Implementasi LDA untuk Pengelompokan Topik Cuitan Akun Bot Twitter bertagar #Covid-19

LDA Implementation for Topic of Bot's Tweets with #Covid-19 Hashtag

Faza Rashif[1], Goldio Ihza Perwira Nirvana[2], Muhammad Alif Noor[3], Nur Aini Rakhmawati[4]
[1,2]Institut Teknologi Sepuluh Nopember; Jalan Teknik Kimia, Keputih, Kec. Sukolilo, Kota Surabaya, Jawa Timur 60111
[3]Jurusan Sistem Informasi, FTEIC ITS, Surabaya
email:[1]**rashiffaza@gmail.com**, [2]nirvanagoldio@gmail.com, [3]malifnoorf@gmail.com, [4]nur.aini@is.its.ac.id

Abstrak

Twitter merupakan media sosial yang sedang mengalami perkembangan yang pesat, karena pengguna dapat berinteraksi satu sama lain menggunakan media komputer atau perangkat mobile. Perubahan tagar trending yang berubah dengan cepat sesuai sesuai dengan intensitas pengguna membicarakan hal tertentu. Sehingga media social twitter ini cocok untuk merumpi membicarakan hal-hal terkini, salah satunya masalah Covid-19. Hal ini tidak menutup kemungkinan ada oknum yang menggunakan predikat ini untuk membuat berita untuk menggiring opini public mengenai Covid-19 mengenai berita baik maupun berita yang tak bersumber yang dapat menyebar dengan cepat. Pada penelitian ini penulis ingin mengetahui macam-macam topik yang dibahas oleh akun bot terhadap penyebaran informasi menggunakan tagar #covid19. Penelitian ini dilakukan dengan menggunakan metode Latent Dirichlet Allocation (LDA). Analisis dilakukan setelah melakukan text mining pada 162 Tweet dari 62 akun bot Twitter. Untuk menentukan jumlah topik yang optimal, yakni dengan melihat nilai perplexity dan topik coherence. Hasil yang didapatkan adalah lima topik teratas antara lain tentang kondisi dan dampak pandemi saat ini, himbauan untuk menjaga jarak agar Kesehatan tetap terjaga, perkembangan penyebaran Covid-19 yang ada di Indonesia, vaksinasi yang terjadi di beberapa wilayah di Indonesia, dan cara menghadapi Covid-19.

Kata kunci—Covid-19, Twitter, Akun Bot, LDA

Abstract

Twitter is a social media that is experiencing rapid development, because users can rely on each other using computers or mobile devices. Trending hashtag that change rapidly according to the intensity of the user talking about a certain thing. So that twitter is suitable for chatting about the latest things, one of them is the Covid-19 topic. There is a possibility that there are people who use this predicate to lead public opinion about Covid-19 regarding good news or news that cannot be trusted which can spread quickly. In this study, the authors wanted to know the kinds of topics discussed by bot accounts for information dissemination using the covid19 hasthag. This research was conducted using the Latent Dirichlet Allocation (LDA) method. The analysis was carried out after text mining on 162 tweets from 62 Twitter bot accounts. To determine the optimal number of topics, namely by looking at the value of perplexity and topic coherence. The results obtained are the top lima topics, including the condition and impact of the

current pandemic, an appeal to health protocol advice on maintaining distance, growth of the spread of Covid-19 in Indonesia, vaccinations that occur in several regions in Indonesia, and how to deal with Covid-19.

Keywords—Covid-19, Twitter, Bot Account, LDA

1. PENDAHULUAN

Twitter merupakan media sosial yang sedang mengalami perkembangan yang pesat, karena pengguna dapat berinteraksi satu sama lain menggunakan media komputer atau perangkat *mobile.* Twitter sendiri memiliki fitur *thread* dan *trending*, yang merupakan *micro blogging* yang cocok untuk dijadikan sebagai tempat berkumpul di dunia maya. Perubahan tagar *trending* yang berubah dengan cepat sesuai sesuai dengan intensitas pengguna membicarakan hal tertentu [1] sehingga media sosial Twitter ini cocok untuk merumpi dan membicarakan hal-hal terkini, salah satunya masalah Covid-19.

