e-ISSN:2714-6529

Sentiment Analysis Dataset on COVID-19 Variant News

[1]Fakhri Muhammad, [2]Nana Mulyana Maghfur, [3]Apriade Voutama
[1,2]Program Studi Teknik Informatika, Universitas Singaperbangsa Karawang
[3]Program Studi Sistem Informasi, Universitas Singaperbangsa Karawang
Email: fakhri.muhammad18054@student.unsika.ac.id

Abstract

The development of technology at this time is getting faster and faster, this is indicated by the number of emerging social media such as Facebook, Instagram, Twitter. Twitter is used as a forum for users to discuss, express opinions, and share stories between users, because many people today often have opinions about the COVID-19 outbreak, plus there are new variants that make people express various types of opinions, both good and bad. good and so on. Therefore, an effort was made to research the covid variant to see labeling sentiment, which in essence is a text mining process that aims to extract sentiment from text using regular expressions so that labels are obtained for each text in the dataset, so a dataset is formed that can be used for further research. The process includes data collection including (scrapping tweets, data tweets), pre-processing (case folding, remove URLs, remove stop words, change into standard words, stemming, tokenization), sentiment labeling (IR in the form of regular expressions, sentiment labeling), and data visualization show pie chart, show word cloud). Of the 8993 tweets that have been analyzed, 2213 positive tweets, 1735 negative tweets, and 5045 neutral tweets were found.

Keywords: *Dataset, Regular expression, Tweets data, Covid variant.*

1. INTRODUCTION

Wabah Covid-19 adalah wabah penyakit terbesar yang terjadi dalam 2 tahun kebelakang ini, awal mula covid-19 ini yaitu datang dari negara China tepatnya di kota Wuhan China. Wabah penyakit ini terus tersebar di beberapa negara maju dan berkembang yang membuat banyak yang terinfeksi penyakit ini, Indonesia menjadi salah satu negara berkembang yang ikut terdampak paparan wabah covid ini. Wabah ini sangat cepat menular umumnya bagi orang yang sudah tua ”manula” rentang umur 50-60 th [1], dan bagi tubuh kita yang sedang kurang sehat membuat virus dengan cepat men-infeksi tubuh seseorang.

Wabah covid-19 membuat banyak kerugian khususnya masyarakat indonesia, para buruh banyak yang mendapat PHK dan para pedagang yang dibatasi jumlah pembelinya agar tidak timbulnya keramaian [2]. Wabah covid-19 ini juga menimbulkan banyak opini berbeda dari masyarakat indonesia di jejaring social media seperti meta dan twiter, Ada masyarakat yang percaya, tidak percaya dan biasa saja dalam menganggapi berita varian covid-19 yang ditampilkan disocial media contohnya twitter. Oleh karena itu harus ada upaya untuk mengumpulkan data tentang reaksi masyarakat di social media terhadap berita varian covid-19 yang kerap kali susah dalam penanggulangan penyebaran covid. Data yang sering digunakan untuk proses analisis sentimen merupakan data subjektif, yaitu seperti opini, yang tidak mempunyai nilai konkret. Ditambah lagi dengan nilai bersumber dari manusia yang

mempunyai pendapat yang berbeda satu sama lain dalam menanggapi berita covid-19 ini. Oleh karena itu dengan kurangnya data-data berkualitas masih terbilang minim dan kurangnya data-data pendukung sangatlah sulit untuk membuat model analisis sentimen.

Twitter adalah salah satu social media populer yang tepat digunakan untuk sumber data pada analisis teks, dilihat pada Tabel 1. Hal ini karena media social twitter cocok digunakan dalam proses analisis sentimen dilihat dari tulisan-tulisan pada media social twitter, atau sering disebut tweet. Mempunyai struktur yang cocok untuk proses analisis. Tidak heran jika banyak penelitian lain sering menggunakan twitter untuk membuat dataset-dataset yang bagus, sebagai contoh Indonesian-Emotion-Twitter Dataset [3]. Media social berpotensi untuk mengubah karakter kehidupan sosial penggunanya secara mendasar [4], media social terdiri dari situs web komunikasi yang memfasilitasi pembentukan hubungan antara pengguna dari berbagai latar belakang, menghasilkan struktur sosial yang kuat [5] [6]. Dengan ini dipilih media social twitter sebagai sumber data untuk proses pembuatan dataset yang nanti akan digunakan untuk proses analisis sentimen terhadap berita varian covid-19. Bahasa pemrograman dalam penelitian ini menggunakan python yang dibuat oleh Guido Van Rossum tahun 1989 dan dirilis pada tahun 1991 [7]. Python dapat digunakan untuk sebuah proyek kecil maupun besar dan dapat digunakan untuk membuat aplikasi standalone(berdiri sendiri) dan pemrograman script [8].

