksudkan untuk membatasi pohon yang ada pada *Random Forest.* Sebagai

gambaran akan operator maka berikut gambar operator yang digunakan dan parameter yang dipakai :



Sumber: (Prakoso, 2019)

Gambar12. Operator-operator *Cross Valdiation 1.2*



Sumber: (Prakoso, 2019)

Gambar13. Parameter yang digunakan untuk operator *Random Forest*

Berdasarkan permodelan tersebut terdapat beberapa skenario atas penerepannya yang dapat diberikan, antara lain dapat terjabar pada tabel dibawah ini :

Tabel 2. Fungsi Komponen Catudaya

No	Model	*Accuracy*	*Recall*	*Precision*
1	*Naive bayes Classifier-Multinomial,*	78.32%	78.32%	79.00%
2	*Random Forest-Feature Selection Information gain*			
3	*Remove Useless Attributes, Naive bayes Classifier-Multinomial*	82.67%	82.67%	82.67%
4	*Remove Useless Attributes, Random Forest-Feature Selection Information gain*	84.33%	84.34%	84.81%
5	*Naive bayes Classifier-Multinomial, Random Forest-Feature Selection Information gain*	81.67%	81.67%	82.74%
6	*Remove Useless Attributes, Naive bayes Classifier-Multinomial, Random Forest-Feature Selection Information gain*	85.67%	85.67%	86.23%

Sumber: (Prakoso, 2019)

KESIMPULAN

Atas hasil penelitian yang didapat dari 6 skenario yang telah diujikan bahwa skenario dengan model *Remove Useless Attributes, Naive bayes Classifier-Multinomial,* dan *Random Forest-Feature Selection Information gain,* mendapatkan hasil evaluasi yang tertinggi dimana hasil tersebut dengan nilai *accuracy* 85.67%, nilai *recall* 85.67%, dan nilai *precision* 86.23%. Dengan adanya perbedaan pada akurasi dengan *precision,* dimana hasil *precision* lebih besar terhadap *accuracy,* terdapat kedekatan nilai terhadap data sudah bagus, tapi ternyata dengan nilai sebenarnya masih kurang. Maka dengan hasil tersebut perlu membuat stemmer bahasa indonesia yang lebih baik dengan *corpus* yang sudah pernah dikumpulkan, karena saat ini untuk *corpus* bahasa indonesia sangat jarang sekali untuk ditemukan.

REFERENSI

Budiman, A. S., Studi, P., Komputer, T., Parandani, X. A., Studi, P., & Informatika, M. (2018). Uji Akurasi Klasifikasi Dan Validasi Data Pada Penggunaan Metode Membership Function Dan Algoritma C4 . 5 Dalam, *9*(1), 565–578.

Buntoro, G. A. (2016). ANALISIS SENTIMEN HATESPEECH PADA TWITTER DENGAN METODE NAÏVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINE. *Jurnal Dinamika Informatika*, *5*(2).

Dewi, N. K., Syafitri, U. D., Mulyadi, S. Y., Statistika, M. D., & Statistika, D. (2011). PENERAPAN METODE RANDOM FOREST DALAM DRIVER (The Application of Random Forest in Driver Analysis), *16*(1), 35–43.

Fanissa, S., Fauzi, M. A., & Adinugroho, S. (2018). Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan

Seleksi Fitur Query Expansion Ranking. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(8), 2766–2770.

Feng, X., Li, S., Yuan, C., Zeng, P., & Sun, Y. (2018). Prediction of Slope Stability using Naive Bayes Classifier. *KSCE Journal of Civil Engineering (2018) 22(3):941-950, PISSN 1226-7988, EISSN 1976-3808*, *22*, 941–950. https://doi.org/10.1007/s12205-018-1337-3

Hadna, N. M. S., Santosa, P. I., & Winarno, W. W. (2016). Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen di Twitter. *Seminar Nasional Teknologi Informasi Dan Komunikasi 2016 (SENTIKA 2016) Yogyakarta, 18-19 Maret 2016*, (March).

Irfan, M. R., & Fauzi, M. A. (2018). Analisis Sentimen Kurikulum 2013 pada Twitter menggunakan Ensemble Feature dan Metode K-Nearest Neighbor, *2*(9), 3006–3014.

Prakoso, B. S. (2019). *Klasifikasi Berita Menggunakan Algoritma Naïve Bayes Classifierss Terintegrasi Pengujian Algoritma Random Forest Menggunakan Seleksi Kriteria Atribut"*. Jakarta.

Pramudita, Y. D., Putro, S. S., Makhmud, N., Olahraga, B., Confix, E., & Stemmer, S. (2018). KLASIFIKASI BERITA OLAHRAGA MENGGUNAKAN METODE NAÏVE BAYES SPORTS NEWS CLASSIFICATION USING NAÏVE BAYES WITH ENHANCED CONFIX STRIPPING STEMMER, *5*(3). https://doi.org/10.25126/jtiik.201853810

Pratama, N. D., Sari, Y. A., & Adikara, P. P. (2018). Analisis Sentimen Pada Review Konsumen Menggunakan Metode Naive Bayes Dengan Seleksi Fitur Chi Square Untuk Rekomendasi Lokasi Makanan Tradisional. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(9), 2982–2988.

Saifudin, A. (2018). Metode Data Mining untuk Seleksi Calon Mahasiswa pada Penerimaan Mahasiswa Baru di Universitas Pamulang. *Jurnal Teknologi*, *Volume 10*(January), 25–35. https://doi.org/10.24853/jurtek.10.1.25-36

Sari, B. N. (2016). Implementasi Teknik Seleksi Fitur Information Gain Pada Algoritma Klasifikasi Machine Learning Untuk Prediksi Performa Akademik Siswa, (March), 6–7.

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48–59. https://doi.org/10.1177/0165551516677946