



Identifikasi Tema Perbincangan Masyarakat Tentang Vaksinasi Covid-19 di Media Sosial

Cici Suhaeni^{1*}, Bagus Sartono¹

¹ IPB University

* Corresponding Author: cici_suhaeni@apps.ipb.ac.id

Article Information

Article History:

Submitted: November, 28 2022

Accepted: December, 14 2022

Published: December, 31 2022

Key Words:

Crawling Data
Data Media Sosial
Text Clustering
Text Mining
Twitter

Abstract

Informasi mengenai vaksin covid-19 dan program vaksinasi pemerintah merupakan isu yang mendapat perhatian besar masyarakat dan menjadi perbincangan utama di media sosial, termasuk twitter. Beragam tema dan sudut pandang telah disampaikan oleh masyarakat, dan penelitian ini berupaya mengidentifikasi opini apa saja yang berkembang. Pengetahuan ini dapat menjadi masukan bagi pemerintah dan pemangku kepentingan lain untuk secara bersama membantu proses pemulihan dampak pandemi. Identifikasi opini masyarakat mengenai vaksin covid-19 dilakukan menggunakan metode text clustering terhadap tweets hasil crawling data di Twitter dalam kurun waktu 1 s.d 7 Agustus 2021. Hasil analisis menunjukkan terdapat enam tema besar yang menjadi isu perbincangan yaitu: (1) Kepercayaan terhadap efek vaksin, (2) keikutsertaan dalam vaksinasi untuk mencegah terpapar covid, (3) vaksinasi sebagai upaya herd immunity, (4) kemampuan vaksin melawan virus, (5) jenis-jenis vaksin (6) riset medis tentang.

1. PENDAHULUAN

Corona virus disease (Covid-19) merupakan pademi global yang telah dirasakan bersama oleh hampir seluruh masyarakat di Indonesia. Menurut laporan Satuan Tugas (Satgas) Penanganan Covid-19 Indonesia per 27 Juni 2021, angka Covid di Indonesia meningkat di awal tahun 2021 yaitu pada bulan Januari dan Februari. Setelah sempat mengalami sedikit penurunan, angka covid kembali melonjak di akhir Juni 2021. Bahkan, pada pada satu minggu terakhir terdapat 29 provinsi yang mengalami kenaikan kasus dengan kenaikan angka kematian sebesar 40,2% (Satgas Covid, 2021). Hal ini tentu menjadi perhatian serius bagi pemerintah.

Berbagai upaya telah dilakukan pemerintah untuk menuntaskan penyebaran Covid di Indonesia, salah satunya adalah dengan vaksinasi. Program vaksinasi Covid-19 di Indonesia yang mulai di laksanakan pada bulan Januari 2021 terus mendapatkan respon dari berbagai kalangan masyarakat. Berbagai sudut pandang disampaikan berbagai pihak tidak hanya melalui forum resmi, namun juga banyak diperbincangkan di media sosial, salah satunya adalah twitter.

Identifikasi opini masyarakat di media sosial Twitter sangatlah perlu dilakukan dalam rangka menjarang isu yang berkembang di masyarakat. Twitter merupakan media sosial terpopuler urutan kelima di Indonesia (Dahono, 2021). Menurut *We Are Social* dan *Hotsuite* (2021), banyaknya pengguna Twitter di Indonesia mencapai 14,05 juta orang. Melalui twitter, masyarakat bebas mengeluarkan pendapatnya terkait apapun dimanapun dan kapanpun. Mengingat popularitas twitter dikalangan masyarakat Indonesia yang cukup tinggi tersebut, mengumpulkan data melalui twitter menjadi cara yang efektif dan efisien di era digital seperti sekarang ini. Menurut Rachman dan Pramana (2020), pemanfaatan data yang bersumber dari media sosial merupakan teknik alternatif pengumpulan data sebagai pengganti survei tradisional.

Penarikan data melalui media sosial dikenal dengan istilah crawling. Menurut Sembodo, Setiawan, dan Baizal (2016) crawling data twitter adalah suatu proses untuk mengambil atau mengunduh data dari server twitter dengan bantuan Application Programming Integration (API) twitter baik berupa data user maupun data tweet. Crawling data Twitter merupakan salah satu metode penarikan data text yang sangat populer

dalam bidang *Natural Language Processing* (NLP). Salah satu metode populer untuk menganalisis data dalam NLP adalah *text clustering*.

