COMSERVA
JURNAL PENELITIAN DAN PENGABDIAN MASYARAKAT
Volume 03 No. 04 Agustus 2023 (1552-1560)
e-ISSN: 2798-5210
p-ISSN: 2798-5652

Pendeteksi Lokasi Covid-19 dari Sosial Media Menggunakan Kombinasi Algoritma *Soundex* dan *Permuterm Index*

COVID-19 Location Detection from Social Media Using a Combination of the Soundex Algorithm and the Permuterm Index

Dwiki Jatikusumo[1], Rahmat Rian Hidayat[2]
Universitas Mercu Buana, DKI Jakarta, Indonesia

*Email: dwiki.jatikusumo@mercubuana.ac.id, Rahmat.rian@mercubuana.ac.id,
**Correspondence: Dwiki Jatikusumo*

DOI:
10.59141/comserva.v3i4.907

ABSTRAK

Pada Desember tahun 2021 sampai Februari 2022 ada varian baru COVID-19, yaitu *omicron* di Indonesia, terdapat beberapa kasus. Dari kasus terakhir ini juga kita tidak mendapatkan informasi secara lengkap di mana lokasi yang terjangkit. Dari *website* pemerintah, juga belum dapat kepastian varian terbaru tersebut. Penelitian ini bertujuan untuk dapat membantu dalam mencari lokasi dengan cara mendapat informasi terkini dari posting sosial media salah satunya yaitu Twitter, dari sini kita bisa tahu titik lokasi post maupun dari kalimat yang ada dalam posting-an tersebut. Metode yang digunakan dalam penelitian ini menggunakan SDLC atau System Development Life Cycle dengan model prototype. Data berdasarkan dari Twitter, akan dikombinasikan dengan algoritma stemming Bahasa Indonesia Berdasarkan kasus tersebut, diharapkan penelitian ini dapat memberikan informasi terkait kasus COVID-19, dari varian terbaru maupun tidak. Dengan adanya *posting*-an dari Twitter, merupakan sumber data yang akan diproses. Selanjutnya tingkat persentase akurasi yang didapat dalam menggabungkan dari algoritma *soundex* dan *permuterm index*.

Kata Kunci: covid-19, sosial media, *soundex*, *permuterm index*

ABSTRACT

From December 2021 to February 2022 there was a new variant of COVID-19, namely omicron in Indonesia, there were several cases. From this last case, we also do not get complete information on where the infected location is. From the government website, there is also no certainty of the latest variant. This study aims to be able to help in finding the location by getting the latest information from social media posts, one of which is Twitter, from here we can know the location of the post and from the sentences in the post. The method used in this study uses SDLC or System Development Life Cycle with a prototype model. Data based on Twitter, will be combined with stemming algorithms Indonesian Based on these cases, it is hoped that this study can provide information related to COVID-19 cases, from the latest variants or not. With posts from Twitter, it is the source of data that will be processed. Furthermore, the percentage level of accuracy obtained in combining the soundex algorithm and permuterm index.

Keywords: *covid-19, sosial media, soundex, permuterm index*

PENDAHULUAN

Menyebarnya wabah COVID-19 ini hingga ke wilayah Indonesia. Seperti dapat dicermati dari pengalaman beberapa negara serta wilayah lain, penanganan covid-19 tidak mungkin dapat dilakukan oleh Pemerintah semata.

Terdapaat temuan baru virus lainnya, ditunjukkan dengan masuknya kasus varian baru yakni SARS-CoV-2 B.1.1.529 atau dikenal dengan Omicron. Sebelumnya, varian Omicron pertama kali dilaporkan kepada WHO dari Afrika Selatan pada tanggal 24 November 2021. Beberapa minggu terakhir ini, Kementerian Kesehatan mengumumkan sebanyak tiga kasus varian Omicron terkonfirmasi masuk ke Tanah Air (UMY, 2021).

Kini teknologi telah memudahkan pekerjaan banyak orang, salah satunya dengan kehadiran banyak media sosial (Fadhiil et al., 2023). Tak hanya digunakan untuk komunikasi pribadi, media sosial memiliki banyak fungsi lainnya yang seiring berjalannya waktu semakin banyak.

