Jurnal Kajian Ilmiah (JKI)
e-ISSN: 2597-792X, ISSN: 1410-9794
Vol. 20 No. 2 (Mei 2020), Halaman: 167 – 176
Terakreditasi Peringkat 5 (SINTA 5) sesuai SK RISTEKDIKTI Nomor. 3/E/KPT/2019

Sentimen Analisis Publik Terhadap Joko Widodo Terhadap Wabah Covid-19 Menggunakan Metode Machine Learning

Sisferi Hikmawan [1,*], Amsal Pardamean [1], Siti Nur Khasanah [2]

[1] Ilmu Komputer; STMIK Nusa Mandiri; Jl. Damai No.8, Warung Jati Barat (Margasatwa), Jakarta Selatan, Telp.(021) 78839513 Fax.(021) 78839421; e-mail: 14002318@nusamandiri.ac.id; e-mail: 14002309@nusamandiri.ac.id
[2] Sistem Informasi, STMIK Nusa Mandiri; ; Jl. Damai No.8, Warung Jati Barat (Margasatwa), Jakarta Selatan, Telp.(021) 78839513 Fax.(021) 78839421; e-mail: siti.skx@nusamandiri.ac.id

* Korespondensi: e-mail: gansisferi@nusamandiri.ac.id

Abstract

Analyzing public sentiment towards a government policy is no longer impossible, the process of analyzing with data mining is a method that is often used. The Data Mining method is always related to the dataset, with the keywords "Jokowi" and "Covid" twitter allowing us to make tweets in it to be used as a dataset. In data mining for sentiment analysis, techniques such as transform, tokenize, stemming, classification, etc. are very influential on its accuracy. Gata Framework is used for preprocessing, and Rapidminer is also used to analyze and compare three classification methods namely Naive Bayes, Support Vector Machine, and k-NN. And the best value is obtained, the Support Vector Machine with an accuracy of 84.58%, precision 82.14% and recall 85.82%.

***Keywords**: Covid, Jokowi, SVM, K-NN, Naive Bayes*

Abstrak

Menganalisa sentimen publik terhadap suatu kebijakan pemerintah merupakan cara yang tidak lagi mustahil, proses analisa dengan data mining merupakan metode yang sering digunakan. Metode Data Mining selalu berkaitan dengan dataset, dengan kata kunci “Jokowi” dan “Covid” twitter memungkinkan kita menjadikan tweet didalamnya untuk dijadikan dataset. Dalam data mining untuk sentimen analisis, dilakukan teknik seperti transform, tokenize, stemming, classification, dan lain-lain sangat berpengaruh pada akurasinya. Gata Framework digunakan untuk preprocessing, dan Rapidminer juga digunakan untuk menganalisa dan membandingkan tiga metode klasifikasi yaitu Naive Bayes, Support Vector Machine, dan k-NN. Dan dihasilkan nilai terbaik yaitu Support Vector Machine dengan accuracy 84.58%, precision 82.14% dan recall 85.82%.

Kata kunci: Covid, Jokowi, SVM, K-NN, Naive Bayes

1. Pendahuluan

Covid-19 menjadi topik yang hangat pada awal 2020. Virus yang bermula dari Wuhan China ini telah menyebar secara cepat ke hampir seluruh dunia. Sejak adanya kasus pertama dengan dua orang positif di Indonesia(Kompas.com, 2020), topik covid-19 ini selalu dibahas dalam berbagai media berita, dan tentu saja media sosial. Joko Widodo sebagai Presiden Republik Indonesia tentu menjadi perhatian masyarakat terutama tentang kebijakan yang diterapkan dalam penanganan Covid-19 di Indonesia.

