Jurnal ICT : Information Communication & Technology
Vol. 21, N0.1, Juli 2022, pp. 99-105
p-ISSN: 2302-0261, e-ISSN: 2303-3363, DOI:

Perbandingan *Naïve Bayes* dan *Support Vector Machine* pada Sentimen Analisis Reputasi Brand Twitter DQLab.id

Renyta Kristianti Dangin[1*], Ferdy Febriyanto[2], Renny Puspita Sari[3]
[1]Program Studi Sistem Informasi, Universitas Tanjungpura, Indonesia
[2,3]Program Studi Sistem Informasi, Universitas Tanjungpura, Indonesia
Email: [1]renytakristianti@student.untan.ac.id, [2]ferdy@sisfo.untan.ac.id, [3]rennysari@sisfo.untan.ac.id

INFORMASI ARTIKEL

Histori artikel:
Naskah masuk, 9 Juni 2022
Direvisi, 12 Juni 2022
Diiterima, 12 Juni 2022

Kata Kunci:

Data Mining,
Analisis Sentimen,
NBR,
Twitter

ABSTRAK

Abstract- The number of companies that use social media such as Twitter to conduct SMM (Social Media Marketing) makes the need for data and information that can help SMM succeed greater. The way to maximize social media marketing is to analyze the company's brand reputation through social media, such as Twitter, to see how their company's reputation is among the public. On condition when there are more positive public opinions, the company's reputation is acceptable to the public, and vice versa. This study calculated the reputation of DQLab.id using the NBR (Net Brand Reputation) formula and collected the data by utilizing Twitter API, then sentiment analysis was applied and compared Naïve Bayes and Support Vector Machine, finally carried out the data validation stage by using the evaluation parameters. The output of this research will be in the form of a dashboard tableau. Based on the NBR calculation, the company DQLab.id has a percentage value of 56%, and the comparison results show that Support Vector Machine is exceed than Naïve Bayes Classifier based on the values of all evaluation parameters.

Abstrak- Banyaknya perusahaan yang memanfaatkan media sosial seperti Twitter untuk melakukan *SMM* (*Social Media Marketing*) membuat dibutuhkannya suatu data/informasi yang dapat membantu kesuksesan *SMM*. Cara untuk memaksimalkan *Social Media Marketing* adalah dengan melakukan analisis reputasi brand perusahaan melalui media sosial seperti Twitter, untuk melihat reputasi perusahaan mereka dikalangan masyarakat dimana jika opini positif masyarakat lebih banyak maka reputasi perusahaan baik di mata publik dan sebaliknya. Penelitian ini melakukan perhitungan reputasi dari DQLab.id menggunakan rumus NBR (*Net Brand Reputation)* dengan pengambilan data menggunakan Twitter API, selanjutnya dilakukan analisis sentimen membandingkan *Naïve Bayes* dan *Support Vector Machine*, setelah itu dilakukan tahap validasi data dengan menggunakan parameter evaluasi dan output dari penelitian ini akan berbentuk dashboard tableau. Berdasarkan pada perhitungan *NBR*, perusahaan DQLab.id memiliki persentase nilai sebesar 56% dan hasil perbandingan kedua metode tersebut didapatkan bahwa *Support Vector Machine* lebih baik daripada *Naïve Bayes Classifier* berdasarkan nilai dari seluruh parameter evaluasi.

Penulis Korespondensi:

Renyta Kristianti Dangin
Program Studi Sistem Informasi,
Universitas Tanjungpura
Jl. Prof.Dr.H. Hadari Nawawi, Pontianak, Indonesia
Email: renytakristianti@student.untan.ac.id

Jurnal ICT : Information Communication & Technology
Vol. 21, N0.1, Juli 2022, pp. 99-105
p-ISSN: 2302-0261, e-ISSN: 2303-3363, DOI:

