ANALISIS PERFORMANSI NAIVE BAYES DAN RANDOM FOREST TERHADAP SENTIMEN KENAIKAN HARGA BBM DI INDONESIA

[1]Muhammad Lutfi Pratama, [2]Yisti Vita Via, [3]Eka Prakarsa Mandyartha
Program Studi Informatika, Fakultas Ilmu Komputer, UPN "Veteran" Jawa Timur
Email: 19081010049@student.upnjatim.ac.id

Abstrak. *Bahan Bakar Minyak (BBM) adalah komoditas penting dalam aktifitas perekonomian masyarakat. Kebijakan kenaikan harga BBM dapat berpengaruh negatif terhadap pertumbuhan ekonomi masyarakat. Namun pemerintah melakukan berbagai upaya baik, seperti Bantuan Langsung Tunai BBM. Fenomena ini menimbulkan beragam sentimen di masyarakat. Beragam sentimen tersebut dapat menjadi tolak ukur pemerintah dalam mengambil keputusan. Oleh karena itu, digunakan algoritma Naïve Bayes Classifier (NBC) dan Random Forest (RF) untuk klasifikasi sentimen masyarakat terhadap kebijakan kenaikan harga BBM melalui data teks Twitter yang berjumlah 250 ribu data tweet. Label kelas sentimen meliputi positif, netral, dan negatif. Analisis performasi dilakukan pada masing-masing algoritma dengan mempertimbangkan nilai accuracy, recall, dan rata-rata nilai kurva AUC-ROC. Kedua algoritma akan melalui proses tuning hyperparameter, untuk NBC yaitu nilai laplace smoothing dan untuk RF yaitu nilai minimum samples split dan minimum samples leaf. Disimpulkan bahwa performa RF lebih unggul dengan nilai akurasi mencapai 85.15% dan rata-rata nilai AUC-ROC sebesar 94.62%., dibandingkan NBC dengan nilai akurasi 79.74% dan rata-rata AUC-ROC sebesar 89.83%.*

Kata Kunci: BBM, Naive Bayes, Random Forest, Analisis sentimen, Twitter

Bahan Bakar Minyak (BBM) adalah komoditas yang sangat mempengaruhi aktifitas perekonomian masyarakat. Kebijakan kenaikan harga BBM mendorong inflasi yang berpengaruh negatif terhadap pertumbuhan ekonomi. Dampak psikologis dapat terjadi dari sisi masyarakat manakala semua berekspektasi bahwa kenaikan BBM diikuti kenaikan harga sektor-sektor lainnya. Kenaikan harga BBM memberikan dampak fluktuasi terhadap harga suplai bahan pokok di pasar tradisional, terutama pada kenaikan harga sembako [1]. Menyikapi hal ini pemerintah bertindak dengan membuat kebijakan baru seperti Bantuan Langsung Tunai BBM (BLT BBM), Bantuan Subsidi Upah (BSU), dan lain-lain. Presiden Joko Widodo menjamin bahwa program telah berjalan secara mudah, cepat, dan tepat sasaran [2]. Fenomena ini menimbulkan banyak opini dari masyarakat, sehingga menimbulkan beragam sentimen. Sentimen masyarakat yang beragam dapat menjadi tolak ukur pemerintah memutuskan kebijakan. Twitter menjadi salah satu media terbaik untuk mengetahui opini seseorang melalui teks *tweet*, dan Twitter telah menjadi salah satu platform media sosial terbesar di dunia sejak tahun 2006 [3].