Submitted: **26 Maret 2020**; Revised: **16 April 2020**; Accepted: **30 April 2020**; Published: **26 Mei 2020**

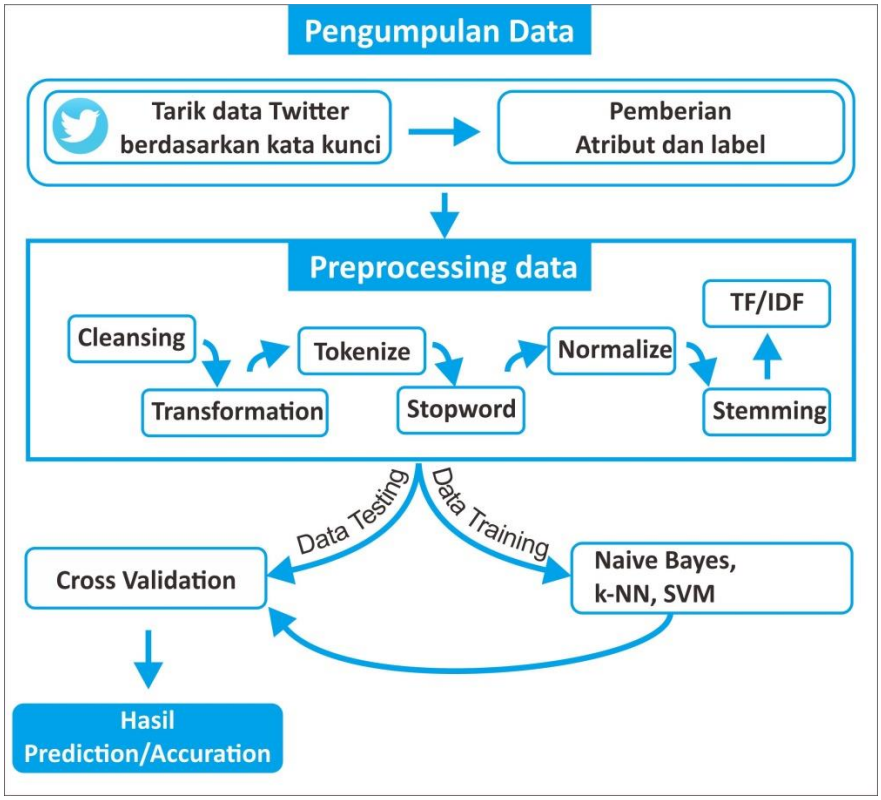
Salah satu media sosial yang dengan bebas masyarakat menuangkan pendapatnya adalah *Twitter*. Dalam beberapa tahun ini *twitter* terakhir, *twitter* memberi banyak pengaruh dalam menghasilkan sumber informasi (Muthia et al., 2019). Dalam *data mining*, banyak hal menarik yang dapat digali pada *twitter* diantaranya bagaimana opini yang terdapat di masyarakat tentang suatu kebijakan pemerintah. Seperti kebijakan pemerintah yang telah dikeluarkan bahkan yang masih dalam wacana terkait Covid-19. Hal ini menjadi sangat penting karena dapat menjadi bahan pertimbangan untuk pemerintah dalam menanggapi sikap publik.

Tujuan dari penelitian ini adalah mengkasifikasikan sentimen positif, netral dan negatif dari twitter terhadap dua kata kunci yaitu “Jokowi” dan “Covid”. Dengan penggunaan dua kata kunci tersebut dapat dihasilkan hasil twit yang fokus hanya pada “Jokowi” dan “Covid”. Dalam pengumpulan data, hingga klasifikasi menggunakan metode *Machine Learning* diantaranya adalah menggunakan algoritma *Naive, Support Vector Machine* dan *k-NN* serta dengan Pembobotan fitur menggunakan Algoritma *Term Frequency Invert Document Frequency* (TF-IDF).