Yogi, Ikhwan, dan Syamsul [9] membuat penelitian tentang analisis log web server pengunjung website dengan teknik regex, penelitian ini bertujuan membuat sebuah sistem yang berfungsi sebagai alat untuk mengetahui informasi ativitas pengunjung pada website menggunakan data acces log dan regex, hasil dari pengujian ini berhasil mendeteksi beberapa aktifitas berbahaya pada website. Pada penelitian yang dilakukan oleh Simon dan Seng [10] tentang cara mengatasi spam filter situs jejaring sosial mahasiswa menggunakan regex, tujuan spam itu sendiri yaitu mengirim sebuah informasi kepada penerima, dimana nanti isi dari pesan terkirim berupa iklan, produk ilegal, atau menyebarkan suatu malware yang dirancang untuk membajak komputer penerima. Oleh karena itu masalah spam ini harus diatasi, hasil penelitian mahasiswa menggunakan regex untuk mencegah spam atau sebagai filtering spam telah berhasil dirancang dan dibangun dengan nama *(social Networking Service UMN)* untuk mengatasi permasalahan terjadinya spam ini di jejaring sosial mahasiswa. Pada penelitian Fenina dan Rena [11] perbandingan metode binary search dan regex, menyelesaikan permasalahan dalam sistem pencarian. Hasil dari penelitian ini yaitu kedua metode ini mempunyai keunggulan masing-masing, metode binary search ini unggul dalam kecepatan dan kesederhanaan dan regex lebih fleksibel dan fungsional. Namun kelebihan regex yaitu tidak membutuhkan data yang sudah terurut dan memungkinkan kita mencari data secara acak.

Dalam penelitian ini dilakukan scrapping data tweets menggunakan IR rule *regular expression* untuk fokus menghasilkan sebuah dataset yang akan digunakan untuk penelitian selanjutnya yaitu analisa pengaruh sentimen varian covid.

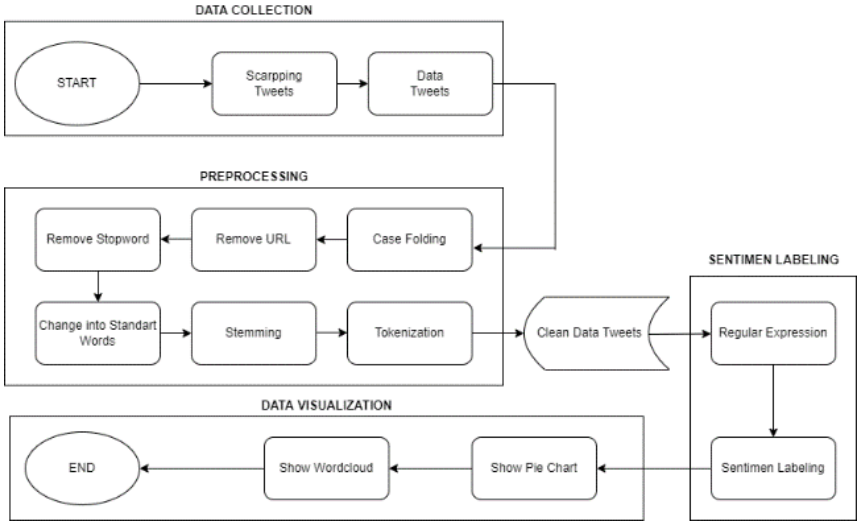
2. METHODS

Penelitian ini menggunakan metodologi flowchart seperti terlihat pada Gambar 1. Merupakan flowchart dari program analisis pembuatan dataset untuk analisis sentimen. Terdapat beberapa proses text mining yang dilakukan dengan tools dan proses yang digunakan sebagai berikut :

1. Python sebagai toolsnya.
2. Mengambil data tweet menggunakan twitter API.
3. Pra-pemrosesan data tweet.
4. Rule IR(information retrieval) berupa *regular expression* dan sentimen labeling
5. Eksplorasi menggunakan pie chart
6. Visualisasi data tweet menggunakan word cloud
7. Menafsirkan dan membuat kesimpulan dari hasil analisis.

