Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 2, Desember 2025

KEYWORD EXTRACTION KOMENTAR TERHADAP KONFLIK INDIA-PAKISTAN PADA PLATFORM YOUTUBE MENGGUNAKAN TF-IDF DAN COSINE SIMILARITY

Nicholas Edison[1*], Kristian Fernando[2], Hafiz Irsyad[3], Abdul Rahman[4]

[1,2,3] Informatika, Universitas Multi Data Palembang, [4] Teknik Elektro, Universitas Multi Data Palembang
[1]nicholasedison_2226250003@mhs.mdp.ac.id, [2]kristianfernando_2226250070@mhs.mdp.ac.id, [3]hafizirsyad@mdp.ac.id, [4]arahman@mdp.ac.id

ABSTRAK

Konflik antara India dan Pakistan merupakan isu geopolitik yang sering menjadi perhatian global dan menimbulkan diskusi luas di media sosial, termasuk platform YouTube. Penelitian ini bertujuan untuk mengekstraksi kata kunci dari komentar-komentar pengguna YouTube mengenai topik konflik India-Pakistan, serta menganalisis kemiripan makna seluruh komentar antar video menggunakan metode TF-IDF dan Cosine Similarity. Data diperoleh dari kolom komentar tiga video YouTube yang relevan dan diproses melalui tahapan pra-pemrosesan teks, perhitungan bobot kata menggunakan TF-IDF, serta pengukuran similaritas menggunakan Cosine Similarity. Hasil ekstraksi kata kunci menggunakan TF-IDF menunjukkan terdapat 20 kata kunci dengan frekuensi tertinggi, dengan 3 kata kunci tertinggi adalah "india", "pakistan" dan "perang". Hasil perhitungan Cosine Similarity menunjukkan bahwa tingkat kemiripan antar komentar video berkisar antara 0,544 hingga 0,695, dimana nilai similarity tertinggi terdapat pada perbandingan Video 1 dan Video 3 (0,695), Video 1 dan Video 2 (0,653), sementara Video 2 dan Video 3 (0,544). Hasil ini menunjukkan bahwa kombinasi metode ini efektif dalam mengidentifikasi topik dominan serta hubungan semantik antar komentar. Visualisasi kata kunci dengan WordCloud juga memperjelas representasi opini publik yang berkembang. Penelitian ini memberikan kontribusi dalam pemetaan diskursus digital secara kuantitatif dan efisien.

Kata Kunci— Analisis Komentar, Cosine Similarity, Ekstraksi Kata Kunci, Konflik India-Pakistan, TF-IDF, WordCloud, YouTube.

ABSTRACT

The conflict between India and Pakistan is a geopolitical issue that frequently garners global attention and sparks widespread discussion on social media platforms, including YouTube. This study aims to extract keywords from YouTube user comments regarding the India-Pakistan conflict topic, as well as to analyze the semantic similarity of all comments across videos using the TF-IDF and Cosine Similarity methods. Data was collected from the comment sections of three relevant YouTube videos and processed through stages including text pre-processing, calculation of word weights using TF-IDF, and similarity measurement using Cosine Similarity. The keyword extraction results using TF-IDF revealed 20 keywords with the highest frequencies, with the top three being "india," "pakistan," and "perang." The Cosine Similarity calculations showed that the similarity levels between video comments ranged from 0.544 to 0.695, with the highest similarity observed between Video 1 and Video 3 (0.695), followed by Video 1 and Video 2 (0.653), and the lowest between Video 2 and Video 3 (0.544). These results indicate that this combination of methods is effective in identifying dominant topics as well as the semantic relationships between comments. The visualization of keywords using a WordCloud also enhances the representation of emerging public opinion. This research contributes to the quantitative and efficient mapping of digital discourse.

Keywords— Comment Analysis, Cosine Similarity, Keyword Extraction, India-Pakistan Conflict, TF-IDF, WordCloud, Youtube.

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 2, Desember 2025

I. PENDAHULUAN

Konflik geopolitik yang berkelanjutan antara India dan Pakistan merupakan isu yang secara konsisten menarik perhatian komunitas global, terutama pada saat terjadinya peningkatan ketegangan bilateral [1]. Dalam era digital kontemporer, respons masyarakat terhadap isu-isu krusial semacam ini tidak lagi terbatas pada pemberitaan oleh media arus utama, melainkan terdistribusi secara ekstensif melalui berbagai platform media sosial, dengan YouTube sebagai salah satu medium daring yang signifikan dalam menyebarkan dan menampung opini publik.