2. Metode Penelitian

2.1. Pengumpulan Data

Berikut Diagram Proses Klasifikasi Sentiment Analysis yang akan dibahas dalam penelitian ini:



Sumber: Hasil Penelitian (2020)

Gambar 1. Diagram Proses Klasifikasi *Sentiment Analysis*

Dalam penelitian ini hal pertama yang dilakukan adalah pengumpulan data dengan melakukan *Web scrapping* dari *twitter*. *Web scrapping* adalah proses yang digunakan untuk mengekstraksi data dari situs web yang diinginkan dengan langsung mengakses *World Wide Web* dengan bantuan HTTP, atau melalui browser web (Jain et al., 2019). *Scrapping web* digunakan untuk mengubah data yang tidak terstruktur pada sebuah web menjadi data terstruktur yang dapat disimpan dan dianalisis dalam database atau spreadsheet (S.C.M. de S Sirisuriya, 2015). Dalam penelitian ini, penarikan data menggunakan aplikasi berbasis python yaitu *twitscraper* (Helmi Satria, 2018). Setelah itu dilakukan pelabelan untuk membuat data training dengan memberi kolom label berisi nilai sentimen “positif”, “netral” dan “negatif”. Hasil keluaran dari proses ini berformat csv agar setelahnya dapat dilakukan preprocessing data.

Tabel 1. Contoh Dataset Yang Telah Memiliki Label

Text	label
"Tolong diam di rumah..."	netral
#COVID19 Beri Dampak Terberat pada Dunia #pariwisata Pariwisata https://www.obsessionnews.com/covid-19-beri-dampak-terberat-pada-dunia-pariwisata/ @jokowi #COVID19indonesia #Covid_19 #COVIDã*ƒ*¼19 #IndonesiaMelawanCovid19 #Indonesia #CoronaIndonesia	negatif
#JumatBerkah INGAT PESAN HABIB RIZIEQ, TOLAK DARURAT SIPIL!!	negatif
#Kamerad_yudian mbok ya ikut mikir, keadaan rakyat: #Gelombang_PHK sdh bergulung-gulung. Biaya hidup Naik.	negatif
[GERAK CEPAT] Anggaran Rp.62,3 Triliun APBN, Menter @KemenkeuRI Sri Mulyani: Siap Dukung Program Prioritas Pemerintah Hadapi Dampak Pandemi Covid-19 http://infokabinet.id/2020/03/21/gerak-cepat-anggaran-rp-623-triliun-apbn-menkeu-sri-mulyani-siap-dukung-program-prioritas-pemerintah-hadapi-dampak-pandemi-covid-19	positif
Indonesia pasti bisa kalahkan covid 19	positif

Sumber: Hasil Penelitian (2020)

2.2. Preprosessing Data

Setelah data terekstraksi dari situs web langkah selanjutnya adalah melakukan pembersihan data, diantaranya adalah:

1. Menghapus tanda baca yang tidak diinginkan dari dataset.
2. Menghapus *Stop-words* atau kata umum yang tidak memiliki makna seperti angka, kata yang / di / ke.
3. Menghapus link url dan tanda baca seperti (@, :, //) dan tanda baca lainnya.
4. Menyeragamkan huruf besar dan kecil (*lowercase, uppercase*)
5. Pada tahapan ini dilakukan penyeragaman seluruh teks menjadi huruf kecil (*lowercase*) dan pembersihan atau penghapusan pada pada semua dokumen yang berisi angka, url

Submitted: **26 Maret 2020**; Revised: **16 April 2020**; Accepted: **30 April 2020**; Published: **26 Mei 2020**

(http://), *username* (@), tanda pagar (#), delimiter seperti koma (,) dan titik (.) dan juga tanda baca lainya.

6. Pemeriksaan ejaan untuk memastikan kalimat yang dihasilkan terdiri dari kata-kata yang relevan dan tidak salah dalam pengejaan, pemeriksaan ejaan ini penting untuk menghitung frekuensi kata secara akurat.