Seperti halnya *clustering* pada data terstruktur, metode *text clustering* juga bertujuan untuk mengelompokkan objek ke dalam beberapa kelompok berdasarkan kemiripan karakteristik objek-objek tersebut. Dalam konteks NLP, karakteristik objek tersebut adalah “kata-kata” yang diperoleh setelah data mentah melewati pra proses data. Menurut Suyal, Panwar, dan Negi (2014), *text clustering* adalah proses untuk membagi konten teks (dokumen) ke dalam kelompok yang berbeda sesuai dengan kesamaan (*similarity*) antar dokumen. Dalam hal ini *text clustering* adalah untuk menemukan dokumen yang memiliki kata-kata paling mirip dengan dokumen lainnya.

Penerapan *text clustering* untuk data *tweet* mengenai vaksin covid dapat membantu mengidentifikasi topik yang ramai diperbincangkan masyarakat. Hasilnya dapat dijadikan sebagai bahan penentuan kebijakan bagi pemerintah dan pemangku kepentingan lainnya untuk bersinergi dalam membantu proses pemulihan dampak pandemi. Oleh karena itu, penelitian ini bertujuan untuk mengidentifikasi tema perbincangan masyarakat dimedia sosial Twitter mengenai vaksin covid menggunakan analisis *text clustering*.

Analisis data twitter mengenai vaksin covid di Indonesia telah banyak dilakukan oleh peneliti terdahulu menggunakan metode yang berbeda. Aditama *et al* (2020) melakukan analisis data twitter mengenai vaksin covid menggunakan analisis sentimen. Rachman dan Pramana (2020) melakukannya menggunakan analisis sentimen berbasis kamus postif-negatif dan pengelompokan topik menggunakan *Latent Dirichlet Allocation* (LDA). Lestandy, Abdurrahim, dan Syafa’ah (2021) melakukan analisis klasifikasi sentimen untuk kasus ini menggunakan *Recurrent Neural Network* dan *Naïve Bayes*. Kemudian, Pristiyono *et al* (2021) melakukan analisis sentimen pada data twitter mengenai vaksin covid ini menggunakan *Naïve Bayes*.

2. METODE

Data yang digunakan dalam penelitian ini diperoleh dari crawling data di Twitter dalam kurun waktu 1 s.d 7 Agustus 2021. Crawling data dilakukan menggunakan Application Programming Interface (API) Twitter dengan *library Tweepy* di Python. Kata kunci yang digunakan untuk menarik data adalah “vaksin covid”.

Setelah data berhasil diperoleh, selanjutnya dilakukan pra proses data, ekstraksi fitur, dan proses clustering sebagai berikut:

1) Praproses Data

Praproses data merupakan tahapan terpenting dalam NLP agar data siap dianalisis. Pada tahap ini, dilakukan beberapa operasi berikut, yaitu menyeragamkan format text ke bentuk *Lower Case*, *Remove Slang Words*, *Remove Punctuation*, *Remove HTML Tag*, *Remove Link*, *Remove username*, *Remove Hashtag*, *Remove Stopword*, dan juga menghapus data duplikat.

2) Ekstraksi Fitur

Pada tahap ini dilakukan transformasi data text menjadi data numerik. Penelitian ini menggunakan model Bag of Words (BoW). BoW adalah sebuah model yang mempelajari sebuah kosakata dari seluruh dokumen, lalu memodelkan tiap dokumen dengan menghitung jumlah kemunculan setiap kata. Dalam model ini, sebuah teks (kalimat ataupun dokumen) direpresentasikan sebagai kantung (bag) multiset dari kata-kata yang terkandung di dalamnya, tanpa memandang urutan kata dan tata bahasa namun tetap mempertahankan keberagamannya (Deepu, Pethuru, & Rajaraajeswari, 2016).

