ISSN : 2527-5836
e-ISSN : 2528-0074

Vol. 6 No. 2, Mei 2021

JISKa

Jurnal Informatika Sunan Kalijaga

Jurusan Teknik Informatika
Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta

Tim Pengelola JISKa Edisi Mei 2021

Mitra Bestari (Reviewer)

Reviewer Internal:

1. Mandahadi Kusuma, M.Eng. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
2. Maria Ulfa Siregar, Ph.D. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
3. Rahmat Hidayat, M.Cs. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
4. Usfita Kiftiyani, M.Sc. (UIN Sunan Kalijaga Yogyakarta, Indonesia)

Reviewer Eksternal (Mitra Bestari):

1. Ahmad Fathan Hidayatullah, M.Cs. (Universitas Islam Indonesia Yogyakarta, Indonesia)
2. Alam Rahmatulloh, M.T. (Universitas Siliwangi Tasikmalaya, Indonesia)
3. Dr. Cahyo Crysdian (UIN Maulana Malik Ibrahim Malang, Indonesia)
4. Dr. Eng. Ganjar Alfian (Dongguk University Seoul, Korea, Republic of)
5. Muhammad Rifqi Maarif, M.Eng. (Universitas Jenderal Achmad Yani Yogyakarta, Indonesia)
6. Mushab Al Barra, M.Kom. (Universitas Ahmad Dahlan Yogyakarta, Indonesia)
7. Dr.Eng. M. Alex Syaekhoni (Dongguk University Seoul, Korea, Republic of)
8. Norma Latif Fitriyani, M.Sc. (Dongguk University Seoul, Korea, Republic of)
9. Oman Somantri, M.Kom. (Politeknik Negeri Cilacap, Indonesia)
10. Puji Winar Cahyo, M.Cs. (Universitas Jenderal Achmad Yani Yogyakarta, Indonesia)
11. Rischan Mafrur, M.Eng. (The University of Queensland Brisbane, Australia)
12. Suhirman, Ph.D. (Universitas Teknologi Yogyakarta, Indonesia)
13. Yunita Ardilla, M.Sc (Institut Teknologi Sepuluh November Surabaya, Indonesia)

ISSN : 2527-5836

e-ISSN: 2528-0074

JISKa

Vol. 6, No. 2, MEI 2021

DAFTAR ISI

JISKa, Vol. 6, No. 2, MEI, 2021, Pp. 70 – 77
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Segmentasi Pelanggan Berdasarkan Perilaku Penggunaan Kartu Kredit Menggunakan Metode *K-Means Clustering*

Fatimah Defina Setiti Alhamdani [(1)], Ananda Ayu Dianti [(2)*], Yufis Azhar [(3)]
Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Malang, Malang
e-mail : {defina.a19,anandadianti8}@gmail.com, yufis@umm.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 6 Juni 2020, direvisi 25 Juni 2020, diterima 1 Juli 2020, dan dipublikasikan 3 Mei 2021.

Abstract

Credit card is one of the payment media owned by banks in conducting transactions. Credit card issuers provide benefits for banks with interest that must be paid. Credit card issuers also provide losses to banks that have agreed to pay not to pay their credit card bills. To request a loan from the bank, a cluster model is needed. This study, proposing a segmentation system in research using credit cards to determine marketing strategies using the K-Means Clustering method and conducting experiments using the 4 methods namely K-Means, Agglomerative Clustering, GMM, and DBSCAN. Clustering is done using 9000 active credit card user data at banks that have 18 characteristic features. The results of cluster quality accuracy obtained by using the K-Means method are 0.207014 with the number of clusters 3. Based on the results obtained by considering 4 of these methods, the best method for this case is K-Means.

Keywords: Credit Card, Cluster, K-Means, Elbow Method, Silhouette Method

Abstrak

Kartu kredit merupakan salah satu media pembayaran yang dimiliki oleh nasabah bank dalam melakukan sebuah transaksi. Penerbitan kartu kredit memberikan keuntungan bagi pihak bank dengan adanya bunga yang harus dibayar. Penerbitan kartu kredit juga memberikan kerugian pada pihak bank apabila nasabah tidak membayar tagihan kartu kreditnya. Untuk mengantisipasi kerugian pada pihak bank diperlukan sebuah model *cluster.* Pada penelitian ini diusulkan sistem segmentasi pelanggan berdasarkan perilaku penggunaan kartu kredit untuk menentukan strategi pemasaran efektif dengan menggunakan metode *K-Means Clustering* dan melakukan sebuah percobaan dengan menguji 4 metode yaitu *K-Means, Agglomerative Clustering,* GMM, dan DBSCAN. *Clustering* dilakukan menggunakan 9000 data pengguna aktif kartu kredit pada sebuah bank yang memiliki 18 fitur karakteristik. Nilai *silhouette coefficient* yang didapatkan dengan menggunakan metode *K-Means* adalah 0.207014 dengan jumlah *cluster* sama dengan 3. Berdasarkan hasil yang didapatkan dengan menguji 4 metode tersebut, metode yang paling baik untuk kasus ini adalah *K-Means*.

Kata Kunci: Kartu Kredit, *Cluster*, *K-Means*, Metode *Elbow*, Metode Silhouette

1. PENDAHULUAN

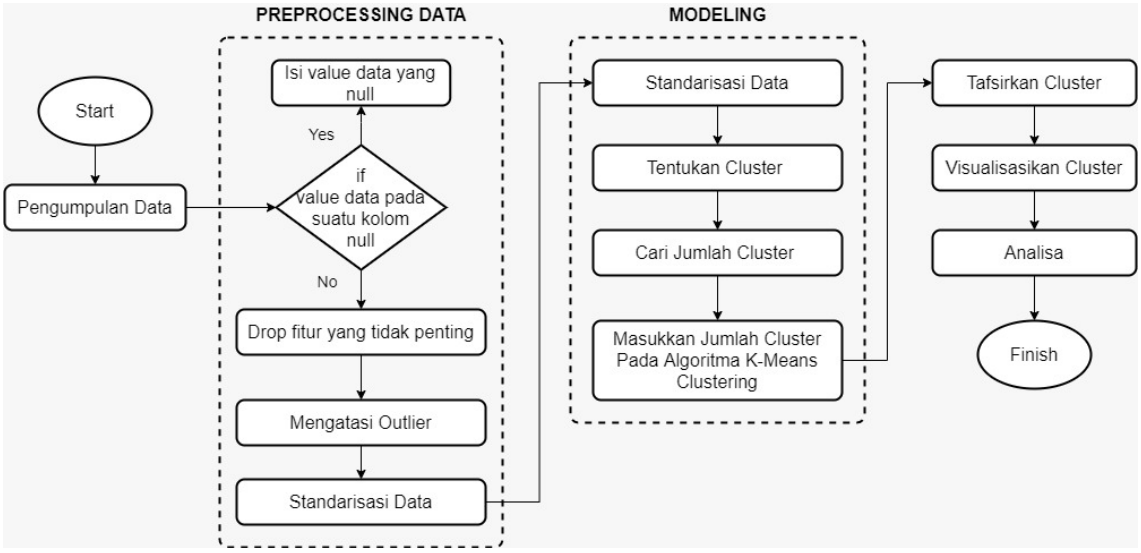
Kartu kredit merupakan salah satu media pembayaran yang dimiliki oleh nasabah bank dalam melakukan sebuah transaksi. Dengan berkembangnya bisnis kartu kredit saat ini fitur-fitur pada kartu kredit semakin beragam sehingga semakin banyak peminat kartu kredit (Sumarto et al., 2012). Namun, pemegang yang bersangkutan juga disertai dengan syarat dan ketentuan yang berlaku. Pada akhir tahun 2013, AKKI mencatat jumlah kartu yang telah beredar sebanyak 15.007.492 buah kartu. Dari informasi pertumbuhan tersebut telah menunjukkan bahwa pertumbuhan penggunaan kartu kredit di Indonesia berkembang sangat pesat (Ramadani, 2019). Penerbitan kartu kredit memberikan keuntungan bagi pihak bank dengan adanya bunga yang harus dibayar oleh nasabah. Penerbitan kartu kredit juga memberikan kerugian pada pihak bank apabila nasabah tidak membayar tagihan kartu kreditnya. Namun, untuk mengantisipasi kerugian pada pihak bank diperlukan sebuah model *cluster* untuk menganalisa pelanggan berdasarkan perilaku penggunaan kartu kredit sehingga pihak bank dapat menentukan strategi pemasaran kartu kredit.

Pada penelitian sebelumnya dalam menganalisis kartu kredit terdapat beberapa topik penelitian salah satunya mendeteksi penipuan transaksi kartu kredit menggunakan pendekatan *clustering* metode *K-Means* dengan melakukan *clustering* menjadi 4 *cluster* yaitu *low, high, risky,* dan *high risky* (Vaishali, 2014). Selain itu, juga telah dilakukan klasifikasi untuk *fraud detection* pada *e-tail merchant* menggunakan metode seperti *random forest, logistic regression,* dan *support vector machine* dan menghasilkan hasil akurasi yang baik (Carneiro et al., 2017). Pada penelitian sebelumnya dilakukan penelitian dalam *behavioral analysis* untuk prediksi penggunaan kartu kredit oleh pemilik kartu dengan beberapa faktor seperti pekerjaan dan kebiasaan pemegang kartu menggunakan kartu kredit. Penelitian ini fokus pada ketertarikan antara pendapatan *customer* dan penggunaan penuh *credit limit* yang dimiliki oleh *customer* (Dewri et al., 2016). Metode yang digunakan adalah *K-Means* untuk *clustering personal credit analysis* untuk menghasilkan analisa *trend* kebiasaan *customer* yang digunakan untuk menganalisa kelompok dari *customer* (Han & Chai, 2012).

Berdasarkan permasalahan di atas, maka dalam penelitian ini diusulkan untuk membuat sistem segmentasi pelanggan berdasarkan perilaku penggunaan kartu kredit untuk menentukan strategi pemasaran efektif dengan menggunakan metode *K-Means Clustering*. Algoritma *K-Means* merupakan suatu metode *clustering* yang dinilai paling sederhana dikarenakan dapat mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputas yang relative cepat dan efisien (Siregar, 2018). Metode yang diusulkan pada penelitian ini adalah melakukan sebuah percobaan dengan menguji 4 metode yaitu *K-Means, Agglomerative Clustering,* GMM (*Gaussian Mixture Model*), dan DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)*. Clustering* dilakukan menggunakan 9000 data pengguna aktif kartu kredit pada sebuah bank. Dengan menggunakan 4 metode *clustering* tersebut dapat menentukan hasil perbandingan yang dilakukan untuk menyakinkan bahwa metode yang diusulkan merupakan metode yang tepat untuk diterapkan.

2. METODE PENELITIAN

Pada penelitian ini menggunakan metode *data mining* dengan beberapa tahapan: pengumpulan data, *preprocessing*, *clustering*, dan analisis. Alur tahapan penelitian tertera pada Gambar 1.



Gambar 1. *Flowchart* Tahapan Penelitian.

2.1. Pengumpulan Data

Pada penelitian ini *dataset* yang digunakan bersifat *public*. Sumber data berasal dari https://www.kaggle.com/arjunbhasin2013/ccdata. Data yang diambil berisi 9000 data dari pengguna aktif kartu kredit pada sebuah bank. Data tersebut berupa angka pada setiap fiturnya kecuali fitur "CUSTID" yang memiiki kombinasi huruf dan angka. *Credit Card Dataset* ini memiliki 18 fitur karakteristik untuk setiap penggunanya, yaitu:

a) **CUSTID**: Identifikasi *customer* kartu kredit.
b) **BALANCE**: jumlah saldo yang tersisa pada akun masing-masing *customer* untuk melakukan pembelian.
c) **BALANCEFREQUENCY**: Seberapa sering saldo pada akun *customer* diperbarui, dengan skor antara 0 dan 1 (1: sering diperbarui, 0: tidak sering diperbarui).
d) **PURCHASES**: jumlah pembelian yang dilakukan dari akun milik *customer.*
e) **ONEOFFPURCHASES**: jumlah pembelian maksimum yang dilakukan dalam sekali jalan.
f) **INSTALLMENTSPURCHASES**: jumlah pembelian yang dilakukan dengan mencicil.
g) **CASHADVANCE**: Uang muka yang diberikan oleh *customer* dalam bentuk tunai.
h) **PURCHASESFREQUENCY**: seberapa sering pembelian dilakukan, dengan skor antara 0 dan 1 (1: sering, 0: tidak sering).
i) **ONEOFFPURCHASESFREQUENCY**: seberapa sering pembelian dilakukan dalamsekali jalan, dengan skor antara 0 dan 1 (1: sering, 0: tidak sering).
j) **PURCHASESINSTALLMENTSFREQUENCY**: seberapa sering pembelian dalam angsuran dilakukan, dengan skor antara 0 dan 1 (1: sering, 0: tidak sering).
k) **CASHADVANCEFREQUENCY**: seberapa sering uang muka dibayarkan.
l) **CASHADVANCETRX**: jumlah transaksi yang dilakukan dengan uang tunai.
m) **PURCHASESTRX**: banyaknya transaksi pembelian yang dilakukan.
n) **CREDITLIMIT**: batas penggunaan kartu kredit.
o) **PAYMENTS**: jumlah pembayaran yang dilakukan oleh *customer.*
p) **MINIMUM_PAYMENTS**: jumlah pembayaran minimum yang dilakukan oleh pengguna.
q) **PRCFULLPAYMENT**: pembayaran penuh yang dibayarkan oleh *customer* dalam bentuk persen.
r) **TENURE**: masa berlaku layanan kartu kredit.

2.2. *Preprocessing* Data

Penelitian ini memiliki beberapa tahapan *preprocessing* yang meliputi:

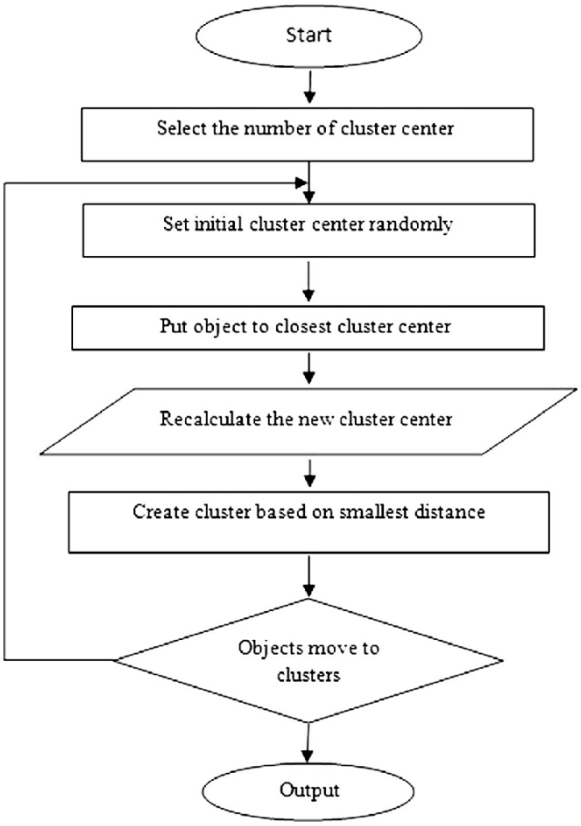
1) *Data Cleansing*
Pada tahapan ini dilakukan pengecekan data yang kosong pada *dataset* yang digunakan. Ketika ditemukan data yang kosong maka data tersebut akan diisi sesuai dengan tipe datanya menggunakan metode yang dipilih. Selain diisi dengan sebuah nilai, pembersihan data juga dapat dilakukan dengan menghapus fitur yang tidak penting atau tidak relevan pada saat pemrosesan *dataset*.

2) Menemukan Korelasi Antar Fitur
Langkah awal yang dilakukan pada tahapan ini adalah dengan menghitung korelasi kolom fitur secara berpasangan. Lalu dilanjutkan dengan memvisualisasikan hasil perhitungan korelasi antar kolom fitur agar lebih mudah untuk diamati dan dipahami. Pada proses ini dapat terlihat keterkaitan antara fitur yang digunakan dan fitur yang tidak digunakan. Fitur yang tidak digunakan tersebut adalah CUST_ID.

3) Mengatasi *Outlier*
Data *outlier* disebut juga dengan data pencilan. Pengertian dari *outlier* adalah data observasi yang memiliki nilai ekstrim secara *univariate* dan *multivariate* (Hidayat, 2016). Nilai ekstrim pada data observasi adalah nilai yang berbeda dengan sebagian nilai lain dalam suatu kelompok. Pada tahap ini dilakukan menghitung nilai absolut pada z *score* dari setiap nilai dalam sampel, *relative* terhadap *mean* sampel dan standar deviasi. Hal ini dilakukan untuk mentransformasikan data agar nilai ekstrim bisa dikurangi jaraknya dengan kelompok yang lain. Lalu filter data menjadi data *outlier free*.

4) Standarisasi Data
Standarisasi fitur dengan menghapus *mean* dan *scaling* ke varian unit. Standarisasi data adalah persyaratan umum bagi banyak penduga pembelajaran mesin. Misalnya banyak elemen yang digunakan dalam fungsi objektif dari algoritma pembelajaran (seperti kernel RBF dari *Support Vector Machines* atau L1 dan L2 regularizer model linier) mengasumsikan bahwa semua fitur berpusat di sekitar 0 dan memiliki varian dalam urutan yang sama. Jika suatu fitur memiliki varians yang urutan besarnya lebih besar dari yang lain, itu mungkin mendominasi fungsi objektif dan membuat estimator tidak dapat belajar dari fitur lain dengan benar seperti yang diharapkan.

2.3. Tahap Clustering



Gambar 2. *Flowchart* Tahapan *Clustering* (Younus et al., 2015).

Dalam penelitian ini dibentuk sebuah model dengan menggunakan algoritma *K-Means Clustering*. *K-Means* dapat diartikan dengan metode *clustering* data non-hirarki yang menggunakan metode partisi (*apportioning strategy*) berbasis *centroid* yang mengelompokkan suatu data menjadi satu atau lebih (Tendean et al., 2020). *Centroid* merupakan sebuah nilai yang digunakan untuk menghitung jarak pada suatu objek data dengan membuat penentuan nilai awalnya dilakukan secara acak dan pada nilai tiap iterasinya menggunakan rumus (Hajar et al., 2020).

$$\bar{v}_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} x_{kj} \qquad (1)$$

Menghitung jarak antara titik *Centroid* dengan titik tiap objek.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \qquad (2)$$

1) Pengelompokan objek untuk menentukan anggota *Cluster* adalah dengan memperhitungkan jarak minimal.

2) Kembali ke tahap 2, melakukan perulangan hingga nilai *Centroid* yang didapatkan tetap dan anggota *Cluster.*

Langkah awal yang perlu dilakukan adalah menentukan *cluster* menggunakan *elbow method*. *Elbow method* adalah suatu metode untuk melihat perbedaan persentase pada jumlah *cluster* (Purnima & Arvind, 2014). Metode ini diperlukan untuk menentukan jumlah *cluster* terbaik yang akan membentuk siku pada suatu titik. *Elbow criterion* adalah suatu *modelling criterion* yang bisa digunakan untuk menentukan jumlah cluster dengan melihat perubahan perbandingan antara nilai RMSSTD (*Root Mean Square Standartd Deviaton* dan RS (*R-Square*). Hal ini dilihat dengan membandingkan persentase tingkat perubahan kedua nilai (RMSSTD dan RS). RMSSTD untuk mengukur kemiripan pada cluster yang harus bernilai rendah sedangkan RS untuk mengukur perbedaan pada cluster yang harus bernilai tinggi (Sharma, 1996).

Jika terdapat suatu kondisi yang berlawanan dengan kondisi sebelumnya, maka titik sebelum terjadinya perubahan tersebut akan dianggap sebagai jumlah *cluster* yang paling tepat. Setelah menentukan *cluster*, proses dilanjutkan dengan mencari jumlah *cluster* menggunakan rata-rata metode *Silhouette Coefficient*. *Silhouette coefficient* digunakan untuk melihat seberapa baik kualitas dan kekuatan *cluster,* seberapa baik suatu objek ditempatkan dalam suatu *cluster* (Anggara et al., 2016). Lalu masukkan jumlah *cluster* yang sudah ditemukan kedalam fungsi *K-Means Clustering*. Setelah itu beri penafsiran menggunakan data cluster yang telah terbentuk dari proses *K-Means Clustering*. Untuk memudahkan proses analisis dilakukan visualisasi terhadap *cluster* yang terbentuk. Visualisasi ini dapat dilakukan dengan mereduksi dimensi data. Hal tersebut dilakukan karena pada sejumlah fitur yang digunakan terdapat kemungkinan fitur yang tidak relevan dan redundan.

Teknik yang digunakan dalam penelitian ini adalah T-SNE (*T-Distributed Stochastic Neighbor Embedding*) Tidak seperti metode seleksi fitur yang mengurangi jumlah fitur dengan cara menghilangkan fitur yang dianggap tidak penting tanpa membentuk fitur baru, "T-SNE mengurangi dimensi data dengan cara meminimalkan dua perbedaan yaitu distribusi yang mengukur kemiripan berpasangan dari objek input dan distribusi yang mengukur kemiripan berpasangan dari titik dimensi rendah yang sesuai dalam penyematan. T-SNE berupaya mengidentifikasi kelompok berdasarkan kesamaan titik data dengan banyak fitur" (Pathak, 2018).

Dalam penelitian ini dilakukan perbandingan *silhouette coefficient score* pada beberapa metode *clustering* lainnya yaitu DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), GMM (*Gaussian Mixture Models*), Aglomerative Clustering. Perbandingan tersebut dilakukan untuk menyakinkan bahwa metode yang diusulkan merupakan metode yang tepat untuk diterapkan.

2.4. Analisis Data

Tahapan analisis data dilakukan dengan mengamati data hasil visualisasi dari pemrosesan *K-Means clustering* yang direduksi dimensinya menggunakan T-SNE. Setelah proses analisis selesai dapat dijadikan sebagai acuan dalam menentukan strategi *marketing* yang dapat diambil sebagai hasil.

3. HASIL DAN PEMBAHASAN

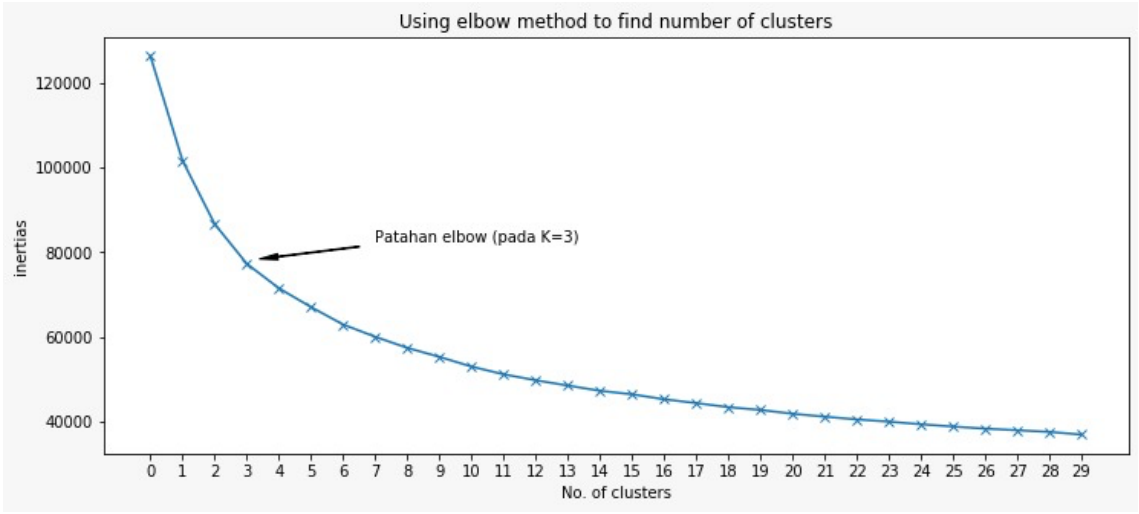
Tabel 1. Hasil Perbandingan *Silhouette Score* dan Jumlah *Cluster*.

No	Metode *Clustering*	*Silhouette Score*	Jumlah *Cluster* yang Dihasilkan
1	DBSCAN	-0.351371	9
2	GMM	0.003558	12
3	*Agglomerative clustering*	0.137499	9
4	*K-Means*	0.207014	3

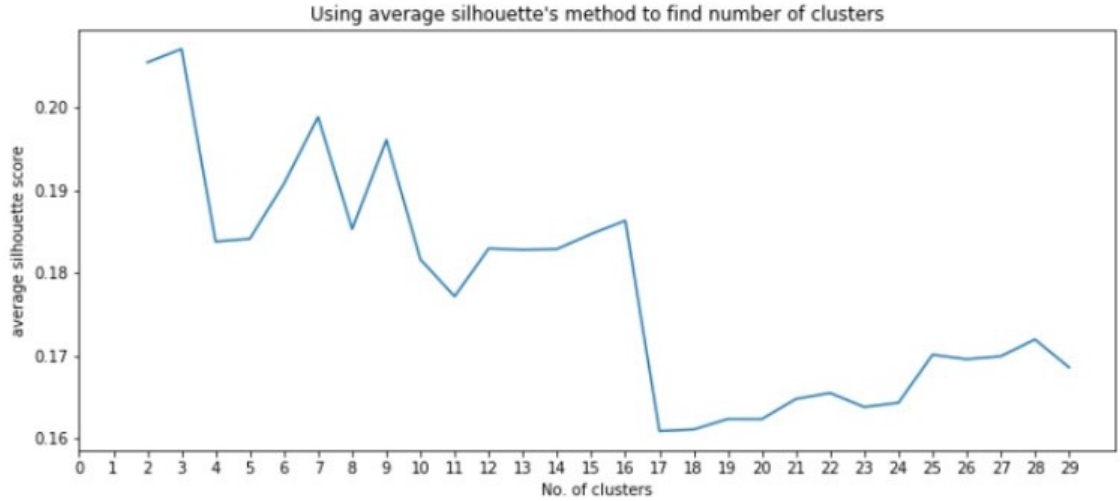
Hasil dari proses perbandingan *silhouette score* pada DBSCAN, GMM, *Aglomerative Clustering*, dan *K-Means* tertera dalam Tabel 1. Berdasarkan hasil perbandingan empat metode *clustering*

tersebut terbukti bahwa metode yang diusulkan pada penelitian ini menghasilkan *silhouette score* terbaik yaitu 0,207014.

3.1. Hasil Uji Elbow Method dan Silhouette Method pada Metode K-Means



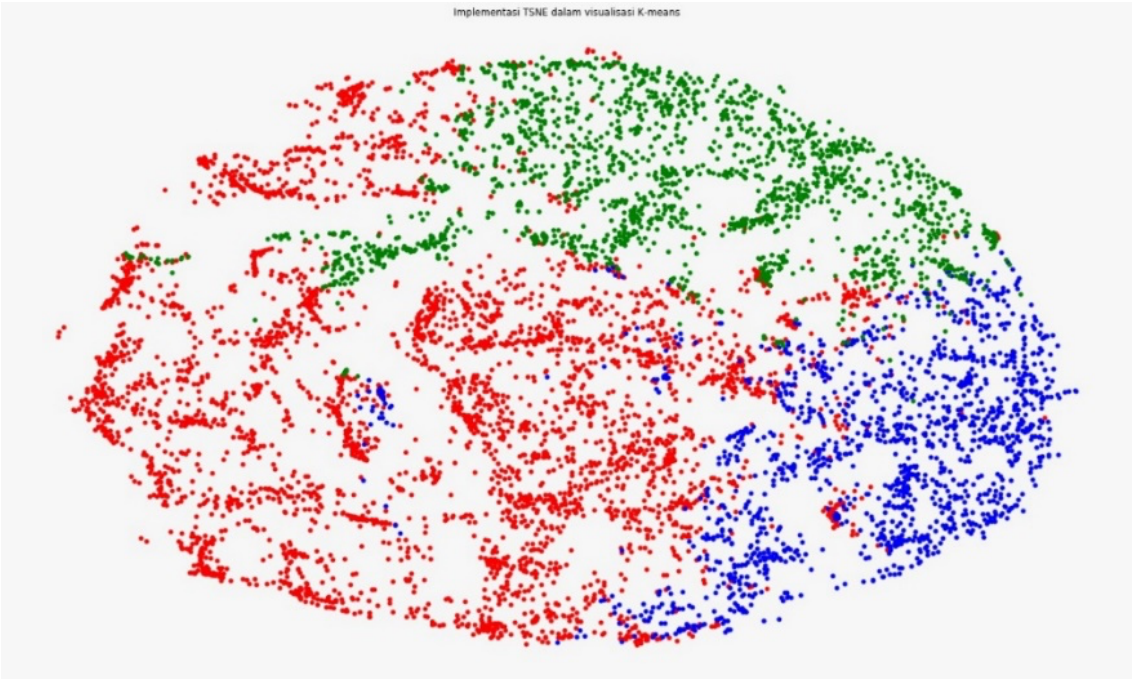
Gambar 3. Hasil Uji *Elbow Method.*



Gambar 4. Hasil Uji *Silhouette Method.*

Dari grafik hasil uji *elbow method* yang telah tersaji pada Gambar 3 terlihat patahan atau siku yang terbentuk terdapat pada nomor *cluster* 3. Lalu pada Gambar 4 grafik hasil uji *silhouette method* terlihat nilai yang paling tinggi adalah nomor *cluster* 3 dengan nilai 0.207014. Penggabungan antara hasil analisa pada Gambar 3 dan Gambar 4 menghasilkan keputusan nilai K terbaik untuk *K-Means* yang cocok untuk digunakan pada penelitian ini adalah 3.