2.3. Pembobotan TF/IDF

Tahap selanjutnya pembobotan terhadap kata berdasarkan frekuensi dari term atau istilah yang muncul pada dokumen dengan metode Pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency*). Pembobotan pada metode ini mengkombinasikan dua konsep yaitu konsep frekuensi kata dan frekuensi dokumen. Yang dimaksud dengan term frequency adalah nilai frekuensi term terhadap satu dokumen, sedangkan document frequency adalah nilai frekuensi dokumen yang terdapat term tersebut. Untuk itu perlu adanya tindakan *preprocessing* agar metode TF-IDF ini dapat optimal. Berikut adalah rumusnya.

$$tf - idf_{t,d} = tf_{t,d} * idf_t$$

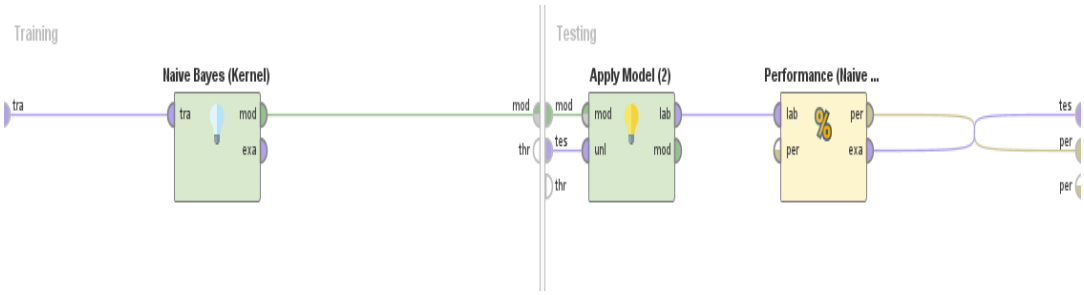
Dimana:

tf = *term frequency* adalah jumlah kemunculan term

t, d = term (t) di dokumen (d)

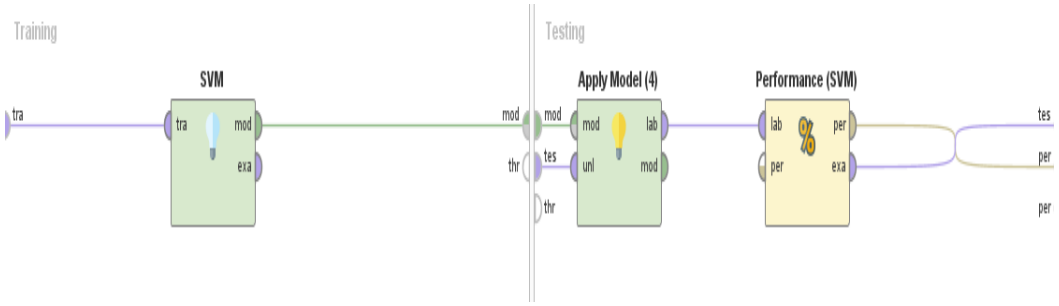
2.4. Klasifikasi Sentimen Analisis

Selanjutnya dilakukan proses pengklasifikasian menggunakan algoritma *Naive Bayes, Support Vector Machine* dan k-NN dan diakhiri dengan pengujian akurasi dengan metode *N-Fold Cross-Validation*. *N-Fold Cross-Validation* adalah salah satu cara untuk resampling data yang paling banyak digunakan untuk memprediksi kesalahan dari model dan untuk mengatur parameter dari model (Berrar, 2018), Dalam pengujian N-*Fold Cross-Validation* dataset dibagi menjadi N buah partisi secara acak. Kemudian dilakukan eksperimen sejumlah N kali, dimana masing-masing eksperimen menggunakan data partisi ke N sebagai data testing dan memanfaatkan sisa partisi lainnya sebagai data training.



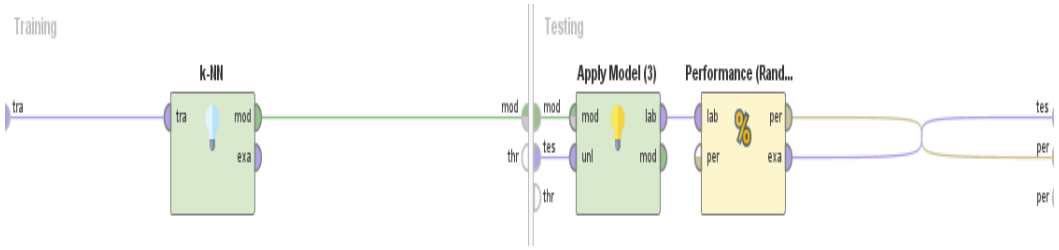
Sumber: Hasil Penelitian (2020)