Untuk menganalisis sentimen jumlah dataset sangat mempengaruhi hasil klasifikasi, semakin besar data yang digunakan pada proses pelatihan semakin bagus performa model klasifikasi yang dihasilkan [4]. Dataset yang besar memiliki karakteristik yang mencakup ukuran data, variasi fitur, dan frekuensi kemunculan [5]. Pendekatan *hybrid* digunakan sebagai metode paling efektif dengan menggabungkan hasil dari *lexicon based* lalu memperkuat konteks permasalahan menggunakan model *machine learning* [6]. Pendekatan hybrid bertujuan agar membuat model analisis sentimen supaya lebih fokus pada topik kebijakan kenaikan harga BBM di Indonesia. Algoritma Naïve Bayes Classifier (NBC) dan Random Forest (RF) digunakan sebagai metode pembuatan model klasifikasi. Algoritma Naïve Bayes adalah algoritma yang sederhana dan cocok digunakan untuk klasifikasi sentimen [7]. Random Forest juga merupakan salah satu algoritma terbaik untuk klasifikasi data yang besar atau memiliki banyak fitur dengan akurasi tinggi [8]. Kedua algoritma ini terkenal dan populer digunakan untuk klasifikasi sentimen melalui *machine learning* [9].

Dataset didapatkan dengan cara *scraping* konten *tweet* di Twitter sebanyak 250 ribu, selanjutnya dilakukan *preprocessing* untuk mengoptimalkan pelatihan model dan pelabelan secara otomatis. Lalu data akan

dibagi menjadi tiga kelas label yaitu positif, negatif, dan netral. Dari hasil klasifikasi model NBC dan RF akan dilakukan analisis performa ketepatan klasifikasi pada masing-masing model dengan menggunakan nilai dari akurasi dan rata-rata AUC-ROC. Penelitian ini juga menguji variabel *hyperparameter* pada kedua algoritma, untuk mengetahui kondisi terbaik dari model penelitian ini. Random Forest dengan variabel *hyperparameter* yaitu *minimum samples split* dan *minimum samples leaf* dan Naïve Bayes dengan nilai *laplace smoothing* atau *alpha.*

Penelitian terkait dilakukan terhadap opini masyarakat dalam pemilihan umum kandidat presiden Indonesia 2019 di media sosial Twitter, menggunakan Naïve Bayes, SVM, dan KNN, diketahui bahwa Naïve Bayes mendapatkan akurasi yang lebih baik sebesar 75.58% [3]. Penelitian lainnya dilakukan pada penilaian kustomer tehadap pembelian produk online di platform digital, menggunakan Random Forest dan SVM untuk mendapatkan nilai *accuracy*, *precision*, *f-measure* dan *recall*, didapatkan bahwa Random Forest lebih baik dengan tingkat akurasi mencapai 97% [8]. Maka, penelitian ini memilih Naïve Bayes dan Random Forest karena dinilai cukup kompetitif. Hasil analisis performasi diharapkan dapat menjadi tolak ukur dalam memilih algoritma yang tepat untuk menganalisis sentimen kebijakan kenaikan harga BBM di Indonesia.

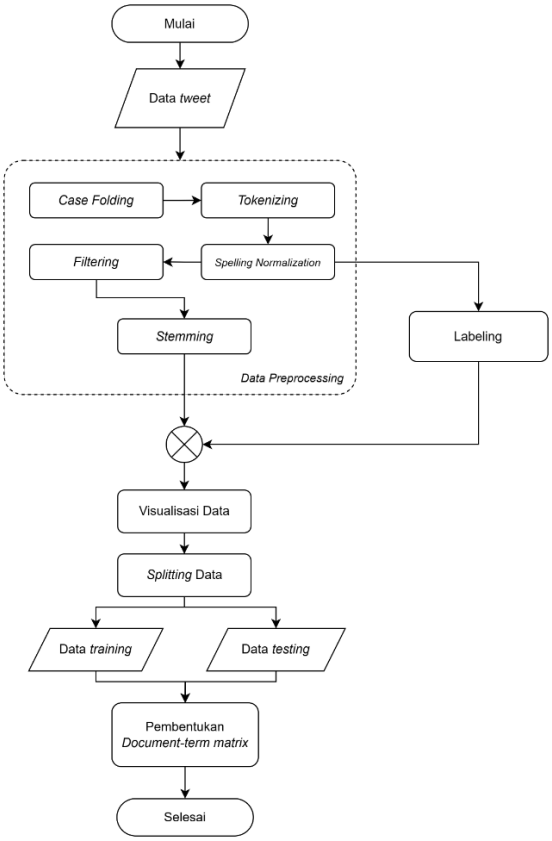
I. Metodologi

Metodologi meliputi langkah-langkah utama yaitu pengumpulan data, pengolahan data, pembuatan model klasifikasi, skema pengujian *tuning hyperparameter* dan interpretasi hasil pengujian.