Tabel 1. Perbandingan Media Social Untuk proses text mining

Pembanding	**Twitter**	**Facebook**	**Instagram**
Jenis Data	Lebih banyak teks, dan sedikit foto	Terdapat teks, foto, video, games, dan lainnya.	Sebagian besar nonteks, teks hanya untuk caption
Isi Data	Lebih banyak opini (tweet)	Terdapat (opini, berita, cerita)	Teks hanya untuk caption suatu gambar
Panjang Data	Jumlah maksimal 280 karakter	Total maksimal 63.206 karakter	Maksimal 2.200 karakter



Gambar 1. Flowchart Metodologi Penelitian Umum

Lakukan scarapping pada tweets, pengambilan data tweets rentan 1 Januari 2021 sampai dengan 31 Oktober 2021, kemudian lakukan langkah pra-pemrosesan meliputi case folding (pengubahan

menjadi huruf kecil), penghilangan simbol (URL, karakter, khusus), Remove stopword (membuang kata yang tidak bermakna) penanganan negasi(standar kata), stemming (penghapusan imbuhan dan sufiks), tokenizing (pemisahan rantai kata dalam kalimat), dan setelah itu lakukan proses *Regular Expression* (regex) adalah cara untuk membedakan data positif, negatif, dan netral. Kemudian lakukan sentimen labeling, Langkah terakhir yaitu visualisasi data, menunjukkan 2 jenis. Ada bentuk pie chart, dan awan kata(word cloud).

2.1 Sumber data

Sumber data yang diperoleh pada penelitian ini adalah berbagai macam tweet dari pengguna media social twitter tentang (varian covid) sebagai kata kunci. Khususnya yang tweet antara tanggal 1 Januari 2021 sampai dengan 31 Oktober 2021, Data tweet dapat diperoleh melalui *API Twitter* (Antarmuka pemrograman aplikasi). Total ada 8993 data diambil yang kemudian dikompilasi menjadi satu file. Struktur data yang digunakan pada penelitian ini menggunakan data setelah diolah terlebih dahulu melalui beberapa proses. Data dari kumpulan teks tweet terdiri dari variabel prediktor yaitu clean tweet dan variabel respon adalah sentimen yaitu diantaranya(Positif, Negatif, Netral). Untuk menentukan tweet menggunakan *regular expression* dan diakhir keputusan, diperkuat dengan sentimen yang didasarkan pada aturan[12].

2.2 Penerimaan Data

Dalam pengumpulan data, peneliti menggunakan Application Programming Interface(API) yang sudah tersedia di twitter developer dan dapat diakses secara gratis [13]. Untuk mengetahui twitter API, seorang peneliti harus sudah mempunyai akun twitter kemudian mendaftar ke twitter developer. Ketika anda berhasil mendaftar sebagai pengembang, Twitter akan memberkan token akses dan kunci konsumen sebagai perantara untuk mengakses twitter API. Selanjutnya token dan kunci tersebut akan digunakan untuk mengambil data yang telah dikodekan menggunakan bahasa python.

2.3 Pra-pemrosesan Data

Tahap Pra-pemrosesan data adalah proses dimana menyeleksi data yang tidak sesuai dan mengubah data tersebut menjadi bentuk yang lebih mudah diproses oleh sistem [14] [15]. Pra pemrosesan data teks tweet yang sudah dikumpulkan, adalah suatu keharusan. Tujuannya adalah untuk mengurangi data teks dengan noise-heavy ke dalam versi yang lebih bersih sehingga menjadi tampilan yang lebih baik di word cloud. Pra-pemrosesan meliputi :

