**Jurnal Ilmiah Teknik Informatika dan Komunikasi**
Halaman Jurnal: http://journal.sinov.id/index.php/juitik/index
Halaman UTAMA Jurnal : https://journal.sinov.id/index.php

# KLASIFIKASI AKUN BUZZER PADA TWITTER MENGGUNAKAN ALGORITMA NAIVE BAYES

**Catur Arpal Perkasa[1], Amalia Andjani Arifiyanti[2], Agus Salim[3]**
[1,2,3]Universitas Pembangunan Nasional "Veteran" Jawa Timur
*Korespondensi penulis:catur.arp@gmail.com*

Jl. Rungkut Madya No.1, Gn. Anyar, Kec. Gn. Anyar, Kota SBY, Jawa Timur 60294

***Abstract***. *Amid Indonesia's e-commerce IPO, there is a suspicious movement of the dissemination of tweets with positive sentiments that involve the use of fabricated content spread by buzzers. In this study, the patterns of accounts involved will be investigated. A classifier will be developed based on the patterns of these accounts to accurately identify and distinguish between buzzer and non-buzzer accounts.The buzzer classifier will have four attributes: the number of followings, the number of followers, the sentiment value of recent tweets, and the age of the account. The data will be processed and cleaned before sentiment analysis is performed in order to provide weight to the data. Then, the data will be labeled according to predetermined characteristics. The Gaussian variant of Naive Bayes is used to classify accounts with and without stamping. The results of the study show that the performance of the model is 80% accurate.*

***Keywords*** *Buzzer, Classification, Gaussian Naïve Bayes, Twitter*

**Abstrak**. Dalam kampanye IPO perusahaan *e-commerce* di Indonesia yang *trending* pada Twitter terdapat dugaan keterlibatan *buzzer* dalam penyebaran konten yang terfabrikasi. Dalam penelitian ini, akan dikaji pola dari akun-akun yang terlibat dalam penyebarluasan. Penelitian ini akan mengembangkan sebuah pengklasifikasi *buzzer* berdasarkan pola dan karakteristik yang didapatkan dari akun-akun tersebut. Pengklasifikasi *buzzer* akan memiliki empat atribut: jumlah *following*, jumlah *follower*, nilai sentimen dari *tweets* terbaru, dan usia akun. Data akan diproses dan dibersihkan sebelum melakukan analisis sentimen untuk memberikan bobot pada data. Kemudian data dilabeli sesuai dengan karakteristik yang telah ditentukan. Algoritma Naive Bayes varian Gaussian akan digunakan untuk melakukan klasifikasi akun *buzzer* dan non-*buzzer*. Hasil dari penelitian ini menunjukkan performa model klasifikasi akun *buzzer* yang memiliki nilai akurasi sebesar 80%.

**Kata kunci**: *Buzzer*, Gaussian Naïve Bayes, Klasifikasi, Twitter

*Received November 30, 2022; Revised Desember 20, 2022; Januari 24, 2023*

## LATAR BELAKANG

Twitter sebagai salah satu platform media sosial yang digemari oleh pengguna di Indonesia memiliki fitur unggulan seperti akses informasi *real-time* dan komunikasi yang mudah dengan teman serta *tweet* yang permanen dan dapat dicari melalui *search engine* secara bebas oleh publik (Boyd & Ellison, 2007). Berdasarkan StatCounter, pengguna Twitter di Indonesia terus bertumbuh pesat, bahkan Indonesia menjadi negara kelima dengan pengguna Twitter terbanyak di dunia (Febriani & Dewi, 2019). Salah satu unsur yang terdapat pada media sosial yang membuat banyak *brand* memutuskan untuk menjalankan kampanye *brand*-nya melalui media sosial adalah *electronic word of mouth* (eWOM) (Ismagilova dkk, 2017). Strategi pemasaran dengan pendekatan eWOM (Electronic Word of Mouth) saat ini menjadi angin segar dalam bidang pemasaran, namun perlu diingat bahwa pengguna tidak memiliki cukup sumber daya waktu untuk menilai keprofesionalan narasumber dan keaslian informasi, sehingga berdampak pada kesalahan pengguna dalam pengambilan keputusan.