1. Pendahuluan

Pengguna media sosial di Indonesia meningkat berdasarkan penelitian Simon Kemp, pada tahun 2021 di Indonesia pengguna aktif media sosial sebanyak 170 Juta dengan persentase sebesar 61,8% populasi masyarakat di Indonesia [1], dan tahun 2022 sebanyak 191 Juta dengan persentase 68,9%. Media Sosial Twitter menempati posisi ke-6 pada urutan "*Most-Used Social Media Platform*" sebesar 58,3% populasi masyarakat Indonesia pengguna aktif media sosial di tahun 2022 [2]. Angka dari jumlah pengguna media sosial di Indonesia menjadi kunci penting dalam kesuksesan *Social Media Marketing* yang dijalankan oleh perusahaan, dimana *Social Media Marketing* merupakan proses bisnis yang dijalankan dengan memanfaatkan Media Sosial [3]. Dengan fenomena ini, banyak perusaahan mulai menarik perhatian konsumen melalui proses *Social Media Marketing* salah satunya DQLab.id. Karena *Social Media Marketing* dijalankan pada media sosial yang didasarkan pada opini masyarakat, maka proses ini sangat bergantung pada reputasi perusahaan.

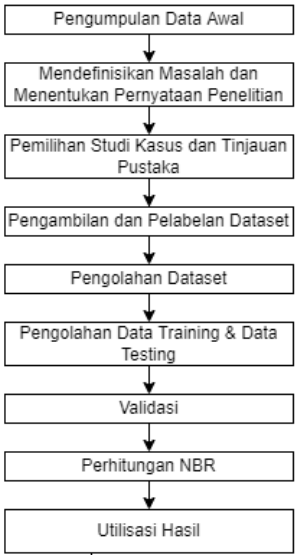
Terdapat berbagai cara untuk menjaga reputasi perusahaan yaitu dengan melakukan Social Media Analytics atau menghitung nilai reputasi brand menggunakan NBR (*Net Brand Reputation*), yang merupakan nilai bersih dari digital *brand* suatu perusahaan yang didapatkan melalui proses Sentimen Analisis [5]. *Social Media Analytics* pada media sosial Twitter DQLab.id dilakukan menggunakan *SocialBakers* namun tidak dapat memberikan informasi mengenai persentasi NBR karena tidak menganalisis jumlah sentimen seperti yang dibutuhkan dalam perhitungan NBR dan hanya menganalisis jumlah likes, retweets, dan following dari DQLab.id, pada proses Sentimen Analisis dapat secara langsung menganalisa opini masyarakat baik positif, negatif atau netral terhadap perusahaan DQLab.id sehingga bisa mendapatkan nilai NBR.

Berdasarkan penelitian terdahulu tentang analisis sentimen twitter terhadap reputasi brand dengan SVM, *Decision Tree* dan NBC, mencapai kesimpulan yaitu SVM memiliki performa terbaik dan diikuti NBC kemudian *Decision Tree* [12]. Analisis sentimen juga digunakan pada penilaian masyarakat Indonesia di Twitter mengenai vaksin COVID-19 dengan metode NBC dan menghasilkan nilai akurasi 93% [13]. Algoritma NBC juga digunakan untuk analisis sentimen data twitter tentang masalah obesitas di Indonesia dengan akurasi NBC mencapai 93% [14]. Ketiga Penelitian tersebut menggunakan *accuracy* dan teknik *cross validation* untuk menentukan algoritma dengan performa terbaik dan pada penelitian ini penentuan performa algoritma tidak hanya berdasarkan pada *accuracy* melainkan berdasarkan pada sejumlah parameter evaluasi yaitu *accuracy, balanced accuracy score, precission, recall dan f1-score*.

Berdasarkan permasalahan tersebut maka dilakukanlah penelitian untuk menghitung NBR dari perusahaan DQLab.id dengan dilakukannya proses sentimen Analisis, selain itu proses ini mengaplikasikan *Naïve Bayes* dan *Support Vector Machine* selanjutnya dibandingkan untuk mencari performa superior dalam kasus data DQLab.id berdasarkan pada parameter evaluasi.