1. Penghapusan simbol(URL dan karakter khusus)
2. penanganan negasi
3. Remove stopword
4. Stemming
5. Tokenizing.

6. Case folding

Berikut ini adalah struktur data tweet sebelum dilakukan pra-pemrosesan atau preprocessing. Tabel 2 menunjukkan tweet yang masih mengandung URL, username, dan simbol ASCII yang tidak diperlukan atau nahkan menghalangi visualisasi, oleh karena itu harus dihapus terlebih dahulu. Langkah selanjutnya adalah penanganan negasi untuk mengubah kata-kata yang disingkat menjadi format yang tepat. Kemudian lakukan remove stopword yaitu membuang kata yang tidak bermakna seperti "yang", "di", "ke". Setelah itu lakukan proses stemming yaitu merubah sebuah kata menjadi kata dasar kemudian untuk meningkatkan akurasi klasifikasi, langkah selanjutnya adalah tokenizing. Tokenizing yaitu mengubah setiap tweet menjadi bagian-bagian kata yang membentuk kalimat. Terakhir, case folding untuk mengubah tweet menjadi huruf kecil.

Tabel 2. Tweet Data sebelum Pra-pemrosesan

Bahasa Indonesia
@liputan6dotcom itulah ajaibnya negeri ini
@liputan6dotcom Agar lebih cepat membunuh rakyat
@cnbcindonesia Bandar gateli
@cnbcindonesia Longsor dulu
@cnbcindonesia iya besok2 aja dah, aing pgn serok bawah

Tabel 3 menunjukkan data tweet setelah *pre-processing* sehingga data lebih mudah diolah menjadi sesuatu yang informatif. Berikut adalah data teks setelah pra-pemrosesan.

Tabel 3. Tweet Data setelah Pra-pemrosesan

Bahasa Indonesia
itulah ajaibnya negeri ini
agar lebih cepat membunuh rakyat
bandar gateli
longsor dulu
iya besok2 aja dah aing pgn serok bawah

2.4 Regular Expression

Regex adalah rule proses yang banyak digunakan untuk pencarian dan memanipulasi suatu teks berdasarkan pola, dan biasa dioperasikan oleh banyak teks editor, *utilities,* dan bahasa pemrograman. Regex yang baik dan kuat dibangun pada syntax, regex mempunyai 2 fungsi utama yaitu mencari dan mengganti, yang utama mencari suatu pola tertentu yang ada di dalam teks lalu menggantinya menjadi suatu pola yang lain[10]. Secara sederhana *regular expression* telah menyediakan cara untuk memanipulasi dan mencocokan string sesuai dengan formula yang telah dibuat[16]. Intinya ini merupakan proses text mining yang bertujuan untuk mengekstrak sentiment dari teks menggunakan *regular expression* sehingga didapatkan label untuk setiap teks dalam dataset, jadi terbentuklah dataset yang bisa digunakan untuk penelitian selanjutnya. Dengan berlandaskan lexicon[12]berisi kumpulan daftar kata positif dan negatif dalam bahasa indonesia, setiap kalimat *tweet* pada data yang dijadikan dat latih akan dicocokan berapa jumlah kata positif dan negatif didalamnya. Kemudian dihitunglah polaritas dari setiap tweet dengan rumus[17].

$$Skor\ Polaritas = \frac{\text{Jumlah Kata Positif} - \text{Jumlah Kata Negatif}}{\text{Jumlah Total Kata Positif Dan Negatif}}$$

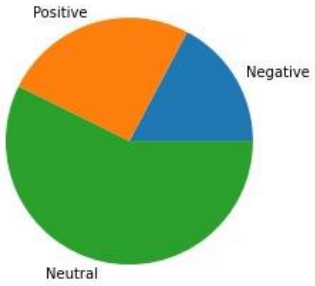
1. jika skor > 0, maka sentimen kalimat secara keseluruhan dianggap “positive”
2. jika skor < 0, maka sentimen kalimat secara keseluruhan dianggap “negative”
3. jika skor = 0, maka sentimen kalimat secara keseluruhan dianggap “netral”, menurut[18].