Hal ini tak menutup kemungkinan adanya oknum yang memanfaatkan twitter untuk menggiring opini tertentu mengenai Covid-19, agar masyarakat dapat terhasut dan mempercayai sebuah informasi. Tercatat Indonesia merupakan negara dengan pengguna aktif terbesar ke delapan saat ini [2]. Hal ini tidak menutup kemungkinan ada oknum yang menggunakan predikat ini untuk membuat berita untuk menggiring opini publik mengenai Covid-19. mengenai berita baik maupun berita yang tak bersumber yang dapat menyebar dengan cepat. Terbukti penelitian telah dilakukan pada 20 Mei 2020 menyebutkan bahwa 4lima persen *Tweet* yang dikirim oleh akun yang berperilaku seperti robot terkomputerisasi daripada manusia [3]. Hal ini seperti mesin propaganda yang menggiring opini publik.

Dengan itu, kami ingin mengadakan penelitian dengan menggunakan Metode LDA untuk mengetahui macam-macam topik yang dibahas oleh akun bot terhadap penyebaran informasi menggunakan tagar #Covid19. Penelitian sebelumnya yang berjudul "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia Tentang Vaksin Covid-19 pada Media Sosial Twitter" juga menggunakan metode *Latent Dirichlet Allocation* (LDA) untuk mengetahui topik pembicaraan yang sering dibahas oleh masyarakat terhadap wacana vaksinasi [4]. Oleh karena itu, kami ingin mencoba untuk mengembangkan LDA dalam Pengelompokan Topik *Tweet* Akun Bot Twitter bertagar #Covid-19.

2. TINJAUAN PUSTAKA

Tahap tinjauan pustaka dilakukan dengan tujuan agar dapat memahami konsep dari topik penelitian yang akan dikerjakan. Tahap studi literatur dilakukan dengan cara menggali informasi berdasarkan pada jurnal, buku, artikel ilmiah, website, dan skripsi mahasiswa yang terpercaya sehingga dapat mendukung berjalanya penulisan ilmiah ini. Beberapa inormasi yang didapatkan meliputi penjelasan mengenai Twitter, Covid-19, Akun bot, dan metode LDA.

2.1. Twitter

Twitter adalah situs web dimiliki dan dioperasikan oleh Twitter, Inc. yang menawarkan jaringan sosial berupa *microblog*. Disebut *microblog* karena situs ini memungkinkan pengguna mengirim dan membaca pesan blog seperti pada umumnya namun terbatas hanya sejumlah 140 karakter yang ditampilkan pada halaman profil pengguna. Twitter memiliki karakteristik dan format penulisan yang unik dengan simbol ataupun aturan khusus. Pesan dalam Twitter dikenal dengan sebutan *Tweet* [5].

2.2 Covid 19

World Health Organization (WHO) menjelaskan bahwa *Coronaviruses* (Cov) adalah virus yang menginfeksi sistem pernapasan. Infeksi virus ini disebut Covid-19. Virus Corona menyebabkan penyakit flu biasa sampai penyakit yang lebih parah seperti Sindrom Pernafasan Timur Tengah (MERS-CoV) dan Sindrom Pernafasan Akut Parah (SARS-CoV). Virus ini menular dengan cepat dan telah menyebar ke beberapa negara, termasuk Indonesia. Di Indonesia, penyebaran virus ini dimulai sejak tanggal 2 Maret 2020, diduga berawal dari salah satu warga negara Indonesia yang melakukan kontak langsung dengan warga negara asing yang berasal dari Jepang. Hal tersebut telah diumumkan oleh Bapak Presiden Jokowi. Seiring dengan berjalanya waktu, penyebaran Covid-19 telah mengalami peningkatan yang signifikan [6].

2.3 Akun Bot

Akun bot merupakan akun di dalam media sosial yang dikendalikan oleh program perangkat lunak tertentu untuk isi konten maupun perilakunya. Akun bot sendiri tidak jarang disalah artikan sebagai akun media sosial individu asli karena di media sosial, akun bot sendiri tidak memiliki tanda khusus sebagai bot media sosial. Akun bot sendiri dapat mendistorsi keriuhan suara di media sosial sehingga ia terlihat seakan-akan benar bahwa banyak orang sedang membicarakannya atau, bahkan lebih buruk, dianggap sebagai suara publik. Dengan menggunakan *machine learning,* akun bot dapat menyerupai perilaku manusia dengan membaca dan belajar konten sosial media melebihi manusia, karena ia dapat aktif 24 jam tanpa henti [7].