Pengumpulan Data

Pengumpulan data dilakukan dengan cara *crawling* dan *scraping* dari halaman website sosial media Twitter menggunakan alat bantu bernama Snscrape dan Tweepy pada pemrograman Python. Melalui *TwitterAPI*, data *tweet* terkait kebijakan kenaikan harga BBM di Indonesia diambil menggunakan kata kunci diantaranya "bbm naik", "kebijakan harga bbm", "subsidi bbm", "subsidi bbm pemerintah", "harga bensin mahal", "harga bensin naik", dan "harga bbm". Data *tweet* yang diambil adalah tweet yang dibuat sejak tanggal 25 Agustus sampai 30 Desember 2022. Pemilihan tanggal awal tersebut karena awal munculnya isu kebijakan kenaikan harga BBM hingga mencapai puncaknya di bulan September. Pemilihan tanggal akhir tersebut karena per September 2022 dan beberapa bulan selanjutnya, kemungkinan besar kebijakan ini masih ramai dibicarakan oleh masyarakat.

Pengolahan Data



Gambar 1. Digram alir metode pengolahan data.

Pre-processing diperlukan untuk mengolah data mentah menjadi lebih terstruktur dan informatif sehingga mempermudah proses klasifikasi serta meningkatkan performa model terutama akurasi [10]. Tahap *Pre-processing* meliputi *case-folding*, *tokenizing*, *spelling normalization*, *filtering*, dan *stemming*. Tahap *labeling* pada 250 ribu data tweet menggunakan *library transformers* dengan model *fine-tuned Indonesian RoBERTa*. Metode ekstraksi fitur yang digunakan adalah *document-term matrix*.

Pembuatan Model Klasifikasi Naive Bayes

Pendekatan Naïve Bayes yang digunakan adalah *Multinomial Naïve Bayes* karena lebih tepat digunakan untuk fitur

bilangan diskrit yaitu frekuensi kemunculan kata dari hasil *document-term matrix* [11]. Secara umum Naïve Bayes dinotasikan dengan persamaan:

$$(C_j|X) = \frac{P(X|C_j)P(C_j)}{P(X)} \quad (1)$$

Diketahui $P(C_j|X)$ adalah probabilitas pengamatan berkategori C_j berdasarkan kondisi X *(posterior probability)*. $P(X|C_j)$ yaitu probabilitas X dengan kemungkinan C_j *(likelihood)*, lalu dilakukan operasi perkalian dengan $P(C_j)$ yaitu nilai probabilitas pada data *training* melalui perhitungan kategori C_j *(prior)*, dan hasilnya dibagi dengan $P(X)$ yaitu total probabilitas untuk semua fitur data training X $(x_1, x_2, \dots, x_k)$ *(evidence)*. Sehingga pendekatan secara Multinomial dapat dinotasikan sebagai berikut.

$$P(x_1 = a_1, \dots, x_k = a_k|C = c_j) \approx \frac{L!}{\prod_{i=1}^{k} a_i!} \prod_{i=1}^{k} P(x_i|C = c_j)^{a_i} \quad (2)$$

Dengan variabel L merepresentasikan banyak kata dalam satu dokumen atau *tweet* dan a_i merepresentasikan jumlah kata ke-i (x_i) dalam satu *tweet.*