2.5 Visualisasi data

Visualisasi data pada penelitian ini menggunakan 2 macam yaitu dalam bentuk pie chart dan awan kata. Untuk pie chart, ini adalah grafik statistik yang terbuat dari lingkaran yang dibagi menjadi beberapa irisan yang luasnya sebanding dengan jumlah data yang dimiliki, jumlah total tweets 8993 dengan masing-masing data tweets positif sebesar(2213), negatif(1735), dan neutral(5045). Terakhir yaitu awan kata, ini adalah salah satu metode visualisasi dokumen teks, yang sering digunakan. Awan kata adalah gambaran grafis dari sebuah dokumen, proses yang dilakukan dengan memplot kata-kata yang sering muncul pada dokumen dalam ruang dua dimensi. Seberapa sering sebuah kata muncul dan ditunjukkan oleh ukuran kata tersebut. Semakin besar, semakin sering muncul dalam sebuah dokumen. Visualisasi word cloud dipilih karena kata-kata yang memiliki jumlah frekuensi kemunculan lebih banyak atau tinggi dapat ditampilkan secara menonjol dari sisi ukuran sehingga lebih intuitif[19].

Gambar 2. Word Cloud Positif

Gambar 3. Word Cloud Negatif



Gambar 4. Visualisasi Pie chart

3. RESULTS AND DISCUSSION

Python adalah bahasa pemrograman dengan banyak paradigma, yang mendukung pemrograman berorientasi objek, prosedural dan fungsional[20]. Bahasa pemrograman python memiliki beberapa modul diantaranya modul re untuk melakukan hal-hal yang berkaitan dengan *regular expression*(regex). Hasil dari program text mining ini ditunjukkan diantaranya sebagai berikut :

Tabel 4. Tweet Data setelah Pra-pemrosesan

No	*Clean Tweet*	Sentiment
	Bahasa Indonesia	
1	tim pantau aja	Neutral
2	asikkk arb	Neutral
3	iri tanda tak punya barang	Positive
4	untung arb masih 7 coba udah simetris hahaha	Positive
5	dah kuning min udah lu serok belum	Negative
6	taii adminnya belom beli ni pasti	Negative
7	seroooooooooookkkkkk	Neutral
8	tenang gabakalan wkwkwk	Positive
9	hahaha wis angel iki	Neutral
10	besok anjlok nih	Negative
…	…	…
8993	Tweet bersih 8993	8993 Sentiment

1. pra pemrosesan sebelum membersihkan data yang terlampir pada metode penelitian, bagian pengolahan data pada tabel 2.
2. pra pemrosesan setelah membersihkan data yang terlampir pada metode penelitian, bagian pengolahan data pada tabel 3.
3. Seluruh hasil sentimen yang sudah menjadi 1 dalam sebuah dataset pada tabel 4.
4. Menampilkan visualisasi pie chart, dengan 3 polaritas yaitu positif, negative, neutral Gambar 4.
5. Menampilkan word cloud dari polaritas yaitu, word cloud positif pada Gambar 2, negative pada Gambar 3, dan neutral.

4. CONCLUSION

Penelitian ini telah berhasil menggunakan *regular expression* untuk data tweets dan telah berhasil menvisualisasikan hasil dari ke tiga polaritas positif, negatif, neutral menggunakan word cloud. Hasil dari analisis yang diperoleh berupa dataset yang terdiri dari 8993 data tweets, ada 2213 positif tweet, 1735 negative tweet, 5045 neutral tweet. Kelebihannya yaitu metode anotasi label sentimen recara rule based menggunakan sintax regex lebih baik untuk bahasa indonesia karena mudah untuk di customize,

cepat dan fleksibel sehingga proses anotasi lebih cepat meski belum terlalu tepat dalam menganalisa sentimen karena hanya berlandaskan *dictionary* kata positif dan negatif dan dihitung polaritasnya. Kekurangannya penelitian ini yaitu pada *regular expression*(regex) masih belum bisa mengambil sentiment kalau ada bahasa selain indonesia seperti bahasa daerah, inggris dan lain-lain, belum sepenuhnya bisa mengambil konteks bahasa gaul, belum bisa tau konteks kalimat saekastik, kalau ada kesalahan pengetikan atau disingkat kata yang tidak bersentimen maka tidak jadi masalah.

REFERENCES

[1] G. Xu *et al.*, "Spatial disparities of self-reported COVID-19 cases and influencing factors in Wuhan, China," *Sustain. Cities Soc.*, vol. 76, no. September 2021, p. 103485, 2022.