Melalui pendekatan *data mining* menggunakan algoritma Naive Bayes varian Gaussian, klasifikasi terhadap akun *buzzer* dapat dilakukan (Ismail dkk, 2020). Atribut yang digunakan dalam pengklasifikasian adalah jumlah *following*, jumlah *follower*, nilai sentimen dari *recent tweets*, dan umur akun. Data yang digunakan adalah akun-akun yang ikut menyebarkan kampanye pada Initial Public Offering perusahaan *e-commerce* di Indonesia yang sempat *trending* pada Twitter. Proses yang dilakukan meliputi pengumpulan data, pelabelan, *training* model terhadap dataset, dan pengujian model klasifikasi.

## KAJIAN TEORITIS

Bagian ini menguraikan teori-teori relevan yang mendasari topik penelitian dan memberikan ulasan tentang beberapa penelitian sebelumnya yang relevan dan memberikan acuan serta landasan bagi penelitian ini dilakukan. Jika ada hipotesis, bisa dinyatakan tidak tersurat dan tidak harus dalam kalimat tanya.

## Dasar Teori

### 1. *Buzzer*

*Buzzer* adalah pengguna media sosial yang memiliki pengaruh dan dapat mempengaruhi pengguna lain (Juzar dan Akbar, 2018), diserap dari istilah *buzz marketing* yang merujuk pada salah satu strategi *marketing* untuk meningkatkan visibilitas melalui pesan *marketing* yang difabrikasi (Tangel dkk, 2019).

### 2. Klasifikasi

Klasifikasi adalah analisis data yang mengekstrak model untuk mendeskripsikan kelas-kelas data penting. *Classifier* digunakan untuk memprediksi label kelas kategorikal (Han, Kamber, dan Pei, 2012).

### 3. Analisis Sentimen

Analisis sentimen adalah proses untuk mengidentifikasi dan mengkategorikan pandangan dalam teks dari *web* untuk menilai perilaku penulis atau narasumber terhadap topik tertentu (D'Andrea dkk, 2015). NLP adalah upaya untuk mengekstrak *arti* dari teks dengan memberikan struktur ke bahasa alami dan mengambil *insights* dari linguistik (Verspoor dan Cohen, 2013)

### 4. Naïve Bayes

Naive Bayes adalah metode klasifikasi yang digunakan untuk memprediksi kelas data berdasarkan probabilitas. Gaussian Naive Bayes adalah varian dari Naive Bayes yang mendukung data bertipe *continuous* (Ismail dkk, 2020).

## Penelitian Terdahulu

Pada penelitian pertama yang berjudul “Klasifikasi Akun *Buzzer* Pemilu Pada Media Sosial Twitter Berdasarkan Data *Tweet* Menggunakan Algoritma C4.5” terdapat kesamaan dalam tujuan pengembangan model klasifikasi, yaitu untuk mengidentifikasi sebuah akun *buzzer*.

Pada penelitian kedua yang berjudul “Twitter Buzzer Detection for Indonesian Presidential Election”, Suciati dkk ingin mengungkap keaslian akun dengan pendekatan

yang berbeda, yaitu melalui nilai Mutual Information (MI). Terdapat kesamaan atribut yang digunakan yaitu *following, follower,* dan tanggal dibuatnya akun Twitter terkait.

Pada penelitian ketiga yang berjudul "Buzzer Detection on Twitter Using Modified Eigenvector Centrality" diusulkan metode Eigenvector Centrality yang telah dimodifikasi untuk melakukan deteksi akun *buzzer* pada Twitter.