Pembuatan Model Klasifikasi Random Forest

Dasar teori algoritma *Random Forest* adalah perhitungan *decision tree* dengan *gini index* atau *gini impurity* sebagai kriteria pemisah [8]. Algoritma CART digunakan sebagai metode perhitungan utama dengan mempertimbangkan nilai *gini impurity* dimana efisien untuk model *ensemble* seperti *Random Forest* [12]. Random Forest menerapkan konsep *bagging* yaitu memperhitungkan semua kelompok pohon keputusan agar dilakukan *voting* dalam memprediksi suatu kelas data. Hal ini mengurangi *variance* saat memprediksi data baru. Agar hasil klasifikasi dapat bervariasi pada setiap *tree*, diperlukan sample data yang *random* yang mendukung data *replacement* baik pada dataset maupun fitur, sehingga dapat mengurangi tingkat *overfitting.* Dalam algoritma RF cara tersebut disebut teknik *bootstrapping*.

Perhitungan Random Forest diawali dengan menentukan banyak estimator, yaitu niai *increment* $b = 1$ sampai B pada pohon atau tree T_b, lalu dibuat sampel bootstrap Z^* dari banyak data *training* yaitu N. Sampel *random* juga digunakan pada pemilihan fitur setiap *tree* T_b. Untuk model klasifikasi, sampel *random* didapatkan dari $\sqrt{p}$, dengan p adalah total fitur *dataset training*. Menghitung setiap tree T_b dilakukan secara rekursif untuk menghasilkan setiap node *tree* sampai minimal node n_{min} yang ditentukan. Menggunakan perhitungan *gini index* untuk memilih *splitting point* terbaik dari banyak fitur di dalam node, sehingga dihasilkan dua pecahan node yang membagi data *training*. Pemilihan kriteria pemisah dilakukan dengan memperhitungkan nilai *average gini impurity* atau *gini-split* terkecil. *Gini-split* dapat dinotasikan sebagai berikut.

$$Gini - split(S \Longrightarrow S_1 \dots S_r) = \sum_{i=1}^{r} \frac{S_i}{S} G(S_i) \quad (3)$$

Diketahui variabel S merupakan *set* data sebelum dipisah. S_i merupakan subset data setelah dipisah, sedangkan r yaitu banyaknya subset data hasil pemisahan. Untuk perhitungan pada gabungan pohon keputusan dapat ditulis $\{T_b\}_1^B$. Pada studi kasus klasifikasi *voting* setiap *tree* menggunakan *majority vote*.

Skema Pengujian Tuning Hyperparameter

Tahap ini menguji menguji kondisi terbaik dari model klasifikasi yaitu Naïve Bayes dan Random Forest berdasarkan hasil *tuning hyperparameter*. Pengujian dilakukan dalam bentuk kombinasi besar rasio pembagian data terhadap nilai-nilai *hyperparameter.* Rasio pembagian data yang diuji meliputi $70:30$, $80:20$, dan $90:10$, dengan angka rasio terbesar untuk data *training* dan sisanya untuk data *testing.*

Naïve Bayes dengan nilai *alpha* atau *laplace smoothing* diantaranya adalah 1, 2, 5, dan 10. Untuk *Random Forest* adalah nilai *minimum sample split* diantaranya adalah 1, 10, 15, 20, dan nilai *minimum sample leaf* diantaranya adalah 1, 5, dan 10. *Minimum sample split* dan *minimum sample leaf* merupakan parameter yang paling mempengaruhi performa model *Random Forest* [13]. Rentang nilai *min sample split* yang direkomendasikan adalah 1 sampai 20, sedangkan untuk *min sample leaf* adalah 1 sampai 10. Pada Random Forest iterasi atau

banyak jumlah *tree* terbaik adalah kurang dari 200 *tree* [13].