Sebagai platform berbagi video dengan jangkauan audiens yang luas secara global, YouTube menyediakan fungsionalitas kolom komentar yang berperan sebagai ruang interaksi virtual bagi pengguna dengan beragam latar belakang sosio-ekonomi dan budaya [2]. Dalam lingkungan digital ini, individu memiliki kemampuan untuk mengekspresikan pandangan personal, memberikan tanggapan terhadap konten yang disaksikan, serta menyajikan informasi tambahan yang relevan. Secara inheren, data yang terkandung dalam komentar-komentar ini menyimpan potensi wawasan yang berharga terkait dengan persepsi kolektif masyarakat. Namun demikian, volume data yang besar dan formatnya yang tidak terstruktur menimbulkan tantangan metodologis dalam pelaksanaan analisis secara manual. Oleh karena itu, implementasi teknik-teknik komputasi menjadi esensial untuk mengekstraksi informasi yang relevan dan bermakna, khususnya dalam mengidentifikasi kata kunci yang merepresentasikan inti dari diskusi yang berlangsung dalam kumpulan komentar tersebut.

Dalam ranah Pengolahan Bahasa Alami (Natural Language Processing/NLP), berbagai metodologi telah dikembangkan untuk mengotomatisasi proses ekstraksi kata kunci dari sejumlah besar dokumen tekstual. Salah satu pendekatan yang mapan dan banyak diimplementasikan adalah TF-IDF (Term Frequency-Inverse Document Frequency). Prinsip operasional metode ini didasarkan pada evaluasi kuantitatif terhadap tingkat kepentingan suatu term dalam konteks dokumen tertentu, yang ditentukan oleh frekuensi kemunculannya dalam dokumen target dan frekuensi inversi dokumen secara keseluruhan dalam korpus yang dianalisis [3]. Keunggulan utama TF-IDF terletak pada kompleksitas komputasi yang relatif rendah dan kemampuannya dalam mengidentifikasi term-term distingtif yang secara efektif merefleksikan fokus tematik suatu teks. Meskipun demikian, metodologi ini memiliki keterbatasan inheren terkait dengan kurangnya sensitivitas terhadap relasi sintaktis antar kata (pemahaman kontekstual) dan efektivitas yang terbatas dalam mengenali variasi leksikal atau konstruksi kalimat yang lebih kompleks [4].

Untuk mengatasi keterbatasan-keterbatasan tersebut, seringkali diterapkan integrasi antara TF-IDF dan metrik similaritas semantik seperti Cosine Similarity. Cosine Similarity berfungsi untuk mengukur tingkat kesamaan semantik antara dua entitas tekstual berdasarkan representasi vektor dalam ruang multidimensional. Metrik ini sangat berguna dalam mengelompokkan komentar-komentar yang memiliki kemiripan makna yang substansial, meskipun menggunakan pilihan kata yang berbeda. Sinergi antara TF-IDF dan Cosine Similarity meningkatkan akurasi sistem dalam mengidentifikasi kata kunci yang relevan dan mengenali hubungan tematik antar komentar [5]. Meskipun tidak mencapai tingkat kerumitan model pembelajaran mesin mendalam (*deep learning*), pendekatan hibrida ini menawarkan efisiensi sumber daya komputasi yang lebih baik dan tingkat aplikabilitas yang tinggi pada himpunan data berskala besar seperti komentar pada platform YouTube.

Berdasarkan latar belakang permasalahan yang telah diuraikan, penelitian ini bertujuan untuk mengaplikasikan metodologi TF-IDF dan Cosine Similarity dalam proses ekstraksi kata kunci dari komentar-komentar pengguna platform YouTube yang berkaitan dengan isu konflik antara India dan Pakistan. Studi ini tidak hanya berfokus pada identifikasi topik-topik utama yang mendominasi diskursus daring, melainkan juga pada evaluasi efektivitas kombinasi metodologis ini dalam konteks wacana publik di ranah digital. Melalui pemahaman terhadap pola kata kunci yang berhasil diidentifikasi, diharapkan hasil penelitian ini dapat memberikan kontribusi signifikan terhadap pemetaan opini publik secara kuantitatif, analisis sentimen berbasis leksikal, serta pengembangan sistem otomatis untuk pemantauan isu-isu geopolitik di lingkungan digital.