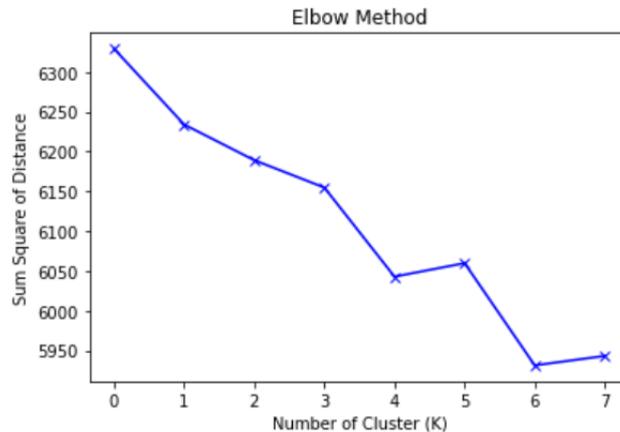
3) Proses Clustering

Setelah data text berbentuk data numerik, selanjutnya dilakukan pengelompokan text menggunakan algoritma *K-Means Clustering*. Ukuran jarak yang digunakan adalah Jarak *Euclid*. Analisis cluster di lakukan dengan menentukan *initial cluster number* mulai dari 2 hingga 10. Selanjutnya, pemilihan banyaknya cluster optimal dilakukan dengan menggunakan metode Elbow. Hasil clustering divisualisasikan menggunakan *Wordcloud*.

3. HASIL DAN PEMBAHASAN

Setelah dilakukan pra proses data, dari 1000 tweet yang ditarik, terdapat 859 tweet dan 2226 fitur (kata-kata). Hal ini disebabkan karena ada beberapa tweet yang tidak memenuhi syarat dan ada beberapa duplikasi. Data hasil pra proses inilah yang selanjutnya masuk pada proses ekstraksi fitur dan *text clustering*.

Pada penelitian ini K-Means Clustering dilakukan dengan menentukan initial cluster number mulai dari 2 sampai dengan 10 cluster. Banyaknya cluster optimal ditentukan berdasarkan metode Elbow. Kriterianya adalah banyaknya cluster optimal dipilih ketika terjadi penurunan grafik yang paling curam. Gambar 1 menampilkan grafik hasil metode Elbow.