Interpretasi Hasil Pengujian

Analisis terhadap performansi ditentukan berdasarkan nilai *accuracy, recall, sensitivity,* dan *specificity* pada *confusion matrix.* Penelitian ini memutuskan model yang lebih baik berdasarkan hasil kurva *Receiver Operating Characteristic* (ROC) dan nilai *Area Under Curve* (AUC). Identifikasi kurva AUC-ROC bertujuan untuk mengidentifikasi keterpisahan kelas dari variasi nilai *thresholds* atau dalam kata lain mengetahui seberapa baik model mengklasifikasi setiap kelas [14]. Untuk mendapatkan hasil evalusi AUC-ROC pada model dengan multi kelas, dilakukan dengan

cara *one-vs-rest* atau OvR, yaitu melabeli sebuah kelas sebagai kelas positif dan lainnya negatif [14]. Di akhir penelitian dilakukan interpretasi dari hasil akhir performa kedua model.

II. Hasil dan Pembahasan

Dapat dikelompokkan berdasarkan hasil dari tahapan utama yang meliputi hasil pengolahan data, hasil pengujian *tuning hyperparameter*, dan interpretasi performa model hasil pengujian.

Pengolahan Data

Data yang berhasil dikumpulkan melalui *crawling* data *scraping* pada platform Twitter dilanjutkan untuk tahap *pre-processing.* Tahapan ini meliputi *case folding*, yaitu mengubah semua kata dengan huruf besar dalam dokumen menjadi huruf kecil, serta menghilangkan tanda baca, *link*, nomor, dan karakter. Dilanjutkan melakukan *tokenizing* yang bertujuan memecah kalimat menjadi kumpulan kata yang biasa disebut token, lalu melakukan *spelling normalization* untuk mengubah data *tweet* yang mengandung kata tidak baku menjadi kata baku, melakukan *filtering* dengan menghilangkan kata yang dianggap tidak penting, dan terakhir melakukan *stemming* yaitu dengan mengonversi kata yang berimbuhan menjadi bentuk kata dasar.

Selanjutnya dilakukan *labeling* dan ekstraksi fitur. Hasil dari ekstraksi fitur menggunakan metode *document-term matrix* ditampilkan pada Gambar 2.

ijani	...	zenix	zero	zet	zh	zhong	zhou	zmn	zona	zonk	zontoloyo
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	1	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0
0	...	0	0	0	0	0	0	0	0	0	0

Gambar 2. Hasil ekstraksi fitur *document-term matrix*

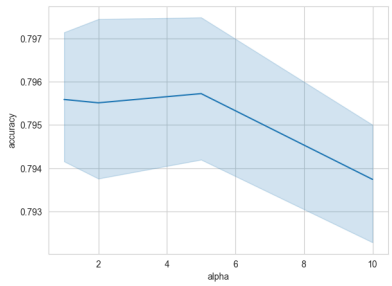
Hasil Pengujian Tuning Hyperparameter Naive Bayes

Hasil pengujian model *Multinomial Naïve Bayes* dengan melibatkan variasi rasio pembagian data dan nilai *laplace smoothing* atau *alpha* dituliskan melalui Tabel 1.

Tabel 1. Tabel hasil pengujian *tuning hyperparameter*.

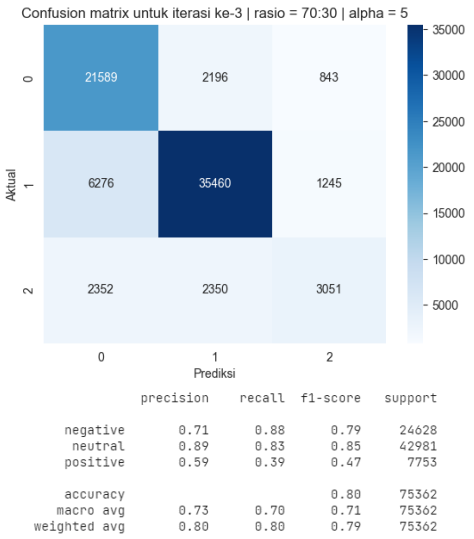
Alpha	**Rasio *Splitting Dataset***		
	70:30	**80:20**	**90:10**
1	79.71%	79.54%	79.415%
2	79.744%	79.53%	79.37%
5	79.748%	79.548%	79.419%
10	79.5%	79.39%	79.2%

Untuk mengetahui pengaruh setiap parameter terhadap perolehan akurasi dapat di visualisasi pada Gambar 3.