II. METODOLOGI PENELITIAN

Pada penelitian ini digunakan pendekatan kuantitatif yang dapat didefinisikan sebagai penelitian yang disusun secara sistematis untuk menemukan kausalitas keterkaitan antara data yang berupa angka [6], dengan metode eksperimen yang sejatinya adalah metode yang dilakukan dalam kondisi terkendali untuk mencari tahu dampak suatu perlakuan terhadap yang lain [7].

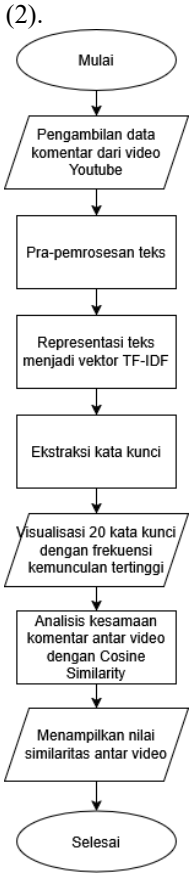
Penelitian ini melibatkan beberapa tahapan, dimulai dari pengambilan data, pre-pemrosesan teks pada data, ekstraksi kata kunci, visualisasi kata kunci, serta perbandingan similaritas terhadap seluruh komentar antar video [8] seperti yang tersaji pada Gambar 1.

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 2, Desember 2025

Penelitian dimulai dengan mengambil data dari kolom komentar 3 video dengan topik perang India-Pakistan. Seluruh komentar dari ketiga video tersebut diambil dengan metode *web crawling* [9], dan didapat sebanyak total 1.190 komentar di saat tahapan *web crawling* dilakukan.

Tahapan pra-pemrosesan teks pada data kemudian dilakukan dengan pertama-tama dilakukan pengubahan terhadap seluruh karakter yang ada pada data menjadi huruf kecil (*case folding*), kemudian penghapusan karakter-karakter angka serta karakter non-alfabet, pengubahan setiap kata menjadi bentuk kata dasar atau yang disebut *stemming*, dan terakhir adalah penghapusan kata-kata yang tidak bermakna dalam data atau yang biasa dikenal dengan *removing stopwords*. Tahapan pra-pemrosesan teks ini dilakukan dengan tujuan untuk 'membersihkan' data agar siap dan lebih mudah untuk dianalisis dan menghasilkan hasil yang lebih akurat [10].

Tahapan ekstraksi kunci kemudian dilakukan dengan pertama-tama merepresentasikan setiap kata ke dalam bentuk numerik dengan metode TF-IDF untuk menilai tingkat bobot kata berdasarkan banyaknya kemunculan kata tersebut dalam satu dan seluruh dokumen. Persamaan TF dan IDF dijabarkan pada persamaan (1) dan (2).



Gambar 1. Visualisasi tahapan penelitan

$$TF(t,d) = \frac{Jumlah\ kemunculan\ kata\ t\ dalam\ dokumen\ d}{Jumlah\ seluruh\ kata\ dalam\ dokumen\ d} \quad (1)$$

$$IDF(t) = log\left(\frac{Total\ jumlah\ dokumen}{Jumlah\ dokumen\ yang\ menaanduna\ kata\ t}\right) \quad (2)$$

Setelah perhitungan TF-IDF dilakukan, kemudian diambil maksimal sebanyak 5 kata dengan nilai TF-IDF tertinggi menjadi kata kunci utama pada setiap komentar.

Lalu diambil 20 kata kunci yang paling banyak muncul dari keseluruhan data dan kemudian divisualisasikan dengan menggunakan WordCloud untuk memberikan gambaran umum dari apa yang dibicarakan oleh pengguna dalam kolom komentar terkait dengan perang India-Pakistan.