Beberapa jenis media sosial yang pada umumnya dimiliki masyarakat Indonesia adalah Facebook, Twitter, Instagram, hingga *TikTok*. Hampir semua masyarakat Indonesia dari berbagai kalangan memiliki media sosial bahkan lebih dari satu.

Adanya media sosial tak hanya membawa dampak baik tetapi juga dampak buruk bila tidak digunakan dengan tepat. Untuk itu penting mengenali fungsi media sosial itu sendiri, agar kamu tidak keluar batas dalam penggunaannya.

Isu COVID-19 untuk menjadikan objek suatu penelitian dengan cara mendapatkan lokasi kejadian menggunakan sosial media sebagai sumber data dan algoritma *soundex* dan teknik *permuterm index* sebagai metode yang digunakan dalam penelitian ini.

Dari hal tersebut, tujuannya dalam penelitian ini mendapatkan titik lokasi kejadian COVID-19 melalui sosial media, dan mendapatkan akurasi yang baik dari hasil proses kombinasi algoritma dan teknik tersebut titik lokasi kejadian COVID-19 sekitar 85% lebih dari data sebanyak 6.000 sampai 10.000 data.

Teknik memperoleh kata-kata menggunakan suara digunakan di Amerika Serikat sensus sejak akhir 1890-an, tetapi solusi konkret untuk ini pertama kali diusulkan dan dipatenkan oleh Robert C. Russell pada tahun 1912 sebagai algoritma *Soundex* (Koneru et al., 2016; Shah & Kumar Singh, 2014). Algoritma Soundex memberikan nilai ke istilah sedemikian rupa sehingga istilah yang terdengar serupa mendapatkan nilai yang sama. Nilai-nilai ini dikenal sebagai pengkode-an *Soundex* (Shah & Singh, 2014). Jika penyandian *Soundex* dari kata apa pun di pos cocok dengan penyandian *Soundex* apa pun di kumpulan data, maka disimpulkan bahwa pos ini berisi konten yang menyinggung. Tabel 1 menyajikan kode fonetik *Soundex* untuk setiap huruf bahasa.

Tabel 1. Kode Numerik Algoritma *Soundex*

Kode	Huruf
Tidak dikodekan	A,I,U,E,O,H,W,Y
1	B,F,P,V
2	C,G,J,K,Q,S,X,Z
3	D,T
4	L

5	M,N

Sumber: (Koneru et al., 2016)

Aturan pengkode-an dengan algoritma *Soundex* dapat dijelaskan seperti langkah berikut:

1. Ubah semua huruf menjadi huruf besar atau uppercase, buang semua huruf vokal, tanda baca yang tidak ada hubungan dengan kata, konsonan H,W, dan Y, serta urutan huruf yang sama (misalnya. sss). Huruf pertama selalu dibiarkan seperti semula.
2. Gabung huruf pertama dengan angka pengganti yang sesuai dengan kode numerik yang ditunjukkan pada Tabel 1.
3. Ambil empat kode terdepan dan selanjutnya kode tersebut menjadi kode *Soundex*.

Kedua kata yang memiliki kode yang sama dapat diklasifikasikan (1) sama, (2) berbeda tetapi setidaknya memiliki satu kode Soundex yang sama, atau (3) tidak berhubungan sama sekali.

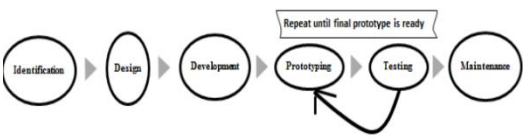
Permuterm Index adalah indeks kata kunci yang mendukung kueri dengan satu simbol *wildcard.* Idenya adalah menyimpan semua rotasi dari kata tertentu yang ditambahkan dengan karakter terminasi, misalnya untuk kata teks, indeks akan terdiri dari kosakata *permuterm* berikut: text$, ext$t, xtte, ttex, $text. Ketika datang untuk mencari, kueri pertama diputar sehingga *wildcard* muncul di akhir, dan selanjutnya awalannya dicari menggunakan indeks. Ini bisa berupa trie atau struktur data lainnya yang mendukung pencarian awalan. Masalah utama dengan *permuterm index* standar adalah penggunaan ruang, karena jumlah string yang dimasukkan ke dalam struktur data adalah jumlah kata dikalikan dengan panjang string rata-rata (Cisłak & Grabowski, 2017; Othman et al., 2022).