Gambar 3. Grafik hasil *tuning hyperparameter* nilai *alpha* Naive Bayes Classifier

Diketahui parameter *alpha*, kondisi terbaik saat mencapai puncak di nilai *alpha* yaitu 5. Penambahan nilai alpha lebih dari 5 memiliki kemungkinan penurunan berketerusan. Pada tabel 1 nilai *alpha* yaitu 5 dan rasio *splitting* sebesar 70:30 memiliki tingkat akurasi tertinggi sebesar 79.748%, sehingga dibuat tabel *confusion matrix* ditunjukkan dalam Gambar 4.



Gambar 4. *Confusion matrix* model terbaik Naive Bayes Classifier

Terlihat bahwa model tersebut kurang bisa memprediksi sentimen positif melalui nilai *recall* yang hanya mencapai 0.3, namun secara keseluruhan rata-rata *recall* mencapai 0.7, ini cukup baik untuk model klasifikasi multi kelas. Maka model NBC dengan rasio nilai *laplace smoothing* sebesar 5 merupakan kondisi model terbaik untuk studi kasus analisis sentimen kebijakan kenaikan harga BBM di Indonesia.

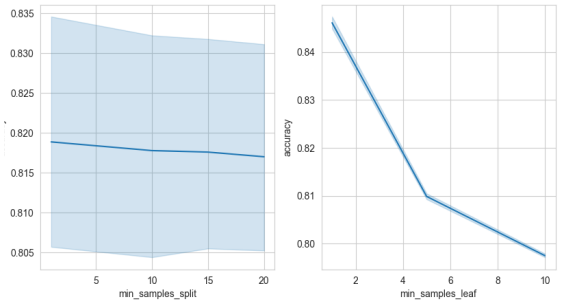
Hasil Pengujian Tuning Hyperparameter Random Forest

Pengujian ini melibatkan kombinasi nilai *minimum sample split* yaitu batas jumlah minimal data pada node untuk dilakukan *splitting* dan *minimum sample leaf* yaitu jumlah minimal data untuk menjadi sebuah *leaf node* terhadap rasio pembagian data.

Tabel 2. Hasil pengujian *tuning hyperparameter* Random Forest

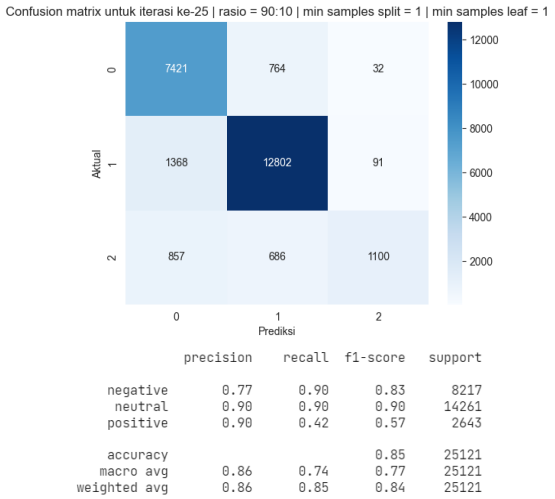
		Min *Split*	Rasio *Splitting Dataset*		
			70:30	80:20	90:10
Min *Leaf*	1	1	84.75%	84.70%	85.15%
		10	84.53%	84.56%	84.88%
		15	84.48%	84.49%	84.59%
		20	84.28%	84.42%	84.41%
	5	1	81.02%	81.01%	81.11%
		10	81.05%	80.90%	80.83%
		15	81.17%	80.83%	80.92%
		20	80.91%	81.05%	80.94%
	10	1	79.74%	79.73%	79.71%
		10	79.82%	79.69%	79.68%
		15	79.69%	79.68%	79.92%
		20	79.72%	79.79%	79.73%

Untuk mengetahui pengaruh setiap variabel parameter, dapat dilakukan visualisasi terhadap perolehan nilai akurasi.