3.2. Hasil Visualisasi T-SNE Menggunakan *K-means*



Gambar 5. Implementasi T-SNE untuk Visualisasi *K-Means.*

Pada Gambar 5. dilakukan visualisasi data hasil pengelompokan menggunakan algoritma *K-Means*. Hal itu dilakukan untuk memudahkan dalam melihat letak persebaran dari masing-masing *cluster* yang telah terbentuk. Dari visualisasi terlihat bahwa *K-Means* menghasilkan tiga *cluster* perilaku *customer credit card* dengan jumlah anggota dari masing-masing *cluster* berbeda-beda tertera pada Tabel 2.

Tabel 2. Keterangan pada Gambar 5.

Warna Titik	Keterangan	Jumlah Titik
Merah	*Customer* dengan penggunaan kartu kredit yang moderat	1696
Biru	*Customer* dengan penggunaan kartu kredit paling sedikit	1341
Hijau	*Customer* dengan lebih banyak menggunakan kartu kredit dan melakukan pembelian produk lebih sering	824

4. KESIMPULAN

Dataset pada penelitian ini memiliki jumlah data yang besar dan memiliki kesamaan pada tiap datanya merupakan hal yang tidak bisa diremehkan dalam menentukan metode *clustering* yang akan digunakan. Maka dari itu dibentuklah beberapa percobaan untuk membandingkan satu sama lain yaitu *K-Means*, *Agglomerative Clustering*, GMM dan DBSCAN dengan melihat *silhouette score* yang dihasilkan oleh masing-masing metode. Setelah melakukan perbandingan dengan 4 metode tersebut, metode terbaik untuk dataset kartu kredit ini adalah *K-Means*. Dari proses *clustering* yang dijalankan dihasilkan 3 *cluster* sehingga bisa digunakan untuk memahami segmentasi perilaku *customer* dalam menggunakan kartu kredit. Nilai *silhouette coefficient* yang didapatkan dengan menggunakan metode *K-Means* adalah 0.207014.

DAFTAR PUSTAKA

Anggara, M., Sujiani, H., & Helfi, N. (2016). Pemilihan Distance Measure Pada K-Means Clustering Untuk Pengelompokkan Member Di Alvaro Fitness. *Jurnal Sistem Dan Teknologi Informasi*, *1*(1), 1–6.

Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems*, *95*(June 2019), 91–101. https://doi.org/10.1016/j.dss.2017.01.002

Dewri, L. V., Islam, M. R., & Saha, N. K. (2016). Behavioral Analysis of Credit Card Users in a Developing Country: A Case of Bangladesh. *International Journal of Business and Management*, *11*(4), 299. https://doi.org/10.5539/ijbm.v11n4p299

Hajar, S., Novany, A. A., Windarto, A. P., Wanto, A., & Irawan, E. (2020). *Penerapan K-Means Clustering Pada Ekspor Minyak Kelapa Sawit Menurut Negara Tujuan*. 314–318.

Han, P., & Chai, J. (2012). The application of K-means in personal credit analysis. *Advanced Materials Research*, *403–408*, 2461–2464. https://doi.org/10.4028/www.scientific.net/AMR.403-408.2461

Hidayat, A. (2016). *Pengertian Data Outlier Univariat dan Multivariat*.

Pathak, M. (2018). *Introduction to t-SNE*.

Purnima, B., & Arvind, K. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, *105*(9), 17–24.

Ramadani, M. (2019). PENGARUH ATTITUDE TOWARD MONEY TERHADAP COMPULSIVE BUYING BEHAVIOUR PENGGUNA KARTU KREDIT. *Jurnal Ekonomi Vokasi*, *2*(9), 1689–1699.

Sharma, S. (1996). *Applied Multivariate Techniques Subhash Sharma* (pp. 1–5).

Siregar, M. H. (2018). Data Mining Klasterisasi Penjualan Alat-Alat Bangunan Menggunakan Metode K-Means (Studi Kasus Di Toko Adi Bangunan). *Jurnal Teknologi Dan Open Source*, *1*(2), 83–91. https://doi.org/10.36378/jtos.v1i2.24

Sumarto, S., Subroto, A., & Arianto, A. (2012). Penggunaan Kartu Kredit Dan Perilaku Belanja Kompulsif: Dampaknya Pada Risiko Gagal Bayar. *Jurnal Manajemen Pemasaran*, *6*(1). https://doi.org/10.9744/pemasaran.6.1.1-7

Tendean, T., Purba, W., & Kom, M. (2020). *Analisis Cluster Provinsi Indonesia Berdasarkan Produksi Bahan Pangan Menggunakan Algoritma K-Means*. *1*(2), 5–11.

Vaishali, V. (2014). Fraud Detection in Credit Card by Clustering Approach. *International Journal of Computer Applications*, *98*(3), 29–32. https://doi.org/10.5120/17164-7225

Younus, Z. S., Mohamad, D., Saba, T., Alkawaz, M. H., Rehman, A., Al-Rodhaan, M., & Al-Dhelaan, A. (2015). Content-based image retrieval using PSO and k-means clustering algorithm. *Arabian Journal of Geosciences*, *8*(8), 6211–6224. https://doi.org/10.1007/s12517-014-1584-7

JISKa, Vol. 6, No. 2, MEI, 2021, Pp. 78 – 89
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisis Sentimen *Review* Halodoc Menggunakan *Nai¨ve Bayes Classifier*

Asep Hendra [(1)*], Fitriyani [(2)]
Sistem Informasi, Fakultas Teknologi Informasi, Universitas Adhirajasa Reswara Sanjaya, Bandung
e-mail : ace.hendra07@gmail.com, fitriyani@ars.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 5 Agustus 2020, direvisi 20 September 2020, diterima 13 Oktober 2020, dan dipublikasikan 3 Mei 2021.

Abstract

Healthcare service has the role to help and serve people to access medical services, i.e. providing medicines, medical consultation, or health control. Healthcare service has been transforming to a digital platform. Halodoc is one of the digital platforms that people can use for free or paid, user can also give reviews of Halodoc's performance and services on Google Play Store to give feedback that Halodoc can use to evaluate and improve the app. The Google Play Store review is increasing every day. Therefore an analysis for the review with sentiment analysis for Halodoc's review is needed, first phase of sentiment analysis for the review is preprocessing which has tokenization, transform to lower cases, filter stopword, dan filter token (by length) processes. The data is divided into two positive and negative classes with cross-validation and a k-fold validation value of 10, using Naïve Bayes Classifier algorithm with 81,68% accuracy and AUC 0.756, categorized as fair classification.

Keywords: Naïve Bayes Classifier, Text Mining, Sentiment Analysis, Classification, Halodoc

Abstrak

Layanan kesehatan yang memiliki peran untuk melayani dan membantu masyarakat dalam memanfaatkan layanan kesehatan, baik itu untuk berobat ketika sakit, konsultasi kesehatan atau untuk kontrol kesehatan saja. Layanan kesehatan kini bertransformasi menjadi layanan kesehatan digital. Halodoc merupakan aplikasi layanan digital yang dapat digunakan oleh masyarakat secara gratis maupun berbayar, penggunanya dapat memberikan ulasan terhadap kinerja atau layanan Halodoc melalui Google Play Store untuk menjadi evaluasi dan peningkatan kinerja Halodoc. Ulasan pada Google Play Store kian hari semakin meningkat, oleh karena itu diperlukan analisis ulasan dengan melakukan analisis sentimen terhadap *review* Halodoc, tahapan awal dalam analisis sentimen adalah *preprocessing* yang di dalam prosesnya terdapat proses *tokenization*, *transform to lower cases, filter stopword, filter token (by length)*. Data dibagi menjadi dua kelas yaitu positif dan negatif dengan validasi *cross validation* dan nilai *k-fold validation* 10, algoritma yang digunakan adalah *Naïve Bayes Classifier* dengan hasil akurasi 81.68% dan AUC 0.756, termasuk ke dalam *fair classification.*

Kata Kunci: *Naïve Bayes Classifier, Text Mining*, Analisis Sentimen, Klasifikasi, Halodoc

1. PENDAHULUAN

Teknologi informasi semakin pesat perkembangannya, teknologi informasi memasuki berbagai bidang seperti: pendidikan, ekonomi, sosial budaya, kesehatan dan lain-lain. Memasuki era 4.0 yang dihadapkan sekarang maka tidak heran masyarakat mau tidak mau harus siap dengan perubahan atau perkembangan zaman. Pengaruh kemajuan teknologi mengantarkan masyarakat akan adanya perubahan dari analog menuju digital, kemajuan teknologi seperti televisi, telepon dan telepon genggam, bahkan internet bukan hanya dapat dinikmati oleh masyarakat kota, namun juga telah dapat dinikmati oleh masyarakat di pelosok desa. Pengguna internet di Indonesia sendiri mencapai 171,71 juta jiwa, dengan penetrasi internet sebesar 64,8% pada tahun 2018. Persentase pertumbuhan pengguna selama satu tahun mencapai 10,12%, kemudian pertumbuhan pengguna internet selama periode 2017-2018 mencapai 27,916 juta lebih dari jumlah penduduk pada tahun 2018 sebesar 264,161 juta jiwa (APJII, 2019).

Di antara berbagai sektor yang terdampak oleh era 4.0, tampaknya sektor kesehatan adalah sektor yang paling mungkin mendapatkan keuntungan dari bergabungnya sistem fisika, digital dan biologi, walaupun sektor ini mungkin juga yang paling tidak siap menerimanya. Hal ini diperkuat dari hasil survei terhadap 622 pemimpin bisnis dari berbagai industri di seluruh dunia oleh *The Economist Intelligence Unit*. Jajak pendapat terhadap para pemimpin bisnis ini menunjukkan bahwa mayoritas yang signifikan dari para eksekutif tersurvei percaya bahwa kesehatan adalah sektor yang akan mendapatkan keuntungan besar dari dampak era 4.0 ini (Tjandrawinata, 2016).

Menurut survei yang dilakukan oleh Ulya (2019), persentase keuntungan dari dampak era 4.0 untuk bidang kesehatan mencapai angka 51,06%, angka ini dinilai cukup besar dan menandakan bahwa masyarakat saat ini dalam memenuhi kebutuhan akan informasi kesehatan melalui *smartphone* masing-masing. Layanan kesehatan digital telah banyak berkembang dalam 2 tahun terakhir, sebut saja Halodoc, Klikdokter, dan beberapa layanan digital yang terintegrasi dengan lembaga kesehatan seperti Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan yang penggunanya mengalami peningkatan. Pertumbuhan pengguna yang tumbuh signifikan ini merupakan bukti layanan kesehatan digital sudah digemari masyarakat, khususnya di era disrupsi teknologi yang membuat kebiasaan hidup dan perilaku seseorang berubah. Hal tersebut juga diperkuat oleh sebuah survei yang dilakukan oleh *Deloitte* Indonesia bekerja sama dengan Bahar dan *Center for Healthcare Policy* and *Reform Studies* (*Chapters*) Indonesia. Berdasarkan survei tersebut, sekitar 84,4% pengguna layanan kesehatan digital mengaku puas dengan layanan yang ada.

Analisis sentimen adalah sebuah metode yang digunakan untuk mengekstrak data opini, memahami serta mengolah tekstual data secara otomatis untuk melihat sentimen yang terkandung dalam sebuah opini (Sari & Wibowo, 2019). Gagasan di balik analisis sentimen adalah untuk terhubung ke ribuan sumber online di internet, mengumpulkan pernyataan tentang merek atau produk, dan analisis dengan cara analisis teks sehubungan dengan sentimen. Hal tersebut menjadi wawasan baru tentang bagaimana cara mengidentifikasi perubahan sentimen dari waktu ke waktu terhadap keberhasilan pemasaran dan bagaimana penyedia layanan dapat meningkatkan reputasi produknya (Hofmann & Klinkenberg, 2013).

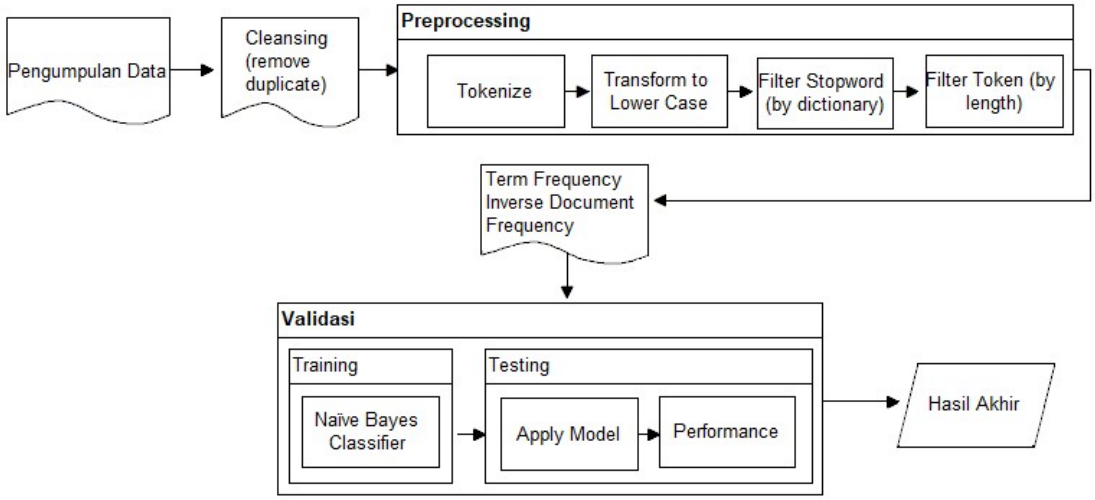
Penggunaan analisis sentimen dalam kategori *Fined-grained sentiment analysis* digunakan untuk menganalisis sesuatu atau produk untuk dapat dilihat reputasi dari suatu produk dan untuk meningkatkan kualitas produk ke depannya. Terdapat banyak ulasan pengguna yang tersedia dari ulasan aplikasi Halodoc, ada ulasan positif ada juga ulasan negatif. Semakin banyak ulasan pengguna yang tersedia maka semakin sulit pula calon pengguna untuk menyimpulkan hasil ulasan, sehingga diperlukan adanya klasifikasi ulasan untuk menyimpulkan ulasan dari aplikasi, untuk dapat membantu calon pengguna dalam mengambil keputusan dari ulasan pengguna aplikasi Halodoc. Oleh karena itu, dalam penelitian ini akan dilakukan analisis sentimen dengan menggunakan algoritma *Naïve Bayes Classifier*, algoritma *Naïve Bayes Classifier* dipilih karena *Naïve Bayes Classifier* merupakan metode klasifikasi yang efisien dan sederhana (Nugroho et al., 2020). Selain itu, algoritma *Naïve Bayes Classifier* adalah algoritma dengan performa yang sangat baik dalam beberapa kasus klasifikasi teks (Taufik, 2017). Tujuan dari penelitian ini adalah untuk mengetahui akurasi pada ulasan Halodoc dan mengetahui kinerja algoritma *Naïve Bayes Classifier* dalam melakukan klasifikasi analisis sentimen Halodoc.

2. METODE PENELITIAN

Penelitian ini mengusulkan metode *Naïve Bayes Classifier* (NBC) sebagai algoritma klasifikasi. Pertama adalah pengumpulan data yang didapat dari laman Google Play Store, kemudian tahap selanjutnya melakukan *preprocessing* dengan *tokenize* ini bertujuan membuat kalimat untuk dijadikan token-token atau pemenggalan kata-kata dengan menggunakan spasi. Kemudian dilakukan *transform to lower case,* tahap ini bertujuan untuk membuat huruf dari suatu kata menjadi seragam, setelah *transform to lower case* selanjutnya menghapus kata yang tidak memiliki relevansi tinggi dengan menggunakan *stopword* dan melakukan penghapusan kata

dalam dokumen yang tidak mencapai batas minimum dan melebihi batas maksimum, dalam penelitian ini menggunakan batas minimum 4 karakter dan batas maksimum 25 karakter.

Setelah tahap *preprocessing* tahap selanjutnya melakukan *term weighting* dengan menggunakan TF-IDF, Kemudian akan dilakukan validasi silang (*Cross Validation*) dengan *k-fold validation* untuk mendapatkan hasil *accuracy* yang maksimal. Desain model yang diusulkan akan melalui pemrosesan data *training* dan data *testing* pada tahap validasi diterapkan algoritma *Naïve Bayes Classifier* pada bagian *training* dan pada bagian *testing* memasukan operator *apply model* dan *performance* untuk menghasilkan performa algoritma yang telah diterapkan, dan yang terakhir adalah evaluasi dan validasi hasil. Berikut gambar desain penelitian pada Gambar 1.



Gambar 1. Desain Penelitian.

2.1. Pengumpulan Data

Penelitian ini menggunakan data *review* Google Play Store pada aplikasi Halodoc sebanyak 950 data sentimen. Data tersebut didapatkan dari laman *review* Google Play Store, kemudian diolah menggunakan Microsoft Excel agar data dapat diolah ke tahap selanjutnya.

2.2. *Data Cleansing (Remove Duplicate)*

Remove duplicate bertujuan untuk menghilangkan data yang sama. Sehingga dapat mencegah adanya duplikasi data dan dapat mengurangi jumlah *term*. *Remove duplicate* bekerja pada semua jenis atribut.

2.3. Naïve Bayes Classifier

Naïve Bayes Classifier merupakan algoritma *machine learning* yang sederhana dan cepat dalam hal klasifikasi, kinerja yang baik dan mudah dalam penerapannya (Fitriyani, 2018). Kaitan antara *Naïve Bayes Classifier* dengan klasifikasi, kolerasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesis dalam *teorema bayes* merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi (Prasetyo, 2012). Konsep dasar yang digunakan oleh *bayes* adalah teorema peluang bersyarat *bayes* aturan *bayes* dapat dinyatakan Samodra et al dalam (Junianto & Riana, 2017):

$$P(C_j) = \frac{P(C_j)P(D|C_j)}{P(D)} \qquad (1)$$

Di mana C_j adalah kategori teks yang akan diklasifikasikan, dan $P(C_j)$ merupakan probabilitas dari kategori teks C_j. Sedangkan *d* merupakan dokumen teks yang dapat direpresentasikan

sebagai himpunan kata (w_1, w_2, … . , w_n), dimana w_1 adalah kata pertama, w_2 adalah kata kedua dan seterusnya. Pada saat proses pengklasifikasian dokumen teks, maka pendekatan *bayes* akan menyeleksi kategori teks yang mempunyai probabilitas tinggi (C_{map}) yaitu:

$$C_{map} = \underset{Cj}{\operatorname{argmax}} \frac{P(C_j)P(D|C_j)}{P(D)} \quad (2)$$

Nilai P(D) dapat diabaikan karena nilainya adalah konstan untuk semua C_j, sehingga persamaan (2) dapat disederhanakan sebagai berikut:

$$C_{map} = \underset{Cj}{\operatorname{argmax}} P(C_j)P(D|C_j) \quad (3)$$

Probabilitas P(C_j) dapat diestimasi dengan cara menghitung jumlah dokumen *training* pada tiap-tiap kategori teks C_j. Kemudian untuk menghitung distribusi P$(D|C_j)$ sangat sulit untuk dilakukan khususnya pada proses pengklasifikasian dokumen teks yang berjumlah besar, karna jumlah *term* P$(D|C_j)$ bisa menjadi sangat besar. Hal ini dikarenakan jumlah *term* tersebut sama dengan jumlah kombinasi posisi kata dikalikan dengan jumlah kategori yang akan diklasifikasikan.

Pendekatan *Naïve Bayes* yang mengasumsikan bahwa setiap kata di dalam setiap kategori adalah tidak bergantung satu sama lain, maka perhitungan dapat lebih di sederhanakan lagi sebagai berikut:

$$P(D|C_j) = \prod i\, P(W_i|C_j) \quad (4)$$

Dengan menggunakan persamaan (4) maka persamaan (3) dapat dituliskan menjadi sebagai berikut:

$$C_{map} = \underset{Cj}{\operatorname{argmax}} P(C_j) \prod i\, P(W_i|C_j) \quad (5)$$

Nilai $P(C_j)$ dan $P(W_i|C_j)$ dihitung ketika proses *training* dijalankan yaitu:

$$P(C_j) = \frac{n(W_j)}{n(Sampel)} \quad (6)$$

$$P(W_i|C_j) = \frac{1+n_j}{|C|+n(kosakata)} \quad (7)$$

Di mana $n(W_j)$ adalah jumlah kata pada kategori j, dan n (sampel) adalah jumlah dokumen sampel yang digunakan dalam proses *training*. Sedangkan n_j adalah jumlah kemunculan dokumen kata W_j pada kategori C_j, $|C|$ adalah jumlah semua kata pada kategori C_j dan $n(kosakata)$ adalah jumlah kata yang unik pada semua data *training*.

2.4. *Preprocessing*

Tahap *preprocessing* merupakan titik awal dan tahap yang penting dalam klasifikasi (Durairaj & Ramasamy, 2016), pada tahap ini adalah tahap proses untuk mempersiapkan data sebelum masuk ke pemodelan. Tahapan ini meliputi:

2.4.1. *Tokenization*

Tahap ini adalah proses memecah kumpulan kalimat-kalimat menjadi kata, sekaligus menghilangkan simbol khusus dan tanda baca yang kemudian terbentuk suatu kumpulan kata yang bersifat unik.

2.4.2. *Transform case*

Setelah tahap mengubah kalimat menjadi pecahan kata, kemudian tahap selanjutnya mengubah semua karakter khususnya huruf akan diubah menjadi huruf non kapital. Tahap ini dimaksudkan agar data yang akan dimasukan, memiliki susunan kata dengan struktur huruf yang sama, misalnya "Dokter", "dOkter", "doKter", "dokTer", "doktEr", "dokteR", "Dokter", "doktER", dan seterusnya akan diubah menjadi kata yang sama yaitu "dokter".

2.4.3. *Filter Token (by length)*

Tahap ini adalah proses dilakukannya pemilihan token sesuai yang kita butuhkan, dalam penelitian ini membatasi token dengan ukuran minimal 4 karakter (huruf) dan batas maksimum yang digunakan 25 karakter (huruf). Contohnya kata seperti: "i", "u", "gw", "km", "sdg", "rsp" adalah kata yang kurang dari 4 huruf, selanjutnya kata semisal akan dihapus dalam tahap ini.

2.4.4. *Filter Stopword*

Tahap ini adalah tahap mengambil kata-kata penting dari hasil *token* dengan membuang kata yang kurang penting atau menyimpan kata yang penting (*wordlist*). *Stopword* biasanya bersifat kata umum yang sering muncul dalam beberapa kalimat dengan jumlah yang besar dan merupakan kata sambung, contohnya: "adalah", "ke", di", "dari", "dan". Inti dari *stopword* adalah menghapus kata-kata yang memiliki nilai informasi yang rendah atau yang tidak memiliki relevansi dengan isi dari dokumen.

2.5. TF-IDF

Setelah tahap *preprocessing* selesai, maka tahap selanjutnya adalah menerapkan perhitungan bobot menggunakan TF-IDF. Tahap *term weighting* merupakan perhitungan *term frequency,* seberapa sering masing-masing kata muncul dalam suatu dokumen. Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Metode ini akan menghitung bobot setiap token t di dokumen d dengan rumus (Maarif, 2015):

$$w = tf_{dt} \times IDF_t \qquad (8)$$

Di mana w merupakan *weight* (bobot) dokumen ke-d terhadap dokumen ke-t sama dengan banyaknya kata yang dicari dalam sebuah dokumen dikali dengan *Inverse Document Frequency* kata ke-t dari kata kunci. Nilai IDF didapatkan dari:

$$IDF(t) = log\frac{D}{df(t)} \qquad (9)$$

Di mana D merupakan total dokumen dibagi df atau banyaknya dokumen yang mengandung kata yang dicari.

2.6. Validasi

Pada tahap ini adalah tahapan untuk melakukan evaluasi menggunakan tabel *confusion matrix* untuk mengetahui apakah model yang telah diterapkan menghasilkan sesuai dengan yang diharapkan. Validasi yang dilakukan menggunakan 10 *fold validation*, yaitu membagi data secara acak ke dalam 10 bagian kemudian proses pengujian dimulai dengan model yang sudah dibangun dengan data pada bagian pertama, dan model yang terbentuk akan diuji pada 9 bagian dari sisa data. Hasil dari tahap ini adalah *precision, recall*, dan nilai *accuracy.* Menurut Bramer, 2007 dalam (Ernawati, 2016). Klasifikasi yang benar diklasifikasikan seperti pada Tabel 1.

Tabel 1. Model Confusion Matrix.

Correct Classification	*Classified As*	
	+	-
+	*TRUE POSITIVE*	*FALSE NEGATIVE*
-	*FALSE POSITIVE*	*TRUE NEGATIVE*

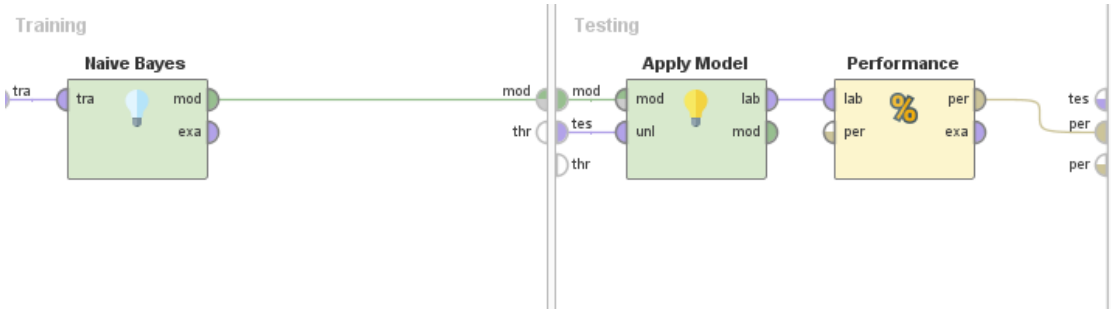
Gorunescu dikutip dalam (Fitriyani & Wahono, 2015) untuk menghitung akurasi digunakan persamaan di bawah ini:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (7)$$

TP merupakan jumlah *true positive* yang berarti dataset positif yang diklasifikasikan ke klasifikasi positif, TN adalah jumlah *true negative* atau jumlah dataset negatif yang diklasifikasikan negatif, FP adalah jumlah *false positive* atau jumlah dataset negatif yang diklasifikasikan positif dan FN adalah jumlah *false negative* atau jumlah dataset positif yang diklasifikasikan negatif.