2. Metode Penelitian

2.1 Kerangka Penelitian



Gambar 1. Kerangka Penelitian

a. Pengumpulan Data Awal

Survey pada posisi dan arah media sosial terhadap brand suatu perusahaan di masa sekarang, membuat peneliti paham akan kondisi dan kebutuhan di masa sekarang sehingga, penelitian dapat benar-benar berhubungan.

b. Mendefinisikan Masalah dan Menentukan Pertanyaan Penelitian

setelah diketahui kondisi saat ini, kemudian didefinisikan permasalahan penelitian dan membuat serangkaian pertanyaan penelitian untuk dijawab oleh hasil penelitian.

c. Proses Pemilihan Studi Kasus dan Tinjauan Pustaka

Tahap ini dipilih studi kasus apa yang akan diambil dan tinjauan pustaka mana yang akan dijadikan acuan selama penelitian

d. Pengambilan dan Pelabelan Dataset

Tahap ini merupakan tahap pengambilan data dari Twitter *API* atau *Twitter Crawling* dengan rentang waktu dari tanggal 1 Januari – 8 Maret 2022. Setelah didapatkan data opini dari perusahaan yang akan dijadikan studi kasus, maka dilakukan pelabelan secara manual untuk menentukan apakah twit tertentu masuk kedalam sentiment positif atau negatif. Pelabelan dilakukan secara manual oleh 3 orang dengan latar belakang dan dipilih pelabelan mana yang paling banyak dipilih masuk kedalam kategori tertentu baik positif/negatif.

e. Pengolahan Dataset

Tahap ini adalah tahap seperti data *filtering, removing duplicate,* Pre-processing untuk menghasilkan data yang optimal. Rangkaian proses yang dilakukan adalah dokumen filtering, tokenization, normalization dari kata gaul/slang word, stopword, stemming

f. Pengolahan Data Testing dan Data Training

Dataset yang telah didapatkan dipilah menjadi data testing dan training dengan rasio 70:30 [10]. kemudian dilakukan proses pemodelan klasifikasi menerapkan *Naïve Bayes* dan *Support Vector Machine* pada dataset *Training*, sedangkan *dataset Testing* diaplikasikan pada hasil model klasifikasi untuk menghasilkan hasil nilai klasifikasi sentimen.

g. Validasi

Tahap ini dilakukan perhitungan nilai validasi dari data di masing-masing hasil metode. Pada proses validasi ini dilakukan perhitungan dengan mencari nilai *Accuracy, Balanced Accuracy Score, Precission, recall, F1-Score*.

h. Perhitungan *Net Brand Reputation*

Proses ini menggunakan inputan berupa dataset yang sudah dilakukan pelabelan, yang kemudian akan dimasukkan ke dalam rumus untuk mengukur *Net Brand Reputation* pada persamaan 1

$$NBR = \frac{(Positive\ Mentions - Negative\ Mentions)}{(Positive\ Mentions + Negative\ Mentions)} \times 100\% \quad (1)$$

i. Utilisasi Hasil

Tahap ini akan dibuat suatu visualiasi data *wordcloud* dan *dashboard* tableau untuk mempermudah pemaparan pada *stakeholder* akan hasil dari penelitian.

2.2 *Naïve Bayes*

Merupakan metode klasifikasi dimana sangat cepat dan sederhana serta cocok untuk data yang bersifat high-dimensional. Karena terkenal sangat cepat dan memiliki parameter yang dapat disesuaikan, algoritma ini sering kali berguna untuk melakukan klasifikasi dengan cepat. Naive Bayes Classifier dibuat berdasarkan Teori Bayessian, dimana merupakan persamaan yang menggambarkan hubungan probabilitas bersyarat besaran statistik [6].