Analisis similaritas dengan menggunakan Cosine Similarity kemudian dilakukan, dengan memanfaatkan nilai TF-IDF pada setiap kata dari setiap komentar yang ada pada satu video, untuk melihat hubungan semantik atau kemiripannya dengan komentar-komentar pada video lain. Adapun persamaan dari Cosine Similarity ditunjukkan pada persamaan (3).

$$Cosine\ Similarity(A,B) = \frac{A.B}{||A||.||B||} \quad (3)$$

Keterangan:
$A.B$ = hasil perkalian titik dari vektor A dan B
$||A||$ = norma dari vektor A
$||B||$ = norma dari vektor B

Hasil dari Cosine Similarity juga divisualisasikan dalam bentuk tabel yang menunjukkan seberapa mirip komentar-komentar dari video satu dengan yang lainnya dalam skala -1 hingga 1.

III. HASIL DAN PEMBAHASAN.

Tahap pertama dalam penelitian ini adalah mengambil seluruh komentar dari 3 video dari platform YouTube dengan tema perang India-Pakistan yang berjudul "Rangkuman Perang India-Pakistan: Gencatan Senjata Dilanggar, Situs Nuklir Dibidik, Rudal China Siaga" yang diunggah oleh kanal Tribunnews, "Adu Kuat Nuklir, Jadi Awal Mula Perang India-Pakistan | OneNews Update" yang diunggah oleh kanal tvOneNews, serta video dengan judul "TRUMP FAILED TO MEDIATE INDIA-PAKISTAN CEASE? Nuclear War Will Happen, Who Will Win?" yang diunggah oleh kanal Auto Populer ID, dengan menggunakan pustaka youtube-comment-downloader sehingga diperoleh sebanyak 1.190 komentar. Komentar-komentar tersebut kemudian di kompilasi ke dalam bentuk *dataframe* dengan

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 2, Desember 2025

menggunakan pustaka pandas, dapat dilihat Pada Tabel 1.

Tabel 1. Dataframe komentar video YouTube

	comment
0	Udah biarin aj.... Perang....
1	Mantap
2	Berhenti berperang kaluan terangga tidak baik ...
3	Ini memicu perang Rusia --- China.
4	Indiakn sodaranya seriwil pasti gk bisa di pag...
...	...
1185	@komanglempongs9035 Kehabisan Referensi yaa,...
1186	@Haryanto-u5t ye. Siapa yg lompat pagar
1187	@komanglempongs9035 dari Kresna ke Baalawi, d...
1188	Sibuk bahas ijasah palsu 😅😅
1189	Berita lama, sekarang mah udah gencatan senjata

Tahap selanjutnya adalah pra-pemrosesan data, dimana teks komentar-komentar yang diambil dibersihkan dengan dilakukan *case folding* dimana setiap karakter pada data teks komentar diubah menjadi huruf non-kapital, penghapusan karakter numerik dan non-alfabet lainnya, *stemming* dimana setiap kata dalam data teks komentar diubah menjadi bentuk kata dasarnya, serta penghapusan *stopwords* atau penghapusan kata-kata yang kurang bermakna dalam setiap data teks komentar. Hasil dari tahap pra-pemrosesan ini kemudian disimpan dalam kolom baru dengan nama pre-processed, dapat dilihat Pada Tabel 2.

Tabel 2. Dataframe sesudah Pra-pemrosesan

	comment	preprocessed
0	Udah biarin aj.... Perang....	udah biarin aj perang
1	Mantap	mantap
2	Berhenti berperang kaluan terangga tidak baik ...	henti perang kaluan angga tidak baik perang ha...
3	Ini memicu perang Rusia --- China.	ini picu perang rusia china
4	Indiakn sodaranya seriwil pasti gk bisa di pag...	indiakn sodaranya seriwil pasti gk bisa di pag...
...	...	...
1185	@komanglempongs9035 Kehabisan Referensi yaa,...	komanglempongs habis referensi yaa lompat paga...
1186	@Haryanto-u5t ye. Siapa yg lompat pagar	haryanto ut ye siapa yg lompat pagar
1187	@komanglempongs9035 dari Kresna ke Baalawi, d...	komanglempongs dari kresna ke baalawi dari baa...
1188	Sibuk bahas ijasah palsu 😅😅	sibuk bahas ijasah palsu
1189	Berita lama, sekarang mah udah gencatan senjata	berita lama sekarang mah udah gencat senjata

Setelah data dibersihkan pada tahap pra-pemrosesan data, tahap ekstraksi kata kunci dilakukan dengan menghitung nilai TF-IDF dari seluruh kata yang terdapat dalam data, kemudian dipilih dari setiap komentar maksimal 5 kata dengan nilai TF-IDF tertinggi sebagai kata kunci yang mewakili komentar tersebut. Kata-kata kunci yang dihasilkan dari tahap ekstraksi kata kunci ini disimpan pada kolom baru dengan nama pre-processed, dapat dilihat Pada Tabel 3.