## METODE PENELITIAN

Dataset yang terdiri dari jumlah following, jumlah follower, umur akun, dan sentimen dari beberapa *tweet* terakhir yang tersedia akan di-*split* untuk *training* dan *testing*. Dataset akan di-*train* menggunakan algoritma Naïve Bayes varian Gaussian, varian Naïve Bayes yang cocok dengan atribut nilai sentimen yang bersifat *continuous*. Hasil dari *training* model akan menghasilkan nilai akurasi yang didapatkan dari prediksi terhadap dataset *test*, serta performa model juga dapat terepresentasi melalui grafik *confusion matrix*, ROC *curve*, dan *Precision Recall Value*.

## HASIL DAN PEMBAHASAN (Sub judul level 1)

Spesifikasi data yang digunakan merupakan data *tweet* yang berkaitan dengan IPO *e-commerce* dan pernah *trending* di Indonesia, yaitu IPO Bukalapak, GoTo, dan BalikBukaAja yang diambil dalam kurun waktu 10 Juni 2021 sampai dengan 9 Desember. Jumlah data sebanyak 12.273 *tweet* dari 5.105 *user* unik. Data lainnya berupa *username,* umur akun, *following, followers,* dan *recent tweets* dari akun terkait didapatkan dengan metode *scrape* dengan menggunakan SNScrape. Data tersebut selanjutnya akan diolah agar sesuai dengan atribut model yang telah ditentukan.

### Preparasi Data Klasifikasi (Sub judul level 2)

Penyiapan dataset untuk klasifikasi diawali dengan *scrape* menggunakan Social Network Scrape untuk mendapatkan data *username* dan *tweet* yang berkaitan dengan topik. Lalu, tahapan *scrape* selanjutnya mengumpulkan informasi akun yang akan digunakan pada model klasifikasi, yaitu *following*, *followers*, umur akun, dan beberapa *tweet* terakhir (hingga 10 *tweet*). Selanjutnya, data *tweet* di atas akan ditimbang nilai

sentimennya menggunakan Vader dan nilai *mean* dari *compound* tiap akunlah yang akan digunakan dalam pengklasifikasian.

| | username | compound_mean | account_age | following | followers |
|---|---|---|---|---|---|
| **0** | MargueriteSavo3 | -0.06861 | 0 | 3 | 4 |
| **1** | AngieYl13520064 | -0.04149 | 0 | 1 | 0 |
| **2** | GraceLeclair4 | -0.14478 | 0 | 5 | 1 |
| **3** | italysingapore | 0.10911 | 7 | 150 | 305 |
| **4** | SriwijayaVerse | 0.14924 | 0 | 11 | 3 |
| **...** | ... | ... | ... | ... | ... |
| **1678** | sagekreyol | 0.36908 | 8 | 416 | 140 |
| **1679** | pecahtelor | 0.18372 | 13 | 713 | 707 |
| **1680** | IamLostTo0 | -0.08717 | 3 | 463 | 510 |
| **1681** | tofukapapa03 | 0.21203 | 0 | 261 | 81 |
| **1682** | NirgunanTiruch1 | -0.04935 | 6 | 31 | 103 |

1683 rows × 6 columns

**Gambar 1. Contoh Data yang Telah Digabungkan dengan Informasi Akun**

### 1. Pelabelan Kelas (Sub judul level 3)

Pelabelan dilakukan menggunakan bahasa pemrograman Python dengan menerapkan konsep seleksi sederhana menggunakan *if statement* terhadap kondisi yang telah ditentukan, yaitu *following* lebih dari 500, *followers* lebih dari 500, umur akun lebih dari 2 tahun, dan sentimen akun lebih dari 0.275. Apabila kriteria sesuai, maka akan dilabeli dengan kelas *non-buzzer*, berlaku sebaliknya. Berikut merupakan contoh hasil data yang telah dilabeli.