2.6.1. *Cross Validation*

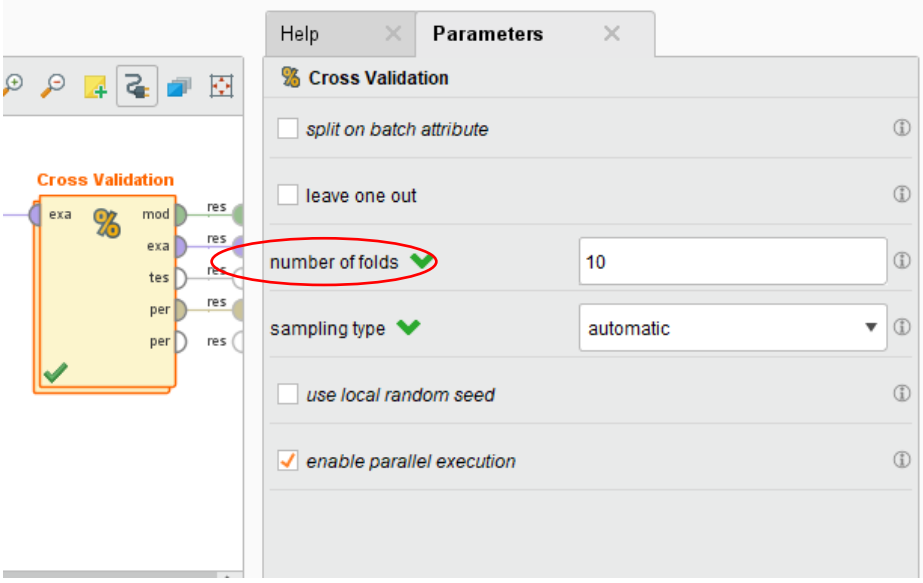
Operator proses pada tahap ini dibagi menjadi dua bagian yaitu *Training* dan *Testing*. Pada bagian *training* diterapkan algoritma yang dipakai yaitu *Naïve Bayes Classifier* sedangkan pada bagian *esting* terdiri dari operator *Apply Model* untuk menerapkan model pada *dataset* dan *Performance* yang bertugas untuk melihat performa dari model yang diterapkan seperti pada Gambar 2.



Gambar 2. Tahap Validasi.

2.6.2. *K-Fold Validation*

Validasi yang dilakukan menggunakan 10 *fold validation*, yaitu membagi data secara acak kedalam 10 bagian kemudian proses pengujian dimulai dengan model yang sudah dibangun dengan data pada bagian pertama, dan model yang terbentuk akan diuji pada 9 bagian dari sisa data.



Gambar 3. K-Fold Validation.

3. HASIL DAN PEMBAHASAN

Naïve Bayes Classifier mempunyai kelebihan, di antaranya Algoritma *Naïve Bayes Classifier* dipakai karena Algoritma *Naïve Bayes Classifier* merupakan metode klasifikasi yang efisien dan sederhana (Nugroho et al., 2020). Selain itu, algoritma *Naïve Bayes Classifier* adalah algoritma dengan performa yang sangat baik dalam beberapa kasus klasifikasi teks (Taufik, 2017). Akan tetapi, *Naïve Bayes Classifier* juga memiliki kelemahan di mana sebuah probabilitas tidak bisa mengukur seberapa besar tingkat keakuratan sebuah prediksi. Selain itu, *Naïve Bayes Classifier* juga memiliki kelemahan pada seleksi atribut sehingga dapat mempengaruhi nilai akurasi. (Muhamad et al., 2017). Pada penelitian ini Algoritma *Naïve Bayes Classifier* diterapkan untuk mengetahui akurasi dari analisis sentimen Halodoc dengan data yang digunakan sebanyak 950 data, dengan pengujian data acak 250 data dan 650 data. Dibagi menjadi dua kelas yaitu kelas positif dan kelas negatif. Sebelumnya data tersebut melalui beberapa tahapan proses, berikut penjelasan mengenai proses penelitian yang telah dilakukan, pada Tabel 1 merupakan contoh data yang telah dihasilkan.

Tabel 2. Contoh Data *Collect*.

No.	Ulasan	Kategori
1.	Aplikasi yg sangat membantu untuk kita yg awam tentang kesehatan, bisa langsung tanya ke dokter.. Harus DI INSTAL Halodoc ini !! Terimakasih Halodoc😘❤️	Positif
2.	KECEWA. RESPON YANG SANGAT LAMA DAN DOKTER YANG SANGAT JUDES😠😡😡😡😡	Negatif
3.	Aplikasi yg tdk di rekomendasikan sekali.	Negatif

3.1. Import Data

Import data adalah tahap awal untuk memproses data pada proses-proses yang terjadi pada *Rapidminer*. Data yang akan dimasukkan menggunakan *operator read excel* untuk mengimpor data yang akan dipakai, kemudian diproses pada tahap selanjutnya.

3.2. Preprocessing

Preprocessing adalah tahapan lanjutan setelah pengumpulan data, tahapan tersebut melibatkan beberapa proses, sebagai berikut:

3.2.1. *Tokenize*

Tahap ini adalah tahap di mana suatu kalimat dipecah menjadi kata-kata yang terpisah, biasanya yang menjadi acuan pemisah antar token adalah spasi atau tanda baca. Contohnya ada pada tabel 2 Ini.

Tabel 3. Tokenisasi.

No.	Ulasan	Kategori
1.	Aplikasi yg sangat membantu untuk kita yg awam tentang kesehatan bisa langsung tanya ke dokter Harus DI INSTAL Halodoc ini Terimakasih Halodoc	Positif
2.	KECEWA RESPON YANG SANGAT LAMA DAN DOKTER YANG SANGAT JUDES	Negatif
3.	Aplikasi yg tdk di rekomendasikan sekali	Negatif

3.2.2. *Transform Case*

Pada tahap ini, semua kata-kata yang didapat kemudian diubah susunan hurufnya menjadi *lower case* atau non kapital, Berikut contohnya seperti pada tabel 3.

Tabel 4. *Transform Case*.

No.	Ulasan	Kategori
1.	aplikasi yg sangat membantu untuk kita yg awam tentang kesehatan bisa langsung tanya ke dokter harus di instal halodoc ini terimakasih halodoc	Positif
2.	kecewa respon yang sangat lama dan dokter yang sangat judes	Negatif
3.	aplikasi yg tdk di rekomendasikan sekali	Negatif

3.2.3. *Filter Token (by length)*

Proses ini akan memfilter token berdasarkan batas minimum dan maksimum yang telah ditentukan yaitu batas minimum yang dipakai adalah 4 karakter dan 25 karakter untuk batas maksimum karakter. Seperti terlihat pada tabel 4 Berikut ini.

Tabel 5. *Filter Token (by length)*.

No.	Ulasan	Kategori
1.	aplikasi sangat membantu untuk kita awam tentang kesehatan bisa langsung tanya dokter harus instal halodoc terimakasih halodoc	Positif
2.	kecewa respon yang sangat lama dokter yang sangat judes	Negatif
3.	aplikasi rekomendasikan sekali	Negatif

3.2.4. *Filter Stopword*

Proses ini bertujuan untuk menghapus kata-kata yang memiliki nilai informasi yang rendah atau yang tidak memiliki relevansi dengan isi dari dokumen, Seperti terlihat pada tabel 6.

Tabel 6. Hasil *Preprocessing*.

No.	Ulasan	Kategori
1.	aplikasi membantu awam tentang kesehatan, bisa langsung tanya dokter harus instal halodoc terimakasih halodoc	Positif
2.	kecewa respon lama dokter judes	Negatif
3.	aplikasi rekomendasikan sekali	Negatif

3.3. Term weigthing TF-IDF

Setelah semua tahapan *preprocessing* selesai, maka tahap selanjutnya adalah *count vector* atau perhitungan kemunculan kata dari suatu kalimat di dalam dokumen dengan menggunakan pembobotan TF-IDF. Tahap *term weighting* adalah tahap menghitung *term frequencies*. Seberapa sering kata tersebut muncul dalam dokumen. Berikut hasil distribusi TF-IDF Seperti terlihat pada tabel 7 dan tabel 8.

Tabel 7. Hasil Distribusi *Term Frequency*.

No.	*Term*	D1	D2	D3	DF	IDF
1	Aplikasi	1	0	1	2	0.176
2	Awam	1	0	0	1	0.477
3	Bisa	1	0	0	1	0.477
4	dokter	1	1	0	2	0.176
5	halodoc	2	0	0	2	0.176
6	harus	1	0	0	1	0.477
7	instal	1	0	0	1	0.477
8	judes	0	1	0	1	0.477
9	langsung	1	0	0	1	0.477
10	rekomendasikan	0	0	1	1	0.477
11	respon	0	1	0	1	0.477
12	sekali	0	0	1	1	0.477
13	tanya	1	0	0	1	0.477
14	tentang	1	0	0	1	0.477
15	terimakasih	1	0	0	1	0.477

Tabel 8. Tabel TF-IDF.

No.	*Term*	TF*IDF		
		D1	D2	D3
1	aplikasi	0.176	0	0.176
2	awam	0.477	0	0
3	bisa	0.477	0	0
4	dokter	0.176	0.176	0
5	halodoc	0.352	0	0
6	harus	0.477	0	0
7	instal	0.477	0	0
8	judes	0	0.477	0
9	langsung	0.477	0	0
10	rekomendasikan	0	0	0.477
11	respon	0	0.477	0
12	sekali	0	0	0.477
13	tanya	0.477	0	0
14	tentang	0.477	0	0
15	terimakasih	0.477	0	0

3.4. Validasi Hasil

Setelah *preprocessing* dilakukan, selanjutnya menerapkan algoritma pada operator *cross validation* untuk validasi dan untuk mengetahui hasil akurasi dari pemodelan. Pada tahap ini dilakukan perbandingan dengan pengujian data acak. Berikut hasil yang didapat dari pemodelan terlihat pada Tabel 9 dan Tabel 10.

Tabel 9. Perbandingan Hasil Akurasi.

Hasil	250 data	650 data	950 data
Akurasi	81.60%	81.65%	81.68%
AUC	0.688	0.746	0.756

Data acak yang digunakan untuk uji dan perbandingan yaitu 250 data dan 650 data, 250 data dengan proporsi 125 sentimen positif dan 125 sentimen negatif. 650 data dengan proporsi 325 sentimen positif dan 325 sentimen negatif.

Tabel 10. *Confusion Matrix*.

	True Positif	***True Negatif***	***Class Precision***
Pred.Positif	387	86	81.82%
Pred.Negatif	88	389	81.55%
Class Recall	81.47%	81.89%	
Accuracy		81.68% +/- 2.54%	
Micro Average		81.68%	

Dengan rincian perhitungan akurasi sebagai berikut:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

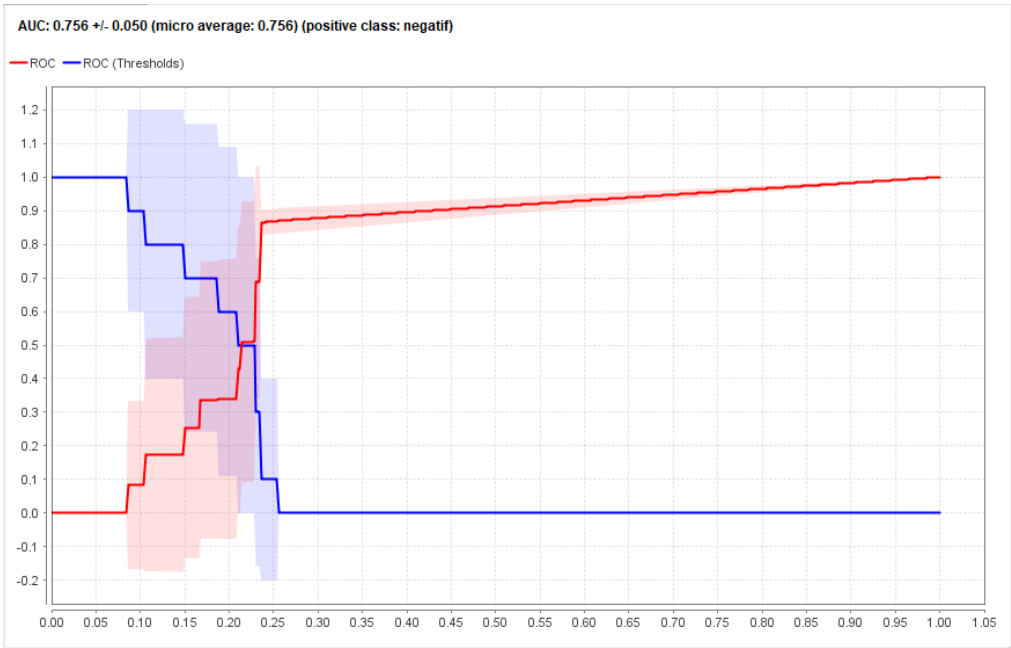
$$accuracy = \frac{387 + 389}{387 + 389 + 88 + 86} \times 100\%$$

$$accuracy = 81.68\%$$

Dari perbandingan hasil akurasi dan AUC yang terlihat pada tabel 10 untuk *Naïve Bayes Classifier*, menunjukkan bahwa pada penelitian ini dengan 950 data memperoleh hasil tertinggi dengan tingkat akurasi sebesar 81.68% dan AUC 0.756.

3.4.1. Kurva ROC

Kurva ROC adalah salah satu cara untuk mengevaluasi hasil akurasi dan klasifikasi dengan visual (Mubarok et al., 2019). Seperti terlihat pada gambar 6, hasil tertinggi menunjukkan nilai AUC sebesar 0.756, dengan artian diagnosa hasil klasifikasi termasuk kedalam *fair classification*.



(sumber: Data Olahan, 2020)

Gambar 4. Kurva ROC.

4. KESIMPULAN

Dari hasil evaluasi dan validasi dapat ditarik kesimpulan bahwa:

1) *Naïve Bayes Classifier* terbukti dapat digunakan untuk pengklasifikasian sentimen *review* Halodoc, hasil tertinggi diperoleh dengan menggunakan jumlah data 950 sentimen. *Naïve Bayes Classifier* menghasilkan AUC 0.756 dalam artian bahwa pengklasifikasian pada penelitian ini termasuk ke dalam kategori *fair classification.*
2) Algoritma *Naïve Bayes Classifier* yang telah diterapkan untuk analisis sentimen Halodoc dari 475 sentimen positif terdapat 387 sentimen yang tepat di kategorikan sebagai sentimen positif dan sisanya 88 masuk ke dalam kategori sentimen negatif, kemudian dari 475 sentimen negatif 389 sentimen yang dikategorikan tepat sebagai sentimen negatif dan sisanya yaitu 86 sentimen yang dikategorikan sebagai sentimen positif, dengan nilai akurasi 81.68%.
3) *Preprocessing* diperlukan untuk mengurangi *term* yang dinilai tidak diperlukan pada saat *count vector.*

Dikarenakan penelitian ini terbatas pada algoritma yang dipakai yaitu *Naïve Bayes Classifier,* dimungkinkan akurasi untuk klasifikasi sentimen *review* Halodoc masih dapat ditingkatkan.

1) Beberapa saran untuk penelitian selanjutnya sebagai berikut:
2) Peningkatan akurasi dengan seleksi fitur untuk optimasi, misalnya *prune methode*, *genetic algorithm*, dan lain-lain.
3) Kemudian pada penelitian selanjutnya dapat digunakan algoritma lain selain *Naïve Bayes Classifier* untuk klasifikasi teks.
4) Penelitian *text mining* lain selain analisis sentimen dengan data yang lebih banyak atau sedikit, dan pengimplementasian analisis sentimen ke dalam program atau aplikasi.

DAFTAR PUSTAKA

APJII. (2019). *Penetrasi & Profil Perilaku Pengguna Internet Indonesia Tahun 2018*.

Durairaj, M., & Ramasamy, N. (2016). A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate. *International Journal of Control Theory and Applications*, *9*(27), 255–260.

Ernawati, S. (2016). Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Analisis Sentimen Review Perusahaan Penjualan Online Menggunakan Naïve Bayes. *Evolusi: Jurnal Sains dan Manajemen*, *4*(1), 45–54. https://doi.org/10.31294/evolusi.v4i1.605

Fitriyani, F. (2018). Metode Bagging Untuk Imbalance Class Pada Bedah Toraks Menggunakan Naive Bayes. *Jurnal Kajian Ilmiah*, *18*(3), 278. https://doi.org/10.31599/jki.v18i3.281

Fitriyani, F., & Wahono, R. S. (2015). Integrasi Bagging dan Greedy Forward Selection pada Prediksi Cacat Software dengan Menggunakan Naïve Bayes. *Journal of Software Engineering*, *1*(2), 101–108.

Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman and Hall/CRC.

Junianto, E., & Riana, D. (2017). Penerapan PSO Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan NBC. *Jurnal Informatika*, *4*(1), 38–45. https://doi.org/10.31294/ji.v4i1.1810

Maarif, A. A. (2015). *Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah*. Universitas Dian Nuswantoro Semarang.

Mubarok, A., Susanti, S., & Handayani, R. N. (2019). *Optimasi Algoritma Support Vector Machine Menggunakan Particle Swarm Optimization Untuk Analisis Sentimen pada Ulasan Produk Tokopedia*. Universitas Adhirajasa Reswara Sanjaya.

Muhamad, H., Prasojo, C. A., Sugianto, N. A., Surtiningsih, L., & Cholissodin, I. (2017). Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *4*(3), 180. https://doi.org/10.25126/jtiik.201743251

Nugroho, K. S., Istiadi, I., & Marisa, F. (2020). Naive Bayes classifier optimization for text classification on e-government using particle swarm optimization. *Jurnal Teknologi dan Sistem Komputer*, *8*(1), 21–26. https://doi.org/10.14710/jtsiskom.8.1.2020.21-26

Prasetyo, E. (2012). *Data Mining : Konsep Dan Aplikasi Menggunakan Matlab*. ANDI.

Samodra, J., Sumpeno, S., & Hariadi, M. (2019). Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naïve Bayes. *SEMINAR NASIONAL ELECTRICAL, INFORMATICS, AND IT'S EDUCATIONS 2019*, 71–74.

Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, *10*(2), 681–686. https://doi.org/10.24176/simet.v10i2.3487

Taufik, A. (2017). Optimasi Particle Swarm Optimization Sebagai Seleksi Fitur Pada Analisis Sentimen Review Hotel Berbahasa Indonesia Menggunakan Algoritma Naive Bayes. *Jurnal Teknik Komputer*, *3*(2), 40–47. https://doi.org/10.31294/jtk.v3i2.1922

Tjandrawinata, R. R. (2016). *Industri 4.0: revolusi industri abad ini dan pengaruhnya pada bidang kesehatan dan bioteknologi*. https://doi.org/10.5281/zenodo.49404

Ulya, F. N. (2019, Agustus 19). *Survei: 84,4 Persen Masyarakat Puas dengan Layanan Kesehatan Digital*. Kompas.com. https://money.kompas.com/read/2019/08/19/134000926/survei--84-4-persen-masyarakat-puas-dengan-layanan-kesehatan-digital?page=all

JISKa, Vol. 6, No. 2, MEI, 2021, Pp. 90 – 97
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Prediksi Barang Keluar TB. Wijaya Bangunan Menggunakan Algoritma *KNN Regression* dengan RStudio

Natcha Kwintarini Suparman [(1)*], Budi Arif Dermawan [(2)], Tesa Nur Padilah [(3)]
Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa, Karawang
e-mail : natcha.16159@student.unsika.ac.id, {budi.arif,tesa.nurpadilah}@staff.unsika.aci.id.
* Penulis korespondensi.
Artikel ini diajukan 19 Agustus 2020, direvisi 25 September 2020, diterima 13 Oktober 2020, dan dipublikasikan 3 Mei 2021.

Abstract

TB. Wijaya Bangunan is a business entity that has weaknesses in managing inventories. This study aims to help TB. Wijaya Bangunan in managing inventory based on existing data reduce the difference between the number of incoming goods and the number of outgoing goods. The methods used are data collection, data preparation, data selection, preprocessing, data transformation, distance calculation, calculation of predictions, evaluation, and display of prediction results using a Shiny framework. This study uses the Time Series KNN Regression algorithm to predict the number of outgoing goods based on time series data with existing data. The most predicted results came out in the 9th week period as much as 22.40%. Based on the process that has been done, it can be concluded that the evaluation value of Root Mean Square Error (RMSE) is at least 3.55, which means it has the best predictive accuracy results.

Keywords: Time Series, Predictions, K-Nearest Neighbor, RStudio, Shiny Framework

Abstrak

TB. Wijaya Bangunan merupakan badan usaha yang memiliki kelemahan dalam mengelola persediaan, sehingga menyebabkan barang masuk tidak sesuai dengan barang keluar. Penelitian ini bertujuan untuk membantu TB. Wijaya Bangunan dalam mengelola persediaan berdasarkan data yang sudah terjadi sebelumnya, agar selisih jumlah barang masuk dan jumlah barang keluar dapat diperkecil. Data yang digunakan merupakan data *time series* penjualan dari bulan Januari sampai Desember 2019. Metode yang dilakukan berupa pengumpulan data, persiapan data, seleksi data, *preprocessing*, transformasi data, perhitungan jarak, perhitungan prediksi, evaluasi, serta tampilan hasil prediksi dengan Shiny *framework.* Penelitian ini menggunakan algoritma *Time Series KNN Regression* untuk memprediksi jumlah barang keluar bedasarkan data deret waktu dengan data yang telah terjadi sebelumnya. Hasil prediksi paling banyak keluar terdapat pada periode minggu ke-9 sebanyak 22.40%. Penerapan algoritma dilakukan menggunakan *software RStudio*. Berdasarkan proses yang telah dilakukan dapat ditarik kesimpulan, bahwa hasil penelitian dengan algoritma *Time Series KNN Regression* menghasilkan nilai evaluasi *Root Mean Square Error* (RMSE) paling kecil 3.55 yang berarti memiliki hasil akurasi prediksi terbaik.

Kata Kunci: *Time Series*, Prediksi, *K-Nearest Neighbor*, RStudio, Shiny *Framework*

1. PENDAHULUAN

Usaha merupakan segala kegiatan ekonomi yang dilakukan oleh manusia dengan mengerahkan tenaga, pikiran, atau badan untuk mencapai suatu maksud dalam rangka mencapai kesejahteraan atau kemakmuran, sehingga dapat membantu meningkatkan taraf hidup seseorang menjadi lebih baik (Indani & Suhairi, 2018). Salah satu usaha yang berkembang pada saat ini, dengan kenaikannya jumlah penduduk yang semakin bertambah serta pembangunan yang semakin meningkat maka usaha properti menjadi usaha yang menjanjikan untuk memenuhi kebutuhan masyarakat akan tempat tinggal (Wijaya & Ananta, 2017).

TB. Wijaya Bangunan merupakan badan usaha yang bergerak di bidang penjualan bahan bangunan di Karawang yang memerlukan adanya pengelolaan yang baik dalam persediaan bahan bangunan. Persediaan bahan bangunan merupakan hal yang paling diperlukan oleh toko bangunan untuk kelancaran pemenuhan permintaan konsumen. Persediaan tersebut perlu

dikelola karena biasanya jumlah barang masuk tidak sesuai dengan jumlah barang keluar, sehingga menyebabkan adanya penumpukan barang. Oleh karena itu, agar selisih jumlah barang masuk dan jumlah barang keluar dapat diperkecil, maka diperlukan adanya prediksi. Tujuan dari penelitian ini untuk membantu TB. Wijaya Bangunan dalam memprediksi barang yang sering keluar, guna mempersiapkan ketersediaan barang di masa mendatang untuk memenuhi persediaan barang di TB. Wijaya Bangunan.

Prediksi memiliki arti yang sama dengan ramalan atau perkiraan, prediksi dapat terjadi berdasarkan metode ilmiah atau bahkan subjektif belaka (Hamdi et al., 2019). Prediksi merupakan suatu usaha untuk meramalkan keadaan di masa mendatang melalui pengujian keadaan di masa lalu (Lestari et al., 2019). *Data mining* merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat serta pengetahuan yang terkait dari berbagai database besar (Nofriansyah, 2014). *Data mining* diperlukan dalam melakukan prediksi untuk menemukan hubungan yang memiliki arti, pola, dan kecenderungan dengan memeriksa sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti statistik dan matematik (Mahena et al., 2015).

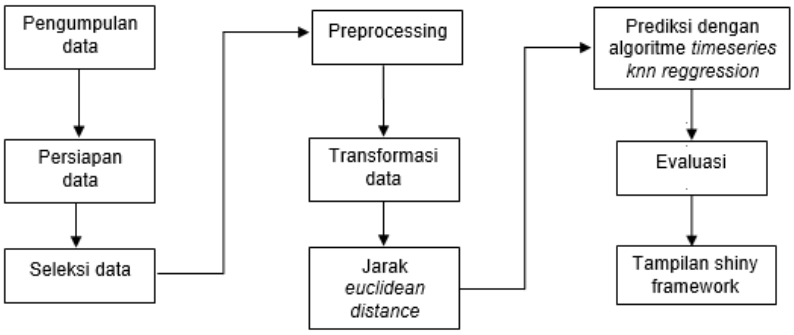
Penelitian yang dilakukan oleh (Sabilla & Putri, 2017), mengenai prediksi ketepatan waktu lulus mahasiswa dengan menggunakan algoritma *K-Nearest Neighbor* dan Naïve Bayes Classifier menghasilkan pengetahuan bahwa algoritma *K-Nearest Neighbor* lebih efektif digunakan untuk melakukan prediksi. Begitu pula penelitian yang dilakukan oleh Putra & Putra (2018), mengenai klasifikasi harga ponsel menggunakan metode *K-Nearest Neighbor* (KNN), bertujuan untuk memprediksi harga ponsel berdasarkan banyaknya fitur. Namun penelitian tersebut belum menggunakan proses pembersihan *missing value*. Pada penelitian prediksi barang keluar pada TB. Wijaya Bangunan peneliti melakukan proses pembersihan *missing value*.

Penelitian yang telah dilakukan oleh Hamdi et al. (2019), memanfaatkan data *time series* untuk prediksi jumlah penjualan minuman menggunakan metode *K-Nearest Neighbor*. Penelitian ini bertujuan untuk mengetahui jumlah pesanan varian minuman yang paling diminati oleh konsumen. Namun penelitian tersebut belum menggunakan proses *cleaning* pada *dataset* sehingga nilai *NA* atau nilai kosong mempengaruhi proses *data mining*. Pada penelitian prediksi barang keluar pada TB. Wijaya Bangunan peneliti melakukan proses *cleaning* agar hasil prediksi lebih maksimal.

Penelitian yang dilakukan oleh Altunsöğüt et al. (2018), mengenai prediksi jumlah sampah pabrik pewarna *textile* dengan enam algoritma J48, Naïve Bayes, KNN, SVM, Regresi Logistik, dan Multi-Layer Perceptron, bertujuan untuk memprediksi jumlah cacat produksi pada kain agar tidak berlebih. Namun pada penelitian tersebut tidak dilakukan perhitungan akurasi untuk hasil prediksi, sehingga peneliti tidak mengetahui apakah hasil prediksi maksimal atau tidak. Pada penelitian prediksi barang keluar pada TB. Wijaya Bangunan dilakukan proses perhitungan akurasi dengan menggunakan metode RMSE (*Root Mean Square Error*). RMSE (Root Means Square Error) merupakan metode yang digunakan dalam mengevaluasi hasil prediksi untuk mengetahui keakuratan hasil peramalan yang telah dilakukan terhadap data yang sebenarnya (Fatkhuroji et al., 2019). Metode RMSE (*Root Mean Square Error*) menghasilkan akurasi yang baik untuk melakukan estimasi besarnya kesalahan pengukuran yang dihasilkan. Nilai RMSE yang kecil memberi petunjuk bahwa nilai yang dihasilkan mendekati nilai observasinya (Sartika, 2019).

Berdasarkan penelitian terdahulu, untuk membantu permasalahan yang dialami, maka peneliti menggunakan metode *time series* dengan algoritma *K-Nearest Neighbor Regression* di mana perbedaanya terletak pada objek penelitian yang diteliti serta langkah dalam melakukan prediksi dan tampilan visualisasi dengan Shiny *framework*. Pada penelitian ini dilakukan proses *cleaning* data untuk mempermudah saat proses *data mining* agar hasil prediksi lebih maksimal, adanya perhitungan akurasi dari hasil prediksi pada seluruh *item*, dan visualisasi hasil *input/output* dengan Shiny *framework* dengan tampilan yang mudah dipahami.

2. METODE PENELITIAN



Gambar 1. Metode Penelitian.

Dalam penelitian ini ada beberapa langkah yang peneliti lakukan, langkah-langkah tersebut dapat dilihat pada Gambar 1, tahap awal pada penelitian ini adalah mengumpulkan data yang diperoleh dari TB. Wijaya Bangunan. Data yang diperoleh berupa data pada tahun 2019 yang dimulai sejak tanggal 1 Januari sampai 31 Desember dalam bentuk *excel*, data ini kemudian diubah menjadi format ".csv" untuk memudahkan dalam pengolahan data menggunakan *software* RStudio.