2.3 *Support Vector Machine*

Merupakan metode *machine learning* klasifikasi. Metode ini memiliki cara kerja untuk memaksmilkan margin / jarak pada data dan memfinalisasikannya kedalam bentuk kuadrat untuk menyelesaikan suatu permasalahan [7]

2.4 Parameter Evaluasi

Untuk mengevaluasi suatu performa dari algoritma *machine learning*, dapat dilakukan dengan beberapa pengukuran seperti *accuracy, balanced accuracy score, precision, recall dan F1-Score* [8].

a. *Accuracy*

Merupakan rasio dari sampel arus suatu data yang diklasifikan dengan benar, dengan jumlah total sampel. *Accuracy* menghasilkan nilai dari performa model dengan dataset yang seimbang.

b. *Precision*

Pengukuran dari suatu rasio positif, yang arus datanya dengan benar memprediksi pada total jumlah prediksi klasifikasi positif. *Precision* menghasilkan nilai yang sebenarnya dengan menjawab pertanyaan “berapa persen mahasiswa yang benar lulus di semester 8 dari keseluruhan mahasiswa yang diprediksi lulus semester 8?”

c. *Recall*

Pengukuran dari rasio positif, yang arus datanya dengan benar memprediksi suatu kelas. *Recall* dapat menjawab pertanyaan “Berapa pertanyaan “Berapa persen mahasiswa yang diprediksi lulus di semester 8 dibandingkan keseluruhan mahasiswa yang sebenarnya lulus di semester 8?”

d. *F1-Score*

Pengukuran nilai rata-rata dari *Precision* dan *Recall. F1-Score* sangat dibutuhkan ketika dataset tidak seimbang

e. *Balanced Accuracy Score*

Pengukuran nilai akurasi seimbang yang digunakan pada *multi-class classification,* dan digunakan untuk menyelesaikan permasalahan mengenai data yang tidak seimbang, dimana salah satu kelas memiliki jumlah data lebih banyak dibandingkan kelas yang lain.

2.5 Tableau

Tableau merupakan *visual analytics platform* yang mengubah cara kita menggunakan data untuk menyelesaikan suatu permasalahan, membantu organisasi untuk memaksimalkan data yang digunakan [11]

3. Hasil dan Pembahasan

3.1 Jumlah Sentimen

Seluruh dataset didapatkan berjumlah 372 Twit, dengan jumlah data tiap sentimen seperti tabel 1

Tabel 1. Data Bersih per Sentimen

No	Sentimen	Jumlah
1	Positif	41
2	Netral	321
3	Negatif	10

namun setelah dilakukan serangkaian proses ditahap pengolahan dataset seperti data filtering yaitu menghapus tweet dari akun *Official* DQLab.id karena mengandung sentimen yang tidak merepresentasikan opini masyarakat, kemudian proses *Removing Duplicate* yaitu menghilangkan twit yang sama karena dari hasil *retweet* kemudian dilakukan serangkaian *preprocessing.* Sehingga menghasilkan jumlah data menjadi 160 twit, hasil akhir dari jumlah data setiap sentimen seperti tabel 2

Tabel 2. Data Bersih Setelah Pemrosesan per Sentimen

No	Sentimen	Jumlah
1	Positif	36
2	Netral	114
3	Negatif	10

Berdasarkan pada tabel 2 diketahui bahwa sebagian besar opini masyarakat terhadap DQLab.id bersentimen netral dengan jumlah 114 twit bersih, kemudian diikuti sentimen positif dengan jumlah twit sebanyak 36 dan sentimen negatif dengan jumlah twit sebanyak 10. Jumlah data tiap sentiment ini yang akan menjadi penentu nilai presentase dari reputasi brand DQLab.id.

3.2 Evaluasi Parameter Algoritma

Tabel 3. Nilai *accuracy* dan *Balanced Accuracy Score*

No	*Algoritma*	*Accuracy*	*Balanced Accuracy Score*
1	SVM	88%	83%
2	NBC	75%	67%

Berdasarkan pada tabel 3 dengan parameter evaluasi *Accuracy* dan *Balanced Accuracy Score* maka dapat dilihat *Support Vector Machine* lebih superior dari *Naïve Bayes*.