2.4 Topic Modelling

Topic modelling terdiri dari entitas-entitas yaitu "kata", "dokumen", dan "korpora". "Kata" dianggap sebagai unit dasar dari data diskrit dalam dokumen, yang didefinisikan sebagai item dari kosakata yang diberi indeks untuk setiap kata unik yang ada dalam dokumen. "Dokumen" merupakan susunan N kata-kata. Sebuah korpus adalah kumpulan M dokumen dan korpora merupakan bentuk jamak dari korpus. *"Topic"* adalah distribusi dari beberapa kosakata yang bersifat tetap. Secara sederhana, setiap dokumen dalam korpus mengandung proporsi yang berbeda dari topik-topik yang dibahas sesuai dengan kata-kata yang ada di dalamnya.

Ide dasar dari *topic modeling* ialah sebuah topik terdiri dari kata-kata tertentu yang menyusun topik tersebut, dan dalam satu dokumen memiliki kemungkinan terdiri dari beberapa topik dengan probabilitas masing-masing. Namun, manusia memahami dokumen-dokumen merupakan sebuah objek yang dapat diamati, sedangkan topik, distribusi topik per-dokumen, dan penggolongan setiap kata pada topik perdokumen merupakan bagian yang tersembunyi. Maka dari itu, *topic modelling* bertujuan untuk menemukan topik dan kata yang yang tersembunyi pada topik tersebut.

Kumpulan dokumen memiliki distribusi probabilitas topik, setiap kata yang dianggap diambil dari salah satu topik tersebut. Dengan distribusi probabilitas topik di setiap dokumen, dapat diketahui seberapa banyak masing-masing topik terlibat dalam sebuah dokumen. Hal ini dapat mengetahui topik mana yang terutama dibicarakan suatu dokumen.

2.5 *Unsupervised learning*

Unsupervised learning adalah jenis pembelajaran mesin yang memungkinkan model untuk belajar sendiri dengan sendirinya dan tidak melibatkan pengawasan supervisor. Berbeda dengan *supervised learning*, *unsupervised learning* lebih bebas dalam proses pendalaman data karena data tidak memiliki label [8]. *Unsupervised learning* memungkinkan pengguna untuk melakukan tugas yang lebih menantang tanpa terikat oleh aturan. Mereka juga dapat mendeteksi kesalahan perilaku dan memberikan hasil yang lebih baik daripada *supervised learning*. Algoritma ini menggunakan titik data sebagai referensi untuk menemukan struktur dan pola yang ada di dalam data set.

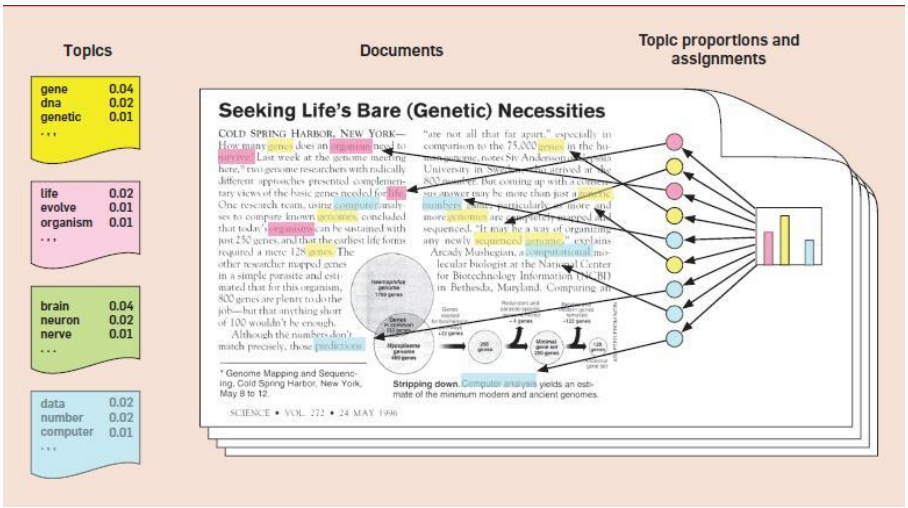
Contoh *unsupervised learning* dalam kehidupan sehari-hari adalah pembelajaran bayi dan seekor anjing peliharaannya. Bayi ini mengetahui detil anjing peliharaanya. Beberapa minggu