Kategori data filter *input* adalah data mentah yang prosesnya lemah di sini dan tantangannya adalah proses pra-filter sebelum melanjutkan teknik filter karena data mentah melibatkan data yang hilang dan bising perlu menghapus nilai ini hasil negatif yang efektif apalagi peta kesalahan untuk mendeteksi kesalahan selama data dimensi tahap terakhir sebelum dalam teknik filter adalah normalisasi (Aliwy & Ameer, 2017; Grechishcheva et al., 2021; Y.Alzyoud & Jum'ah Al_Zyadat, 2016).

Tokenisasi adalah teknik dalam NLP (*Natural Language Processing*) yang membagi dokumen menjadi *token* yang adalah kata atau frase. Token ini digunakan untuk lebih lanjut pengolahan. Tokenisasi berguna untuk mengidentifikasi kata kunci yang berarti. *Token* dipisahkan dengan warna putih spasi, jeda baris, atau tanda baca. Tanda baca tidak termasuk dalam *token* yang dihasilkan (Cong-Cuong, 2019; Joseph & Jeba, 2019).

METODE

Tahapan pengembangan yang digunakan penulis dalam riset ini yaitu menggunakan model SDLC (System Development Life Cycle) dengan menggunakan model/metode MADLC (Mobile Application Development Lifecycle).



Gambar 1. Kerangka Kerja Pengembangan Sistem Informasi (MADLC)
Sumber: (Kaur & Kaur, 2015)

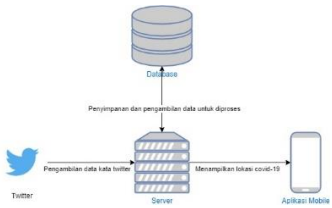
1. Fase Identifikasi: Pada fase pertama, ide-ide dikumpulkan dan dikategorikan. Ide bisa datang dari pelanggan atau dari pengembang.

2. Fase Perancangan: Pada fase ini, ide dari tim aplikasi *mobile* dikembangkan menjadi desain awal aplikasi.
3. Fase Pengembangan: Pada fase ini, aplikasi dikodekan. Pengkodean untuk modul yang berbeda dari prototipe yang sama dapat dilanjutkan secara paralel.
4. Fase *Prototyping*: Pada fase ini, kebutuhan fungsional dari setiap prototipe dianalisis; prototipe diuji dan dikirim ke klien untuk umpan balik. Setelah umpan balik, perubahan yang diperlukan diimplementasikan melalui fase pengembangan. Ketika prototipe kedua siap, itu diintegrasikan dengan prototipe pertama, diuji dan kemudian dikirim ke klien. Tahap pengembangan, pembuatan prototipe dan pengujian diulang sampai prototipe akhir siap.
5. Tahap Pengujian: Pengujian jenis prototipe dilakukan pada *emulator/simulator* diikuti dengan pengujian pada perangkat.

HASIL DAN PEMBAHASAN

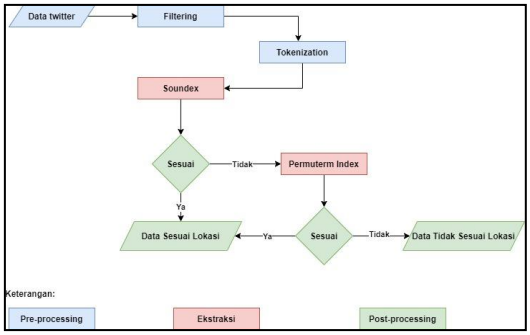
Perancangan

Pembahasan pada gambar di bawah yang dibutuhkan pada pendeteksi lokasi kejadian covid-19 adalah data dari twitter, kita ambil dari internet kemudian disimpan di dalam *database* dan akan diperlihatkan peta dari lokasi kejadian covid-19.