Tabel 4. Nilai Parameter Evaluasi

No	*Algoritma*	*Precission*	*Recall*	*F1-Score*
1	SVM	89%	88%	88%
2	NBC	81%	75%	72%

Berdasarkan pada tabel 4 dengan parameter evaluasi dari *Precission, Recall,* dan *F1-Score* dapat dilihat *Support Vector Machine* lebih superior performanya ketimbang *Naïve Bayes,* dimana dari ketiga parameter tersebut metode *Support Vector Machine* selalu bernilai lebih tinggi. Sehingga dapat diambil kesimpulan bahwa pada perbandingan menggunakan data DQLab, *Support Vector Machine* lebih superior performanya ketimbang *Naïve Bayes.*

3.3 Net Brand Reputation

Perhitungan NBR menggunakan rumus pada persamaan 1

$$NBR = \frac{(36 - 10)}{(36 + 10)} \times 100\%$$

$$NBR = \frac{26}{46} \times 100\%$$

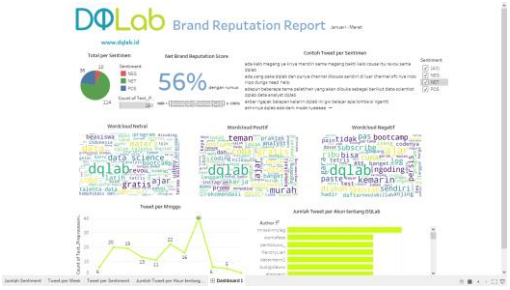
$$NBR = 0.56 \times 100\%$$

$$NBR = 56\%$$

Sehingga, menghasilkan nilai NBR sebesar 56%, dengan reputasi brand DQLab.id dalam rentang waktu 1 Januari – 8 Maret 2022.

3.4 Visualisasi Data

Visualisasi Data dilakukan dengan pembuatan wordcloud berdasarkan kelas sentimen dan dashboard Tableau, sehingga tampilannya adalah seperti gambar 2

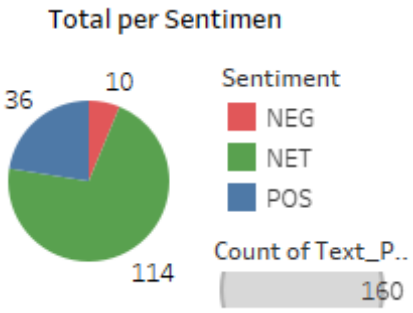


Gambar 2. Tampilan Dashboard Tableau

Gambar 2, berisi poin penting yang dapat menjelaskan kondisi reputasi brand yang didapatkan dari hasil proses sentiment analisis.

a. Visualisasi Jumlah per Sentimen

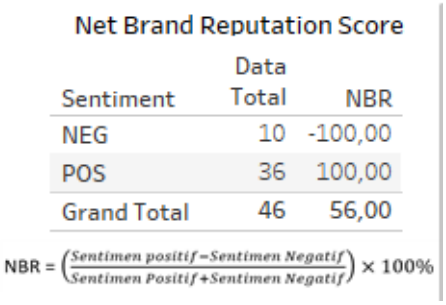
Memberikan visualisasi secara mudah mengenai jumlah berapa banyak sentiment yaitu Positif, Netral dan Negatif dalam tampilan *pie chart*, seperti pada gambar 3



Gambar 3. Visualiasi Jumlah per Sentimen

b. Visualisasi tampilan nilai NBR (*Net Brand Reputation*)

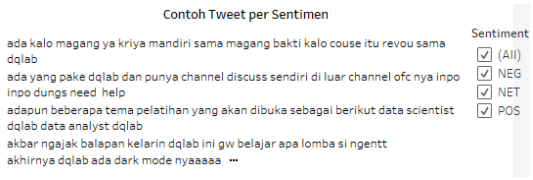
Pada bagian ini, menampilkan tampilan nilai persentase dari NBR seperti gambar 4