Tahapan selanjutnya merupakan tahapan persiapan data dengan melakukan beberapa langkah seperti identifikasi masalah, serta melakukan studi literatur sebagai bahan perbandingan dengan penelitian lain. Tahap ketiga merupakan tahap seleksi data, pada tahap ini peneliti melakukan pemahaman terhadap data yang telah diberikan mengenai kebutuhan untuk penelitian lebih lanjut kemudian dilakukan pemilihan atribut yang digunakan untuk melakukan penelitian selanjutnya.

Tahap keempat merupakan tahap *preprocessing*. Pada tahap ini peneliti melakukan pembersihan data agar data yang diolah benar-benar relevan. Data dapat dikatakan tidak relevan apabila ada atribut dalam *dataset* yang kosong atau tidak terisi nilai. Tahap ini, peneliti melakukan *cleaning* data dengan mengubah nilai *NA* atau *dataset* kosong menjadi nol. Tahap kelima merupakan tahap transformasi data, tahap ini dilakukan untuk mempermudah prediksi perbarang dalam kurun waktu perminggu. Tahap ini merupakan proses perubahan tipe data dengan merubah format data yang sebelumnya didapat dari TB. Wijaya Bangunan, data harus dirubah berdasarkan data *time series* dengan format tanggal, bulan, dan tahun yang disusun secara vertikal agar dapat dengan mudah diproses pada RStudio, pada tahap ini juga dilakukan pembagian data menjadi data *training* dan data *testing*.

Tahap keenam merupakan tahap penentuan jarak terdekat antar *data training* dengan menggunakan metode *Euclidean distance* (1).

$$d = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \qquad (1)$$

Di mana x_1 merupakan sampel data yang akan digunakan untuk menghitung jarak, x_2 merupakan data *testing*, variabel i merupakan variabel data, sedangkan variabel d merupakan variabel jarak, dan p merupakan dimensi data (Mustakim & Oktaviani, 2016).

Tahap ketujuh merupakan tahapan perhitungan prediksi dengan menerapkan algoritma *time series K-Nearest Neighbor Regression* pada RStudio. Tahap ini diolah dengan menggunakan *package* "tsfknn" yang ada pada Rstudio. Fungsi dari *package* ini untuk melakukan prediksi dengan data yang berbentuk *time series* menggunakan penggabungan antara algoritma *K-Nearest Neighbor* dengan *regression* hanya dengan satu *function* atau kode-kode yang disusun untuk melakukan suatu tugas dengan menggabungkan beberapa perintah dalam satu kode pemrograman pada RStudio (Martínez et al., 2019).

Tahapan kedelapan merupakan tahapan untuk menentukan hasil prediksi mana yang lebih baik. Untuk penentuan ini peneliti menggunakan metode evaluasi RMSE (*Root Mean Square Error*), yang di mana pada metode ini dilakukan pengukuran kesalahan dengan mengukur nilai rata-rata kesalahan prediksi dengan mengkuadratkan nilai kesalahan dan mencari nilai akarnya seperti pada Pers. 2.

$$RMSE = \sqrt{\frac{\sum_{t-1}^{n}(x_t - F_t)^2}{n}} \qquad (2)$$

Di mana x_t merupakan nilai aktual pada periode ke-t, $x_t - F_t$ merupakan nilai kesalahan (*error*) pada periode ke-t, dan n merupakan jumlah data (Bode, 2017).

Tahapan terakhir merupakan tahapan *shiny framework,* pada tahap ini dilakukan penerapan visualisasi *input*/*ouput* dengan *shiny framework* untuk menampilkan hasil prediksi serta plot dari hasil prediksi barang keluar.

3. HASIL DAN PEMBAHASAN

Data yang peneliti dapatkan berupa data selama 1 tahun pada 2019. Dalam penelitian ini, peneliti menggunakan data bahan bangunan yang dapat keluar persatuan yang kemudian dilakukan proses seleksi data dengan memilih atribut yang digunakan untuk proses selanjutnya. Adapun atribut yang digunakan dalam penentuan prediksi diantaranya nama *item*, dan tanggal keluar yang terdiri dari data *time series* berupa data harian.

3.1. *Preprocessing*

Pada tahap ini peneliti melakukan data *cleaning* agar data yang diolah benar-benar relevan dengan mengganti nilai *NA* menjadi nol menggunakan *software* RStudio, yang di mana pada proses ini menggunakan kode pemrograman “is.na” pada RStudio yang berfungsi untuk mengganti nilai *NA* atau nilai kosong menjadi nol. Pada Tabel 1 merupakan data yang belum dilakukan *cleaning*, sedangkan pada Tabel 2 merupakan data yang telah selesai dilakukan *cleaning* di mana nilai *NA* sudah berubah menjadi nol.

Tabel 1. Data Awal.

No.	Nama Item	Tanggal Keluar									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**…**
1	Afur BCP	7	6	4	2	1	*NA*	4	*NA*	*NA*	…
2	BCP 1 Lubang	2	*NA*	1	*NA*	3	2	*NA*	*NA*	8	…
3	BCP Royal	1	*NA*	*NA*	*NA*	2	*NA*	*NA*	*NA*	*NA*	…
4	Cat YOKO	10	2	*NA*	*NA*	2	*NA*	4	6	*NA*	…
5	Closet Jongkok	3	*NA*	3	*NA*	1	*NA*	*NA*	2	*NA*	…
6	DOP Rucika	4	*NA*	2	1	3	4	5	1	4	…
7	Engsel Arnita	1	*NA*	4	*NA*	4	*NA*	*NA*	5	3	…
8	Engsel Ferza	6	*NA*	4	1	2	*NA*	3	*NA*	4	…
…	…	…	…	…	…	…	…	…	…	…	…

Tabel 2. Data Setelah Proses *Cleaning*.

No.	Nama Item	Tanggal Keluar									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**…**
1	Afur BCP	7	6	4	2	1	0	4	0	0	…
2	BCP 1 Lubang	2	0	1	0	3	2	0	0	8	…
3	BCP Royal	1	0	0	0	2	0	0	0	0	…
4	Cat YOKO	10	2	0	0	2	0	4	6	0	…
5	Closet Jongkok	3	0	3	0	1	0	0	2	0	…
6	DOP Rucika	4	0	2	1	3	4	5	1	4	…

7	Engsel Arnita	1	0	4	0	4	0	0	5	3	…
8	Engsel Ferza	6	0	4	1	2	0	3	0	4	…
…	…	…	…	…	…	…	…	…	…	…	…

3.2. Transformasi Data

Tahap ini dilakukan proses perubahan format data, data harus diubah berdasarkan data *time series* dengan format tanggal, bulan, dan tahun yang disusun secara vertikal, kemudian dilakukan transformasi bentuk data keluar per-hari menjadi data keluar per minggu, serta pembagian data *training* dan data *testing*. Hasil dari transformasi data dapat dilihat pada Tabel 3 di mana data telah berubah menjadi data keluar per minggu dari setiap bahan bangunan, sedangkan hasil untuk pembagian data *training* dan data *testing* dapat dilihat pada Tabel 4 di mana data dibagi secara manual dengan data *training* yang digunakan berupa data dari bulan Januari sampai Oktober, sedangkan data *testing* yang digunakan berupa data pada bulan November sampai Desember.

Tabel 3. Tabel Hasil Transformasi Per Minggu.

Minggu ke	Afur BCP	BCP 1 Lubang	BCP Royal	Cat YOKO	Closet Jongkok	DOP Rucika	Engsel Arnita	Engsel Ferza	…
1	20	8	3	14	7	14	9	13	…
2	17	10	5	28	7	19	14	14	…
3	12	4	8	20	20	19	15	8	…
4	3	10	2	15	9	13	22	10	…
5	0	10	10	19	17	10	8	11	…
…	…	…	…	…	…	…	…	…	…

Tabel 4. Tabel Data *Training* dan Data *Testing*.

Data *Training*

Afur BCP	BCP 1 Lubang	BCP Royal	Cat YOKO	Closet Jongkok	DOP Rucika	Engsel Arnita	Engsel Ferza	…
20	8	3	14	7	14	9	13	…
17	10	5	28	7	19	14	14	…
12	4	8	20	20	19	15	8	…
3	10	2	15	9	13	22	10	…
0	10	10	19	17	10	8	11	…
…	…	…	…	…	…	…	…	…

Data *Testing*

Afur BCP	BCP 1 Lubang	BCP Royal	Cat YOKO	Closet Jongkok	DOP Rucika	Engsel Arnita	Engsel Ferza	…
9	0	0	13	7	7	17	14	…
0	0	16	9	11	7	2	9	…
32	7	1	6	9	4	17	18	…
12	5	13	6	10	10	20	0	…
11	10	2	16	18	6	1	3	…
7	25	11	29	26	5	12	8	…
9	15	0	15	8	7	3	9	…
3	0	6	6	5	0	6	0	…

3.3. Jarak *Euclidean Distance*

Pada tahap ini dilakukan proses perhitungan jarak perbarang pada data *training* menggunakan rumus jarak *Euclidean distance.* Berikut merupakan hasil yang telah diurutkan berdasarkan jarak terdekat.

$d_{1,31}$ = 8.37, $d_{1,2}$ = 11.14, $d_{1,9}$ = 11.83, $d_{1,23}$ = 12.08, $d_{1,12}$ = 12.49, $d_{1,13}$ = 12.69, $d_{1,44}$ = 12.77, $d_{1,24}$ = 12.88, $d_{1,29}$ = 13.60, $d_{1,25}$ = 13.75, $d_{1,11}$ = 13.86, $d_{1,40}$ = 13.96, $d_{1,8}$ = 15.10, $d_{1,30}$ = 15.20, $d_{1,22}$ =

15.56, $d_{1,21}$ = 15.81, $d_{1,7}$ = 16.12, $d_{1,39}$ = 16.16, $d_{1,43}$ = 16.43, $d_{1,10}$ = 16.85, $d_{1,32}$ = 16.91, $d_{1,14}$ = 17.83, $d_{1,35}$ = 17.94, $d_{1,42}$ = 20.25, $d_{1,6}$ = 20.88, $d_{1,3}$ = 20.98, $d_{1,27}$ = 21.10, $d_{1,34}$ = 21.40, $d_{1,5}$ = 22.72, $d_{1,17}$ = 22.85, $d_{1,33}$ = 23.37, $d_{1,18}$ = 23.98, $d_{1,28}$ = 24.06, $d_{1,20}$ = 24.39, $d_{1,4}$ = 24.74, $d_{1,26}$ = 25.28, $d_{1,38}$ = 25.42, $d_{1,41}$ = 28.16, $d_{1,19}$ = 30.59, $d_{1,16}$ = 31.70, $d_{1,37}$ = 33.38, $d_{1,45}$ = 34.21, $d_{1,36}$ = 35.01, $d_{1,15}$ = 38.77.

Pada hasil perhitungan di atas, didapatkan nilai jarak terdekat dari seluruh data *training* yang telah dihitung dan dapat ditarik kesimpulan bahwa jarak yang paling dekat terletak pada data 1 terhadap data 31 dengan nilai sebesar 8.37.

3.4. Hasil Prediksi dengan Algoritma *K-Nearest Neighbor Regression*

Pada proses prediksi peneliti menggunakan nilai k yang telah ditentukan dengan nilai k=5. Pada Tabel 5 merupakan hasil prediksi selama 9 periode perminggu ditahun 2020, hasil prediksi paling tinggi pada periode minggu ke-1 dengan jumlah prediksi sebanyak 22%, pada periode minggu ke-2 paling tinggi didapat dengan jumlah prediksi sebanyak 22.20%, periode minggu ke-3 paling tinggi didapat dengan jumlah prediksi sebanyak 17.60%, periode minggu ke-4 paling tinggi didapat dengan jumlah prediksi sebanyak 15.20%, periode minggu ke-5 paling tinggi didapat dengan jumlah prediksi sebanyak 20.40%, periode minggu ke-6 paling tinggi didapat dengan jumlah prediksi sebanyak 19.40%, periode minggu ke-7 paling tinggi didapat dengan jumlah prediksi sebanyak 21.20%, periode minggu ke-8 paling tinggi dengan jumlah prediksi sebanyak 21%, sedangkan periode minggu ke-9 paling tinggi dengan jumlah prediksi sebanyak 22.40%.

Tabel 5. Hasil Prediksi.

Minggu ke	Afur BCP	BCP 1 Lubang	BCP Royal	Cat YOKO	Closet Jongkok	DOP Rucika	Engsel Arnita	Engsel Ferza	…
1	10	7.4	4	22	10.2	8.6	4	8.6	…
2	11.6	8	9.8	19.8	5.2	11.2	4.8	11.2	…
3	10	8	9	10.2	8.4	12	6.6	14.4	…
4	13.4	10	12.2	4	12.4	12.4	14	7.4	…
5	9.6	9.2	13.2	3.4	9	10.6	15.4	9.6	…
6	5.2	9	14.4	6.4	10.8	11.4	9	11.2	…
7	11.4	9.2	15.8	19.2	10.4	13	8.4	8.6	…
8	11.6	11.8	13.8	18.2	13.2	13.6	4.4	5.4	…
9	15	9.6	14	10	12.6	13.8	3.4	9	…

3.5. Evaluasi

Pada tahap ini hasil prediksi yang telah didapat kemudian dievaluasi menggunakan metode RMSE (*Root Mean Square Error*) untuk mendapatkan hasil akurasi yang terbaik, dengan kriteria semakin kecil nilai *error* maka tingkat akurasi semakin baik (Nanja & Purwanto, 2015). Nilai evaluasi didapat dari mengkuadratkan nilai *error* (nilai aktual-nilai hasil prediksi) dibagi dengan jumlah data sehingga menghasilkan nilai rata-rata yang kemudian diakarkan. Hasil untuk evaluasi dapat dilihat pada Tabel 6 hasil tingkat akurasi terkecil yang didapat bernilai 3,56 yang berarti memiliki hasil akurasi prediksi terbaik.

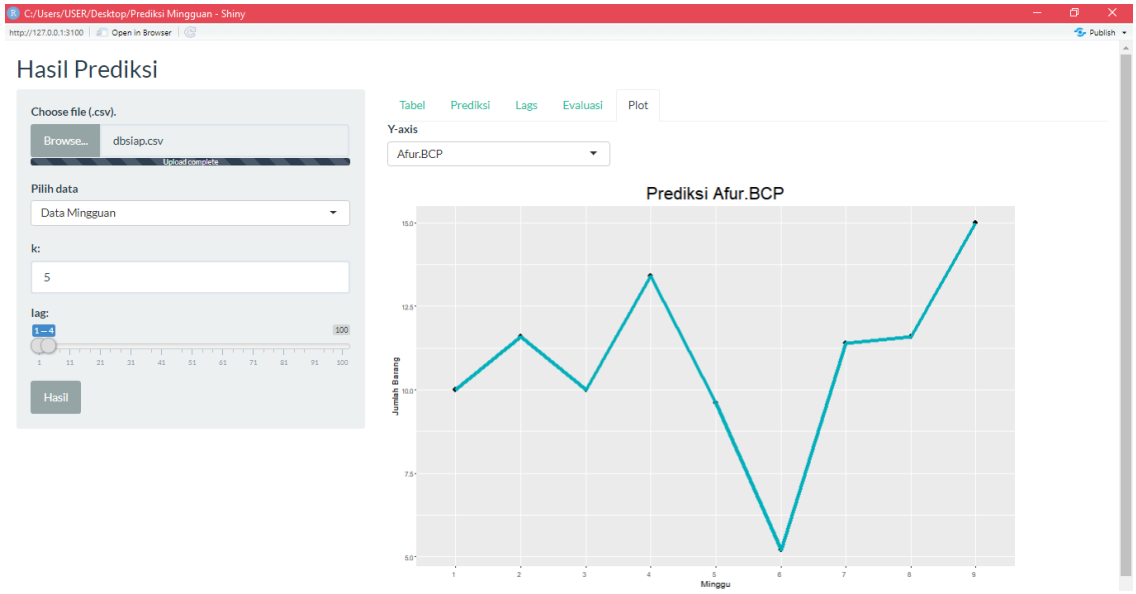
Tabel 6. Hasil Evaluasi.

Nama Bahan Bangunan	RMSE (*Root Means Square* Error)
Afur BCP	8.58
BCP 1 Lubang	7.72
BCP Royal	8.93
Cat YOKO	6.27
Closet Jongkok	6.69
DOP Rucika	7.56
Engsel Arnita	6.23
Engsel Ferza	6.70

… …

3.6. Penerapan Shiny *Framework*

Data yang telah selesai dilakukan prediksi, selanjutnya ditampilkan dalam bentuk visualisasi dari hasil prediksi dengan menggunakan Shiny *framework*. Hasil untuk tampilan plot pada Shiny *framework* dapat dilihat pada Gambar 2.



Gambar 2. Tampilan Plot.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilaksanakan maka didapat kesimpulan hasil prediksi penjualan bahan bangunan yang paling sering keluar selama 9 periode perminggu di tahun 2020 sebanyak 5 produk bahan bangunan dari 60 produk bahan bangunan, yaitu cat yoko, mangkok listrik, kunci kaca faster, lis 20cm, dan klem selang. Serta ditampilkan hasil visualisasi berupa grafik plot hasil prediksi yang telah didapat dengan menggunakan Shiny *framework* untuk mempermudah dalam menampilkan hasil prediksi. Hasil evaluasi perhitungan tingkat akurasi dengan menggunakan metode RMSE (*Root Means Square Error*) dengan hasil bahwa Engsel Sendok Cobra Sherlock mendapat hasil evaluasi paling kecil dengan nilai *error* 3.55 yang berarti memiliki hasil akurasi terbaik. *Package* "tsfknn" pada RStudio juga cukup sederhana digunakan untuk melakukan prediksi dengan algoritma *time series k-nearest neighbor regression* karena *package* ini cukup menggunakan satu *function* atau kode-kode yang disusun untuk melakukan suatu tugas dengan menggabungkan beberapa perintah dalam satu kode pemrograman pada RStudio.

DAFTAR PUSTAKA

Altunsöğüt, Ö., Uçar, E., & Kılıçaslan, Y. (2018). PREDICTING THE AMOUNT OF WASTAGE OF FINISHED GOODS IN TEXTILE DYEING FACTORIES. *International Scientific Conference "UNITECH 2018,"* 248–288.

Bode, A. (2017). K-NEAREST NEIGHBOR DENGAN FEATURE SELECTION MENGGUNAKAN BACKWARD ELIMINATION UNTUK PREDIKSI HARGA KOMODITI KOPI ARABIKA. *ILKOM Jurnal Ilmiah*, *9*(2), 188–195. https://doi.org/10.33096/ilkom.v9i2.139.188-195

Fatkhuroji, F., Santosa, S., & Pramunendar, R. A. (2019). PREDIKSI HARGA KEDELAI LOKAL DAN KEDELAI IMPOR DENGAN METODE SUPPORT VECTOR MACHINE BERBASIS FORWARD SELECTION. *Jurnal Teknologi Informasi*, *15*(1), 61–76.

Hamdi, A., Indriani, F., & Muliadi, M. (2019). METODE TIMESERIES K-NEAREST NEIGHBOR

REGRESSION DALAM PREDIKSI BARANG KELUAR PADA GUDANG PT PUTRA PRENUER BANJARBARU. *Seminar Nasional Ilmu Komputer (SOLITER)*, *2*, 37–45.

Indani, & Suhairi, L. (2018). *Pengelolaan Usaha Boga Edisi II* (2 ed.). Syiah Kuala University Press.

Lestari, S. I. P., Andriani, M., GS, A. D., Subekti, P., & Kurniawati, R. (2019). *Peramalan Stok Spare Part Menggunakan Metode Least Square*. SEFA BUMI PERSADA.

Mahena, Y., Rusli, M., & Winarso, E. (2015). Prediksi Harga Emas Dunia Sebagai Pendukung Keputusan Investasi Saham Emas Menggunakan Teknik Data Mining. *Kalbiscentia Jurnal Sains dan Teknologi*, *2*(1), 36–51.

Martínez, F., Frías, María, P., Charte, F., & Rivera, Antonio, J. (2019). Time Series Forecasting with KNN in R: the tsfknn Package. *The R Journal*, *11*(2), 229. https://doi.org/10.32614/RJ-2019-004

Mustakim, M., & Oktaviani, G. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. *Jurnal Sains, Teknologi, dan Industri*, *13*(2), 195–202. https://doi.org/10.24014/sitekin.v13i2.1688

Nanja, M., & Purwanto, P. (2015). METODE K-NEAREST NEIGHBOR BERBASIS FORWARD SELECTION UNTUK PREDIKSI HARGA KOMODITI LADA. *Pseudocode*, *2*(1), 53–64. https://doi.org/10.33369/pseudocode.2.1.53-64

Nofriansyah, D. (2014). *Konsep Data Mining vs Sistem Pendukung Keputusan* (1 ed.). Deepublish.

Putra, S. H., & Putra, B. T. (2018). Klasifikasi Harga Cell Phone menggunakan Metode K-Nearest Neighbor (KNN). *Prosiding Annual Research Seminar*, *4*(1), 242–245.

Sabilla, W. I., & Putri, T. E. (2017). Prediksi Ketepatan Waktu Lulus Mahasiswa dengan k-Nearest Neighbor dan Naïve Bayes Classifier (Studi Kasus Prodi D3 Sistem Informasi Universitas Airlangga). *Jurnal Komputer Terapan*, *3*(2), 233–240.

Sartika, E. (2019). Analisis Metode K Nearest Neighbor Imputation (KNNI) Untuk Mengatasi Data Hilang Pada Estimasi Data Survey. *Jurnal TEDC*, *12*(3), 219–227.

Wijaya, A., & Ananta, W. P. (2017). *Hukum Bisnis Properti Indonesia*. Grasindo.

JISKa, Vol. 6, No. 2, MEI, 2021, Pp. 98 – 105
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Metode *Accumulative Difference Images* untuk Mendeteksi Berhentinya Putaran Kincir Air

Adri Priadana [(1)*], Aris Wahyu Murdiyanto [(2)]
[1] Informatika, Fakultas Teknik dan Teknologi Informasi, Universitas Jenderal Achmad Yani, Yogyakarta
[2] Sistem Informasi, Fakultas Teknik dan Teknologi Informasi, Universitas Jenderal Achmad Yani, Yogyakarta
e-mail : {adripriadana3202,ariswahyumurdiyanto}@gmail.com.
* Penulis korespondensi.
Artikel ini diajukan 6 September 2020, direvisi 14 Oktober 2020, diterima 28 Oktober 2020, dan dipublikasikan 3 Mei 2021.

Abstract

Vannamei shrimp is one of Indonesia's fishery commodities with great potential to be developed. One of the essential things in shrimp farming is a source of dissolved oxygen (DO) or a sufficient amount of oxygen content, which can be maintained by placing a waterwheel driven by a generator set engine called a generator. To keep the waterwheel running, the cultivators must continue to monitor it in real-time. Based on these problems, we need a method that can be used to detect the cessation of waterwheel rotation in shrimp ponds that focuses on the rotation of the waterwheel. This study aims to analyze the performance of the Accumulative Difference Images (ADI) method to detect the stopped waterwheel-spinning. This method was chosen because compared with the method that only compares the differences between two frames in each process, the ADI method is considered to reduce the error-rate. After all, it is taken from the results of the value of several frames' accumulated movement. The ADI method's application to detect the stopped waterwheel-spinning gives an accuracy of 95.68%. It shows that the ADI method can be applied to detect waterwheels' stop in shrimp ponds with a very good accuracy value.

Keywords: Motion Detection, Accumulative Difference Images, ADI Method, Waterwheel Spin, Cultivating Vannamei Shrimp

Abstrak

Udang *vannamei* merupakan salah satu komoditas perikanan Indonesia yang memiliki potensi besar untuk dikembangkan. Salah satu hal penting dalam budidaya udang adalah sumber *Disolved Oksigen* (DO) atau jumlah kadar oksigen yang cukup di mana dapat dijaga dengan menempatkan kincir air yang digerakkan dengan mesin generator set yang disebut Genset. Dalam memastikan agar kincir air tetap menyala, para pembudidaya harus terus memantau secara *real-time*. Berdasarkan permasalahan tersebut, diperlukan sebuah cara yang dapat digunakan untuk mendeteksi berhentinya putaran kincir air pada tambak udang yang berfokus pada putaran kincir air. Penelitian ini bertujuan untuk menganalisis kinerja metode *Accumulative Difference Images* (ADI) untuk mendeteksi berhentinya putaran kincir air. Metode ini dipilih karena jika dibandingkan dengan metode yang hanya membandingkan perbedaan antara dua *frame* pada setiap prosesnya, metode ADI dinilai dapat mengurangi *error-rate* karena diambil dari hasil nilai akumulasi pergerakan dari beberapa *frame*. Penerapan metode ADI untuk mendeteksi berhentinya gerak kincir air memberikan hasil akurasi sebesar 95,68% di mana menunjukkan bahwa metode ADI dapat diterapkan untuk mendeteksi berhentinya gerak kincir air pada tambak udang dengan nilai akurasi yang sangat baik.

Kata Kunci: Deteksi Gerak, *Accumulative Difference Images*, Metode ADI, Putaran Kincir Air, Budidaya Udang Vannamei

1. PENDAHULUAN

Udang *vannamei*, atau yang biasa dikenal masyarakat dengan sebutan udang vaname merupakan salah satu jenis udang yang dibudidayakan oleh masyarakat Indonesia saat ini. Udang vaname telah menjadi salah satu komoditas perikanan Indonesia yang memiliki potensi besar untuk dikembangkan. Sejak tahun 2013 mulai terjadi peningkatan permintaan udang

vaname di pasar internasional (Musrowati Lasindrang, 2015). Hal ini menyebabkan banyak pembudidaya udang vaname yang terus meningkatkan usaha budidayanya. Salah satu hal penting dalam budidaya udang adalah sumber *Disolved Oksigen* (DO) atau jumlah kadar oksigen yang cukup di dalam air agar udang mendapatkan pasokan oksigen yang cukup (H.Kordi & Tancung, 2007). Peningkatan kadar DO tersebut dapat dilakukan dengan menggunakan kincir air pada tambak atau kolam (Mardhiya et al., 2018). Hal ini membuat peran kincir air pada tambak atau kolam udang menjadi hal yang utama di mana dapat membantu meningkatkan kadar oksigen di area sekitar perairan tambak (Nugraha et al., 2017).

Mesin *generator set* bertenaga diesel, atau yang sering disebut Genset, merupakan salah satu perangkat pembangkit daya listrik yang dimanfaatkan sebagai mesin penggerak kincir air. Dalam memastikan agar kincir air tidak mati, para pembudidaya harus terus memantau secara *real-time* untuk memastikan mesin Genset agar terus menyala. Tidak sedikit terjadi kasus di mana Genset masih menyala akan tetapi kincir air tidak bergerak. Hal ini dapat disebabkan karena patahnya saluran penggerak yang menghubungkan antara mesin Genset dengan kincir air. Jika kincir air tidak menyala, maka kadar oksigen di dalam air akan berkurang di mana akan mengancam keberlangsungan hidup dan meningkatkan jumlah kematian atau mortalitas udang vaname. Berdasarkan permasalahan tersebut, diperlukan sebuah cara yang dapat digunakan untuk mendeteksi berhentinya putaran kincir air pada tambak udang yang berfokus pada putaran kincir air.

Penelitian ini bertujuan untuk menganalisis kinerja penggunaan salah satu metode pada pemrosesan citra digital untuk mendeteksi berhentinya putaran kincir air. Pemanfaatan pemrosesan citra digital untuk deteksi dipilih karena dilakukan berdasarkan pemantauan visual langsung pada objek utamanya yaitu putaran kincir air, bukan pada mesin Genset yang menjadi mesin penggeraknya. Penelitian ini akan menerapkan metode *Accumulative Difference Images* (ADI) untuk mendeteksi gerakan kincir air yang berputar pada suatu tambak atau kolam. Meskipun terdapat beberapa metode untuk mendeteksi objek bergerak (Saubari et al., 2019), metode ADI ini dipilih karena jika dibandingkan dengan metode yang hanya membandingkan perbedaan antara dua *frame* pada setiap prosesnya, metode ADI dinilai dapat mengurangi *error-rate* karena diambil dari hasil nilai akumulasi pergerakan dari beberapa *frame* (Priadana & Harjoko, 2017).