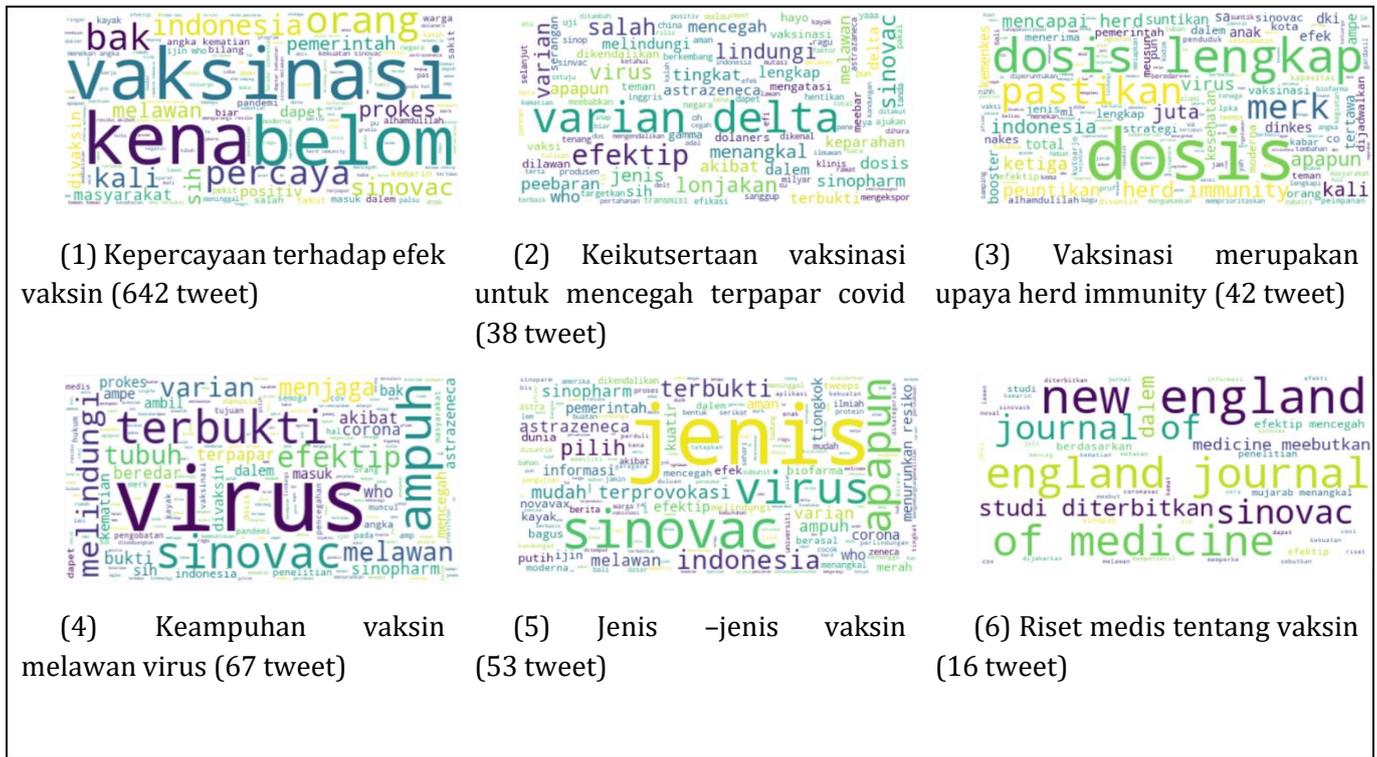


Gambar 1. Hasil Metode Elbow

Pada Gambar 1 terlihat bahwa penurunan grafik terjadi pada banyaknya cluster (k) sama dengan 6. Oleh karena itu, banyaknya cluster optimal yang dipilih dalam penelitian ini adalah 6 cluster. Banyaknya cluster ini mengindikasikan ada sebanyak 6 tema perbincangan mengenai vaksin covid. Hal ini karena dalam setiap cluster dapat diidentifikasi kata-kata yang sering muncul yang kemudian itu menjadi ciri dari cluster tersebut. Dari kata-kata pencari tersebut dapat diinterpretasikan sebuah tema atau topik dalam cluster. Sebuah *tweet* akan berada dalam satu cluster dengan *tweet* lainnya apabila kata-kata yang sering muncul di *tweet* tersebut memiliki banyak kemiripan dengan kata-kata pada *tweet* di dalam satu cluster.

Hasil clustering menggunakan algoritma K-Means ditampilkan menggunakan *wordcloud* pada Gambar 2. Kata-kata yang memiliki frekuensi kemunculan tinggi ditandai dengan ukuran huruf yang lebih besar di dalam *wordcloud*. Melalui kata-kata tersebut dapat disimpulkan topik pembicaraan yang ramai diperbincangkan. Dalam menginterpretasikan hasil text clustering ini diperlukan kehati-hatian agar tidak bias dalam menyimpulkannya. Pada gambar 2 juga terdapat informasi banyaknya anggota setiap cluster, yang berarti disini merupakan banyaknya tweet.

Berdasarkan kata-kata yang sering muncul pada wordcloud di cluster 1 yaitu sebanyak 642 *tweet*, dapat disimpulkan bahwa mayoritas tweet membicarakan tentang kepercayaan terhadap efek vaksin. Hal ini ditandai dengan kata-kata “vaksinasi”, “kena”, “belum”, “percaya”, “orang”, “Indonesia”. Cluster 2 yang berisi 38 *tweet* banyak berkomentar tentang keikutsertaan vaksinasi untuk mencegah terpapar covid. Hal ini ditandai dengan kemunculan kata “varian delta”, “varian”, “efektif”, “lindungi”, “sinovac” yang lebih sering. Pada cluster 3 yang terdiri atas 42 *tweet*, banyak muncul kata-kata “dosis lengkap”, “dosis”, “pastikan”, “merk”, dan juga ada kata “herd immunity”. Oleh karena itu, dapat disimpulkan bahwa mayoritas tweet pada cluster 3 ini membicarakan tentang vaksinasi merupakan upaya *herd immunity*.



Gambar 2. Hasil Clustering

Selanjutnya, pada cluster 4 yang berisi 67 tweet, dapat disimpulkan bahwa topik yang banyak dibicarakan adalah mengenai keampuhan vaksin melawan virus. Hal ini dicirikan dengan kata-kata “virus”, “terbukti”, “ampuh”, “sinovac”, “melindungi”, “sinopharm”, “efektif” dan lain-lain. Pada cluster ini, vaksin sinovac merupakan vaksin yang banyak diperbincangkan. Cluster 5 terdiri atas 53 tweet, merupakan kumpulan tweet yang banyak membicarakan tentang jenis-jenis vaksin. Hal ini dapat dilihat dari kata-kata “jenis”, “virus”, “sinovac”, “apapun”, “astrazeneca”, “sinopharm” yang berukuran dominan. Jika diperhatikan, kata-kata yang muncul pada cluster 4 dan 5 masih terlihat banyak yang tumpang tindih, artinya kata tersebut muncul di cluster 4 dan juga di cluster 5. Terakhir, cluster 6 yang terdiri atas 16 tweet banyak membicarakan riset medis tentang vaksin covid.