Gambar 2. Arsitektur lokasi kejadian covid-19

Dalam pengesktrasian data dari data mentah menjadi data yang diolah dengan menentukan lokasi adalah sebagai berikut:



Gambar 3. Alur processing pendeteksi lokasi kejadian covid-19

1. *Filtering*, proses *pemfilteran* meningkatkan efisiensi algoritme *Soundex* dan *permuterm index* dengan secara rasional mengurangi ukuran file teks yang diimpor. Ini menghilangkan kata-kata berulang yang tidak mengubah arti kalimat dan tidak memiliki nilai apa pun (misalnya, kata depan dan kata berhenti). Selain itu, menghapus semua *hyperlink*, gambar, rekaman audio, dan video.

2. *Tokenization*, membagi teks input menjadi token dan menghasilkan serangkaian token. Ini juga menghilangkan spasi sehingga setiap kata hanya dapat dipisahkan oleh satu spasi putih. Langkah ini diperlukan untuk mengubah teks input yang memiliki bentuk tidak terstruktur menjadi bentuk yang sesuai untuk diproses. Untuk melakukan tugas ini dan memastikan pemisahan teks *input* menjadi token, proses tokenisasi menggunakan metode *String.Split*().
3. Kemudian diproses menggunakan algoritma *soundex*, dan jika tidak sesuai akan dilakukan *permuterm index* sampai data sesuai dengan lokasi.

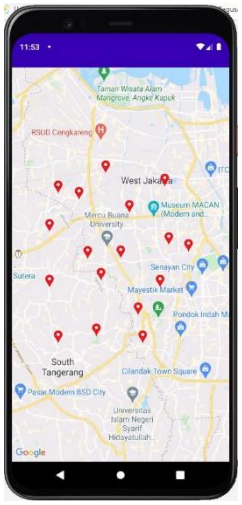
Dari desain sistem yang sudah disiapkan sebelumnya, berikut proses *preprocessing* yang dilakukan. Dalam pengesktrasian data dari data mentah menjadi data yang diolah dengan menentukan lokasi adalah sebagai berikut:

1. *Filtering*
2. *Tokenization*
3. Tidak menggunakan algoritma soundex dan teknik permuterm index
4. Menggunakan algoritma soundex dan teknik permuterm index

Setelah *filtering*, dan dilakukan tokenization, selanjutnya menggunakan *soundex* dan teknik *permuterm index* yang dilakukan untuk mencari kata-kata lokasi yang sesuai dengan hasil pencarian data dari *twitter*.

Hasil

Berikut contoh hasil dari pemetaan dengan menggunakan api *google maps* di *Android*. Jadi dari sini mendapatkan letak posisi kejadian covid-19 dari cuitan *twitter* yang ada dari bulan Januari sampai Maret 2022.



Gambar 4. Hasil dari data proses menggunakan kombinasi *Soundex* dan *Permuterm Indeks*

Data diproses menggunakan *preprocessing* (Ardhana et al., 2019; Shakeel et al., 2019)(Ardhana et al., 2019; Shakeel et al., 2019) seperti tahap pada metodologi riset ini, data yang diolah sebanyak 8843 data dari cuitan *twitter* sebelumnya sudah disimpan di dalam *database* untuk keperluan riset ini, dari bulan Januari sampai Maret 2022 hasilnya adalah 5729 yang terdapat lokasi kejadian berupa kata lokasi yang dicari. Selanjutnya adalah perbandingan yang merupakan kejadian atau tidak dengan ujicoba tanpa dan menggunakan algoritma *soundex* serta teknik *permuterm index*. Di sini membedakan dari dua kata

dengan “covid” dan “covid 19”. Dari beberapa referensi untuk mengetahui akurasinya (Alksasbeh et al., 2021; Chen et al., 2020; Sivapriya et al., 2020).