Gambar 4. Visualisasi tampilan nilai NBR

c. Visualisasi Contoh Twit per Sentimen

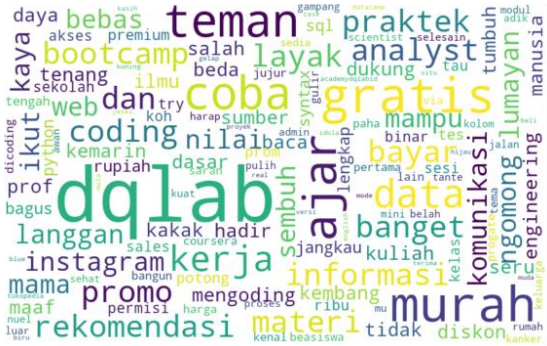
Pada bagian ini menampilkan tampilan contoh tweet dari tiap sentimen, dengan tampilan seperti gambar 5



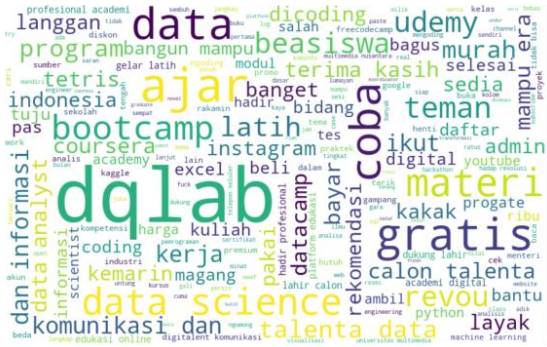
Gambar 5. Visualisasi contoh twit per sentimen

d. Visualisasi *Wordcloud*

Menampilkan visualisasi *wordcloud* yang dibagi kedalam 3 *wordcloud*, yaitu *wordcloud* seluruh sentimen, *wordcloud* Positif dan *wordcloud* negatif. *Wordcloud* menampilkan variasi kata yang terdapat pada data, dimana semakin sering kata tersebut disebutkan maka semakin besar tampilan katanya dan sebaliknya sehingga *wordcloud* mampu memberikan informasi mengenai kondisi suatu perusahaan berdasarkan kalimat apa yang sering dibicarakan. Tampilan *wordcloud* pada *dashboard* tableau seperti gambar 6 sampai gambar 8.



Gambar 6. Visualisasi Wordcloud Positif

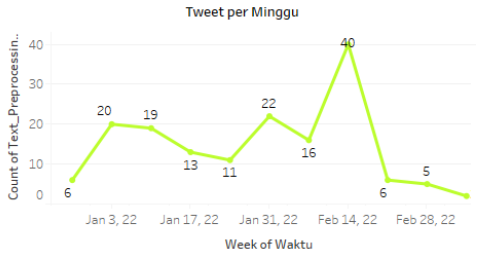


Gambar 7. Visualisasi Wordcloud Netral

Gambar 8. Visualisasi Wordcloud Negatif

e. Visualiasi Jumlah Twit tentang DQLab per minggu

Pada bagian ini, menampilkan tampilan grafik data jumlah twit yang membicarakan tentang DQLab per minggu dengan hasil visualisasinya seperti pada gambar 8

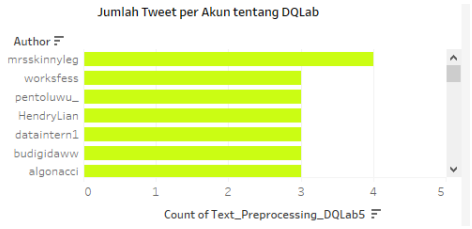


Gambar 9. Visualisasi jumlah twit tentang DQLab per minggu

Maka dari visualisasi di gambar 9 dapat dilihat bahwa jumlah twit tiap minggunya mengalami kenaikan dan penurunan yang signifikan. Contohnya pada periode tertentu seperti pada minggu ke1 bulan Februari terjadi kenaikan signifikan terhadap jumlah twit karena adanya program yang sedang dijalankan DQLab bekerja sama dengan KOMINFO.

f. Visualisasi Jumlah Twit per Akun

Pada bagian ini, menampilkan hasil visualisasi Jumlah Twit per akun, untuk melihat akun mana yang sering membahas mengenai DQLab lewat twitnya, dengan visualisasi seperti pada gambar 10