Terdapat beberapa penelitian yang menerapkan teknologi dalam pengembangan budidaya udang. Penerapan teknologi seperti *Internet of Things* (IoT) telah diterapkan oleh peneliti (Sneha & Rakesh, 2017) di mana diterapkan untuk memantau secara otomatis dan mengendalikan sistem budidaya udang dan sawah. Teknologi ZigBee *network* juga telah banyak diterapkan pada beberapa penelitian (Nguyen Tang Kha Duy, Tran Trong Hieu, et al., 2015; Rerkratn & Kaewpoonsuk, 2015) untuk sistem pemantauan dan kontrol. Selain itu, penerapan teknologi seperti *embedded system* dan *wireless sensor network* juga pernah dilakukan peneliti (Nguyen Tang Kha Duy, Nguyen Dinh Tu, et al., 2015) untuk pemantauan secara otomatis dan sistem kontrol untuk tambak udang.

Terdapat beberapa penelitian yang menerapkan teknologi dalam upaya menurunkan mortalitas udang pada suatu tambak, di mana dilakukan dengan cara merancang sistem *monitoring* kualitas air. Penerapan teknologi seperti Arduino dan *Internet of Things* (IoT) telah diterapkan oleh beberapa peneliti sebelumnya (Maulana et al., 2017; Multazam & Hasanuddin, 2017; Pratama et al., 2019) di mana diterapkan untuk memantau kualitas air seperti suhu, pH, temperatur, dan *Disolved Oksigen* (DO). Pada beberapa penelitian tersebut memanfaatkan beberapa sensor seperti pH *sensor*, *temperature sensor*, *salinity sensor*.
Pada penelitian ini, peneliti akan menerapkan teknologi pemrosesan citra digital untuk membangun sistem deteksi berhentinya putaran kincir air di mana dilakukan dalam upaya menurunkan mortalitas udang pada suatu tambak. Jika dibandingkan dengan penerapan teknologi dalam upaya menurunkan mortalitas udang pada penelitian-penelitian sebelumnya di mana dilakukan dengan cara memantau kadar oksigen pada air, pada penelitian ini pemantauan langsung berfokus pada putaran kincir air sebagai salah satu komponen yang dapat meningkatkan kadar oksigen pada air. Oleh sebab itu, pada saat kincir air berhenti berputar

karena suatu hal, penanganan dapat segera dilakukan tanpa harus menunggu kadar oksigen dalam air menurun.

Berdasarkan dari studi literatur pada penelitian sebelumnya, dapat disimpulkan bahwa belum terdapat penelitian yang menerapkan teknologi pemrosesan citra digital untuk mendeteksi berhentinya putaran kincir air. Penelitian ini memberikan kebaruan yaitu (a) dalam hal objek penelitian, deteksi gerak pada putaran kincir air berbasis pemrosesan citra digital, dan (b) dalam metode penelitian, yaitu penggunanan metode ADI untuk mendeteksi berhentinya putaran kincir air. Metode ADI yang pada dasarnya berfungsi sebagai deteksi gerak dapat mendeteksi gerakan putaran kincir air melalui kamera, sehingga metode ADI tentunya juga dapat digunakan untuk mendeteksi berhentinya gerakan kincir air.

Metode ADI juga telah dimanfaatkan oleh beberapa penelitian untuk mendeteksi gerak. Priadana dan Harjoko di 2017 (Priadana & Harjoko, 2017), menerapkan metode ADI untuk mendeteksi gerak pada sistem perubahan citra pada video di mana menghasilkan nilai akurasi 95.12%. Nurhopipah dan Harjoko di 2018 (Nurhopipah & Harjoko, 2018), menerapkan metode ADI untuk sistem pengawasan melalui video CCTV di mana menghasilkan nilai akurasi deteksi gerak 92.655%. Kholid et.al., di 2020 (Mohammad Faisal Kholid et al., 2020), menerapkan metode ADI untuk mendeteksi gerakan manusia pada video CCTV di mana menghasilkan nilai akurasi 95.23%. Selain itu, berdasarkan hasil perbandingan dari beberapa metode deteksi gerak seperti *Background Subtraction*, *Sobel*, *Adaptive Motion Detection*, dan *Frame Differences*, metode ADI merupakan metode yang digunakan untuk mendeteksi gerak yang memiliki nilai akurasi tertinggi (Ramadhan et al., 2018).

2. METODE PENELITIAN

Tahapan pada penelitian ini terdiri dari tiga tahapan utama, yaitu pengumpulan data video putaran kincir air, penerapan metode ADI, dan pengujian metode.

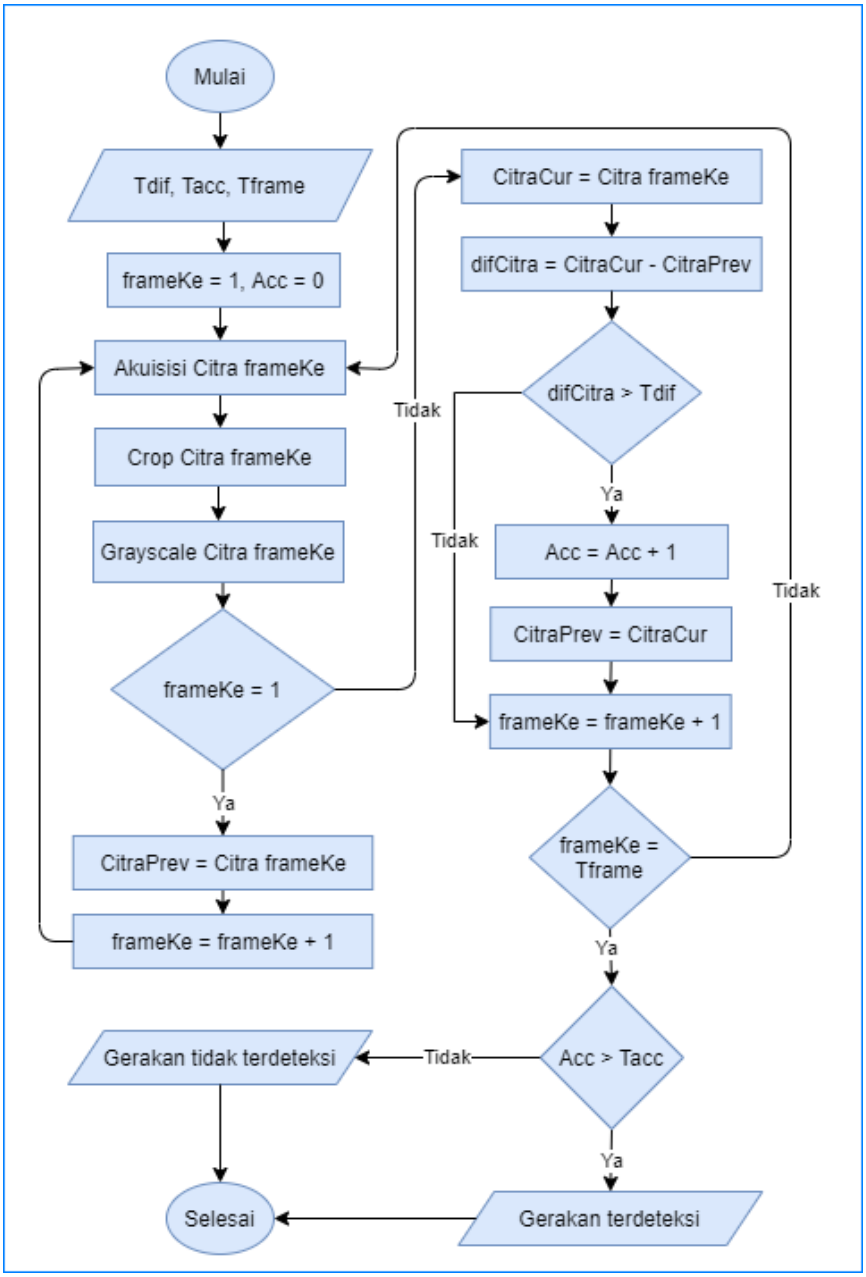
2.1. Data Video Putaran Kincir Air

Data yang digunakan pada penelitian ini adalah video yang berisi perubahan gerak kincir air baik dari berhenti menjadi bergerak ataupun sebaliknya. Video tersebut diambil dengan memanfaatkan kamera GoPro Hero 7 Black. Pengambilan video putaran kincir air tersebut dilakukan pada saat malam hari. Hal ini dikarenakan pada malam hari, kincir air harus selalu menyala untuk dapat menyuplai oksigen pada tambak udang.

2.2. Metode Accumulative Difference Images (ADI)

Penerapan metode ADI pada penelitian ini dilakukan dengan memanfaatkan bahasa pemrograman Python. *Flowchart* deteksi gerak dengan metode ADI pada penelitian ini ditunjukan dalam diagram pada Gambar 1. Deteksi gerak dengan metode ADI diawali dengan memberikan masukkan untuk variabel $Tdif$ yang menunjukkan nilai *threshold* perbedaan (*absolute*) antara *frame* saat ini dengan *frame* sebelumnya, Tframe untuk menentukan jumlah *frame* sebagai dasar satu sesi deteksi gerak, dan masukkan variabel Tacc menunjukkan nilai *threshold* akumulasi.

Dalam penelitian ini, nilai variabel $Tframe$ bernilai sama dengan jumlah *frame* per detik dari video masukkan. Hal ini bertujuan agar hasil deteksi gerak dapat diketahui pada setiap detiknya. Sedangkan untuk nilai variabel $Tacc$ didapatkan dari 50% dari jumlah *frame* per detik dari video masukkan. Hal ini dapat diartikan bahwa dalam satu detik dapat dideteksi adanya gerakan jika terdapat lebih dari 50% dari jumlah *frame* dalam setiap detiknya mengalami perubahan. Nilai variabel $Tdif$ pada penelitian ini bernilai nol. Hal ini berarti bahwa suatu *frame* dapat disimpulkan adanya perubahan jika terdapat lebih dari nol *pixel* perbedaan.



Gambar 1. Flowchart deteksi gerak dengan metode ADI.

Tahap selanjutnya adalah inisialisasi pada beberapa variabel seperti variabel *frameKe* yang menunjukkan urutan perulangan yang diberi nilai awal yaitu satu dan variabel *Acc* menunjukkan nilai akumulasi yang diberi nilai awal yaitu 0. Selanjutnya akuisisi citra akan dilakukan pada *frame* saat ini dan sebelumnya di mana keduanya dilakukan *cropping* untuk mengambil area lokasi kincir air pada *frame* atau sering disebut sebagai *region of interest* (ROI) serta dikonversi ke dalam model warna *grayscale* dengan Pers. (1) (Dawson-Howe, 2014).

$$Y = 0.299R + 0.587G + 0.1114B \quad (1)$$

Apabila citra yang diakuisisi merupakan *frame* pertama, maka citra tersebut akan dianggap sebagai citra sebelumnya dan kemudian akan dilakukan akuisisi citra pada *frame* selanjutnya atau dapat disebut sebagai *frame* saat ini. *Frame* saat ini hasil proses *cropping* dan *grayscale* akan dibandingkan dengan *frame* sebelumnya hasil proses *cropping* dan *grayscale* dengan teknik

image subtraction. Metode tersebut merupakan teknik yang digunakan untuk menghitung perbedaan antara dua citra $f(x,y)\ dan\ h(x,y)$ di mana dihitung dengan Pers. (2) (Gonzalez & Woods, 2018).

$$g(x,y) = f(x,y) - h(x,y) \qquad (2)$$

Nilai $g(x,y)$ atau $difCitra$ diperoleh dengan menghitung selisih antara semua pasangan pixels yang sesuai dari $f(x,y)$ atau $CitraCur$ dan $h(x,y)$ atau $CitraPref$. Kemudian nilai perbandingan antara kedua *frame* tersebut akan dibandingkan dengan nilai $Tdif$. Jika nilai perbandingannya lebih besar dari pada $Tdif$, maka nilai akumulasi pada *frame* sebelumnya akan bertambah satu $Acc = Acc + 1$. Jika nilai perbandingannya lebih kecil dari pada $Tdif$, maka nilai akumulasi pada frame sebelumnya tidak akan berubah. Proses perbandingan antara *frame* saat ini dengan *frame* sebelumnya tersebut akan dilakukan pada *frame* selanjutnya sejumlah $Tframe$ di mana nilai akumalasi Acc akan bertambah jika nilai perbandingannya lebih besar dari pada $Tdif$.

Nilai akumulasi Acc tersebut kemudian dibandingkan dengan nilai $TAcc$. Jika nilai akumulasi Acc lebih besar dari $TAcc$ maka dapat disimpulkan bahwa terdapat objek yang bergerak pada kumpulan *frame* yang berurutan dalam jangka waktu $Tframe$ tersebut. Jika nilai akumulasi Acc lebih kecil dari $TAcc$ maka dapat disimpulkan bahwa tidak terdapat objek yang bergerak pada kumpulan *frame* yang berurutan dalam jangka waktu $Tframe$ tersebut atau dapat disimpulkan bahwa kincir air telah berhenti berputar.

2.3. Pengujian Metode

Pengujian penerapan metode ADI untuk deteksi berhentinya kincir air dilakukan dengan menghitung nilai *precision*, *recall*, dan *The Percentage Correct Classification* (PCC) *of system*, atau dapat disebut akurasi. *Precision* adalah kemampuan dari sistem untuk tidak mendeteksi kondisi yang tidak benar. Sedangkan *recall* adalah kemampuan sistem untuk mendeteksi kondisi yang benar. Nilai *precision* dan *recall* dilakukan dengan menggunakan Pers. (3) dan Pers. (4) serta pengukuran akurasi deteksi dilakukan dengan menggunakan Pers. (5) (Martín & Pobil, 2012),

$$precision = \frac{TP}{TP + FP} \text{ x } 100\% \qquad (3)$$

$$recall = \frac{TP}{TP + FN} \text{ x } 100\% \qquad (4)$$

$$akurasi = \frac{TP+TN}{TP + TN + FP + FN} \text{ x } 100\% \qquad (5)$$

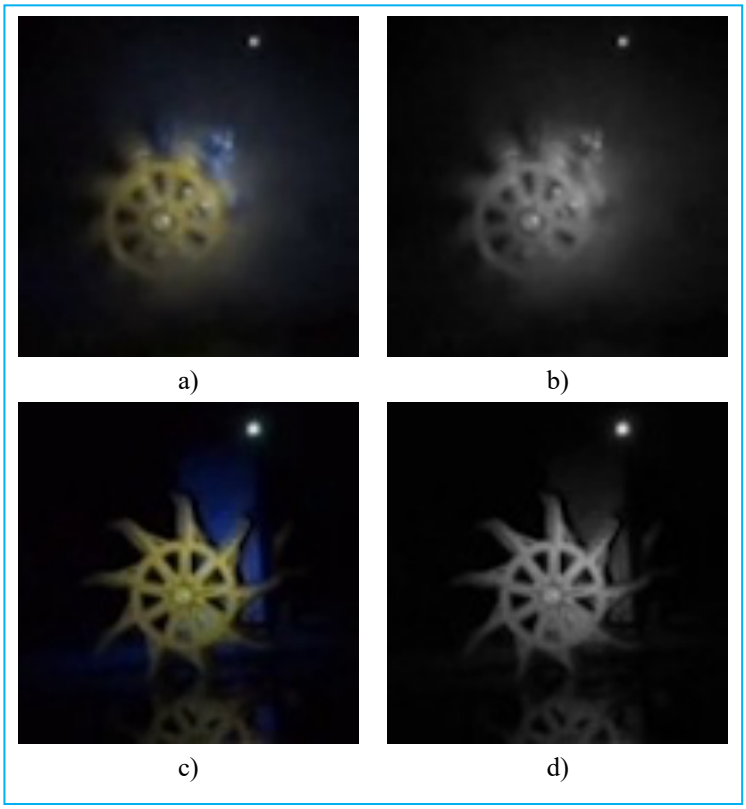
di mana *true positive* (TP) merupakan gerakan yang terdeteksi di mana gerakan ini memang ada berdasarkan kenyataan. *False positive* (FP) merupakan gerakan yang terdeteksi namun tidak ada berdasarkan kenyataan. *False negative* (FN) merupakan gerakan yang tidak terdeteksi namun ada berdasarkan kenyataannya. *True negative* (TN) merupakan gerakan yang tidak terdeteksi di mana gerakan atau perubahan citra ini memang tidak ada berdasarkan kenyataan.

3. HASIL DAN PEMBAHASAN

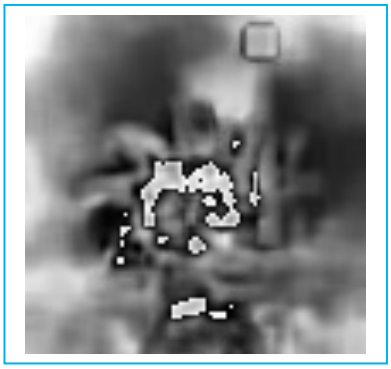
Pada penelitian ini, data video yang digunakan adalah 10 video yang masing-masing berisi perubahan gerakan kincir air di mana lima video berisi perubahan gerak dari bergerak menjadi berhenti dan lima video berisi perubahan gerak dari berhenti menjadi bergerak. Kesepuluh video tersebut memiliki resolusi 640p atau nHD di mana ukurannya tepat sepersembilan (*one ninth*) dari *Full High Definition* (FHD) yang memiliki ukuran 1080p. Jumlah *frames per second* (FPS) dari masing-masing video adalah 30 FPS. Berdasarkan jumlah FPS tersebut maka nilai variabel *Tframe* menjadi bernilai 30 dan nilai variabel *Tacc* bernilai 50% dari 30 yaitu 15 *frames*. Setelah mendapatkan kedua nilai masukkan tersebut, proses selanjutunya adalah melakukan *cropping* dan konversi *grayscale* pada *frame* masukkan baik *frame* saat ini dan juga *frame* sebelumnya di mana ditunjukkan pada Gambar 2. *Frame* saat ini hasil proses *cropping* dan *grayscale* akan

dibandingkan dengan *frame* sebelumnya hasil proses *cropping* dan *grayscale* dengan teknik *image subtraction* di mana hasilnya ditunjukkan pada Gambar 3.

Dari Gambar 3 tersebut dapat dilihat nilai perbedaan di mana menunjukkan adanya gerak pada *frame* tersebut. Proses perbandingan tersebut akan dilakukan pada *frame* selanjutnya sejumlah *Tfrmae* yang bernilai 30 di mana nilai akumalasi *Acc* akan bertambah jika nilai perbandingannya lebih besar dari pada *Tdif* yang bernilai 0. Nilai akumulasi *Acc* tersebut kemudian dibandingkan dengan nilai *TAcc* yang bernilai 15. Jika nilai akumulasi *Acc* lebih kecil dari *TAcc* maka dapat disimpulkan bahwa tidak terdapat objek yang bergerak pada kumpulan *frame* yang berurutan dalam jangka waktu *Tframe* tersebut atau dapat disimpulkan bahwa kincir air telah berhenti berputar. Hasil pengujian metode ADI pada data video yang telah dikumpulkan, ditunjukkan pada Tabel 1.



Gambar 2. *Frame* kincir air a) *cropping frame* sebelumnya, b) *grayscale frame* sebelumnya, c) *cropping frame* saat ini, b) *grayscale frame* saat ini.



Gambar 3. *Frame* hasil *image subtraction*

Berdasarakan dari hasil pengujian pada Tabel 1, didapatkan hasil perhitungan *precision* dan *recall* dari deteksi gerak kincir air di mana didapatkan nilai *precision* sebesar 96,64% dan nilai *recall* 92,90%. Nilai akurasi yang didapat dari deteksi gerak kincir air dengan metode ADI secara keseluruhan adalah sebesar 95,82%. Jika dibandingkan dengan metode lain yang digunakan untuk mendeteksi gerak seperti metode *background subtraction* dengan algoritma *gaussian mixture model* yang mendapatkan hasil akurasi sebesar 88,3% (Harry et al., 2017), dan dengan metode *spektral residual* yang mendapatkan hasil akurasi sebesar 90% (Rahmawati & Nugroho, 2018), metode ADI untuk mendeteksi gerak kincir air pada penelitian ini medapatkan hasil akurasi yang lebih baik yaitu sebesar 95,82%.

Lebih lanjut, berdasarkan dari hasil pengujian pada Tabel 1, dapat dilihat nilai *precision* dan *recall* dari deteksi berhentinya gerak kincir air di mana didapatkan nilai *precision* sebesar 93,24% dan nilai *recall* sebesar 98,57%. Nilai akurasi yang didapat dari deteksi berhentinya gerak kincir air dengan metode ADI adalah sebesar 95,68%. Hal ini menunjukkan bahwa metode ADI dapat diterapkan untuk mendeteksi berhentinya gerak kincir air pada tambak udang dengan nilai akurasi yang sangat baik.

Tabel 1. Hasil Deteksi Gerak dan Berhentinya Kincir Air.

No	Nama File	Keterangan	Durasi (detik)	TP	TN	FP	FN	Precision (%)	Recall (%)	Akurasi (%)
1.	V1.mp4	Berisi perubahan gerakan kincir air dari berhenti menjadi berputar	29	18	11	0	0	100	100	100
2.	V2.mp4		32	15	15	0	2	100	88,23	93,75
3.	V3.mp4		30	13	16	0	1	100	92,86	96,67
4.	V4.mp4		30	19	11	0	0	100	100	100
5.	V5.mp4		27	10	14	0	3	100	58,82	88,89
6.	V6.mp4	Berisi perubahan gerakan kincir air dari berputar menjadi berhanti	31	15	16	0	0	100	100	100
7.	V7.mp4		30	14	16	0	0	100	100	100
8.	V8.mp4		29	16	13	0	0	100	100	100
9.	V9.mp4		29	16	13	0	0	100	100	100
10.	V10.mp4		20	8	6	5	1	61,54	88,89	70
Jumlah				136	132	0	19	96,64	92,90	95,82
Jumlah (deteksi berhentinya gerak kincir air dari video nomor 6 sampai 10)				69	64	5	1	93,24	98,57	95,68

4. KESIMPULAN

Penerapan metode ADI untuk mendeteksi gerak kincir air secara keseluruhan memiliki nilai akurasi yang lebih baik jika dibandingkan dengan metode lain yang digunakan untuk mendeteksi gerak seperti metode *background subtraction* dengan algoritma *gaussian mixture model* dan dengan metode *spektral residual*. Lebih lanjut, penerapan metode ADI untuk mendeteksi berhentinya gerak kincir air memberikan hasil akurasi sebesar 95,68% di mana menunjukkan bahwa metode ADI dapat diterapkan untuk mendeteksi berhentinya gerak kincir air pada tambak udang dengan nilai akurasi yang sangat baik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Kementerian Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia atas dukungan yang diberikan kepada penulis berupa bantuan dana penelitian dalam skema Penelitian Dosen Pemula (PDP) tahun pelaksanaan 2020.

DAFTAR PUSTAKA

Dawson-Howe, K. (2014). *A Practical Introduction to Computer Vision with OpenCV.* Wiley.

Gonzalez, R. C., & Woods, R. E. (2018). *Digital Image Processing* (4th ed). Pearson.

H.Kordi, M. G., & Tancung, A. B. (2007). *Pengelolaan Kualitas Air Dalam Budidaya Perairan*. PT Rineka Cipta.

Harry, M., Pratama, B., Hidayatno, A., Ajub, D., & Zahra, A. (2017). APLIKASI DETEKSI GERAK PADA KAMERA KEAMANAN MENGGUNAKAN METODE BACKGROUND SUBTRACTION DENGAN ALGORITMA GAUSSIAN MIXTURE MODEL. In *Transient: Jurnal Ilmiah Teknik Elektro* (Vol. 6, Nomor 2). Universitas Diponegoro.

https://doi.org/10.14710/TRANSIENT.6.2.246-253
Mardhiya, I. R., Surtono, A., & Suciyati, S. W. (2018). Sistem Akuisisi Data Pengukuran Kadar Oksigen Terlarut Pada Air Tambak Udang Menggunakan Sensor Dissolved Oxygen (DO). *Jurnal Teori dan Aplikasi Fisika*, *6*(1), 133–140. https://doi.org/10.23960/JTAF.V6I1.1836
Martín, E. M., & Pobil, Á. P. del. (2012). *Robust Motion Detection in Real-Life Scenarios* (1 ed.). Springer-Verlag London. https://doi.org/10.1007/978-1-4471-4216-4
Maulana, Y. Y., Wiranto, G., & Kurniawan, D. (2017). Online Monitoring Kualitas Air pada Budidaya Udang Berbasis WSN dan IoT. *INKOM Journal*, *10*(2), 81–86. https://doi.org/10.14203/J.INKOM.456
Mohammad Faisal Kholid, Jian Budiarto, Ahmad Ashril Rizal, & Gibran Satya Nugraha. (2020). HUMAN MOVEMENT DETECTION DENGAN ACCUMULATIVE DIFFERENCES IMAGE. *TEKNIMEDIA: Teknologi Informasi dan Multimedia*, *1*(1), 1–7. https://doi.org/10.46764/teknimedia.v1i1.7
Multazam, A. E., & Hasanuddin, Z. B. (2017). Sistem Monitoring Kualitas Air Tambak Udang Vaname. *JURNAL IT Media Informasi STMIK Handayani Makassar*, *8*(2), 118–125.
Musrowati Lasindrang, L. S. N. K. (2015). KAJIAN SEBARAN POTENSI EKONOMI SUMBER DAYA. *Jurnal Teknosains*, *4*(2), 101–198. https://doi.org/10.22146/teknosains.7953
Nguyen Tang Kha Duy, Nguyen Dinh Tu, Tra Hoang Son, & Luong Hong Duy Khanh. (2015). Automated monitoring and control system for shrimp farms based on embedded system and wireless sensor network. *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1–5. https://doi.org/10.1109/ICECCT.2015.7226111
Nguyen Tang Kha Duy, Tran Trong Hieu, & Luong Hong Duy Khanh. (2015). A versatile, low poweron monitoring and control system for shrimp farms based on NI myRIOand ZigBee network. *2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, 0282–0287. https://doi.org/10.1109/ICCPEIC.2015.7259476
Nugraha, N. P. A., Agus, M., & Mardiana, T. Y. (2017). REKAYASA KINCIR AIR PADA TAMBAK LDPE UDANG VANNAMEI (Litopenaeus vannamei) DI TAMBAK UNIKAL SLAMARAN. *Pena Akuatika : Jurnal Ilmiah Perikanan dan Kelautan*, *16*(1). https://doi.org/10.31941/PENAAKUATIKA.V16I1.527
Nurhopipah, A., & Harjoko, A. (2018). Motion Detection and Face Recognition for CCTV Surveillance System. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *12*(2), 107. https://doi.org/10.22146/ijccs.18198
Pratama, A. S., Efendi, A. H., Burhanudin, D., & Rofiq, M. (2019). Simkartu (Sistem Monitoring Kualitas Air Tambak Udang) Berbasis Arduino dan SMS Gateway. *Jurnal SITECH : Sistem Informasi dan Teknologi*, *2*(1), 121–126. https://doi.org/10.24176/sitech.v2i1.3498
Priadana, A., & Harjoko, A. (2017). Deteksi Perubahan Citra Pada Video Menggunakan Illumination Invariant Change Detection. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *11*(1), 89–98. https://doi.org/10.22146/ijccs.17526
Rahmawati, L., & Nugroho, H. (2018). Deteksi Gerak Pada Citra Objek Video Surveillance Dengan Menggunakan Metode Spektral Residual. *INTEGER: Journal of Information Technology*, *3*(1). https://doi.org/10.31284/j.integer.2018.v3i1.219
Ramadhan, D. I., Sari, I. P., & Sari, L. O. (2018). COMPARISON OF BACKGROUND SUBTRACTION, SOBEL, ADAPTIVE MOTION DETECTION, FRAME DIFFERENCES, AND ACCUMULATIVE DIFFERENCES IMAGES ON MOTION DETECTION. *SINERGI*, *22*(1), 51. https://doi.org/10.22441/sinergi.2018.1.009
Rerkratn, A., & Kaewpoonsuk, A. (2015). ZigBee based wireless temperature monitoring system for shrimp farm. *2015 15th International Conference on Control, Automation and Systems (ICCAS)*, 428–431. https://doi.org/10.1109/ICCAS.2015.7364953
Saubari, N., Gazali, M., & Ansari, R. (2019). Metode HLF untuk Deteksi Objek Terapung pada Permukaan Sungai Martapura. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *4*(2), 43. https://doi.org/10.14421/jiska.2019.42-06
Sneha, P. S., & Rakesh, V. S. (2017). Automatic monitoring and control of shrimp aquaculture and paddy field based on embedded system and IoT. *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 1085–1089. https://doi.org/10.1109/ICICI.2017.8365307

JISKa, Vol. 6, No. 2, MEI, 2021, Pp. 106 – 112
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisis *Hashtag* pada Twitter untuk Eksplorasi Pokok Bahasan Terkini Mengenai *Business Intelligence*

Arif Himawan [(1)], Muhammad Rifqi Ma'arif [(2)*], Ulfi Saidata Aesyi [(3)]
[1,3] Program Studi Sistem Informasi, Fakultas Teknik dan Teknologi Informasi, Universitas Jenderal Achmad Yani, Yogyakarta
[2] Program Studi Teknik Industri, Fakultas Teknik dan Teknologi Informasi, Universitas Jenderal Achmad Yani, Yogyakarta
e-mail : {reef1881,muhammad.rifqi,ulfiaesyi}@gmail.com.
* Penulis korespondensi.
Artikel ini diajukan 11 September 2020, direvisi 7 Desember 2020, diterima 8 Desember 2020, dan dipublikasikan 3 Mei 2021.