1. **Tidak menggunakan *Soundex* dan Permuterm Index**

Tabel 2. Tanpa algoritma *soundex* & *permuterm index*

No	Kata	Kejadian	Tidak Kejadian	Total	Persentase Akurasi
1	“covid”	1973	639	2612	75,54 %
2	“covid 19”	2629	488	3117	84,34 %

2. **Menggunakan *Soundex***

Tabel 3. Menggunakan algoritma *soundex*

No	Kata	Kejadian	Tidak Kejadian	Total	Persentase Akurasi
1	“covid”	2032	496	2528	80,38 %
2	“covid 19”	2846	355	3201	88,91 %

3. **Menggunakan Teknik *Permuterm Index***

Tabel 4. Menggunakan algoritma *soundex*

No	Kata	Kejadian	Tidak Kejadian	Total	Persentase Akurasi
1	“covid”	2194	499	2693	81,47 %
2	“covid 19”	2713	323	3036	89,36 %

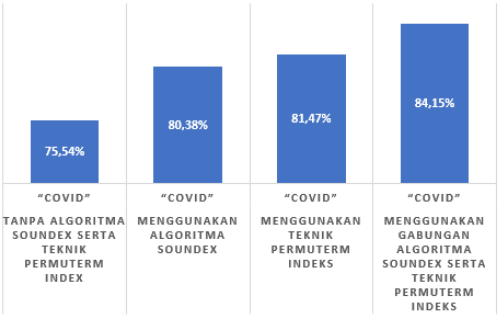
4. Menggunakan Gabungan *Soundex* dan *Permuterm Index*

Tabel 5. Menggunakan gabungan algoritma *soundex* serta teknik *permuterm index*

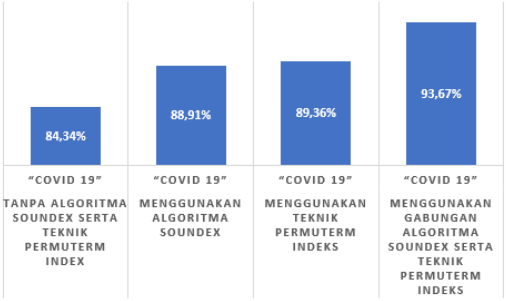
No	Kata	Kejadian	Tidak Kejadian	Total	Persentase Akurasi
1	"covid"	2315	436	2751	84,15 %
2	"covid 19"	2798	189	2987	93,67 %

5. Perbandingan

Berikut dari hasil perbandingan dari empat percobaan yang dilakukan:



Gambar 5. Hasil dari tabel ke grafik batang untuk kata "*covid*"



Gambar 6. Hasil dari tabel ke grafik batang untuk kata "*covid 19*"

Dari hasil grafik di atas, dengan perbandingannya, memang paling besar persentase untuk mendapatkan kata "covid" dan "covid 19" adalah yang menggunakan algoriitma soundex dan teknik permuterm index sebesar rata-rata 88,91%.

SIMPULAN

Kesimpulan dari penelitian yang telah dilakukan, dari pengolahan data covid-19 didapat lokasi kejadiannya yang tidak menggunakan algoritma *soundex* dan teknik *permuterm index* terdapat besaran

akurasi 79,94%. Selanjutnya untuk data yang diproses menggunakan algoritma *soundex* rata-rata 84,65%. Kemudian untuk data yang diproses menggunakan teknik *permuterm index* rata-rata 85,42%. Dan untuk data yang diproses menggunakan algoritma *soundex* dan teknik *permuterm index* rata-rata 88,91%. Dalam perhitungan yang sudah dilakukan bisa dikatakan setidaknya lebih besar menggunakan algoritma *soundex* dan teknik *permuterm index* dibandingkan tidak menggunakan algoritma dan teknik tersebut, serta yang masing-masing dipisah dengan algoritma dan teknik tersebut. Kedepannya, pengembangan penelitian selanjutnya dengan kombinasi dari algoritma *soundex* dan *permuterm index* tersebut, memungkinkan untuk membuat algoritma terbaru khususnya untuk algoritma pendeteksi lokasi dengan sosial media.

DAFTAR PUSTAKA

Aliwy, A. H., & Ameer, E. H. A. (2017). Comparative Study of Five Text Classification Algorithms with their Improvements. In *International Journal of Applied Engineering Research* (Vol. 12).