Tabel 3. Dataframe sesudah pemrosesan

	comment	preprocessed	top_keywords
0	Udah biarin aj.... Perang....	udah biarin aj perang	[aj, biarin, udah, perang]

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 2, Desember 2025

1	Mantap	mantap	[mantap]
2	Berhenti berperang kaluan terangga tidak baik ...	henti perang kaluan angga tidak baik perang ha...	[kaluan, angga, perang, henti, rugi]
3	Ini memicu perang Rusia --- China.	ini picu perang rusia china	[picu, china, rusia, perang]
4	Indiakn sodaranya seriwil pasti gk bisa di pag...	indiakn sodaranya seriwil pasti gk bisa di pag...	[indiakn, sodaranya, pagang, seriwil, janji]
...	...	...	...
1185	@komangl empongs90 35 Kehabisan Referensi yaa,...	komanglemp ongs habis referensi yaa lompat paga...	[meluluuuuu, yaa, lompat, pagar, referensi]
1186	@Haryanto -u5t ye. Siapa yg lompat pagar	haryanto ut ye siapa yg lompat pagar	[ye, lompat, pagar, haryanto, ut]
1187	@komangl empongs90 35 dari Kresna ke Baalawi, d...	komanglemp ongs dari kresna ke baalawi dari baa...	[baalawi, referensi, dna, waras, odgj]
1188	Sibuk bahas ijasah palsu 😅😅	sibuk bahas ijasah palsu	[ijasah, bahas, sibuk, palsu]
1189	Berita lama, sekarang mah udah gencatan senjata	berita lama sekarang mah udah gencat senjata	[mah, sekarang, lama, udah, berita]

Tahap visualisasi data kemudian dilakukan, dengan pertama-tama dillakukan visualisasi terhadap 20 kata kunci dengan frekuensi kemunculan terbanyak dengan menggunakan pustaka WordCloud, dimana 3 kata kunci dengan kemunculan terbanyak diantara 20 kata kunci tersebut adalah "india", "pakistan" dan "perang" seperti yang terpapar pada Gambar 2, dengan jumlah kemunculan kata kunci "india" sebanyak 79 kali, "pakistan" sebanyak 61 kali dan "perang" sebanyak 53 kali seperti yang dipaparkan pada Gambar 3.

Gambar 2. Visualisasi 20 kata kunci terbanyak

```
Jumlah Kemunculan 20 Keyword Teratas di Semua Komentar:
india        79
pakistan     61
perang       53
negara       32
sama         32
israel       32
dukung       27
yg           27
nuklir       23
indonesia    22
aja          20
china        19
hancur       18
amerika      18
rusia        18
lanjut       18
banyak       16
senjata      16
islam        15
damai        15
```

Gambar 3. Jumlah kemunculan 20 kata kunci terbanyak

Analisis lebih lanjut kemudian dilakukan dengan menghitung nilai similaritas komentar-komentar antar video dengan menggunakan metode Cosine Similarity berdasarkan representasi vektor TF-IDFnya. Tahap visualisasi kemudian dilanjutkan dengan dibuatnya visualisasi terhadap seluruh nilai similaritas antar komentar yang dihitung dengan persamaan Cosine Similarity, sehingga didapat hasil dimana seluruh komentar pada Video 1 yang adalah video berjudul "Rangkuman Perang India-Pakistan: Gencatan Senjata Dilanggar, Situs Nuklir Dibidik, Rudal China Siaga" memiliki kemiripan sebesar 0,652779 terhadap seluruh komentar pada Video 2

yang berjudul "Adu Kuat Nuklir, Jadi Awal Mula Perang India-Pakistan | OneNews Update", dan kemiripan sebesar 0,695318 terhadap seluruh komentar pada Video 3 yang berjudul "TRUMP FAILED TO MEDIATE INDIA-PAKISTAN CEASE? Nuclear War Will Happen, Who Will Win?". Komentar-komentar pada Video 2 dan Video 3 memiliki kemiripan sebesar 0,544673 seperti yang tersaji pada Tabel 4. Nilai-nilai similaritas ini menunjukkan bahwa komentar-komentar pada Video 1 dan Video 3 memiliki kemiripan yang paling tinggi, dengan kemiripan isi komentar sebesar sekitar 69,5%, diikuti dengan Video 1 dan Video 2 dengan kemiripan isi komentar yang cukup mirip, dengan kemiripan sebesar sekitar 65,2%, dan isi komentar-komentar pada Video 2 dan Video 3 memiliki kemiripan konten sebesar 54,4%.