Abstract

The main purpose of this paper is to examine the dominant topics about Business Intelligence in micro-blogging Twitter. There are 7.153 tweets collected from Twitter API. Text mining and natural language processing are used to analyze the dominant topics among those tweets. Computational method used to count the most frequent hashtag that appears together with Business Intelligence hashtag. Twitter users are large and scattered around the world with a diverse range of skills (expertise) that can give a new perspective on a subject that may not be predicted before. For example, for topics related to Business Intelligence, the very dominant general topic discussed in the scientific literature are about data management, as well as for analytics and machine learning data. The result contributes to understanding dominant topics about Business Intelligence that can help researchers to level their research.

Keywords: Business Intelligence, Twitter, Social Media, Hashtag Analysis, Exploratory Analysis

Abstrak

Tujuan dari penelitian ini adalah untuk memeriksa topik yang dominan tentang Business Intelligence di *micro-blogging* Twitter. Ada 7,153 tweets dikumpulkan menggunakan Twitter API. *Text mining* dan pengolahan bahasa alami yang digunakan untuk menganalisis topik yang dominan di cuitan tersebut. Metode komputasi digunakan untuk menghitung *hashtag* yang paling sering muncul bersama dengan *hashtag business intelligence*. Pengguna Twitter yang berjumlah sangat banyak dan tersebar di seluruh penjuru dunia dengan beragam keahlian (*expertise*) dapat memberikan perspektif baru atas suatu bahasan yang mungkin tidak diperkirakan sebelumnya. Sebagai contoh, untuk topik yang terkait dengan *business intelligence*, topik umum yang sangat dominan dibahas dalam literatur ilmiah adalah mengenai manajemen data, serta data analytics dan *machine learning*. Hasil dari penelitian ini berkontribusi untuk memahami topik yang dominan tentang *business intelligence* yang dapat membantu para peneliti untuk meningkatkan penelitian mereka.

Kata Kunci: *Business Intelligence*, Twitter, Media Sosial, Analisis Hashtag, Analisis Eksploratif

1. PENDAHULUAN

Pada era globalisasi sekarang ini, setiap aspek berhubungan dengan teknologi dengan tujuan mempermudah pekerjaan aspek termasuk aspek bisnis. Informasi yang dibutuhkan menjadi lebih penting untuk mendukung dalam pengambilan keputusan. Untuk menghasilkan informasi yang penting, maka dibutuhkan *database* yang memiliki relasi yang optimal dengan menggunakan kosep *business intelligence* (BI). *Business intelligence* meliputi semua aktivitas dalam pengolahan data termasuk di dalamnya pengumpulan, analisis, dan visualisasi dalam pengoperasian dan pengambilan keputusan (Mariani et al., 2018).

Business intelligence merupakan suatu bagian dari proses manajemen strategi yang telah diterapkan pada sektor publik terutama pada bagian perencanaan (Hellström & Ramberg, 2019).

Penerapan *business intelligence* dalam suatu organisasi dapat memberi dampak positif pada *sharing* pengetahuan, distribusi informasi, dan mampu menjadi alat untuk menyelesaikan masalah yang dihadapi organisasi (Eidizadeh et al., 2017; Scholtz et al., 2018). Dalam penerapannya, *business intelligence* digunakan dalam berbagai bidang seperti pariwisata, manajemen untuk pengambilan keputusan, manajement informasi pendidikan tinggi, perencanaan sumber daya perusahaan, hingga perusahaan tambang (Chongwatpol, 2016; Labonte-LeMoyne et al., 2017). Hal tersebut dimungkinkan karena *business intelligence* yang awalnya hanya berupa pertanyaan strategis menjadi suatu tugas operasional (Vujošević et al., 2019). Beberapa keuntungan menggunakan BI seperti menghilangkan pengulangan dan redudansi data pada saat pengolahan data. Dalam bidang kesehatan, BI dapat memberi dampak pada peningkatan kualitas pengambilan keputusan, performa dan proses klinis, serta efisiensi biaya (Ratia et al., 2018). Dalam perencanaan strategis, BI dapat mengurangi biaya perawatan, menyediakan informasi secara langsung (*real-time*), serta meningkatkan kualitas laporan dan kualitas perencanaan (Scholtz et al., 2018).

Business intelligence merupakan tema yang sangat luas dan memiliki banyak relasi dengan tema bahasan yang lain. Bagi praktisi di bidang *business intelligence* maupun *data science*, sangat penting untuk bisa terus mengikuti topik terkini mengenai *business intelligence*. Salah satu cara yang dapat ditempuh adalah mengikuti perkembangan pembicaraan topik *business intelligence* di media sosial, salah satunya adalah platform Twitter. Sosial media merupakan suatu tempat di mana masyarakat berkumpul atau bersosialisasi secara virtual. Tanpa disadari sosial media juga mewakili informasi dari tingkah laku manusia yang berorientasi bisnis, lokasi, multimedia, dan lainnya (Garg & Kumar, 2016). Hal tersebut memungkinkan aktivitas bersosial media dapat mempengaruhi performa seseorang dalam bekerja (Yingjie e al., 2019). Saat ini pengguna sosial media semakin banyak dan mengakibatkan data yang diproduksi oleh pengguna semakin meningkat, sehingga analisis sosial media berkembang menjadi metode baru untuk menginvestigasi tren dan pola (Park et al., 2016).

Analisis media sosial melingkupi *natural language processing* (NLP), *text mining*, *sentiment analysis*, dan algoritma *data mining* lainnya (Kim et al., 2016). Kemampuan metode-metode analisis media sosial dalam menginvestigasi tren dan pola telah diterapkan pada industri besar maupun kecil seperti industri film, perhotelan, kuliner, bisnis ritel, bisnis teknologi, maupun industri makanan dalam menentukan arah pengembangan industri dan penelitian (Cluley & Green, 2019; Kim et al., 2016; Park et al., 2016).

2. METODE PENELITIAN

Penelitian ini adalah penelitian deskriptif kuantitatif. Penelitian deskriptif kuantitatif merupakan penelitian yang memiliki tujuan untuk menggambarkan suatu fenomena yang muncul melalui ukuran-ukuran numerik (angka) untuk memberikan penjelasan mengenai karakteristik suatu kelompok maupun individu (Yusuf, 2014). Dalam artikel ini, yang menjadi objek penelitian adalah unggahan-unggahan pada platform Twitter yang bertagar #businessintelligence. Sebagai sampel, diambil unggahan-unggahan yang diunggah selama satu pekan mulai tanggal 26 September – 3 Oktober 2019.

Data dari platform Twitter tersebut diambil secara *real-time* menggunakan *Application Programming Interface* (API) yang disediakan oleh Twitter. Dalam rentang waktu kurang lebih satu pekan tersebut terkumpul kurang lebih berjumlah 7.153 unggahan. Data yang terkumpul kemudian dibersihkan (*preprocessing*) untuk memudahkan proses analisis. Proses pembersihan data yang diperoleh dari platform Twitter menggunakan metode yang dirancang oleh Hidayatullah dan Ma'arif (2017). Setelah dibersihkan, metode komputasional digunakan untuk menghitung tagar yang paling sering muncul secara bersama dengan tagar #businessintelligence dalam satu unggahan, serta akun-akun yang paling sering menggunakan tagar tersebut.

3. HASIL DAN PEMBAHASAN

Analisis atas data yang diperoleh selama waktu pengumpulan data pada pada platform jejaring sosial Twitter terbagi menjadi dua kategori yaitu analisis konten (*content analysis*) dan analisis

aktor (*actor analysis*). Analisis konten bertujuan untuk memberikan *insight* atas konten-konten yang terkait dengan topik *business intelligence*, sementara itu analisis aktor untuk mendapatkan insight mengenai aktor-aktor yang berpengaruh dalam topik pembicaraan terkait dengan *business intelligent*.

Analisis konten dilakukan dengan dengan memperhatikan tagar yang paling sering digunakan secara berdampingan (*co-occurence*) dengan tagar #businessintelligence. Tagar/hashtag yang disimbolkan dengan karakter "#" seringkali digunakan oleh pengguna Twitter untuk memperlihatkan penekanan topik pada unggahan yang dibuat (Hidayatullah & Ma'arif, 2017). Penggunaan *hashtag* ini bertujuan untuk mempermudah pembaca dalam mencari unggahan-unggahan dengan topik pembicaraan yang serupa. Tabel 1 menunjukkan 10 kelompok tagar yang paling sering digunakan untuk menyertai tagar #businessintelligence.

Tabel 1. Kelompok tagar/*hashtag* yang paling sering muncul bersama dengan tagar #businessintelligence.

No.	Tagar	Jumlah Kemunculan
1	#Crypto, #Cryptocurrency, #Blockchain	1,630
2	#BigData, #Data, #DataScience, #Analytic	1,458
3	#AI, #ArtificialIntelligence, #MachineLearning	773
4	#PowerBI, #SAP, #SQLServer, #Tableau, #Python	314
5	#DataWarehouse, #DataVault, #CloudComputing	261
6	#DigitalTransformation, #DigitalMarketing	254
7	#BusinessAnalytics, #DataDriven, #DataAnalytic	219
8	#PredictiveAnalytics, #DeepLearning	87
9	#InformationGovernance, #GDPR	66
10	#Organization, #CDO	56

Tagar yang menempati urutan pertama dari jumlah kemunculannya adalah #Crypto, #CryptoCurrency dan #Blockchain. Teknologi Blockchain yang menjadi basis bagi teknologi *cryptocurrency* pada dasarnya tidak secara langsung terkait dengan *business intelligence*. Awal kemunculan teknologi Blockchain adalah untuk meningkatkan keamanan (*security*) atas transaksi-transaksi penting melalui internet, khususnya transaksi finansial. Terkaitan dengan topik *business intelligence*, teknologi *blockhain* akan menyediakan mekanisme untuk melakukan akuisisi dan penyimpanan data secara terdistribusi yang lebih aman untuk aplikasi-aplikasi *business intelligence* khususnya yang terkait dengan domain finansial (Swan, 2018).

Tagar selanjutanya yang sering muncul terkait dengan topik *business intelligence* adalah tagar yang terkait dengan teknologi data. Teknologi data sebagai tulang punggung (*backbone*) utama dalam *business intelligence* terbagi kedalam dua kategori, takni teknologi untuk manajemen dan rekayasa data (*data engineering*) dan teknologi untuk anailik dan sain data (*data science and analytics*). Unggahan yang terkait dengan teknologi rekayasa data tercermin dari tagar-tagar yang bearada dalam kemlompok 5 yakni #DataWarehouse, #DataValut dan #CloudComputing. Konsep *data warehouse* dan *Cloud Computing* merupakan dua konsep yang sudah lama diketahui. Konsep yang relatif baru dalam teknik rekayasa data adalah *data vault*. *Data vault* merupakan salah satu konsep pemodelan dalam basis data. Pemodelan *data vault* merupakan metode pemodelan basis data yang didesain untuk menyediakan penyimpanan data dalam jangka panjang atas data yang berasal dari berbagai sumber. *Data vault* sangat erat kaitannya dengan *business intelligence* karena memberikan spektrum data yang lebih luas untuk dianalisis dan dicari polanya untuk mendapatkan deskripsi yang lebih luas atas suatu data serta untuk membuat model prediksi yang lebih akurat (Nogueira et.al, 2018).

Untuk teknologi yang terkait dengan analitik dan sains data, tagar yang sering muncul adalah tagar yang berada dalam kelompok 2, 3 dan 8. Tagar-tagar dalam kelompok 2 dan 3 merepresentasikan unggahan-unggahan yang terkait dengan konsep umum dalam sains data yakni #BigData, #MachineLearning, #ArtificialIntelligence dst. Sementara itu tagar dalam kelompok 8 yakni #DeepLearning dan #PredictiveAnalytics merepresentasikan porsi kecil unggahan-unggahan yang membahas dua teknologi yang saat ini cukup banyak diadopsi dalam

sains data. Implementasi *predictive analytics* saat ini banyak mengalami peningkatan dalam sisi performa dan akurasi dengan munculnya *deep learning* (Muniasamy et al., 2019). *Deep learning* sendiri merupakan teknologi yang muncul untuk memanfaatkan melimpahnya jumlah data yang dikelola dengan teknologi *big data* (Shi et.al, 2017).

Pokok bahasan lain terkait dengan #businessintelligence adalah *tools* atau perangkat lunak yang banyak digunakan dalam mengimplementasikan *business intelligence*. Dari unggahan-unggahan yang dikumpulkan, terdapat 5 tagar yang masing-masing merepresentasikan perangkat lunak *business intelligence* yang populer digunakan. Kelima tagar tersebut adalah #PowerBI, #SAP, #SQLServer, #Tableau dan #Python. Porsi lain dari unggahan-unggahan di Twitter yang terkait dengan *business intelligence* yang memiliki prosentase relatif kecil membahas hal-hal yang bersifat non-teknis. Isu yang diangkat terepresentasikan dalam tagar #InformationGovernance dan #GDPR yang terkait dengan konsep pengelolaan data dan informasi untuk keperluan bisnis dan organisasi.

GDPR atau *General Data Protection Regulation* itu sendiri merupakan sebuah regulasi dari hukum yang berlaku di Uni Eropa (EU) mengenai proteksi dan privasi data untuk semua warga EU dan EEA (*European Economic Area*). Regulasi tersebut juga mengatur mekanisme transfer data personal yang masuk maupun keluar dari wilayah EU dan EEA (Voigt & von dem Bussche, 2017). Area lain yang tercover dalam unggahan di Twitter mengenai *business intelligence* adalah hal-hal yang terkait dengan isu organisasi dan transformasi digital. Tagar-tagar dalam kelompok 6 dan 10 yang berisi tagar seperti #DigitalTransformation, #DigitalMarketing, #DataDriven, #Organization dlsb berisi unggahan-unggahan yang berisi pesan akan pentingnya bagi organisasi untuk merubah arah organisasi ke budaya digital.

Tabel 2. Akun yang paling perpengaruh dalam unggahan bertagar #BusinessIntelligence.

No.	Nama Akun	*Enggagement Rate*
1	@RobertBeadles	561
2	@MicroStrategy	221
3	@Ronald_vanLoon	98
4	@jamilahmed_16	93
5	@wiomax	82
6	@DataVault_UK	44
7	@BigDataLove	42
8	@karla_redhead	37
9	@Reactionpower	23
10	@iebschool	19

Analisis selanjutnya adalah analisis aktor. Aktor yang dimaksudkan dalam artikel ini direpresentasikan oleh sebuah akun dalam Twitter. Analisis aktor dimaksudkan untuk mengetahui akun-akun yang paling berpengaruh pada persebaran informasi mengenai topik *business intelligence* di Twitter. Pengaruh akun atas persebaran unggahan tagar #businessintelligence dilihat dari *enggagement rate* masing-masing akun atas unggahan #businessintelligence yang dibuat. Adapun *enggagement rate* tersebut dihitung dengan menjumlahkan jumlah *retweet* dengan *reply* pada setiap postingan bertagar #businessintelligent yang dibuat.

Dari Tabel 2, nampak bahwa akun yang paling berpengaruh dan memiliki *enggagement rate* yang relatif jauh lebih tinggi dari yang lain adalah akun @RobertBeadles dan @MicroStrategy. @RobertBeadles merupakan akun dari seorang praktisi *bitcoin*/*cryptocurrency* yang merupakan pengelola laman http://cryptobeadles.com. Unggahan-unggahan dari akun @RobertBeadles mengenai *business intelligence* melalui tagar #businessintelligence hampir semuanya bertema *cryptocurrency* maupun *blockchain* yang merupakan salah satu fundamen dalam pengembangan *business intelligence*. Akun kedua yang memiliki *enggagement rate* tinggi adalah @MicroStrategy. Akun tersebut merupakan akun resmi dari perusahaan analitik Micro Strategy (http://microstrategy.com). Unggahan bertagar #businessintelligence dari akun ini membahas

topik terkait yang cukup luas. Topik yang cukup sering dibahas yang terkait dengan *business intelligence* dari akun @MicroStrategy adalah peran penting *business intelligence* dalam mencapai *competitive advantages* suatu organisasi. Contoh unggahan dari akun @MicroStrategy yang menggunakan tagar #businessintelligence dapat dilihat pada Gambar 1.

Gambar 1. Contoh unggahan bertagar #businessintelligence dari akun Twitter @MicroStrategy

Dari contoh unggahan akun @MicroStrategy di Twitter pada Gambar 1, unggahan bertagar #businessintelligence tersebut membahas mengenai peranan analitik data (#analytics) dalam meningkatkan *competitive advantages* suatu organisasi melalui 3 hal yakni: (1) peningkatan efisiensi dan produktifitas, (2) pengambilan keputusan yang lebih cepat dan (3) performa finansial yang lebih baik.

4. KESIMPULAN

Dari hasil penelitian yang dilakukan, unggahan-unggahan di Twitter dapat memberikan wawasan (*insight*) yang lebih luas mengenai topik-topik pembicaraan yang terkait dengan suatu bahasan tertentu. Pengguna Twitter yang berjumlah sangat banyak dan tersebar di seluruh penjuru dunia dengan beragam keahlian (*expertise*) dapat memberikan perspektif baru atas suatu bahasan yang mungkin tidak diperkirakan sebelumnya. Sebagai contoh, untuk topik yang terkait dengan *business intelligence*, topik umum yang sangat dominan dibahas dalam literatur ilmiah adalah mengenai manajemen data, serta *data analytics* dan *machine learning*. Namun dengan melakukan analisis pada unggahan di Twitter, dapat diperoleh perspektif yang sangat beragam atas tema *business intelligence*.

Penelitian ini masih sebatas melakukan analisis kuantitatif dasar yakni penghitungan sederhana (*counting*) atas objek tertentu yakni topik pembicaraan dan aktor yang mengunggah suatu bahasan dalam platform media sosial Twitter. Untuk penelitian selanjutnya, dapat dilakukan analisis yang lebih kompleks untuk mendapatkan *insight* yang lebih mendalam. Pendekatan komputasional seperti *text mining* dan *natural language processing* dapat digunakan untuk melakukan analisis yang lebih komprehensive atas data tekstual yang didapatkan dari platform media sosial, termasuk Twitter.

UCAPAN TERIMA KASIH

Ucapan terima kasih diberikan kepada Program Studi Sistem Informasi, Universitas Jenderal Achmad Yani Yogyakarta yang telah memberikan pendanaan untuk penelitian ini.

DAFTAR PUSTAKA

Chongwatpol, J. (2016). Managing big data in coal-fired power plants: a business intelligence framework. *Industrial Management & Data Systems*, *116*(8), 1779–1799. https://doi.org/10.1108/IMDS-11-2015-0473

Cluley, R., & Green, W. (2019). Social representations of marketing work: advertising workers and social media. *European Journal of Marketing*, *53*(5), 830–847. https://doi.org/10.1108/EJM-12-2016-0682

Eidizadeh, R., Salehzadeh, R., & Chitsaz Esfahani, A. (2017). Analysing the role of business intelligence, knowledge sharing and organisational innovation on gaining competitive advantage. *Journal of Workplace Learning*, *29*(4), 250–267. https://doi.org/10.1108/JWL-07-2016-0070

Garg, M., & Kumar, M. (2016). Review on event detection techniques in social multimedia. *Online Information Review*, *40*(3), 347–361. https://doi.org/10.1108/OIR-08-2015-0281

Hellström, M., & Ramberg, U. (2019). Senior public leaders' perceptions of business intelligence. *International Journal of Public Leadership*, *15*(2), 113–128. https://doi.org/10.1108/IJPL-11-2018-0055

Hidayatullah, A. F., & Ma'arif, M. R. (2017). Pre-processing Tasks in Indonesian Twitter Messages. *Journal of Physics: Conference Series*, *801*(1), 012072. https://doi.org/10.1088/1742-6596/801/1/012072

Kim, Y., Dwivedi, R., Zhang, J., & Jeong, S. R. (2016). Competitive intelligence in social media Twitter: iPhone 6 vs. Galaxy S5. *Online Information Review*, *40*(1), 42–61. https://doi.org/10.1108/OIR-03-2015-0068

Labonte-LeMoyne, E., Leger, P.-M., Robert, J., Babin, G., Charland, P., & Michon, J.-F. (2017). Business intelligence serious game participatory development: lessons from ERPsim for big data. *Business Process Management Journal*, *23*(3), 493–505. https://doi.org/10.1108/BPMJ-12-2015-0177

Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, *30*(12), 3514–3554. https://doi.org/10.1108/IJCHM-07-2017-0461

Muniasamy, A., Tabassam, S., Hussain, M. A., Sultana, H., Muniasamy, V., & Bhatnagar, R. (2020). Deep Learning for Predictive Analytics in Healthcare. In *Advances in Intelligent Systems and Computing* (Vol. 921, hal. 32–42). Springer Verlag. https://doi.org/10.1007/978-3-030-14118-9_4

Nogueira, I. D., Romdhane, M., & Darmont, J. (2018). Modeling Data Lake Metadata with a Data Vault. *Proceedings of the 22nd International Database Engineering & Applications Symposium on - IDEAS 2018*, 253–261. https://doi.org/10.1145/3216122.3216130

Park, S. B., Jang, J., & Ok, C. M. (2016). Analyzing Twitter to explore perceptions of Asian restaurants. *Journal of Hospitality and Tourism Technology*, *7*(4), 405–422. https://doi.org/10.1108/JHTT-08-2016-0042

Ratia, M., Myllärniemi, J., & Helander, N. (2018). The new era of business intelligence. *Meditari Accountancy Research*, *26*(3), 531–546. https://doi.org/10.1108/MEDAR-08-2017-0200

Scholtz, B., Calitz, A., & Haupt, R. (2018). A business intelligence framework for sustainability information management in higher education. *International Journal of Sustainability in Higher Education*, *19*(2), 266–290. https://doi.org/10.1108/IJSHE-06-2016-0118

Shi, S., Wang, Q., Xu, P., & Chu, X. (2016). Benchmarking State-of-the-Art Deep Learning Software Tools. *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, 99–104. https://doi.org/10.1109/CCBD.2016.029

Swan, M. (2018). Blockchain for Business: Next-Generation Enterprise Artificial Intelligence Systems. In *Advances in Computers* (Vol. 111, hal. 121–162). Academic Press Inc. https://doi.org/10.1016/bs.adcom.2018.03.013

Voigt, P., & von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). In *Information Governance Alliance*. Springer International Publishing.

https://doi.org/10.1007/978-3-319-57959-7
Vujošević, D., Kovačević, I., & Vujošević-Janičić, M. (2019). The learnability of the dimensional view of data and what to do with it. *Aslib Journal of Information Management*, *71*(1), 38–53. https://doi.org/10.1108/AJIM-05-2018-0125
Yingjie, L., Deng, S., & Pan, T. (2019). Does usage of enterprise social media affect employee turnover? Empirical evidence from Chinese companies. *Internet Research*, *29*(4), 970–992. https://doi.org/10.1108/INTR-03-2018-0140
Yusuf, M. (2014). *Metode Penelitian Kuantitatif, Kualitatif & Penelitian Gabungan* (4 ed.). Kencana.

JISKa, Vol. 6, No. 2, MEI, 2021, Pp. 113 – 119
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Deteksi Dini Mahasiswa *Drop Out* Menggunakan C5.0

Ulfi Saidata Aesyi [(1)*], Alfirna Rizqi Lahitani [(2)], Taufaldisatya Wijatama Diwangkara [(3)], Riyanto Tri Kurniawan [(4)]
[1,3,4] Sistem Informasi S-1, Fakultas Teknik dan Teknologi Informasi, Universitas Jenderal Achmad Yani, Yogyakarta
[2] Sistem Informasi D-3, Fakultas Teknik dan Teknologi Informasi, Universitas Jenderal Achmad Yani, Yogyakarta
e-mail : {ulfiaesyi,alfirnarizqi,t.wijatama.d,riyantotri14}@gmail.com.
* Penulis korespondensi.
Artikel ini diajukan 7 Oktober 2020, direvisi 18 November 2020, diterima 19 November 2020, dan dipublikasikan 3 Mei 2021.

Abstract

The decline in the number of active students also occurred at the Faculty of Engineering and Information Technology, Universitas Jenderal Achmad Yani. This greatly affects the profile of study program graduates. So it is necessary to have a system that can detect students who are threatened with dropping out early. In this study, the attributes chosen were the student's GPA and the percentage of attendance. This attribute is used to classify students who are predicted to drop out. The research data use student data from the Faculty of Engineering and Information Technology, Universitas Jenderal Achmad Yani. This study uses the C5.0 algorithm to build a decision tree to assist data classification. The decision tree that was built with 304 data as training data resulted from a C5.0 decision tree which had an error rate of 5%. The accuracy results obtained from the 76 test data are 93%.

Keywords: Drop Out, Prediction, C5.0, Classification, Decision Tree

Abstrak

Penurunan jumlah mahasiswa aktif juga terjadi pada Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani. Hal ini sangat mempengaruhi profil lulusan program studi. Sehingga perlu adanya sebuah sistem yang mampu mendeteksi mahasiswa yang terancam drop out secara dini. Pada penelitian ini, atribut yang dipilih adalah IPK mahasiswa dan persentase kehadiran. Atribut tersebut digunakan untuk mengklasifikasikan mahasiswa yang terprediksi drop out. Data penelitian menggunakan data mahasiswa Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani. Penelitian ini menggunakan algoritma C5.0 untuk membangun pohon keputusan untuk membantu klasifikasi data. Pohon Keputusan yang dibangun dengan 304 data sebagai data latih menghasilkan pohon keputusan C5.0 yang memiliki tingkat error sebesar 5%. Hasil akurasi yang diperoleh dari 76 data uji adalah 93%.

Kata Kunci: *Drop Out,* Prediksi*, C5.0,* Klasifikasi*, Decision Tree*

1. PENDAHULUAN

Akreditasi program studi menjadi daya tarik dalam proses penerimaan mahasiswa. Oleh karena itu, program studi selalu berusaha untuk meningkatkan akreditasinya supaya dapat menjaring mahasiswa baru (Sutanto, 2017). Beberapa faktor terpenting dalam sebuah akreditasi program studi di perguruan tinggi adalah mahasiswa dan lulusan.

Di beberapa perguruan tinggi terutama di Universitas Jenderal Achmad Yani Yogyakarta pada Fakultas Teknik dan Teknologi Informasi mengalami penurunan jumlah mahasiswa aktif. Terdapat beberapa mahasiswa yang tidak melakukan Registrasi Ulang tanpa cuti selama lebih dari 2 Semester, sehingga dari pihak Program Studi harus melakukan tindakan berupa pemberian surat peringatan atau *drop out*. Pada proses perkuliahan terdapat beberapa mahasiswa yang kehadirannya sangat kurang sehingga jumlah kehadirannya tidak cukup untuk memenuhi minimal syarat mengikuti Ujian Tengah Semester dan Ujian Akhir Semester. Hal ini mempengaruhi hasil IPK mahasiswa dan rata- rata IPK lulusan yang masuk dalam salah satu elemen penilaian (Cahyo, 2018).

Ketika seorang mahasiswa hanya melakukan KRS tanpa mengikuti perkuliahan dan ujian, hasil IPS pada akhir semester adalah 0.0 sehingga mahasiswa tersebut membutuhkan waktu lebih lama untuk lulus. Jumlah mahasiswa aktif dan lama masa studi mahasiswa ini juga menjadi salah satu komponen penilaian akreditasi (Khasanah & Harwati, 2017).