kemudian, seorang teman keluarga membawa serta seekor anjing dan mencoba bermain dengan bayi itu. Bayi itu belum pernah melihat anjing ini sebelumnya. Tapi ia mengenali banyak kesamaan seperti anjing peliharaannya (2 telinga, mata, berjalan dengan 4 kaki). Dia mengidentifikasi hewan baru itu sebagai anjing. Ini adalah pembelajaran tanpa pengawasan, di mana Anda tidak diajarkan, tetapi Anda belajar dari data (dalam hal ini data tentang seekor anjing.) Jika ini adalah supervised learning, teman keluarga akan memberi tahu bayi itu bahwa itu adalah anjing seperti yang ditunjukkan di atas.

Unsupervised learning memiliki kelebihan dan kekurangan [9]. Kelebihan dari *unsupervised learning* adalah bisa mendeteksi apa yang tidak dipahami mata manusia, potensi pola tersembunyi bisa sangat kuat untuk bisnis atau bahkan mendeteksi fakta yang sangat menakjubkan, dan deteksi penipuan, dan output dapat menentukan wilayah yang belum dijelajahi dan menentukan usaha baru untuk bisnis. Kekurangan dari *unsupervised learning* adalah pembelajaran yang lebih sulit dibanding dengan supervised learning, biaya lebih mahal, dan kegunaan hasil sulit diidentifikasi karena tidak ada label.

2.6 Metode Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) adalah contoh metode *unsupervised learning* yang digunakan untuk mengelompokkan data kedalam beberapa kelas, meringkas, dan memproses data yang berukuran besar. LDA berkerja menggunakan *Gibbs sampling* [10]. LDA dipilih karena dapat melakukan analisis pada data serta dokumen yang berukuran besar. LDA menggunakan metode *bag of words* untuk mengidentifikasi informasi topik tersembunyi dalam kumpulan dokumen besar. Metode ini memperlakukan setiap dokumen sebagai vektor jumlah kata dan mewakili distribusi probabilitas beberapa topik, dan setiap topik direpresentasikan sebagai distribusi probabilitas beberapa kata. Mekanisme kerja LDA dibagi menjadi dua bagian yaitu penalaran dan realisasi. Inferensi adalah proses LDA yang digunakan untuk menentukan bobot setiap kata dalam setiap dokumen dalam korpus. Implementasi merupakan tahapan dimana aplikasi LDA selanjutnya memenuhi kebutuhan temu kembali informasi. Implementasi merupakan tahap penerapan LDA untuk kebutuhan temu kembali informasi selanjutnya [11].



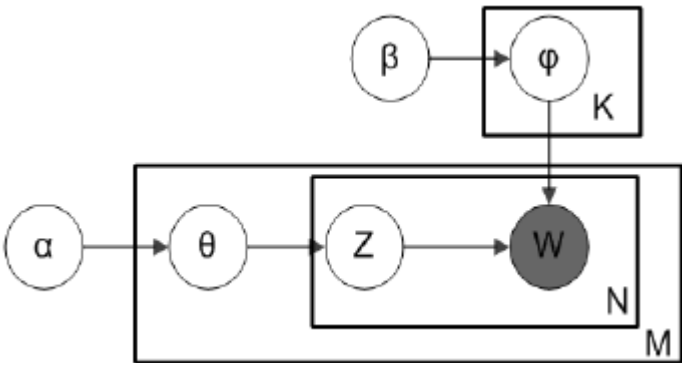
Gambar 1 Ilustrasi LDA

Cara kerja model LDA adalah pertama mengasumsikan bahwa entri ditentukan sebelum diambil dan dokumen dalam gambar adalah baris subjek di sebelah kiri. Untuk setiap dokumen dalam kumpulan dilakukan:

1. Distribusi secara acak subjek yang dipilih (dari angka yang diwakili oleh grafik distribusi item di sebelah kanan)
2. Untuk setiap kata dalam dokumen:

a. Subjek dipilih secara acak dari distribusi subjek pada langkah 1 (ditunjukkan oleh hubungan dalam grafik yen pada Gambar).
b. Distribusi kata dipilih secara acak dari distribusi kosakata yang sesuai. (Pada gambar direnderdengan memilih warna lingkaran.)