Gambar 2. *Cross Validation* dengan Sub-proses Naive Bayes



Sumber: Hasil Penelitian (2020)

Gambar 3. *Cross Validation* dengan Sub-proses SVM



Sumber: Hasil Penelitian (2020)

Gambar 4. *Cross Validation* dengan Sub-proses k-NN

Untuk lebih jelasnya tentang perbedaan metode klasifikasi yang digunakan berikut adalah penjelasannya:

1. Naïve Bayes Classifier

Metode klasifikasi *Naïve Bayes* adalah metode *machine learning* ysang sering digunakan untuk klasifikasi teks. *Naïve Bayes* mengasumsikan bahwa fitur (kata) memiliki nilai yang independen pada posisi kata(Dewi et al., 2018). *Naïve Bayes* menggunakan metode probabilitas dan statistik untuk memprediksi peluang di masa depan dari data masa lalu. Namun, *Naïve Bayes* juga memiliki kelemahan, yaitu kesalahan dalam pemilihan *feature* dapat mengurangi akurasi, di beberapa kasus terlalu banyak *feature* membuat metode ini memiliki akurasi yang rendah dan juga membuat waktu perhitungan semakin bertambah.

2. K-NN

K-NN merupakan metode *machine learning* yang berdasarkan data pembelajaran untuk mengklasifikasikan suatu objek berdasarkan k-tetangga terdekat (Bayhaqy et al., 2018). K-NN merupakan salah satu metode *machine learning* yang cukup populer dan simpel(Wibawa et al., 2018).

3. Support Vector Machine

Support Vector Machine (SVM) sejak awal sudah digunakan dalam *machine learning* untuk mengklasifikasi dari data yang dianalisa(Herlawati, 2020). SVM sangat memungkinkan digunakan untuk *text mining* (Kristiyanti et al., 2019). SVM berfungsi untuk memisahkan antara beberapa class yang berbeda. SVM memiliki proses yang efektif dan efisien untuk klasifikasi. Namun SVM memiliki kelemahan yaitu pemilihan parameter atau feature dalam beberapa kasus sangat mempengaruhi akurasi secara signifikan.

Submitted: **26 Maret 2020**; Revised: **16 April 2020**; Accepted: **30 April 2020**; Published: **26 Mei 2020**

2.5. Evaluasi

Setelah proses klasifikasi sentimen analisis telah selesai, langkah selanjutnya adalah mengevaluasi dengan mengukur keakurasian dan kualitas dari hasil tersebut. Evaluasi yang dilakukan adalah dengan pengujian performa dan akurasi sehingga menghasilkan *nilai accuracy, precision*, dan *recall.*

Accuracy (A) adalah total nilai *True Positif* dan *True Negatif* dibagi dengan jumlah keseluruhan data.

$$A = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})} x\ 100\%$$

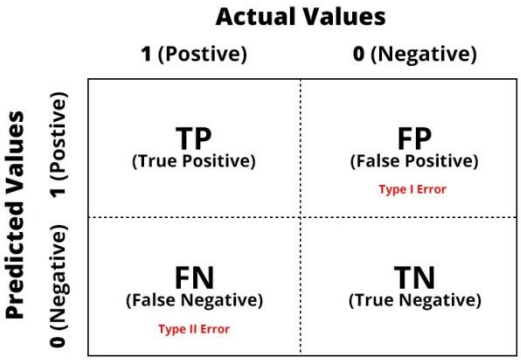
Precision (P) adalah prosentase nilai True Positif dari seluruh nilai Positif yang diprediksi.

$$P = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} x\ 100\%$$

Recall (R) adalah persentase prediksi Positif dibandingkan dengan True Positif.

$$R = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} x\ 100\%$$

Parameter TP, FP, FN, TN berdasarkan *Confusion Matrix* seperti pada Gambar 5.