[2] A. Voutama, I. Maulana, and N. Ade, "Interactive M-Learning Design Innovation using Android-Based Adobe Flash at WFH (Work From Home)," *Sci. J. Informatics*, vol. 8, no. 1, pp. 127–136, 2021, doi: 10.15294/sji.v8i1.27880.

[3] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, "Dataset Indonesia untuk Analisis Sentimen," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 334, 2019.

[4] C. Wang and X. Gu, "Influence of adolescents' peer relationships and social media on academic identity," *Asia Pacific J. Educ.*, vol. 39, no. 3, pp. 357–371, 2019.

[5] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in Social Media Research: Past, Present and Future," *Inf. Syst. Front.*, vol. 20, no. 3, pp. 531–558, 2018, doi: 10.1007/s10796-017-9810-y.

[6] A. Voutama and E. Novalia, "Perancangan Aplikasi M-Magazine Berbasis Android Sebagai Sarana Mading Sekolah Menengah Atas," *J. Tekno Kompak*, vol. 15, no. 1, p. 104, 2021, doi: 10.33365/jtk.v15i1.920.

[7] N. Nyoman, P. Pinata, I. M. Sukarsa, N. Kadek, and D. Rus jayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *J. Ilm. Merpati*, vol. 8, no. 3, pp. 188–196, 2020.

[8] D. N. Zuraidah, M. F. Apriyadi, A. R. Fatoni, M. Al Fatih, and Y. Amrozi, "Menelisik Platform Digital Dalam Teknologi Bahasa Pemrograman," *Teknois J. Ilm. Teknol. Inf. dan Sains*, vol. 11, no. 2, pp. 1–6, 2021.

[9] I. Ruslianto, S. Bahri, J. Rekayasa Sistem Komputer, and J. H. Hadari Nawawi, "Analisa Log Web Server Untuk Mengetahui Pola Perilaku Pengunjung Website Menggunakan Teknik Regular Expressions," *Coding J. Komput. dan Apl.*, vol. 07, no. 01, pp. 120–130, 2019.

[10] S. Salomon and S. Hansun, "Spam Filter Situs Jejaring Sosial Mahasiswa Menggunakan Regular Expression," *J. Ultim. InfoSys*, vol. 8, no. 2, pp. 69–73, 2018.

[11] F. Tobing and R. Nainggolan, "Analisis Perbandingan Penggunaan Metode Binary Search Dengan Regular Search Expression," *METHOMIKA J. Manaj. Inform. dan Komputerisasi*

Akunt., vol. 4, no. 1, pp. 168–172, 2020, doi: 10.46880/jmika.v4i2.202.

[12] D. H. Wahid and A. SN, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 10, no. 2, p. 207, 2016.

[13] A. A. Chaudhri, S. S. Saranya, and S. Dubey, "Implementation Paper on Analyzing COVID-19 Vaccines on Twitter Dataset Using Tweepy and Text Blob," *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 3, pp. 8393–8396, 2021.

[14] R. Wati *et al.*, "Analisis Sentimen Persepsi Publik Mengenai PPKM Pada Twitter Berbasis SVM Menggunakan Python," vol. 06, pp. 240–247, 2021.

[15] A. Voutama, "Perancangan Aplikasi M-Discussion Berbasis Android Sebagai Wadah Diskusi Sekolah," *Syntax J. Inform.*, vol. 7, no. 2, pp. 116–124, 2018.

[16] D. Arisandi, Z. Indra, and K. Kartini, "Mengidentifikasi Hoax Pada Hasil Pencarian Berita Online Dengan Teknik Web Scraping Dan Algoritma C4.5," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 6, no. 2, pp. 130–137, 2021.

[17] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *J. Big Data*, vol. 5, no. 1, pp. 1–10, 2018.

[18] Z. Wu and D. C. Ong, "Context-Guided BERT for Targeted Aspect-Based Sentiment Analysis," 2020.

[19] A. Priyanto and M. R. Ma'arif, "Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik)," *Indones. J. Inf. Syst.*, vol. 1, no. 1, pp. 25–33, 2018.

[20] A. Javed, M. Zaman, M. M. Uddin, and T. Nusrat, "An analysis on python programming language demand and its recent trend in bangladesh," *PervasiveHealth Pervasive Comput. Technol. Healthc.*, pp. 458–465, 2019.