LDA diwakili oleh model grafis menggunakan notasi pelat seperti yang ditunjukkan pada Gambar 2.



Gambar 2 Plate Notation LDA

dimana :
- β adalah dirichlet parameter atas distribusi kata terhadap topik.
- φ adalah distribusi kata terhadap topik dalam corpus.
- K adalah kumpulan topik.
- W adalah kata.
- N adalah kumpulan kata.
- M adalah kumpulan dokumen.
- Z adalah topik *index assignment*.
- θ adalah dokumen.
- α adalah dirichlet parameter atas distribusi topik terhadap dokumen.

LDA dirumuskan sebagai berikut :

$$p(w,z,\vartheta,\varphi|\alpha,\beta) = p(\varphi|\beta)p(\vartheta|\alpha)p(z|\vartheta)p(w|\varphi_k) \quad (1)$$

Implementasi tidak dapat menerapkan LDA karena variabel Z disembunyikan/tidak diketahui. Juga sulit untuk menemukan hubungan antara Z dan W karena kata tersebut mungkin mengandung lebih dari satu elemen. Juga menghasilkan p (Z | W) :

$$p(\vec{z}|\vec{w}) = \frac{p(\vec{z},\vec{w})}{p(\vec{w})} = \frac{\prod_{i=1}^{W} p(zi,wi)}{\prod_{i=1}^{W}\sum_{k=1}^{K} p(zi=k,wi)} \quad (2)$$

Pembagi diperlukan untuk menjumlahkan klausa KW dan karenanya tidak dapat dikurangi atau dikelola. Salah satu cara untuk mengatasi masalah ini adalah dengan menggunakan algoritma *Gibbs Sampling* [12].

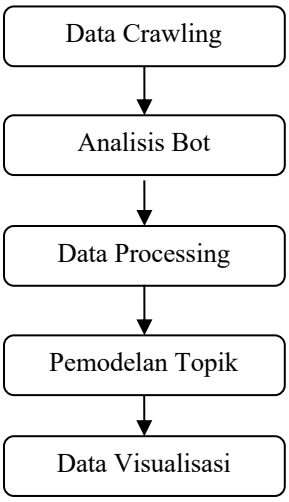
2.7 Algoritma Gibbs Sampling

Gibbs sampling diperkenalkan oleh Geman dan Geman pada tahun 1984. Algoritma ini adalah kasus khusus dari algoritma *MetropolisHastings* satu komponen yang menggunakan kepadatan untuk menunjukkan bahwa ini adalah distribusi target bersyarat penuh. Popularitas tawaran ini menciptakan peluang untuk diterima. Oleh karena itu, perpindahan yang diusulkan dapat diterima untuk semua iterasi. Salah satu keuntungan dari *Gibbs sampling* adalah kebutuhan

untuk menghasilkan nilai acak dari distribusi satu arah dimana alat komputasi yang berbeda tersedia pada setiap langkah. Karena dalam banyak kasus distribusi bersyarat ini memiliki bentuk yang diketahui, maka mudah untuk mensimulasikan beberapa nilai arbitrer menggunakan fungsi standar dalam perangkat lunak statistik dan komputasi. Pengambilan sampel *Gibbs* selalu beralih ke nilai baru, dan yang terpenting tidak memerlukan spesifikasi distribusi yang diusulkan.

3. METODE PENELITIAN

Pada bab ini akan dijelaskan metode penelitian agar pengerjaan lebih terarah dan sistematis dan untuk mendapatkan hasil akhir yang sesuai dengan tujuan pada bab latar belakang. Adapun urutan metode penelitian dapat dilihat pada gambar 3



Gambar 3 Diagram Metode Penelitian

3.1 Data Crawling

Crawling data dilakukan untuk mengumpulkan data pada sebuah website yakni Twitter. Tahap pertama Crawling data dilakukan dengan menghimpun terlebih dahulu ciutan dengan tagar #Covid-19 berbahasa Indonesia dengan menggunakan *query* bahasa *Python*. Selanjutnya akan dilakukan penambangan data untuk mendapatkan data berupa index, ciutan, tanggal jumlah *tweet* kembali, dan nama pengguna yang akan dikumpulkan untuk tahap selanjutnya. Penelitian ini membutuhkan *API key* dan *customer id* sebagai bahan untuk melakukan scraping dimana scraping adalah suatu cara untuk mengumpulkan data dari sebuah situs atau website secara otomatis.