Universitas menentukan beberapa kebijakan sebagai upaya untuk mengatasi penurunan akreditasi dengan melakukan *drop out* terhadap mahasiswa-mahasiswa yang nilai IPK nya dalam beberapa semester sangat rendah dan mahasiswa yang hanya melakukan KRS tanpa mengikuti perkuliahan (Gustian & Hundayani, 2017), namun kebijakan ini juga berpengaruh terhadap jumlah mahasiswa aktif. Oleh karena itu, diperlukan upaya untuk pencegahan mahasiswa *drop out* dengan mendeteksi secara dini mahasiswa *drop out* untuk mendapatkan penangan dan perhatian khusus.

Mahasiswa akan diklasifikasikan berdasarkan hasil studi dan kehadirannya, sehingga dapat membantu pihak Program Studi untuk menentukan tindakan terhadap mahasiswa-mahasiswa yang terindikasi masuk ke klasifikasi mahasiswa yang terancam *drop out*. Pada deteksi dini mahasiswa *drop out* ini akan menggunakan algoritma C5.0. Di mana pengklasifikasian berdasarkan hasil studi dan kehadiran Algoritma C5.0 ini merupakan Algoritma modifikasi dari algoritma ID3 dan C4.5 (Mutrofin et al., 2019). Dengan menggunakan algoritma ini diharapkan akan dapat membantu dalam mendeteksi secara dini mahasiswa *drop out* di Universitas Jenderal Achmad Yani.

1.1. Tinjauan Pustaka

1.1.1. Mahasiswa dan *Drop Out*

Mahasiswa memiliki pengaruh yang cukup tinggi dalam akreditasi, sedangkan tidak sedikit Perguruan Tinggi yang menerapkan *drop out* untuk mengatasi permasalahan mahasiswa, misalnya IPK rendah, kurang serius dalam perkuliahan, hingga lama lulus. Hal ini yang membuat banyak peneliti yang melakukan penelitian terhadap faktor yang menyebabkan mahasiswa *drop out*. Penelitian yang sejenis dengan menggunakan algoritma *data mining* diterapkan pada solusi prediksi mahasiswa *drop out* dengan melihat dari sisi SKS perkuliahan, IPK dan jumlah semester yang telah dilalui (Utari et al., 2020). Dengan melakukan prediksi mahasiswa *drop out*, program studi dapat memantau mahasiswanya yang terprediksi *drop out* sehingga mahasiswa mendapatkan bimbingan secepatnya dan mahasiswa yang bersangkutan dapat lulus tepat waktu (Putra, 2017).

1.1.2. Algoritma C5.0

Algoritma C5.0 adalah salah satu algoritma *data mining* yang merupakan penyempurnaan dari algoritma ID3 dan C4.5 (Ahmadi et al., 2018). Algoritma ini cukup banyak diterapkan dalam beberapa penelitian, seperti contohnya pada penelitian mengenai diagnosa penyakit *Disk Hernia* dan *Spondilolisthesis* menggunakan algoritma C5.0. Penelitian ini menggunakan 310 data dari UCI *Machine Learning*, di mana terdapat tiga kelas klasifikasi yaitu Normal, *Disk Hernia*, dan *Spondylolisthesis*. Hasil penelitian menunjukkan bahwa algoritma C5.0 mampu melakukan identifikasi dengan akurasi sebesar 79%. Kemudian pohon keputusan yang dihasilkan Algoritma C5.0 dimaksimalkan dengan menggunakan algoritma *AdaBoost*, sehingga akurasi meningkat menjadi 83% (Aesyi et al., 2020).

Penelitian serupa dengan melakukan kombinasi beberapa algoritma diaplikasikan pada sistem deteksi intrusi (IDS) dalam jaringan komputer dengan tujuan untuk meningkatkan kualitas deteksi. Sulit untuk membedakan koneksi yang tidak sah dari yang resmi karena penyusup bertindak mirip dengan pengguna normal. Dalam algoritma yang diusulkan, integrasi *Tree Augmented Naive Bayes* (TAN) dalam *Bayesian Network* (BN) dan *Boosting* dalam struktur pohon keputusan C5.0 digunakan untuk mengambil keuntungan dan menghindari kelemahan. Hasil percobaan menunjukkan bahwa algoritma yang diusulkan tidak hanya mencapai hasil yang memuaskan dalam akurasi dan mampu mengurangi tingkat alarm palsu (presentasi kejadian yang

diklasifikasikan sebagai jenis serangan dalam jaringan komputer tetapi tingkat serangan masih termasuk normal), tetapi juga meningkatkan pekerjaan yang ada (Nia & Khalili, 2015).

Algoritma C5.0 pernah digunakan untuk penilaian Kinerja Pegawai Negeri Sipil dengan hasil akurasi 98.08% (Kastawan et al., 2018). Kedudukan dan peranan pegawai negeri sipil sebagai abdi masyarakat mengharuskan pegawai untuk memberikan pelayanan yang adil kepada masyarakat. Peran tersebut membuat kinerja pegawai menjadi hal yang penting. Penggunaan algoritma C5.0 untuk memproses data kinerja pegawai, mampu menghasilkan prediksi atau masukan dalam memberikan rekomendasi jabatan, kepangkatan maupun pemberian tunjangan gaji.

Selain itu, C5.0 juga dapat digunakan untuk menemukan atribut penting dalam kumpuan data. Kumpulan data mungkin mengandung banyak atribut yang bekerja semua. Akan tetapi dengan semua atribut kemungkinan hasil dapat tidak efisien, karenanya hanya atribut penting yang harus dipertimbangkan untuk dipilih dalam proses klasifikasi dan pengelompokkan pola data yang besar. Studi kasus dalam karya ini adalah eksplorasi atribut dengan menemukan akurasi dalam *dataset* kanker payudara (Ojha et al., 2017).

Tahap penentuan *root node* algoritma C5.0 (Kastawan et al., 2018):

1) Menghitung *entropy* yang digunakan untuk mengklasifikasikan label kelas dari *tuple* acak di D. Pada variable pi merupakan peluang yang bukan nol dari dari *tuple* acak di D. Penggunaan log basis 2 dalam menghitung *entropy* adalah untuk pengkodean ke dalam bit.

$$Info(D) = -\sum_{i=1}^{m} pi\ log_2(pi) \quad (1)$$

2) Hasil perhitungan *entropy* digunakan untuk menghitung *information gain* pada atribut A.

$$Info_A(D) = \sum_{J=I}^{Y} \frac{|Dj|}{D} \times Info(Dj) \quad (2)$$

3) Setelah itu dilanjutkan dengan menghitung *gain* pada atribut A.

$$GAIN\ (A) = Info(D) - Info(DJ) \quad (3)$$

Hasil *information gain* tertinggi akan dipilih menjadi *node*.

2. METODE PENELITIAN

Tahap awal penelitian ini adalah dengan melakukan analisis permasalahan yang ada kemudian melakukan pemetaan solusi dengan menggunakan data yang sesuai dan yang tersedia (Morales et al., 2017). Pada metode penelitian dalam penelitian ini berupa pembuatan aplikasi dan pengujian terhadap aplikasi tersebut. Adapun tujuan yang dicapai untuk membantu mengklasifikasikan mahasiswa *drop out* secara dini berdasarkan hasil studi dan kehadiran menggunakan algoritma C5.0.

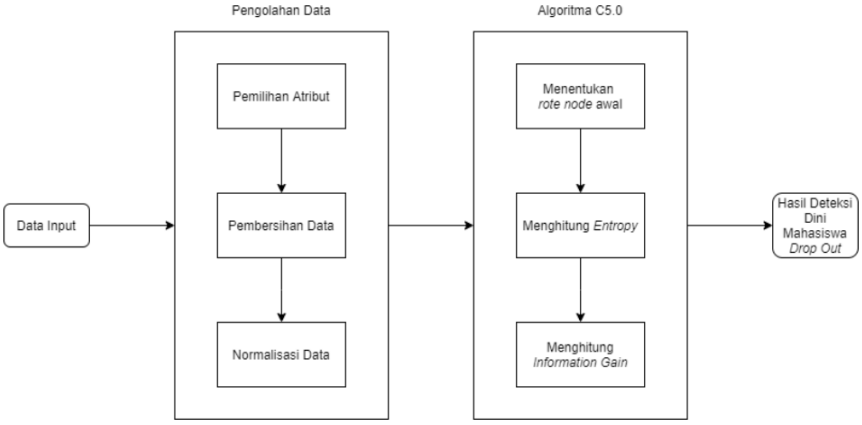
Dengan melihat penelitian yang sebelumnya, deteksi *drop out* banyak menggunakan metode *Naïve Bayes*, ID3 dan C4.5. Sedangkan dalam beberapa pengujian pohon keputusan, algoritma C5.0 mencapai hasil akurasi yang paling baik dibanding ID3 dan C4.5 (Rajeswari & Suthendran, 2019). Oleh karena itu penelitian ini mengakomodasi algoritma C5.0 sebagai solusi pohon keputusan untuk melakukan deteksi dini mahasiswa *drop out* dengan menggunakan fitur diantaranya adalah hasil studi dan jumlah kehadiran.

Penelitian dilakukan pada Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani Yogyakarta. Tahap awal peneliti akan mengumpulkan data terkait hasil studi dan presensi mahasiswa. Selanjutnya data akan diolah dengan dilakukan pemilihan atribut yang sesuai dengan penelitian yang ditunjukkan oleh Gambar 1. Kemudian data yang atributnya sudah terpilih akan dibersihkan yang nantinya data akan dinormalisasikan.

Setelah data dinormalkan, maka data akan masuk ke proses pengklasifikasian menggunakan algoritma C5.0. Langkah awal dari algoritma ini adalah dengan menentukan *root node* awal dari atribut yang ada. Kemudian setelah *root node* sudah ditentukan, maka menghitung *entropy* dari atribut-atribut yang akan dicari *node* selanjutnya. Setelah semua atribut telah dihitung *entropy*-nya, maka langkah selanjutnya adalah dengan menghitung *Information Gain*. Nilai *information gain* tertinggi akan dijadikan *node* selanjutnya. Kemudian kembali lagi dengan menghitung *entropy* dan *information gain* dari atribut tersisa. Proses ini berlangsung hingga *node* akhir (Susanti et al., 2019).

Kemudian tahap berikutnya adalah melakukan analisis terhadap data dengan melakukan klasifikasi dengan menggunakan metode algoritma C5.0. Hasil dari tahap analisis ini akan dilakukan pembahasan terhadap hasil klasifikasi dari data yang ada yang nantinya akan mengelompokkan mahasiswa yang masuk dalam kelas *Drop out* (Alban & Mauricio, 2019).

Setelah hasil klasifikasi keluar, maka data mahasiswa yang terprediksi DO akan diperoleh. Tahap akhir dari penelitian ini adalah melakukan pengujian dan evaluasi sistem. Hal ini dilakukan untuk melihat performa sistem yang dibangun.



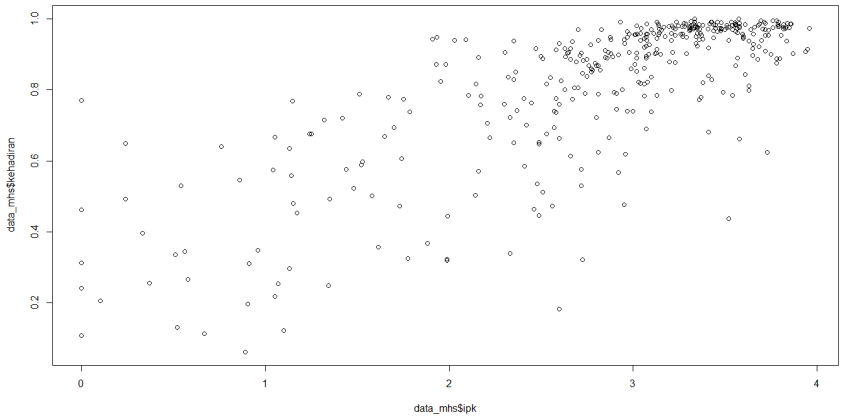
Gambar 1. Desain Sistem.

3. HASIL DAN PEMBAHASAN

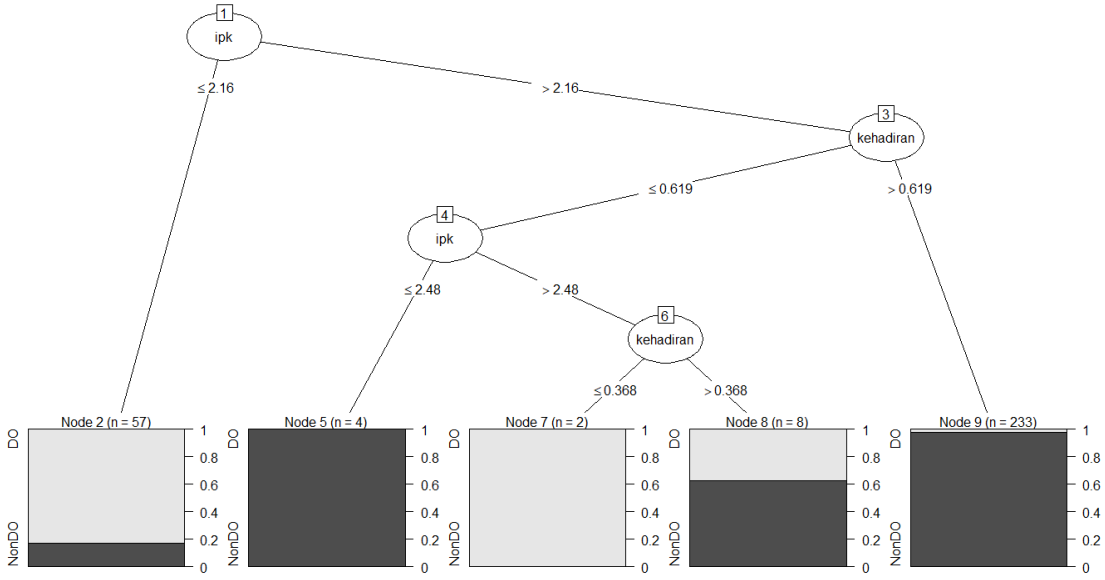
Penelitian ini mengambil data pada Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani Yogyakarta. Data yang diperoleh adalah data mahasiswa angkatan 2012, angkatan 2013, angkatan 2015, angkatan 2017, dan angkatan 2018. Pengarsipan data yang belum rapi menjadi kendala dalam pengumpulan data, sehingga hanya terkumpul data sebanyak 380 data.

Atribut yang digunakan dalam penelitian ini adalah IPK dan kehadiran. Dari seluruh data yang diperoleh, dilakukan proses pembersihan data sehingga data yang dapat digunakan untuk penelitian sebanyak 380 data. Dari data yang diperoleh kemudian dilakukan proses normalisasi data termasuk pengelompokan data berdasarkan label DO dan NonDo yang digunakan sebagai hasil prediksi. Hasil dari normalisasi dan pelabelan data menghasilkan 308 data mahasiswa NonDO dan 72 data mahasiswa DO. Distribusi data berdasarkan atribut IPK dan kehadiran dari 380 data yang dapat dilihat pada Gambar 2.

Proses pembuatan pohon keputusan dimulai dengan pembagian data menjadi data latih dan data uji secara *random* dengan perbandingan 80% dan 20% yaitu 304 data latih dan 76 data uji. Dari data latih tersebut, dibangun model pohon keputusan dengan menentukan *root node* terlebih dahulu. *Root note* ditentukan berdasarkan *information gain* tertinggi dari atribut. Pada penelitian *information gain* tertinggi adalah atribut IPK. Kemudian *node* lainnya akan terbentuk sehingga membentuk pohon keputusan seperti pada Gambar 3.



Gambar 2. Distribusi Data.



Gambar 3. Pohon Keputusan Menggunakan Algoritma C5.0.

Pohon keputusan yang terbentuk memiliki *error* sebesar 5.9%. Klasifikasi data dari pohon keputusan dibagi menjadi 2, yaitu Do dan NonDo. Penggunaan atribut IPK sebesar 100% dan atribut kehadiran sebesar 81.25%. Hasil klasifikasi ditunjukkan pada Tabel 1.

Tabel 1. Hasil Klasifikasi

	DO	**NonDO**
DO	49	8
NonDO	10	237

Pada tahap selanjutnya dilakukan untuk mengetahui seberapa tepat sistem melakukan prediksi. Pengujian menggunakan data uji sebanyak 76 data. Dari hasil pengujian diperoleh data 71 terprediksi benar yang dapat dilihat pada Tabel 2.

Tabel 2. Hasil Prediksi

	DO	NonDO
DO	11	3
NonDO	2	60

Dari hasil prediksi dapat dihitung dengan melakukan perbandingan jumlah diagnosis benar (x) oleh sistem dengan jumlah data (n) (Aesyi & Wardoyo, 2019):

$$Akurasi = \frac{x}{n} X\ 100\% \qquad (4)$$

Perhitungan akurasi menggunakan rumus tersebut menghasilkan akurasi sebesar 93%.

4. KESIMPULAN

Berdasarkan hasil penelitian tersebut, maka dapat disimpulkan:

1) Penelitian ini telah membangun sistem yang mampu menghasilkan pohon keputusan mengunakan algoritma C5.0 untuk membantu mengklasifikasikan mahasiswa yang terprediksi *drop out* secara dini. Pohon keputusan tersebut memiliki tingkat *error* sebesar 5%.
2) Penggunaan data latih 80% dan data uji 20% secara *random* menghasilkan akurasi sebesar 93%. Hal ini menunjukkan bahwa performa dari pohon keputusan tersebut cukup baik untuk melakukan prediksi dini mahasiswa *drop out*.

UCAPAN TERIMA KASIH

Ucapan terimakasih ini ditujukan kepada Direktorat Penelitian dan Pengabdian kepada Masyarakat Direktorat Jenderal Penguatan Penelitian dan Pengembangan (DRPM) Kementerian Riset dan Pendidikan Tinggi (Kemristekdikti) Republik Indonesia yang telah memberikan kesempatan kepada tim peneliti untuk melakukan penelitian pada tahun 2020 skema Penelitian Dosen Pemula (PDP). Tim penelitian juga mengucapkan terima kasih kepada Fakultas Teknik dan Teknologi Informasi Universitas Jenderal Achmad Yani Yogyakarta yang telah membantu tim peneliti dalam penyediaan data penelitian.

DAFTAR PUSTAKA

Aesyi, U. S., Diwangkara, T. W., & Kurniawan, R. T. (2020). Diagnosa Penyakit Disk Hernia Dan Spondylolisthesis Menggunakan Algoritma C5. *Telematika*, *16*(2), 81. https://doi.org/10.31315/telematika.v16i2.3181

Aesyi, U. S., & Wardoyo, R. (2019). Prediction of Length of Study of Student Applicants Using Case Based Reasoning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(1), 11. https://doi.org/10.22146/ijccs.28076

Ahmadi, E., Weckman, G. R., & Masel, D. T. (2018). Decision making model to predict presence of coronary artery disease using neural network and C5.0 decision tree. *Journal of Ambient Intelligence and Humanized Computing*, *9*(4), 999–1011. https://doi.org/10.1007/s12652-017-0499-z

Alban, M., & Mauricio, D. (2019). Predicting University Dropout trough Data Mining: A systematic Literature. *Indian Journal of Science and Technology*, *12*(4), 1–12. https://doi.org/10.17485/ijst/2019/v12i4/139729

Cahyo, P. W. (2018). Klasterisasi Tipe Pembelajar Sebagai Parameter Evaluasi Kualitas Pendidikan Di Perguruan Tinggi. *Teknomatika*, *11*(1), 49–55.

Gustian, D., & Hundayani, R. D. (2017). Combination of AHP Method with C4.5 in the level classification level out students. *2017 International Conference on Computing, Engineering, and Design (ICCED)*, 1–6. https://doi.org/10.1109/CED.2017.8308098

Kastawan, P. W., Wiharta, D. M., & Sudarma, M. (2018). Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil. *Majalah Ilmiah Teknologi Elektro*, *17*(3), 371. https://doi.org/10.24843/MITE.2018.v17i03.P11

Khasanah, A. U., & Harwati. (2017). A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques. *IOP Conference Series: Materials Science and Engineering*, *215*, 012036. https://doi.org/10.1088/1757-899X/215/1/012036

Morales, A. C., Amir, O., & Lee, L. (2017). Keeping It Real in Experimental Research—Understanding When, Where, and How to Enhance Realism and Measure Consumer Behavior. *Journal of Consumer Research*, *44*(2), 465–476. https://doi.org/10.1093/jcr/ucx048

Mutrofin, S., Khalimi, A. M., Kurniawan, E., Ginardi, R. V. H., Fatichah, C., & Sari, Y. A. (2019). Detection of Potentially Students Drop Out of College in Case of Missing Value Using C4.5. *2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC)*, 349–354. https://doi.org/10.1109/ICSECC.2019.8907014

Nia, F. Y., & Khalili, M. (2015). An efficient modeling algorithm for intrusion detection systems using C5.0 and Bayesian Network structures. *2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, 1117–1123. https://doi.org/10.1109/KBEI.2015.7436203

Ojha, U., Jain, M., Jain, G., & Tiwari, R. K. (2017). Significance of important attributes for decision making using C5.0. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–4. https://doi.org/10.1109/ICCCNT.2017.8204031

Putra, A. (2017). SOLUSI PREDIKSI MAHASISWA DROP OUT PADA PROGRAM STUDI SISTEM INFORMASI FAKULTAS ILMU KOMPUTER UNIVERSITAS BINA DARMA. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, *8*(1). https://doi.org/10.24176/simet.v8i1.893

Rajeswari, S., & Suthendran, K. (2019). C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. *Computers and Electronics in Agriculture*, *156*, 530–539. https://doi.org/10.1016/j.compag.2018.12.013

Susanti, Y., Respatiwulan, Handajani, S. S., Pratiwi, H., Slamet, I., Hartatik, & Istiqomah, F. (2019). Classification of teak wood production in Central Java using the C5.0 algorithm. *AIP Conference Proceedings*, *2202*(1), 020094. https://doi.org/10.1063/1.5141707

Sutanto, E. M. (2017). The influence of organizational learning capability and organizational creativity on organizational innovation of Universities in East Java, Indonesia. *Asia Pacific Management Review*, *22*(3), 128–135. https://doi.org/https://doi.org/10.1016/j.apmrv.2016.11.002

Utari, M., Warsito, B., & Kusumaningrum, R. (2020). Implementation of Data Mining for Drop-Out Prediction using Random Forest Method. *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 1–5. https://doi.org/10.1109/ICoICT49345.2020.9166276

JISKa, Vol. 6, No. 2, MEI, 2021, Pp. 120 – 129
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Perbandingan Algoritma Klasifikasi Sentimen Twitter Terhadap Insiden Kebocoran Data Tokopedia

Nadhif Ikbar Wibowo [(1)], Tri Andika Maulana [(2)], Hamzah Muhammad [(3)], Nur Aini Rakhmawati [(4)*]
Sistem Informasi, Fakultas Teknologi Elektro dan Informatika Cerdas, Institut Teknologi Sepuluh Nopember, Surabaya
e-mail : {nadhif.18052,tri.18052,hamzah.18052}@mhs.its.ac.id, nur.aini@is.its.ac.id
* Penulis korespondensi.
Artikel ini diajukan 7 November 2020, direvisi 10 Januari 2021, diterima 24 Januari 2021, dan dipublikasikan 3 Mei 2021.

Abstract

Public responses, posted on Twitter reacting to the Tokopedia data leak incident, were used as a data set to compare the performance of three different classifiers, trained using supervised learning modeling, to classify sentiment on the text. All tweets were classified into either positive, negative, or neutral classes. This study compares the performance of Random Forest, Support-Vector Machine, and Logistic Regression classifier. Data was scraped automatically and used to evaluate several models; the SVM-based model has the highest f1-score 0.503583. SVM is the best performing classifier.

Keywords: Model Performance Analysis, Sentiment Classification, Logistic Regression, Random Forest, Support Vector Machine

Abstrak

Twit respon masyarakat terhadap insiden kebocoran data Tokopedia di Twitter dimanfaatkan sebagai *dataset* untuk melakukan perbandingan performa tiga *classifier* berbeda dengan pemodelan *supervised learning* untuk melakukan klasifikasi sentimen pada teks. Setiap *twit* diklasifikasikan dalam salah satu dari tiga kelas, yaitu positif, negatif, atau netral. Penelitian ini membandingkan performa dari *classifier Random Forest*, *Support-Vector Machine*, dan *Logistic Regression*. Data *twit* diambil secara otomatis menggunakan *scrapper* dan digunakan untuk melakukan evaluasi model. Model *classifier Support Vector Machine* memiliki performa terbaik dengan *f1-score* sebesar 0.503583. SVM adalah *classifier* dengan performa terbaik.

Kata Kunci: Analisis Performa Model, Klasifikasi Sentimen, Logistic Regression, Random Forest, Support Vector Machine

1. PENDAHULUAN

Data merupakan catatan atas kumpulan fakta (Vardiansyah, 2008). Saat menggunakan internet, keamanan data menjadi aspek penting yang perlu diperhatikan. Khususnya ketika berurusan dengan data-data pribadi yang bersifat sensitif, kelalaian akan keamanan data pribadi dapat menimbulkan masalah-masalah yang berkaitan dengan privasi seseorang yang dapat menimbulkan berbagai macam kerugian dengan skala yang beragam. Data pribadi menjadi salah satu incaran utama penjahat siber. Tren serangan terhadap data bukan hanya sekedar pencurian, tetapi juga jual-beli data. Serangan tidak hanya ditujukan kepada individu namun juga industri atau organisasi (Librianty, 2016).

Salah satu insiden besar kebocoran data di Indonesia adalah bocornya 91 juta akun pengguna dan 7 juta akun pedagang pada *marketplace* Tokopedia yang terjadi pada bulan Mei tahun 2020. Data-data yang bocor mulai dari nama lengkap, tanggal lahir, nomor ponsel, lokasi, hingga jenis kelamin (CNN Indonesia, 2020). Insiden ini menuai berbagai respon dari masyarakat yang dapat diamati melalui cuitan-cuitan yang ditulis pada media sosial Twitter.

Pada penelitian ini, penulis memanfaatkan respon-respon masyarakat di Twitter untuk membuat model-model klasifikasi sentimen yang dapat menentukan apakah respon terhadap insiden

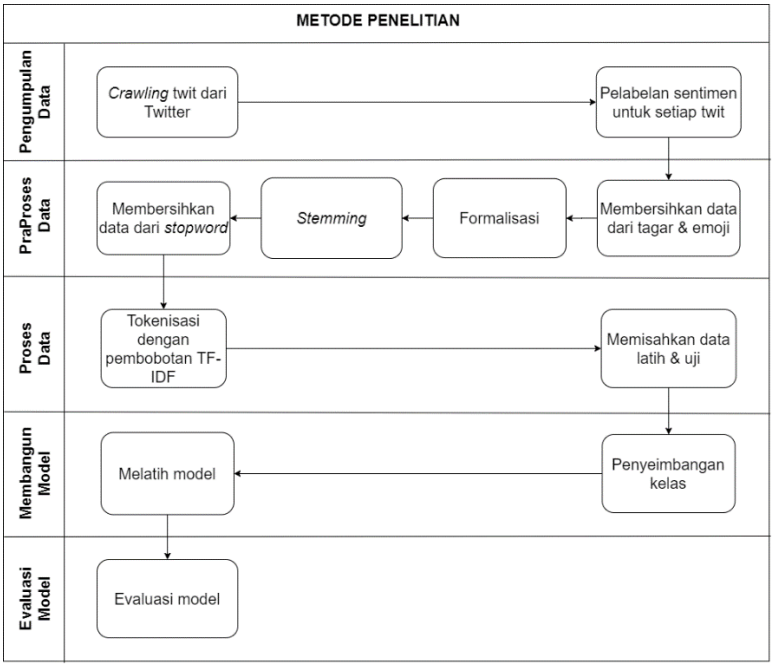
tersebut memiliki sentimen positif, negatif, atau netral. Model-model tersebut akan digunakan untuk melakukan analisis sentimen.