Sumber: Hasil Penelitian (2020)

Gambar 5. Tabel *Confusion Matrix* (Nugroho, 2019)

Ketika proses pengumpulan data telah selesai, maka dipisahkan data menjadi *data training* dan *data testing.* Pembagian data tersebut dilakukan pada petode N-*Fold Cross-Validation.* N-Fold Cross Validation adalah metode untuk memisahkan data dengan sebagian terdapat label sebagai data training dan sebagian dihapus labelnya sebagai data training lalu membagi data secara acak sebanyak N untuk dilakukan pengujian sebanyak N kali. Ini adalah langkah terakhir untuk mengetahui hasil dan performa dari *machine learning* yang telah dibuat.

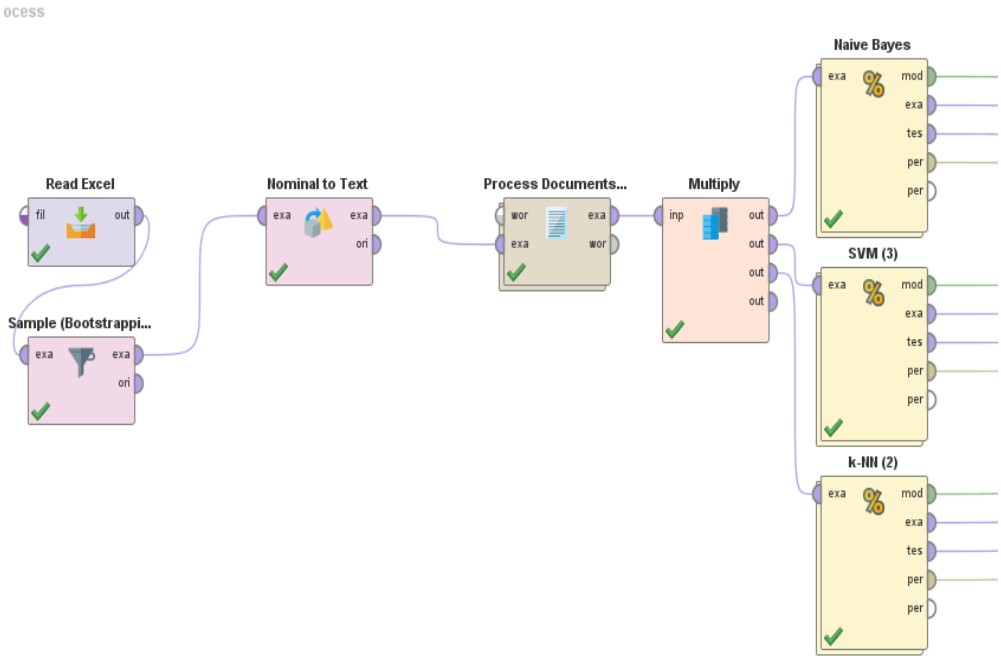
3. Hasil dan Pembahasan

Pada bagian ini akan dijelaskan hasil dari percobaan yang telah dilakukan yaitu menganalisa hasil performa dan akurasinya. Dalam *preprocessing* penelitian ini menggunakan Gata *Framework* (*Gata Framework Website*, n.d.) dengan melakukan pengaturan pada teknik seperti pada Gambar 6.

Sumber: Hasil Penelitian (2020)

Gambar 6. Gata Framework – *Preprocessing Menu*

Penggunaan Gata Framework dikarenakan tweet text yang diambil merupakan text berbahasa Indonesia. Dan hasil keluaran ini digunakan pada rapidminer untuk dijadikan dataset untuk dilakukan pengujian.



Sumber: Hasil Penelitian (2020)

Gambar 7. Proses utama di Rapidminer

Pada gambar 7 operator *"Read Excel"* berfungsi untuk membaca file Excel hasil dari keluaran twitterscraper yang telah di preprocessing dengan Gataframework sebelumnya, dan dilakukan *"Process Documents*" untuk melakukan preprocessing kembali agar data benar-benar bersih. Di tahap selanjutnya dengan menggunakan operator "Cross Validation" ditambahkan di dalamnya operator untuk klasifikasi dan evaluasi dari sentimen analisis dengan nilai N adalah 10 (10-*Fold Cross Validation*).