Gambar 10 Visualisasi Jumlah Twit per Akun

4. Kesimpulan

Maka berdasarkan penjelasan penelitian sebelumnya, kesimpulannya yaitu:

a. Dalam rentang waktu 1 Januari 2022 – 8 Maret 2022 ditemukan bahwa sebagian besar hasil sentimen yang didapatkan dari data Twit masyarakat Indonesia mengenai DQLab adalah 372 twit dengan sentimen Netral dengan jumlah sebesar 321 Twit, Positif berjumlah 41 Twit dan Negatif berjumlah 10 Twit, dan setelah dibersihkan data twit menjadi berjumlah 160 twit, dengan sentimen netral berjumlah 114, sentiment positif berjumlah 36 dan sentimen netral berjumlah 10.

b. Berdasarkan hasil perbandingan didapatkan *Support Vector Machine* lebih superior performanya ketimbang *Naïve Bayes* di seluruh parameter validasi.

c. Pada perhitungan *NBR (Net Brand Reputation)* maka diketahui DQLab memiliki nilai reputasi brand sebesar 56%.

d. Pada penelitian selanjutnya sebaiknya dilakukan proses klasifikasi yang tidak hanya terpaku dengan metode yang sama.

Daftar Pustaka

[1] S. Kemp, “Digital 2021 Indonesia,” Hootsuite(We are Social), 2021.

[2] S. Kemp, “Digital 2022 Indonesia,” Hootsuite(We are Social), 2022.

[3] Buffer, “What is Social Media Marketing?,” *Buffer*, 2022. [Online]. Available: https://buffer.com/social-media-marketing. [Accessed 7 June 2022]

[4] N. A. Vidya, M. I. Fanany, and I. Budi, “Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers,” in *Procedia Computer Science*, 2015, vol. 72, pp. 519–526.

[5] A. Iva “Brand Reputation 101: Monitoring, Analysis and Management Tools”*Mediatoolkit*,.2022.[Online].Available: https://www.mediatoolkit.com/blog/brand-reputation-monitoring-analysis-management-tools/. [Accessed 8 June 2022]

[6] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*. O’Reilly Media, 2016. [Online]. Available: https://books.google.co.id/books?id=6omNDQAAQBAJ

[7] B. Gaye, D. Zhang, and A. Wulamu, “Improvement of Support Vector Machine Algorithm in Big Data Background,” *Mathematical Problems in Engineering*, vol. 2021, p. 5594899, 2021, doi: 10.1155/2021/5594899.

[8] R. M. AlZoman and M. J. F. Alenazi, “A Comparative Study of Traffic Classification Techniques for Smart City Networks,” *Sensors*, vol. 21, no. 14, 2021, doi: 10.3390/s21144677.

[9] P. K. Singh, B. K. Panigrahi, N. K. Suryadevara, S. K. Sharma, and A. P. Singh, *Proceedings of ICETIT 2019: Emerging Trends in Information Technology*. Springer International Publishing, 2019. [Online]. Available: https://books.google.co.id/books?id=kIixDwAAQBAJ

[10] A. Gholamy, V. Kreinovich, and O. Kosheleva, "A Pedagogical Explanation A Pedagogical Explanation Part of the Computer Sciences Commons," 2018. [Online]. Available: https://scholarworks.utep.edu/cs_techrephttps://scholarworks.utep.edu/cs_techrep/1209

[11] Tableau, "What is Tableau?," *Tableau*, 2021. [Online] Available: https://www.tableau.com/why-tableau/what-is-tableau [Accessed 12 Desember 2021]

[12] N. A. Vidya, M. I. Fanany, and I. Budi, "Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers," in *Procedia Computer Science*, 2015, vol. 72, pp. 519–526. doi: 10.1016/j.procs.2015.12.159.

[13] W. Yulita *et al.*, "Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier," *JDMSI*, vol. 2, no. 2, pp. 1–9, 2021.

[14] F. Fridom Mailo *et al.*, "Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia," 2019.