| compound_mean | account_age | following | followers | buzzer |
|---|---|---|---|---|
| 0.52308 | 2 | 1788 | 1050 | 0 |
| 0.1736 | 1 | 8 | 60 | 1 |
| 0.28391 | 3 | 36 | 12 | 1 |
| 0.0814199999999999 | 0 | 3 | 0 | 1 |
| 0.75917 | 13 | 106 | 68390 | 1 |
| -0.04245 | 7 | 22739 | 22257 | 1 |
| 0.46403 | 0 | 5 | 3 | 1 |
| 0.6077899999999999 | 0 | 6 | 9 | 1 |
| 0.46315 | 0 | 7 | 2 | 1 |
| 0.60273 | 0 | 6 | 6 | 1 |

1 2 3 4 **5** 6 10 30 40

**Gambar 2. Contoh Data Setelah Dilabeli**

Data yang telah dilabeli berjumlah 3548 dengan kelas *buzzer* sebanyak 2874 dan *non-buzzer* sebanyak 374 untuk kelas *buzzer*. Setelah dataset disiapkan, dataset terindikasi *imbalance class* atau ketidakseimbangan kelas, yang merupakan kondisi sebuah dataset tidak imbang antar satu kelas dan kelas lainnya, yang dapat menyebabkan model *overfit*.

**2. Penyeimbangan Kelas (Sub judul level 3)**

Dataset yang nilai kelas *buzzer*-nya berjumlah 2874 dikurangi sebanyak 2500 baris agar dapat seimbang dengan kelas *non-buzzer*. Jumlah dataset yang telah dilakukan penyimbangan kelas adalah 374 untuk *buzzer* dan 363 untuk *non-buzzer*. Dataset yang telah dilakukan penyeimbangan inilah yang akan digunakan dalam tahapan *training* model.

**Perancangan Model Klasifikasi (Sub judul level 2)**

Data yang telah terkumpul sebelumnya, digunakan untuk melatih model *machine learning* yang dirancang, yaitu Naive Bayes. Proses pemodelan terdiri dari tiga tahap utama, yaitu perancangan model, pelatihan model Naive Bayes, dan evaluasi performa model melalui pembobotan klasifikasi. Berikut merupakan tahapan-tahapan dalam proses klasifikasi.

**1. Perancangan Model (Sub judul level 3)**

Dalam penelitian ini, penggunaan model klasifikasi naive bayes yang dipilih yaitu Naïve Bayes dengan spesifikasi gaussianNB. Gaussian Naïve Bayes digunakan karena terdapat atribut nilai sentimen akun yang datanya merupakan tipe data *continuous*.

```
[ ]  # Split 75:25
     x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)

     # Initialize Gaussian NB
     gnb = GaussianNB()
```

**Gambar 3. Membuat Instance Gaussian Naive Bayes dan *Splitting* Dataset**

**2. Training dan Performa Model (Sub judul level 3)**

Setelah dilakukan *training*, model mendapatkan hasil akurasi sebesar 0.80 atau 80%.

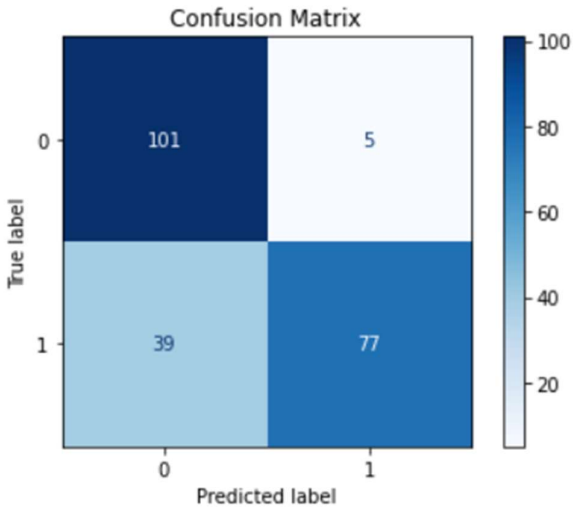
```
[14] y_pred = gnb.predict(x_test)

# calculate the accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy:.2f}')

Accuracy: 0.80
```