Gambar 5. Grafik hasil t*uning hyperparameter* nilai *min sample split* dan *min sample leaf* Random Forest.

Dapat diketahui bahwa pengurangan nilai *min samples split* cukup memberikan peningkatan sekitar 1%. Pada *min samples leaf* sangat mempengaruhi hasil akurasi prediksi, semakin kecil nilainya akurasi akan semakin naik secara signifikan. Pada Tabel 2 dan gambar 5 diketahui bahwa kombinasi nilai *min samples split* sebesar 1, *min samples leaf* sebesar 1, mendapat perolehan akurasi tertinggi hingga mencapai 85.15% untuk rasio 90:10.



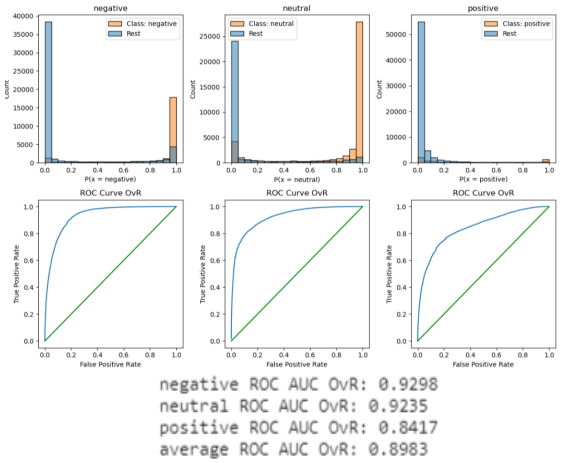
Gambar 6. *Confusion matrix* model terbaik Random Forest.

Diketahui bahwa model kurang bisa memprediksi sentimen positif dengan *recall* yang didapat hanya mencapai 0.42, namun rata-rata ketepatan klasifikasi untuk *recall* mencapai 0.74, yang mana nilai tersebut sudah cukup baik untuk model klasifikasi multi kelas.

Maka dapat simpulkan bahwa *hyperparameter min samples split* sebesar 1 dan nilai *min samples leaf* sebesar 1, merupakan kondisi terbaik model RF untuk penelitian ini.

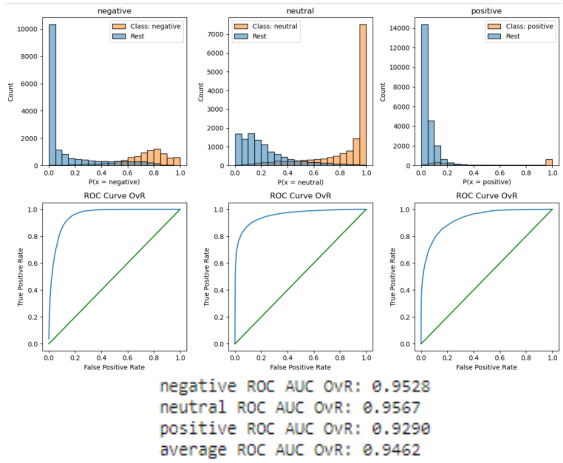
Interpretasi Performa Model Hasil Pengujian

Pada model *Naïve Bayes* dengan rasio *splitting dataset* 70:30 dan nilai *alpha* sebesar 5, dihasilkan kurva ROC dan nilai AUC ditunjukkan dalam Gambar 7.



Gambar 7. Kurva ROC dan nilai AUC model Naive Bayes.

Pada model *Random Forest* dengan *rasio splitting dataset* yang sama yaitu 70:30, *n-estimator* sebesar 150, *min samples split* 1, *min samples leaf* 1 yang ditampilkan sebagai berikut.



Gambar 8. Kurva ROC dan nilai AUC model Random Forest.