Mencermati banyaknya tweet pada setiap cluster, dapat dilihat bahwa cluster 1 memiliki paling banyak anggota, yaitu 642 tweet. Hal ini berarti mayoritas tweet yang diperoleh pada penelitian ini memperbincangkan tentang kepercayaan masyarakat terhadap efek vaksin. Sedangkan, cluster 6 memiliki paling sedikit anggota, artinya hanya sedikit tweet yang membicarakan riset medis tentang vaksin.

4. SIMPULAN

Penelitian ini telah menerapkan *text clustering* dengan algoritma K-Means pada data *tweet* mengenai vaksin covid di Indonesia. Sebelum masuk pada proses clustering, dilakukan pra proses data sehingga dihasilkan data bersih berupa kata-kata penting yang sudah siap ditransformasi ke dalam vektor numerik. Ekstraksi fitur yang dipilih dalam penelitian ini adalah Bag of Words (BoW). Berdasarkan grafik Elbow, diperoleh banyaknya cluster optimal sebanyak 6 cluster. Cluster ini mengindikasikan tema perbincangan masyarakat di media sosial Twitter. Tema pertama adalah mengenai kepercayaan masyarakat terhadap efek vaksin; kedua tentang kekhawatiran baru virus covid; ketiga tentang vaksinasi merupakan upaya herd immunity; keempat tentang keampuhan vaksin melawan virus; kelima tentang jenis-jenis vaksin; dan keenam adalah riset medis tentang vaksin.

Penelitian ini masih banyak keterbatasan, sehingga perlu dilakukan beberapa hal untuk studi lanjutan. Pertama, data yang diambil dari twitter hanya terbatas 7 hari, karena menggunakan akun student untuk *twitter developer*. Bagi peneliti yang ingin melakukan crawling data twitter dengan volume yang lebih besar

dan jangka waktu yang lebih panjang, dapat mendaftar akun *twitter developer* untuk riset akademik. Kedua, mengenai metode yang digunakan dalam penelitian ini hanya menggunakan K-Means clustering dan metode penentuan cluster optimum adalah Elbow. Penelitian selanjutnya dapat mencobakan metode clustering lain dan metode penentuan cluster optimum atau evaluasi kebaikan cluster yang lainnya. Ketiga, untuk mengidentifikasi topik pembicaraan di media sosial, selain *text clustering*, dapat juga digunakan *topic modelling*.

REFERENSI

- [1] Satgas Covid-19. (2021). Analisis Data Covid-19 Indonesia : Update Per 27 Juni 2021.
- [2] Dahono, Y. (2021). Data: Ini Media Sosial Paling Populer di Indonesia 2020-2021. [Internet]. Tersedia pada: <https://www.beritasatu.com/digital/733355/data-ini-media-sosial-paling-populer-di-indonesia-20202021>. Diakses pada tanggal 27 September 2021
- [3] We Are Social dan Hootsuite. Digital 2021 Indonesia.
- [4] Rachman, F. F., & Pramana, S. (2020). Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter. *Indonesian of Health Information Management Journal*. Vol.8, No.2 p.100-109.
- [5] Sembodo, J.E., Setiawan, B.S., & Baizal, Z. K. A. (2016). Data Crawling Otomatis pada Twitter. *Ind Symposium on Computing*. No ISSN 2460-3295. Pp 11-16. doi:10.21108/indosc.2016.111.
- [6] Suyal, H., Panwar, A., & Negi, A. S. (2014). Text Clustering Algorithms: A Review. *International Journal of Computer Applications*. Vol. 96 – No 24.
- [7] Aditama, M. I., Pratama, R. I., Wiwaha, K. H. U., & Rakhmawati, N. A. (2020). Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin Covid-19. *Journal Information Engineering and Educational Technology*. Volume 04 Nomor 02.
- [8] Lestandy, M., Abdurrahim, A., & Syafa'ah, L. (2021). Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(4), 802 - 808. <https://doi.org/10.29207/resti.v5i4.3308>.
- [9] Pristiyono, Ritonga, M., Ihsan, M. A. A, Anjar, A., & Rambe, F. H. (2021). Sentiment Analysis of Covid-19 Vaccine in Indonesia Using Naive Bayes Algorithm. *Annual Conference on Computer Science Engineering (AC2SET) 2020*. 1088 (2021) 012045. doi:10.1088/1757-899X/1088/1/012045.
- [10] Deepu, S., Raj, P., & Rajaraajeswari, S. *International Journal of Advanced Networking & Applications (IJANA)*. ISSN: 0975-0282