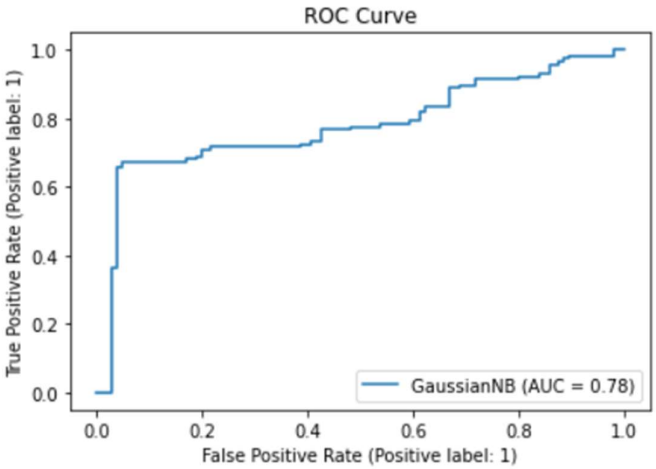
**Gambar 4. Hasil Akurasi Model**

### a) Grafik Performa Model (Sub judul level 4)



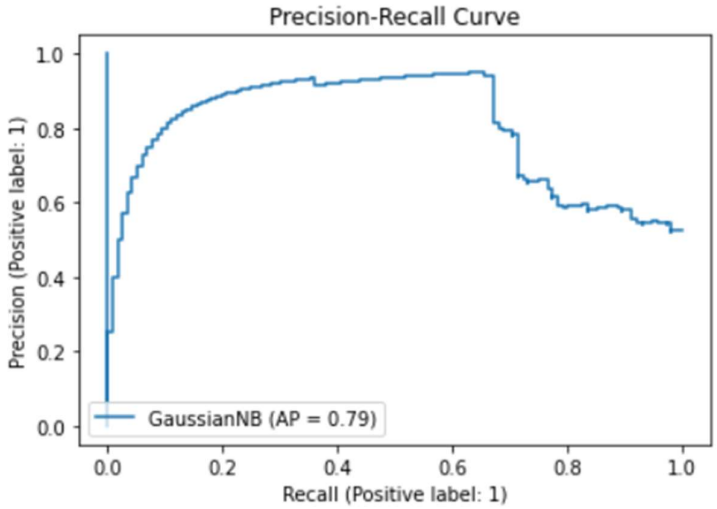

**Gambar 5. Grafik Confusion Matrix**

Grafik menunjukkan bahwa model memprediksi *non-buzzer* sebanyak 140, 39 dari tebakannya salah, sedangkan penebakan pada kelas *buzzer* sebanyak 77 dari 82 tebakannya benar.




**Gambar 6. Grafik ROC Curve**

ROC Curve dari model terlampir pada Gambar 6. Selain itu, model juga memiliki skor AUC sebesar 0.78.




**Gambar 7. Grafik Precision Recall Value**

Grafik menunjukkan ketidakstabilan pada nilai precision dan recall, serta nilai AP yaitu 0.79.

## KESIMPULAN DAN SARAN

Setelah menyelesaikan seluruh tahapan yang ada diperoleh beberapa poin sebagai berikut:

1. Dengan pendekatan *data mining*, yaitu klasifikasi dengan menggunakan Algoritma Naïve Bayes varian Gaussian yang sesuai dengan tipe data *continuous*, didapatkan pola atau *pattern* dari akun-akun *buzzer*.
2. Atribut yang digunakan dalam penelitian ini antara lain, umur akun, nilai *mean* dari nilai *compound* akun sebagai sentimen akun, jumlah *following*, dan jumlah *followers*.
3. Model Naïve Bayes yang dibuat telah diuji dengan metode pengujian yang sesuai. Hasil pengujian menunjukkan bahwa model Naïve Bayes yang dibuat memiliki nilai *accuracy* sebesar 0.80 atau 80%.
4. Berdasarkan hasil grafik ROC *curve*, *confusion matrix*, dan *precision recall value* visualisasi hasil pengujian model, ditentukan bahwa model yang di-*train* pada dataset yang telah dilakukan penyeimbangan kelas memiliki performa model cukup baik, walaupun masih memerlukan peningkatan.