3.2 Analisis Bot

Data yang telah didapatkan akan disaring kembali menggunakan website *https://botometer.osome.iu.edu* untuk mengindentifikasi apakah suatu akun pengguna merupakan akun *bot*. Pada penelitian ini, akan dicari 62 akun bot dari hasil *crawling* data sebelumnya Website Botometer akan memunculkan skor yang menjadi indikator dengan kisaran antara 0 hingga lima. Skor tertinggi mengindikasikan bahwa akun Twitter tersebut kemungkinan besar adalah akun bot, sementaraa skor rendah menandakan akun tersebut adalah akun asli atau dikendalikan manusia sebagai pengguna langsung [13].

3.3 Data Processing

Data yang telah didapatkan kemudian dilakukan proses *cleaning* dengan menghilangkan tanda baca dan karakter yang dirasa tidak perlu sehingga data menjadi lebih akurat dalam pengelompokkan data. Setiap data yang sejenis akan dimasukkan ke dalam suatu indek yang bertujuan untuk membedakan dokumen satu dengan dokumen lainnya. Dari sekumpulan data yang telah didapatkan kemudian dilakukan proses tokenisasi yang bertujuan memisahkan deretan kata di dalam kalimat atau paragraf menjadi potongan kata tunggal.

3.4 Pemodelan Topik

Tahapan pemodelan topik dilakukan dengan metode LDA dengan tujuan untuk mendeteksi topik-topik yang ada pada koleksi dokumen beserta proporsi kemunculan topik tersebut. Tahapan pemodelan topik dibagi menjadi dua sub tahap yaitu pembentukan model topik dan tahap validasi topik.

Tahap membentuk model topik bertujuan untuk menghasilkan model topik yang paling tepat untuk dokumen. Model topik dikatakan tepat apabila mampu menghasilkan luaran yang baik pada tahap validasi model topik. Untuk menghasilkan model topik yang tepat, hal yang dilakukan adalah dengan melakukan eksperimen pada nilai input parameter.

Tahap validasi topik bertujuan untuk memastikan model topik yang dihasilkan dari hasil model topik model yang dilakukan pada dokumen adalah benar, baik luaran berupa topik maupun kata-kata dalam topik. Dalam hal ini, tingkat kebenaran topik dapat dilakukan secara otomatis dengan *Perplexity*. *Perplexity* adalah ukuran statistik tentang seberapa baik model probabilitas memprediksi sampel, kemudian memberikan distribusi kata yang diwakili oleh topik dan membandingkan dengan distribusi kata dalam dokumen [14].

3.5 Visualisasi Data

Data yang didapatkan dari penerapan LDA akan menjadi data mentah yang kemudian dilakukan visualisasi dengan memanfaatkan grafik dan bar-chart untuk mempermudah proses analisis data berikutnya. Data visualisasi ini akan dilakukan dengan menunjukkan seberapa besar perbandingan persebaran data yang satu dengan data yang lainnya.