Masing-masing model menggunakan *classifier* yang berbeda dengan pemodelan *supervised learning*, selanjutnya model-model tersebut dibandingkan untuk mencari *classifier* terbaik. Penelitian ini membandingkan performa dari 3 *classifier* yang digunakan untuk membuat model yaitu *Random Forest, Support-Vector Machine,* dan *Logistic Regression.*

Beberapa studi sebelumnya yang membahas tentang analisis sentimen terhadap suatu topik dengan memanfaatkan data *twit* telah dilakukan. Berdasarkan (Faradhillah et al., 2016) penulis menggunakan data *twit* warga Surabaya untuk membangun model klasifikasi sentiment terhadap kinerja Pemkot Surabaya yang hasilnya ditampilkan secara interaktif dengan aplikasi berbasis web yaitu R Shiny. Didapatkan model SVM memiliki akurasi lebih baik dibanding Naïve Bayes dengan nilai akurasi 78,66% dengan menggunakan kernel RBF. Berdasarkan (Deviyanto & Wahyudi, 2018) penulis menggunakan data *twit* mengenai pemilihan gubernur DKI Jakarta tahun 2017 untuk membangun model klasifikasi sentiment menggunakan algoritma KNN. Penelitian tersebut menggunakan 2000 data *twit* dan didapatkan hasil akurasi model KNN yaitu 67,2%. Berdasarkan (Hasan et al., 2018) penelitian dilakukan analisis stentimen dengan mamanfaatkan *classifier* Naïve Beyes dengan hasil akurasi 62% pada klasifikasi sentimen akun Twitter dengan bahasa Urdu yang diterjemahkan terlebih dahulu.

2. METODE PENELITIAN

Metode penelitian secara garis besar terdiri atas 5 fase meliputi pengumpulan data, pra proses data, proses data, membangun model, dan evaluasi model. Secara lebih jelas, metode penelitian ini digambarkan oleh Gambar 1.



Gambar 1. Diagram Metode Penelitian.

2.1. Pengumpulan Data

Data diambil dengan memanfaatkan fitur *Explore* yang terdapat pada Twitter, fitur ini mendukung kueri pencarian *twit* yang sederhana, di mana hanya memanfaatkan serangkaian kata kunci, mapun kueri tingkat lanjut untuk membuat aturan pencarian yang menerapkan kondisi tertentu.

Kueri sederhana yang digunakan: “Tokopedia bocor”, “Kebocoran data Tokopedia”, “Tokopedia diretas”. Kueri tingkat lanjut kemudian dimanfaatkan untuk mendapatkan *twit* balasan dari *thread* yang memiliki banyak interaksi *reply*. Kueri tingkat lanjut yang digunakan: “Tokopedia bocor to:secgron”, “Tokopedia bocor to:tokopedia”. Contoh *twit* dengan banyak interaksi *reply* bisa dilihat pada Gambar 2.

Gambar 2. Contoh *Twit* dengan Banyak Interaksi *Reply*.

Pengambilan data dilakukan secara otomatis menggunakan *crawler* yang merupakan *fork package* Python TweetScrapper oleh jonbakerfish, yang kemudian dikembangkan lebih lanjut oleh penulis. Sumber kode *crawler* yang digunakan pada penelitian ini dapat diakses di https://github.com/nadhifikbarw/ep-scrapper (Wibowo, 2020).

Untuk menjalankan *crawler* digunakan perintah sebagai berikut yang akan diulang untuk semua kueri yang telah ditetapkan.

scrapy crawl TweetScrapper -a query="Tokopedia bocor"

Data masing-masing *twit* yang diambil oleh *scrapper* disimpan pada file *plaintext* dengan format *JSON* menggunakan nama sesuai dengan *conversation_id* dari *twit* tersebut (tanpa ekstensi file), sehingga tidak ada data duplikat yang tersimpan. Seluruh data *twit* tersimpan pada folder./Data/tweets.

Setelah *scrapping* selesai dilakukan untuk setiap kueri, seluruh data diagregasi menjadi satu file CSV untuk memudahkan pemrosesan lebih lanjut. Data yang digunakan pada penelitian ini berjumlah 1060 data *twit*. Masing-masing *twit* diberi label positif, negatif, netral, atau tidak relevan sesuai dengan sentimennya. Jika data tidak relevan maka data *twit* tersebut dikeluarkan dari *dataset* (Maulana et al., 2020). Setelah data yang tidak relevan dikeluarkan data yang tersisa berjumlah 494 *twit*. Rekapitulasi data *twit* yang telah diberi label dapat dilihat pada Tabel 1.

Tabel 1. Jumlah Data *Twit* Teragregasi.

Data *Twit*	Jumlah
Twit sentimen positf	15
Twit sentimen negatif	318
Twit sentimen netral	161
Total	494

Adapun contoh *twit* untuk masing-masing kelas dapat dilihat pada Tabel 2.

Tabel 2. Contoh Data *Twit* Teragregasi.

No	*Twit*	Label
1	@TokopediaCare bagaimana bisa perusahaan sebesar Tokopedia datanya bisa bocor gw jadi ragu belanja di tokped	Negatif
2	Enggak ada salahnya mengucapkan terima kasih dan apresiasi. Untuk isu kebocoran data ini, ya jangan panik namun tetap waspada. @tokopedia big thanks for your service. Panjang umur.	Positif
3	15 Juta Data Pengguna Tokopedia Diinformasikan Bocor, Ini Daftar Email & Password Sebagian Korban https://t.co/fkIFEfgng0	Netral
4	Kementerian Komunikasi dan Informatika (Kemenkominfo) akan segera memanggil Direksi Tokopedia. Pemanggilan ini terkait dugaan kebocoran data pribadi 91 juta akun pengguna layanan ecommerce itu.	Netral
5	@ezash @tokopedia Terima kasih bang Eca atas infonyaa	Tidak Relevan

2.2. Pra-Proses Data

Pada fase ini penulis membuat program Python bernama Data Processor (Wibowo, 2020), yang dapat ditemukan pada *repository* yang sama dengan program *crawler*, untuk melakukan proses ETL (*extraction, transformation, load)*. Pada tahapan *transformation* setiap *twit* dibersihkan dari tagar, *username*, alamat tautan, angka, tanda baca, dan emoji setelah itu dilakukan *case folding*. Proses pembersihan data ini memiliki tujuan untuk menghilangkan *noise* yang terdapat pada data mentah dari proses *scrapping*. Menurut (Tang et al., 2005) proses ini dapat mempengaruhi performa dari model yang dihasilkan terutama pada metrik akurasi.

Keseluruhan *twit* kemudian diformalisasi dari kata yang tidak baku, singkatan. Lalu data dibersihkan kembali dari *stopword* karena kata-kata tersebut tidak memberikan kontribusi pada sentimen. Setelah itu data *twit* ditransformasi *stemming* untuk mendapatkan kata dasar dari setiap kata yang ada di data *twit*.

2.3. Proses Data

Data *twit* dilakukan tokenisasi dalam bentuk *unigram* (satu kata) lalu dilakukan pembobotan dengan metode TF-IDF. *Term Frequency* (TF) mengambarkan banyaknya kemunculan suatu kata pada suatu dokumen, dalam hal ini suatu *twit*. *Inverse Document Frequency* (IDF) mengambarkan prioritas suatu kata dari keseluruhan dokumen, dalam hal ini keseluruhan *twit* (Beel et al., 2017). Selanjutnya seluruh *twit* yang telah dilakukan tokenisasi tersebut dibagi dua menjadi data latih dan data uji dengan perbandingan 80% data latih dan 20% data uji.

2.4. Membangun Model

Menyeimbangkan jumlah ketiga kelas dengan menggunakan teknik SMOTE pada data latih. Metode SMOTE melakukan *oversampling* pada kelas minoritas, dengan membuat data sintetis dari kelas minoritas sehingga jumlahnya sama dengan jumlah data kelas mayoritas. Penelitian sebelumnya menunjukan bahwa teknik SMOTE berhasil meningkatkan performa akurasi dari model (Chawla et al., 2002).

Pembentukan model dilakukan dengan memanfaatkan 3 algoritma *classifier*:

2.4.1. Random Forest

Random Forest adalah algoritma non-parametrik. Algoritma ini merupakan bentuk dari metode *ensemble*, yaitu metode yang mengabungkan (*voting*) hasil dari beberapa model yang lebih sederhana. Dengan banyaknya model yang digabungkan maka hasil klasifikasi dapat lebih baik (VanderPlas, 2016).

2.4.2. Logistic Regression

Logistic Regression merupakan metode statisik yang mirip dengan *Linear Regression* karena metode ini menemukan persamaan bersifat logistik yang memprediksi hasil untuk variabel biner (Y) dari satu atau lebih variabel respon (X). Namun, variabel respon (X) dapat bersifat kategorikal atau kontinu (Hoffman, 2019).

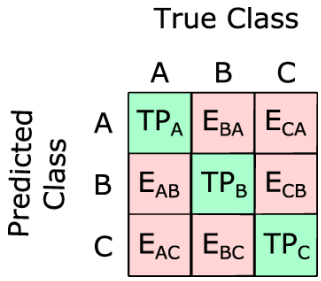
2.4.3. Support Vector Machine

Support Vector Machine merupakan mesin pembelajaran universal yang bisa diterapkan pada regresi maupun pengenalan pola (*pattern recognition*). SVM menggunakan perangkat yang disebut pemetaan kernel (*kernel mapping*) untuk memetakan data dalam ruang input ke ruang fitur berdimensi tinggi di mana masalah menjadi dapat dipisahkan secara linier (Zhang et al., 2004).

2.5. Evaluasi Model

Mengukur performa klasifikasi dari model-model yang dibangun dilakukan dengan memanfaatkan metrik-metrik pengukuran performa yang diturunkan dari pemetaan *confusion matrix.* Berdasarkan (Tharwat, 2020) performa klasifikasi suatu model dapat diwakili oleh nilai skalar seperti metrik-metrik turunan *confusion matrix* layaknya *precision*, *recall*, dan *f1-score*.

Confusion matrix adalah tabel *cross-tabulation* yang memetakan ukuran seberapa baik prediksi model klasifisikasi dibandingkan dengan hasil prediksi model (Lanham & Bedinelli, 2015). Model *confusion matrix* dengan banyak kelas dapat dilihat pada Gambar 3.



Gambar 3. *Confusion Matrix* Banyak Kelas (Tharwat, 2020).

Sumbu diagonal yang berwarna hijau merepresentasikan jumlah prediksi benar sedangkan sel berwarna pink mengindikasikan jumlah prediksi salah yang dihasilkan oleh model. Ketika sampel positif diklasifikasikan dengan kelas positif maka prediksi tersebut *true positive* (TP). Ketika sampel positif diklasifikasikan dengan kelas negatif maka prediksi tersebut *false negative* (FN) atau disebut *Type II error*. Apabila sampel negatif diklasifikasikan dengan kelas positif maka prediksi tersebut *false positive* (FP) yang merupakan *false alarm* atau *Type I error*. Ketika sampel negatif diklasifikasikan dengan kelas negatif maka prediksi tersebut *true negative* (TN). *Confusion matrix* ini digunakan untuk menghitung berbagai macam metrik performa model.

Dalam melakukan perhitungan performa model klasifikasi multi kelas terdapat dua metode dalam melakukan perhitungan metrik, dengan menghitung rata-rata dari metrik yang sama yang dihitung untuk seluruh *classifier* atau yang disebut *macro-average*. Metode kedua dengan menentukan nilai TP, FN, dan TN kumulatif dan kemudian menghitung ukuran kinerja, metode ini disebut dengan metode *micro-averaging.* Metode *macro-averaging* memperlakukan semua kelas secara setara sementara *micro-averaging* lebih menguntungkan kelas yang memiliki sampel yang lebih besar (Sokolova & Lapalme, 2009). Pada studi ini digunakan metode *macro-averaging* untuk mengantisipasi ketidakseimbangan data pada tiap kelas sehingga sehingga nilai metrik akan merefleksikanan performa model ketika memiliki performa buruk dalam klasifkasi suatu kelas.

Sensitivitas, *true positive rate* (TPR), *hit rate*, atau *recall*, adalah metrik mewakili sampel yang diklasifikasikan dengan benar positif dari seluruh jumlah total sampel positif (Tharwat, 2020), dan metrik ini memiliki formula sebagai berikut:

$$Recall_M = \frac{\sum_{i=1}^{i} \frac{TP_i}{(TP_i + FN_i)}}{i} \quad (1)$$

Metrik komplemen lain yang dapat digunakan adalah presisi atau yang juga disebut Nilai Prediksi Positif (PPV) mewakili proporsi sampel positif yang diklasifikasikan dengan benar ke jumlah total sampel yang diprediksi positif seperti yang dihitung menggunakan formula sebagai berikut.

$$Precison_M = \frac{\sum_{i=1}^{i} \frac{TP_i}{(TP_i + FP_i)}}{i} \quad (2)$$

F-measure atay juga disebut *f1-score*, merupakan nilai yang menunjukkan rata-rata harmonik presisi dan *recall* yang dihitung menggunakan persamaaan berikut

$$F1\ Score_M = \frac{2 \times Recall_M \times\ Precison_M}{Recall_M + Precison_M} \quad (3)$$

Nilai *f-measure* berkisar dari nol hingga satu, dan nilai *f-measure* yang tinggi menunjukkan kinerja klasifikasi yang tinggi. Metrik ini sensitif terhadap perubahan dalam distribusi data (Tharwat, 2020).

3. HASIL DAN PEMBAHASAN

Dari 494 data yang relevan terdiri dari 318 data *twit* bersentimen negatif, 161 data *twit* bersentimen netral, dan 15 data *twit* bersentimen positif. Tabel 3 merupakan contoh data setelah dibersihkan dari tanda baca, emoji, dan angka.

Tabel 3. Contoh *Twit* Setelah Pra-Proses.

No	*Twit*	Label
1	bagaimana bisa perusahaan sebesar tokopedia datanya bisa bocor gw jadi ragu belanja di tokped	Negatif
2	enggak ada salahnya mengucapkan terima kasih dan apresiasi untuk isu kebocoran data ini ya jangan panik namun tetap waspada big thanks for your service panjang umur	Positif
3	juta data pengguna tokopedia diinformasikan bocor ini daftar email password sebagian korban	Netral
4	kementerian komunikasi dan informatika kemenkominfo akan segera memanggil direksi tokopedia pemanggilan ini terkait dugaan kebocoran data pribadi juta akun pengguna layanan ecommerce itu	Netral

Setelah itu *twit* dilakukan formalisasi dengan maksudkan untuk mendapatkan kata yang formal dari suatu akronim dan kata yang tidak baku. Misal, ‘yg’ diubah menjadi ‘yang’, ‘abis’ menjadi ‘habis’, dan khusus kasus ini ‘tokped’ menjadi ‘tokopedia’. *Twit* yang telah melalui proses formalisasi dapat dilihat pada Tabel 4.

Tabel 4. Contoh *Twit* Setelah Proses Formalisasi.

No	*Twit*	Label
1	bagaimana bisa perusahaan sebesar tokopedia datanya bisa bocor saya jadi ragu belanja di tokopedia	Negatif
2	tidak ada salahnya mengucapkan terima kasih dan apresiasi untuk isu kebocoran data ini iya jangan panik namun tetap waspada big terima kasih for your service panjang umur	Positif
3	juta data pengguna tokopedia diinformasikan bocor ini daftar email password sebagian korban	Netral
4	kementerian komunikasi dan informatika kemenkominfo akan segera memanggil direksi tokopedia pemanggilan ini terkait dugaan kebocoran data pribadi juta akun pengguna layanan ecommerce itu	Netral

Setelah data *twit* dilakukan *stemming* untuk mendapatkan kata dasar dari tiap kata. Misal, kata ‘mencuri’ dan ‘dicuri’ akan menjadi bentuk dasarnya yaitu ‘curi’. *Stemming* dilakukan untuk memperkecil ukuran data karena setiap kata diperoses pada bentuk dasarnya tanpa mengurangi sentimen yang terkandung pada data tersebut. Contoh *Twit* setelah *stemming* dapat dilihat pada Tabel 5.

Tabel 5. Contoh *Twit* Setelah Proses *Stemming*.

No	*Twit*	Label
1	bagaimana bisa usaha besar tokopedia data bisa bocor saya jadi ragu belanja di tokopedia	Negatif
2	tidak ada salah ucap terima kasih dan apresiasi untuk isu bocor data ini iya jangan panik namun tetap waspada big terima kasih for your service panjang umur	Positif
3	juta data guna tokopedia informasi bocor ini daftar email password bagi korban	Netral
4	menteri komunikasi dan informatika kemenkominfo akan segera panggil direksi tokopedia panggil ini kait duga bocor data pribadi juta akun guna layan ecommerce itu	Netral

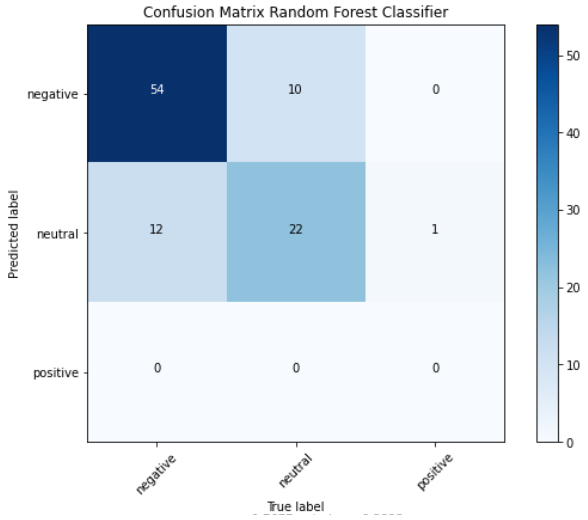
Setelah itu semua *stopword* seperti ”dan”, “itu”, “yang”, dan yang lainnya dihapus dari setiap *twit*. Hal tersebut dilakukan untuk menghilangkan kata yang tidak bermakna bagi sentimen suatu kalimat (*twit*) sehingga dapat memperkecil ukuran data *twit* dan dapat meningkatkan akurasi (Silva & Ribeiro, 2003). Tabel 6 merupakan contoh *Twit* setelah menghilangkan *stopword*.

Tabel 6. Contoh *Twit* Setelah Menghilangkan *Stopword.*

No	*Twit*	Label
1	usaha tokopedia data bocor ragu belanja tokopedia	Negatif
2	salah terima kasih apresiasi isu bocor data iya panik waspada big terima kasih for your service umur	Positif
3	juta data tokopedia informasi bocor daftar email password korban	Netral
4	menteri komunikasi informatika kemenkominfo panggil direksi tokopedia panggil kait duga bocor data pribadi juta akun layan ecommerce	Netral

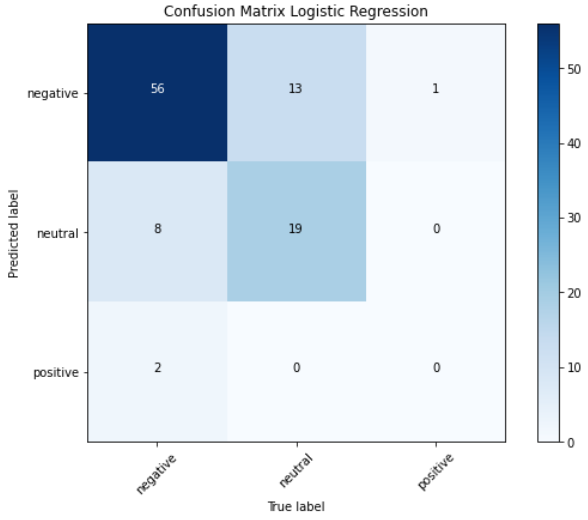
Setelah itu seluruh data *twit* dilakukan transformasi *unigram* dengan pembobotan TF-IDF. Tokenisasi *unigram* merubah *twit* yang mulanya “menteri komunikasi informatika kemenkominfo panggil direksi tokopedia panggil kait duga bocor data pribadi juta akun layan ecommerce” menjadi kumpulan kata seperti “mentri” “komunikasi” “informatika” “kemenkominfo” “panggil” “direksi” “tokopedia” “panggil” “kait” “duga” “bocor” “data” “pribadi” “juta” “akun” “layanan” “ecommerce”. Setelah itu kumpulan data tersebut diberi bobot menggunakan perhitungan TF-IDF. Sehingga diperoleh sebanyak 1329 kata yang menjadi kolom.

Setelah itu data dibagi menjadi data latih dan data uji, data uji terdiri dari *twit* negatif 66.66%, *twit* positif 1.01%, dan *twit* netral 32.32%.



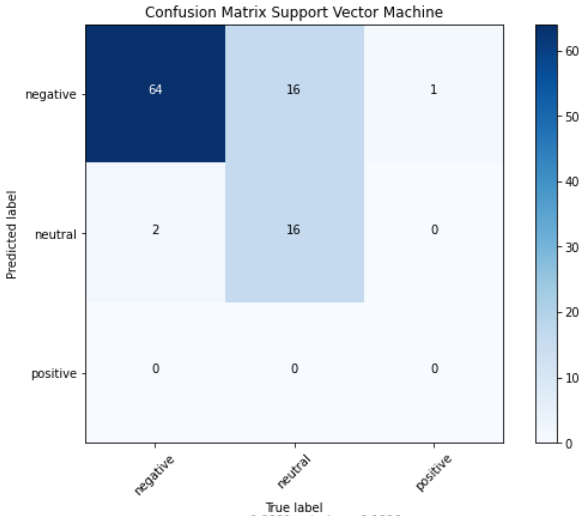
Gambar 4. *Confusion Matrix* Model *Random Forest*.

Berdasarkan Gambar 4 *Random Forest* diketahui bahwa prediksi *twit* negatif dan aktualnya negatif sebanyak 54, prediksi *twit* negatif namun aktualnya netral sebanyak 10, serta prediksi *twit* negatif namun aktualnya positif tidak ada. Selanjutnya, prediksi *twit* netral namun aktualnya negatif sebanyak 12, prediksi *twit* netral dan aktualnya netral sebanyak 22, serta prediksi *twit* netral yang aktualnya positif sebanyak 1. Sementara itu, *confusion matrix* jenis ini tidak bisa memprediksi *twit* bersentimen positif.



Gambar 5. *Confusion Matrix* Model *Logistic Regression*.

Berdasarkan Gambar 5 prediksi *twit* negatif dan aktualnya negatif dari Logistic Regression sebanyak 56, prediksi *twit* negatif namun aktualnya netral sebanyak 13, serta prediksi *twit* negatif namun aktualnya positif hanya ada 1. Kemudian, prediksi *twit* netral namun aktualnya negatif sebanyak 8, prediksi *twit* netral dan aktualnya netral sebanyak 19, serta tidak ada prediksi *twit* netral namun aktualnya positif. Sedangkan untuk prediksi *twit* positif namun aktualnya negatif sebanyak 2, dan tidak terdapat prediksi *twit* positif yang aktualnya netral maupun positif.



Gambar 6. *Confusion Matrix* Model *Support Vector Machine*.

Berdasarkan Gambar 6 hasil prediksi untuk *confusion matrix* ini yaitu, prediksi *twit* negatif dan aktualnya negatif sebanyak 64, prediksi *twit* negatif namun aktualnya netral sebanyak 16, serta prediksi *twit* negatif namun aktualnya positif hanya ada 1. Selanjutnya, prediksi *twit* netral namun aktualnya negatif hanya ada 2, prediksi *twit* netral dan aktualnya netral sebanyak 16, serta tidak terdapat predikti *twit* netral yang aktualnya positif. Berdasarkan gambar, *Confusion matrix* ini tidak mampu membuat prediksi *twit* bersentimen positif.

Setelah diamati dari ketiga kelas yang ada dan dari ketiga *Confusion Matrix* yang telah dibuat tidak ada satupun yang dapat memprediksi *twit* bersentimen positif dengan benar. Ketepatan klasifikasi dipengaruhi olah beberapa faktor, diantaranya jumlah teks atau *term* yang diidentifikasi, jumlah data latihan yang digunakan, fitur klasifikasi, algoritma yang digunakan, dan kemiripan kata yang ada pada saat proses klasifikasi (Faradhillah et al., 2016).

Tabel 7 memberi informasi mengenai *precision, recall*, dan *f1-score* untuk setiap model.

Tabel 7. Hasil Prediksi Model.

Model	Macro Precision	Macro Recall	Macro F1
Random Forest	0.500316	0.501578	0.500829
Logistic Regression	0.501235	0.480745	0.489199
Support Vector Machine	0.559671	0.489899	0.503583

4. KESIMPULAN

Berdasarkan pengamatan yang telah dilakukan terhadap tiga *classifier* berbeda untuk membuat model analisa sentimen *twit* tentang insiden kebocoran data Tokopedia, dapat disimpulkan bahwa *Support Vector Machine* merupakan *classifier* dengan performa terbaik karena dari total 494 *twit* yang dianalisa, *classifier* ini memberikan *f1-score* tertinggi sebesar 0.503583.

Penulis menyarankan penelitian selanjutnya untuk membuat daftar *stopword* yang spesifik untuk suatu *dataset*, karena menggunakan *stopword* yang umum dapat berdampak negatif pada performa model yang diamati (Silva & Ribeiro, 2003). Selain itu penulis menyarankan penelitian selanjutnya untuk mengamati tolok ukur yang lain untuk menilai performa model, seperti AUC pada grafik ROC.

UCAPAN TERIMA KASIH

Dalam pembuatan paper ini, penulis mendapatkan banyak bantuan dari berbagai pihak baik secara langsung maupun tidak langsung sehingga paper ini bisa diselesaikan oleh penulis. Oleh

karena itu, penulis ingin mengucapkan terima kasih kepada semua pihak yang terlibat, diantaranya:
1) Orang Tua yang selalu memberi dukungan moril dan materiil.
2) Dr. Mudjahidin, S.T, M.T. selaku Kepala Departemen Sistem Informasi.
3) Nur Aini Rakhmawati S.Kom., M.Sc.Eng., Ph.D. selaku Dosen Pembimbing dan Pengampu mata kuliah Etika Profesi.

DAFTAR PUSTAKA

Beel, J., Langer, S., & Gipp, B. (2017). TF-IDuF: A Novel Term-Weighting Sheme for User Modeling based on Users' Personal Document Collections. *Proceedings of the iConference 2017*, 1–7.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

CNN Indonesia. (2020). *Deretan Peristiwa Kebocoran Data Warga RI Sejak Awal 2020*. CNN Indonesia. https://www.cnnindonesia.com/teknologi/20200623160834-185-516532/deretan-peristiwa-kebocoran-data-warga-ri-sejak-awal-2020

Deviyanto, A., & Wahyudi, M. D. R. (2018). PENERAPAN ANALISIS SENTIMEN PADA PENGGUNA TWITTER MENGGUNAKAN METODE K-NEAREST NEIGHBOR. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *3*(1), 1. https://doi.org/10.14421/jiska.2018.31-01

Faradhillah, N. Y. A., Kusumawardani, R. P., Hafidz, I., Informasi, J. S., & Informasi, F. T. (2016). Eksperimen Sistem Klasifikasi Analisa Sentimen Twitter Pada Akun Resmi Pemerintah Kota Surabaya Berbasis Pembelajaran Mesin. *Seminar Nasional Sistem Informasi Indonesia*, 15–24.

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications*, *23*(1), 11. https://doi.org/10.3390/mca23010011

Hoffman, J. I. E. (2019). Basic Biostatistics for Medical and Biomedical Practitioners. In *Biostatistics for Medical and Biomedical Practitioners*. Elsevier. https://doi.org/10.1016/C2018-0-02190-8

Lanham, M., & Bedinelli, R. (2015). *Evaluating Stochastic Cost-Benefit Classification Measures for A Retailer's Assortment Mix Decision*.

Librianty, A. (2016, Maret). *Data Jadi Incaran Utama Penjahat Cyber*. Liputan6. https://www.liputan6.com/tekno/read/2466293/data-jadi-incaran-utama-penjahat-cyber

Maulana, T., Rakhmawati, N., Wibowo, N., & Muhammad, H. (2020). *Data Set Sentimen Twit Terhadap Insiden Kebocoran Data Tokopedia* (1.0). Zenodo. https://doi.org/10.5281/ZENODO.4430588

Silva, C., & Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, *3*, 1661–1666. https://doi.org/10.1109/IJCNN.2003.1223656

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tang, J., Li, H., Cao, Y., & Tang, Z. (2005). Email data cleaning. *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, 489. https://doi.org/10.1145/1081870.1081926

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. In *O'Reilly* (1 ed.). O'Reilly Media.

Vardiansyah, D. (2008). *Filsafat Ilmu Komunikasi Suatu Pengantar*. Indeks.

Wibowo, N. (2020). *Program Scrapper Twit Tanpa API dan Pemroses Data* (1.0). Zenodo. https://doi.org/10.5281/zenodo.4231819

Zhang, L., Zhou, W., & Jiao, L. (2004). Wavelet Support Vector Machine. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *34*(1), 34–39. https://doi.org/10.1109/TSMCB.2003.811113

9 772527 583007