Didapatkan rata-rata skor AUC dari model klasifikasi NBC berdasarkan Gambar 7 adalah 89.83%, lalu pada RF dari Gambar 8 adalah 94.62%. Jadi model NBC dengan menggunakan nilai *hyperparameter alpha* yaitu 5 dan rasio *splitting dataset* 70:30, mendapatkan akurasi sebesar 79.71% dan nilai rata-rata AUC-ROC sebesar 89.83%. Untuk model RF dengan raiso *splitting* 70:30, nilai *hyperparameter min samples split* yaitu 1, *min samples leaf* sebesar 1, dan jumlah estimator 150, didapatkan akurasi sebesar 84.75%, lalu untuk nilai rata-rata AUC-ROC sebesar 94.62%.

III. Kesimpulan

Disimpulkan bahwa *Random Forest* lebih unggul dengan selisih akurasi mencapai 5.41%, dan untuk selisih rata-rata nilai AUC-ROC mencapai 4.79%. Random Forest mendapatkan akurasi mencapai 84.75% untuk *splitting* data 70:30 dan rata-rata skor *AUC-ROC* sebesar 94.62%. Lebih bagus dibandingkan *Naïve Bayes* untuk rasio yang sama didapatkan akurasi 79.71% dan skor *AUC-ROC* sebesar 89.83%.

Membuktikan bahwa algoritma *Random Forest* memiliki ketelitian yang tinggi, karena sifatnya yang *ensemble learning*. Diperkuat juga karena *dataset* yang digunakan adalah data teks, sehingga dapat menghasilkan banyak variasi kata dari hasil ekstraksi fitur yang mempengaruhi proses komputasi. Maka, algoritma *Random Forest* lebih cocok digunakan pada studi kasus analisis sentimen kebijakan kenaikan harga BBM menggunakan data teks *tweet* Twitter dibandingkan *Naïve Bayes*.

IV. Daftar Pustaka

[1] Latif, A. (2015). Dampak Fluktuasi Harga Bahan Bakar Minyak Terhadap Suplai Sembilan bahan Pokok di pasar Tradisional. *Al-Buhuts*, *11*(1), 91-116.

[2] "Presiden: BLT BBM dan BSU Lakukan secara Mudah, Cepat, dan Tepat Sasaran." https://www.kemenkeu.go.id/informasi-publik/publikasi/berita-utama/Presiden-BLT-BBM-dan-BSU-Lakukan-secara-Mudah (accessed Feb. 28, 2023).

[3] Wongkar, M., & Angdresey, A. (2019, October). Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter. In *2019 Fourth International Conference on Informatics and Computing (ICIC)* (pp. 1-5). IEEE.

[4] Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221.

[5] Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

[6] S. Kannan *et al.*, "Big Data Analytics for Social Media," *Big Data: Principles and Paradigms*, pp. 63–94, Jan. 2016, doi: 10.1016/B978-0-12-805394-2.00003-9.

[7] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

[8] Karthika, P., Murugeswari, R., & Manoranjithem, R. (2019, April). Sentiment analysis of social media network using random forest algorithm. In *2019 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)* (pp. 1-5). IEEE.

[9] Bonta, V., & Janardhan, N. K. N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, *8*(S2), 1-6.

[10] Damaratih, D. A. (2021, October). Sentiment analysis of online lecture opinions on twitter social media using naive bayes classifier. In *2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)* (pp. 24-28). IEEE.

[11] Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019, April). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 593-596). IEEE.

[12] Aggarwal, C. C. (2015). *Data mining: the textbook* (Vol. 1). New York: springer.

[13] Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. P. D. L. F. (2018). An empirical study on hyperparameter tuning of decision trees. *arXiv preprint arXiv:1812.02207*.

[14] Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., ... & Holzinger, A. (2021). Deep ROC analysis and AUC as balanced average accuracy to improve model selection, understanding and interpretation. *arXiv preprint arXiv:2103.11357*.