Submitted: **26 Maret 2020**; Revised: **16 April 2020**; Accepted: **30 April 2020**; Published: **26 Mei 2020**

Tabel 2. Hasil prediksi nilai *Confidence (positif, netral, negatif)* terhadap nilai label

Row No.	label	prediction(label)	confidence(positif)	confidence(netral)	confidence(negatif)
1	positif	netral	0.095	0.729	0.176
2	positif	positif	0.996	0.000	0.004
3	positif	positif	0.991	0.009	0.000
4	positif	negatif	0.000	0.000	1.000
5	positif	positif	1.000	0.000	0.000
6	positif	positif	1.000	0.000	0.000
7	positif	positif	0.971	0.002	0.027
8	positif	netral	0.103	0.704	0.193
9	positif	netral	0.103	0.704	0.193
10	positif	positif	1.000	0.000	0.000
11	positif	positif	1.000	0.000	0.000
12	positif	positif	0.955	0.038	0.007

Sumber: Hasil Penelitian (2020)

Pada tabel 2 ditampilkan hasilnya, yaitu label dengan sentimen sesungguhnya dan prediksi label.

3.1. Perbandingan Pengujian

Dalam penerapan metode Naive Bayes, SVM dan k-NN diperlukan tuning dan percobaan untuk mengoptimalisasi hasil agar lebih baik. Pada k-NN membutuhkan penentuan nilai k untuk mendapatkan akurasi tertinggi, dan pada kajian ini didapatkan nilai terbaik k=3. Serta Naive Bayes dan SVM sangat sensitif terhadap dataset yang telah dibuat, yaitu menghapus missing value pada dataset. *Missing value* muncul dikarenakan pada saat pre-processing terdapat text pada twitter yang tidak berisi tulisan yang memiliki *term.*

Tabel 3 adalah hasil *confusion matrix* dari *Rapid Miner:*

Tabel 3. *Confusion matrix* hasil klasifikasi

Metode	True Positif	True Negatif	True Netral	False Positif	False Negatif	False Netral
Naive Bayes	224	476	161	19	18	120
SVM	227	572	147	5	66	1
k-NN	246	465	141	77	23	66

Sumber: Hasil Penelitian (2020)

Tabel 4 adalah nilai rata-rata dari *accuracy, precision* dan *recall* pada setiap metode yang digunakan.

Tabel 4. Hasil *Accurace*, *Precision* dan *Recall*

Metode	Accuracy	Precision	Recall
Naive Bayes	84.58%	82.14%	85.82%
SVM	92.93%	95.70%	89.17%
k-NN	83.70%	80.66%	84.13%

Sumber: Hasil Penelitian (2020)

Dari hasil yang ditampilkan pada Tabel 4, *accuracy* dari Naive Bayes sebesar 84.58%, Support Vector Machine sebesar 92.93%, dan k-NN sebesar 83.70%. Hasil *precision* dari Naive Bayes

sebesar 82.14%, SVM sebesar 95.70% dan k-NN sebesar 80.66%. Juga hasil *recall* dari Naive Bayes sebesar 85.82%, SVM sebesar 89.17%, dan k-NN sebesar 84.13%. Jadi dapat dilihat bahwa pengklasifikasian Support Vector Machine yang terbaik di antara Naive Bayes dan Random Forest jika digunakan untuk dataset sentimen analisis yang menggunakan bahasa Indonesia karena memiliki akurasi dan presisi tertinggi. Perbedaan hasil ini karena pengaruh karakteristik dataset dan juga proses lainnya.