Dalam Penelitian ini terdapat beberapa hal yang dapat dijadikan bahan pembelajaran, hal tersebutlah yang dapat dilakukan agar hasil dari penelitian ini menjadi lebih baik. Berikut adalah beberapa saran yang direkomendasikan dari penelitian kali ini.

1. Bagi Penelitian Selanjutnya

Penambahan atribut lain seperti frekuensi *tweet* per hari, jumlah URL dalam *tweet*, dan jumlah *retweet* dari 100 *tweet* terakhir dapat meningkatkan akurasi model (Ibrahim dkk, 2015). Atribut yang lebih beragam akan memberikan informasi yang lebih banyak dan membuat model lebih mampu menangani data yang lebih kompleks. Selain itu, variasi dataset *buzzer* dan *non-buzzer* dari topik lain juga dapat digunakan dalam proses *training* untuk mencegah *overfitting* pada satu topik tertentu.

2. Bagi organisasi bisnis

Penelitian yang sudah dilakukan menunjukkan bahwa *buzzer* dapat dikenali melalui karakteristik atau perilakunya dalam *posting* pada Twitter. Sehingga, organisasi bisnis harus lebih bijak untuk mempertimbangkan penggunaan strategi *buzz marketing* dalam menjalankan kampanye organisasi bisnisnya, alih-alih digemari masyarakat, bisa saja masyarakat malah meninggalkan *brand* karena tingkat kepercayaannya menurun.

## DAFTAR REFERENSI

*Referensi berisi daftar jurnal, buku, atau referensi lain yang diacu dalam naskah yang terbit dalam 5 tahun terakhir dengan jumlah minimal 75% dari seluruh referensi yang digunakan. Mayoritas referensi adalah sumber primer yaitu jurnal ilmiah/prosiding. Jumlah referensi secara keseluruhan yang diacu* ***minimal 20 buah****, dan sebanyak* ***75%nya berasal dari publikasi jurnal ilmiah/prosiding*** *hasil penelitian. Penulisan referensi secara alfabetis dan mengikuti gaya penulisan* ***American Psychological Association (APA) 6th Edition****. Manajemen penulisan referensi (dan kutipan) sangat disarankan menggunakan aplikasi Mendeley. Contoh penulisan referensi berdasarkan APA 6th Edition sebagai berikut:*

### Artikel Jurnal (satu, dua, atau lebih dari dua penulis)

Alasmari, S. F., & Dahab, M. (2017). Sentiment Detection, Recognition, and Aspect Identification. *International Journal of Computer Applications*, 177(2), 31- 38. http://dx.doi.org/10.5120/ijca2017915675

Aqlan, A. A. Q., Manjula, B., & Naik, R. L. (2019). A Study of Sentiment Analysis: Concepts, Techniques, and Challenges. *Proceedings of International Conference on Computational Intelligence and Data Engineering*, 147-162.

https://doi.org/10.1007/978-981-13-6459-4_16

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *IEEE Symposium Series on Computational Intelligence*, 159-167.
https://doi.org/10.1109/SSCI.2015.33

Etaiwi, W., & Naymat, G. (2017). The Impact of applying Different Preprocessing Steps on Review Spam Detection. *The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks*, 274-279.
http://dx.doi.org/10.1016/j.procs.2017.08.368

Handini, V. A., & Dunan, A. (2019). Buzzer as the Driving Force for Buzz Marketing on Twitter in the 2019 Indonesian Presidential Election. *International Journal Of Science, Technology & Management*, 479-491.
https://doi.org/10.46729/ijstm.v2i2.172

Ismail, M., Hassan, N., & Bafjaish, S. S. (2020). Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task. *Journal of Soft Computing and Data Mining*, 1(2), 1-10.
http://dx.doi.org/10.30880/jscdm.2020.01.02.001