Tabel 4. Hasil cosine similarity antar video

	Video 1	Video 2	Video 3
Video 1	1.000000	0.652779	0.695318
Video 2	0.652779	1.000000	0.544673
Video 3	0.695318	0.544673	1.000000

PENUTUP

Penelitian ini berhasil mengimplementasikan metode TF-IDF untuk mengekstraksi kata kunci dari komentar-komentar pengguna YouTube yang membahas konflik India–Pakistan, serta menganalisis kemiripan antar komentar menggunakan algoritma Cosine Similarity. Hasil dari penelitian ini menunjukkan bahwa proses ekstraksi kata kunci dapat mengidentifikasi topik utama yang sering dibicarakan oleh pengguna, dan pengukuran similaritas antar komentar mampu memperlihatkan hubungan semantik antar diskusi yang terjadi di beberapa video berbeda.

Visualisasi menggunakan WordCloud mempermudah dalam memahami dominasi kata dan topik, sedangkan nilai Cosine Similarity menunjukkan sejauh mana keterkaitan opini atau respon antar video.

Diharapkan hasil penelitian ini dapat menjadi pijakan awal dalam studi analisis opini publik digital, serta dapat diterapkan pada berbagai isu lain yang berkembang di platform sosial media. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi integrasi model pembelajaran mesin berbasis neural network guna meningkatkan pemahaman kontekstual.

REFERENSI

[1] Thelwall, M. (2017, September). *"Social media analytics for YouTube comments: potential and limitations." 21*. 10.1080/13645579.2017.1381821

[2] Ganguly, S., Smetana, M., Abdullah, S., & Karmzin, A. (2018, Agustus). *"India, Pakistan, and the Kashmir dispute: unpacking the dynamics of a South Asian frozen conflict."* 10.1007/s10308-018-0526-5

[3] J. Li, "A Comparative Study of Keyword Extraction Algorithms for English Texts," *J. Intell. cyst*., Vol.30, No.1, pp. 808-815, 2021, doi: 10.1515/jisys-2021-0040.

[4] Bouazizi, M., & Ohtsuki, T. (2017). "*A pattern-based approach for sarcasm detection on Twitter"*. Vol. 4, IEEE Access, 4, 5477–5488

[5] Nugraha, K. A., & Sebastian, D. (2018, Desember). Pembentukan DatasetTopikKata Bahasa Indonesia pada Twitter Menggunakan TF-IDF & Cosine Similarity Vol 4, No.3, https://journal.maranatha.edu/index.php/jutisi/article/view/1473/1146

[6] Z. Afif, D. S. Azhari, M. Kustati, and N. Sepriyanti, "Penelitian ilmiah (kuantitatif) beserta paradigma, pendekatan, asumsi dasar, karakteristik, metode analisis data dan outputnya," *Innovative: Journal of Social Science Research*, vol. 3, no. 3, pp. 682–693, 2023.

[7] M. Ramdhan, *Metode Penelitian*, A. A. Effendy, Ed., Surabaya, Indonesia: 2021.

[8] John, M., Marbach, E., Lohmann, S., Heimerl, F., & Ertl, T. (2018, Mei). *MultiCloud: Interactive Word Cloud Visualization for the Analysis of Multiple Texts*. Proceedings of Graphics Interface 2018, 34–41.

[9] Mitchell, R. (2018). *Web Scraping with Python: Collecting More Data from the Modern Web* (2nd ed.). O'Reilly Media.

[10] M. Z. Haq, C. S. Octiva, Ayuliana, U. W. Nuryanto, and D. Suryadi, "Algoritma Naïve Bayes untuk mengidentifikasi hoaks di media sosial," *J. Minfo Polgan*, vol. 13, no. 1, pp. 1079–1084, Jul. 2024