4. Kesimpulan

Pada penelitian ini, upaya untuk mengetahui pendapat publik dilakukan berupa sentimen positif, negatif dan netral di media *twitter* terhadap tindakan pencegahan Covid-19 oleh Pemerintahan di Indonesia dapat menjadi masukan bagi indikator keberhasilan pemerintah. Untuk merepresentasikannya maka dilakukan *text mining* dengan menggunakan metode SVM, Naive Bayes dan k-NN untuk mengklasifikasikan label sentimen dari dataset. Dari hasil pengujian menunjukkan , *accuracy* dari Naive Bayes sebesar 84.58%, Support Vector Machine sebesar 92.93%, dan k-NN sebesar 83.70%. Hasil *precision* dari Naive Bayes sebesar 82.14%, SVM sebesar 95.70% dan k-NN sebesar 80.66%. Juga hasil *recall* dari Naive Bayes sebesar 85.82%, SVM sebesar 89.17%, dan k-NN sebesar 84.13%. Dapat disimpulkan bahwa Support Vector Machine yang terbaik karena memiliki akurasi dan presisi tertinggi. Untuk kedepannya kita perlu menggunakan dataset yang lebih besar dan kompleks lagi serta penyempurnaan preprocessing untuk bahasa Indonesia yang tidak baku.

Daftar Pustaka

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018). Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes. *2018 International Conference on Orange Technologies, ICOT 2018*, 1–6. https://doi.org/10.1109/ICOT.2018.8705796

Berrar, D. (2018). Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Dewi, Y. N., Riana, D., & Mantoro, T. (2018). Improving Naïve Bayes performance in single image pap smear using weighted principal component analysis (WPCA). *3rd International Conference on Computing, Engineering, and Design, ICCED 2017*, *2018-March*, 1–5. https://doi.org/10.1109/CED.2017.8308130

Gata Framework Website. (n.d.). http://www.gataframework.com

Helmi Satria. (2018). *Cara Mendapatkan Data (Tweet) dari Twitter*. https://medium.com/@helmisatria/cara-mendapatkan-data-tweet-dari-twitter-e0ce79cdeed4

Herlawati, H. (2020). COVID-19 Spread Pattern Using Support Vector Regression. *PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic*, *8*(1), 67–74.

Submitted: **26 Maret 2020**; Revised: **16 April 2020**; Accepted: **30 April 2020**; Published: **26 Mei 2020**

https://doi.org/10.33558/PIKSEL.V8I1.2024

Jain, M., Vaish, S., Patil, M., & Anant, G. M. (2019). Data extraction and sentimental analysis from “twitter” using web scrapping. *International Journal of Engineering and Advanced Technology*. https://doi.org/10.35940/ijeat.A2226.109119

Kompas.com. (2020). *Fakta Lengkap Kasus Pertama Virus Corona di Indonesia*. https://nasional.kompas.com/read/2020/03/03/06314981/fakta-lengkap-kasus-pertama-virus-corona-di-indonesia?page=all

Kristiyanti, D. A., Umam, A. H., Wahyudi, M., Amin, R., & Marlinda, L. (2019). Comparison of SVM Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter. *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, *Citsm*, 1–6. https://doi.org/10.1109/CITSM.2018.8674352

Muthia, D. A., Putri, D. A., Rachmi, H., & Surniandari, A. (2019). Implementation of Text Mining in Predicting Consumer Interest on Digital Camera Products. *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, *Citsm*, 1–7. https://doi.org/10.1109/CITSM.2018.8674063

Nugroho, K. S. (2019). *Confusion Matrix untuk Evaluasi Model pada Supervised Learning*. https://medium.com/@ksnugroho/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f

S.C.M. de S Sirisuriya. (2015). A Comparative Study on Web Scraping. *8th International Research Conference KDU*.

Wibawa, D. W., Nasrun, M., & Setianingsih, C. (2018). Sentiment Analysis on User Satisfaction Level of Cellular Data Service Using the K-Nearest Neighbor (K-NN) Algorithm. *Proceedings - 2018 International Conference on Control, Electronics, Renewable Energy and Communications, ICCEREC 2018*, 235–241. https://doi.org/10.1109/ICCEREC.2018.8711992