Juzar, M. T., & Akbar, S. (2018). Buzzer Detection on Twitter Using Modified Eigenvector Centrality. *International Conference on Data and Software Engineering*, 1-5. https://doi.org/10.1109/ICODSE.2018.8705788

Leung, C. K., Chen, Y., Hoi, C. S. H., Shang, S., & Cuzzocrea, A. (2020). Machine Learning and OLAP on Big COVID-19 Data. *International Conference on Big Data (Big Dat*a), 5118-5127.

https://doi.org/10.1109/BigData50022.2020.9378407

Luo, H., Huang, W., Chen, C., Kangqiang, X., & Fan, Y. (2018). An Empirical Study on the Impact of Negative Online Word-of-mouth on Consumer's Purchase Intention. *International Conference on Service Systems and Service Management*, 1-6. https://doi.org/10.1109/ICSSSM.2018.8465093

Manjari, K. U., Rousha S., Sumanth D., & Devi J. S. (2020). Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm. *International Conference on Trends in Electronics and Informatics*, 648- 652. https://doi.org/10.1109/ICOEI48184.2020.9142938

Maulana, A., & Kuswayati, Sri. (2021). Klasifikasi Akun Buzzer Pemilu Pada Media Sosial Twitter Berdasarkan Data Tweet Menggunakan Algoritma C4.5. *Jurnal Ilmiah Nasional Riset Aplikasi dan Teknik Informatika*, 3(2), 30-35. https://doi.org/10.53580/naratif.v3i02.132

Nongthombam, K., & Sharma, D. (2021). Data Analysis using Python. *International Journal of Engineering Rresearch & Technology*, 10(7), 463- 468. https://doi.org/10.17577/IJERTV10IS070241

Ping, H., & Qin, S. (2018). A Social Bots Detection Model Based on Deep Learning Algorithm. International Conference on Communication Technology, 1435- 1439. https://doi.org/10.1109/ICCT.2018.8600029

Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019). Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. *International Conference on System Modeling & Advancement in Research Trends*, 266-270. http://dx.doi.org/10.1109/SMART46866.2019.9117512

Ramadhani, R. A., Indriani, F., & Nugrahadi, D. T. (2016). Comparison of Naive Bayes Smoothing Methods for Twitter Sentiment Analysis. *International Conference on Advanced Computer Science and Information Systems*, 287- 292. https://doi.org/10.1109/ICACSIS.2016.7872720

Stancin, I., & Jovic, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. *International Convention on Information and Communication Technology, Electronics and Microelectronics*, 977-982.

Suciati, S., Wibisono, A., & Mursanto, P. (2019). Twitter Buzzer Detection for Indonesian Presidential Election. *International Conference on Informatics and Computational Sciences*, 1-5. https://doi.org/10.1109/ICICoS48119.2019.8982529

**Artikel Prosiding**

Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019). Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest. Proceedings of the Third International Conference on Computing Methodologies and Communication, 679-684. https://doi.org/10.1109/ICCMC.2019.8819654

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of SocialMediaText. Proceedings of the Eighth International AAAI Conference on Weblogs and SocialMedia, 216-225. https://doi.org/10.1609/icwsm.v8i1.14550

**Buku Teks**

Chai, C. P. (2020). *The importance of data cleaning: Three visualization exampl*es. CHANCE, 33(1), 4-9

Verspoor, K., & Cohen, K. B. (2013). *Natural Language Processing. Encyclopedia of Systems Biology*, 1495–149

**Sumber dari internet dengan nama penulis**

Beri, A. (2020). SENTIMENTAL ANALYSIS USING VADER. Available at: https://towardsdatascience.com/sentimental-analysisusing-vader-a3415fef7664, diakses tanggal 10 November 2022

Gupta, S. (2018). Sentiment Analysis: Concept, Analysis and Applications. Available at: https://towardsdatascience.com/sentimentanalysis-concept-analysis-and-applications-6c94d6f58c17, Diakses tanggal 05 November 2022.