Alksasbeh, M. Z., Alqaralleh, B. A. Y., Abukhalil, T., Abukaraki, A., Al Rawashdeh, T., & Al-Jaafreh, M. (2021). Smart detection of offensive words in social media using the soundex algorithm and permuterm index. *International Journal of Electrical and Computer Engineering*, *11*(5), 4431–4438. https://doi.org/10.11591/ijece.v11i5.pp4431-4438

Ardhana, A. P., Cahyani, D. E., & Winarno. (2019). Classification of Javanese Language Level on Articles Using Multinomial Naive Bayes and N-Gram Methods. *Journal of Physics: Conference Series*, *1306*(1). https://doi.org/10.1088/1742-6596/1306/1/012049

Chen, Y., Pei, S., Liu, Y., Liu, Y., & Lin, Y. (2020). Summary of Application Research of Deep Learning in Operational Inspection of Transmission and Distribution Equipment. *Journal of Physics: Conference Series*, *1570*(1). https://doi.org/10.1088/1742-6596/1570/1/012057

Cisłak, A., & Grabowski, S. (2017). A practical index for approximate dictionary matching with few mismatches. *Computing and Informatics*, *36*(5), 1088–1106. https://doi.org/10.4149/cai_2017_5_1088

Cong-Cuong, L. (2019). Text Classification: Naïve Bayes Classifier with Sentiment Lexicon. *IAENG International Journal of Computer Science*, *46*(2), 1–8.

Fadhiil, M., Syarifah, S., & Simanjuntak, E. R. (2023). Application of Technology Acceptance Model (TAM) in Telemedicine Application During Covid-19 Pandemic. *Journal of World Science*, *2*(7), 909–921.

Grechishcheva, S., Lenivtceva, I., Kopanitsa, G., & Panfilov, D. (2021). Filtering free-text medical data based on machine learning. *Procedia Computer Science*, *193*, 82–91. https://doi.org/10.1016/j.procs.2021.10.009

Joseph, J., & Jeba, J. R. (2019). Information Extraction Using Tokenization And Clustering Methods. *International Journal of Recent Technology and Engineering*, *8*(4), 3690–3692. https://doi.org/10.35940/ijrte.D7943.118419

Kaur, A., & Kaur, K. (2015). Suitability of Existing Software Development Life Cycle (SDLC) in Context of Mobile Application Development Life Cycle (MADLC). *International Journal of Computer Applications*, *116*(19), 1–6. https://doi.org/10.5120/20441-2785

Koneru, K., Pulla, V. S. V., & Varol, C. (2016). Performance evaluation of phonetic matching algorithms on english words and street names comparison and correlation. *DATA 2016 - Proceedings of the 5th International Conference on Data Management Technologies and Applications*, *Data*, 57–64. https://doi.org/10.5220/0005926300570064

Othman, A. U., Moses, T., Aisha, U. Y., & Ya, A. (2022). *Big Data Indexing : Taxonomy , Performance Evaluation , Challenges and Research Opportunities*. *3*(2), 71–94.

Shah, R., & Kumar Singh, D. (2014). Analysis and Comparative Study on Phonetic Matching

Techniques. *International Journal of Computer Applications*, *87*(9), 14–17. https://doi.org/10.5120/15236-3771

Shah, R., & Singh, D. K. (2014). Improvement of Soundex algorithm for Indian language based on phonetic matching. *International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol*, *4*.

Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I. (2019). Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. *Measurement: Journal of the International Measurement Confederation*, *145*, 702–712. https://doi.org/10.1016/j.measurement.2019.05.027

Sivapriya, G., Kavipirathipa, M., Roshini, R., Vidhyalakshmi, N., & Student, U. G. (2020). Automatic Detection and Segmentation of Brain Tumors Using Convolutional Neural Network. *International Journal of Disaster Recovery and Business Continuity*, *11*(1), 94–101.

UMY. (2021). *Temuan Virus Omicron di Indonesia, Alarm Kesiapsiagaan untuk Civitas Akademika*. Https://Www.Umy.Ac.Id/.

Y.Alzyoud, F., & Jum'ah Al_Zyadat, W. (2016). The classification filter techniques by field of application and the results of output. *Australian Journal of Basic and Applied Sciences*, *August*, 68–77.