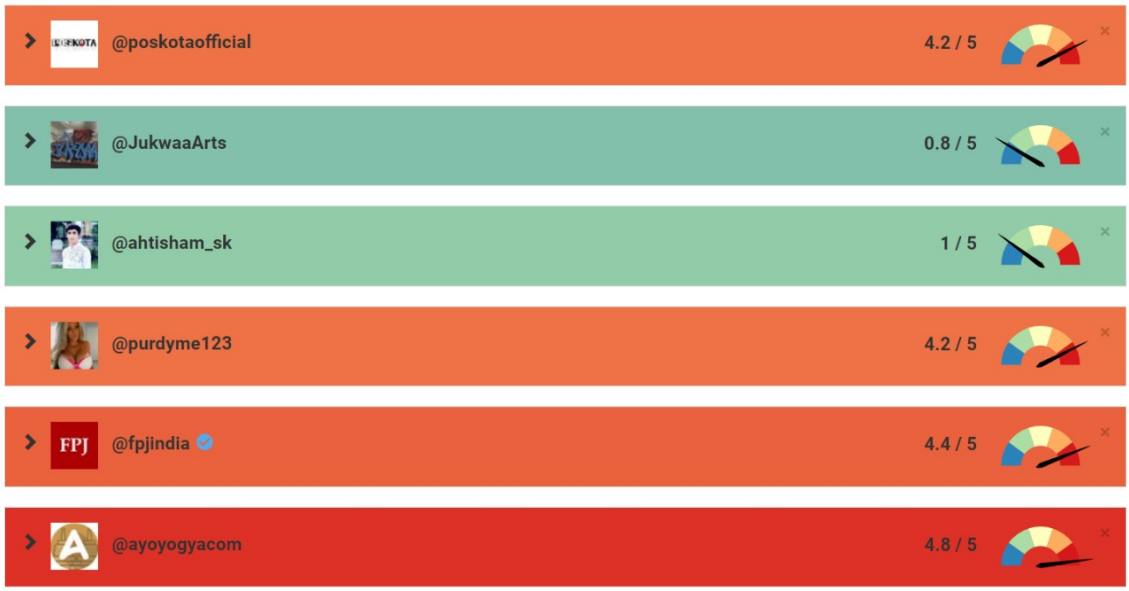
4. HASIL DAN PEMBAHASAN

4.1 Hasil Crawling Data

Hal pertama yang dilakukan yaitu melakukan *crawling data* dimana data yang digunakan dalam penelitian ini adalah data twitter. Setelah memperoleh *API key*, didapatkan data dengan variabel indek, tanggal *tweet* ,*tweet*, nama akun, dan jumlah *tweet* kembali. Semua data tersebut dihimpun dalam bentuk tabel yang disimpan dalam file berekstensi *.csv* sebagai bahan pada tahap analisis bot.

4.2 Hasil Analisis Bot

Data yang telah dihimpun akan disaring kembali untuk menemukan apakah akun twitter yang ada pada data termasuk akun bot dengan menggunakan website https://botometer.osome.iu.edu. Skor tertinggi mengindikasikan bahwa akun Twitter tersebut kemungkinan besar adalah akun bot, sementara skor rendah menandakan akun tersebut adalah akun asli atau dikendalikan manusia sebagai pengguna langsung.



Gambar 4 Salah Satu Hasil Analisis Akun Bot

Pada gambar 4, bisa dilihat bahwa akun yang memiliki *background* berwarna oren atau merah (bernilai 4 hingga 5) menandakan bahwa akun tersebut adalah akun bot sementara akun yang memiliki *background* hijau atau biru (bernilai 1 hingga 2) menandakan akun bukan bot. Untuk akun dengan *background* kuning/bernilai 3 tidak dimasukkan kedalam kriteria akun bot. Penelitian ini mengambil 62 sampel akun bot berbeda yang nantinya sebagai bahan pada tahap *data processing*. Apabila data belum mencapai 62 sampel akun bot, dilakukan *crawling data* kembali untuk mencari akun bot yang berbeda. Hasil crawling data diunggah melalui akun Zenodo .

4.3 Hasil Data Processing

Data akun bot yang telah dihimpun akan memasuki tahap preprocessing untuk mendapatkan hasil yang maksimal. Dari sekumpulan variabel data yaitu indek, *tanggal tweet ,tweet*, nama akun, dan jumlah *tweet* kemudian dipilih kembali variabel *tweet*. Dengan menggunakan program yang ada di Zenodo , dilakukan proses tokenisasi dengan memanfaatkan spasi sebagai sebagai pemisah kata dalam sebuah kalimat judul video. Adanya fungsi *lower()* ini untuk mengubah karakter yang berbentuk *uppercase* menjadi bentuk *lowercase*. Kemudian dilakukan tahapan *stopword* untuk menghilangkan beberapa kata yang tidak informatif dan tidak diperlukan dalam penelitian sehingga data yang didapatkan berupa kata yang bisa dimasukkan dalam proses LDA . Kata-kata yang dimasukkan dalam *stopword* ini didefinisikan secara manual. Salah satu bentuk *stopword* adalah kata hubung, kata sapaan, kata sifat, kata ajakan, dan kata tanya.

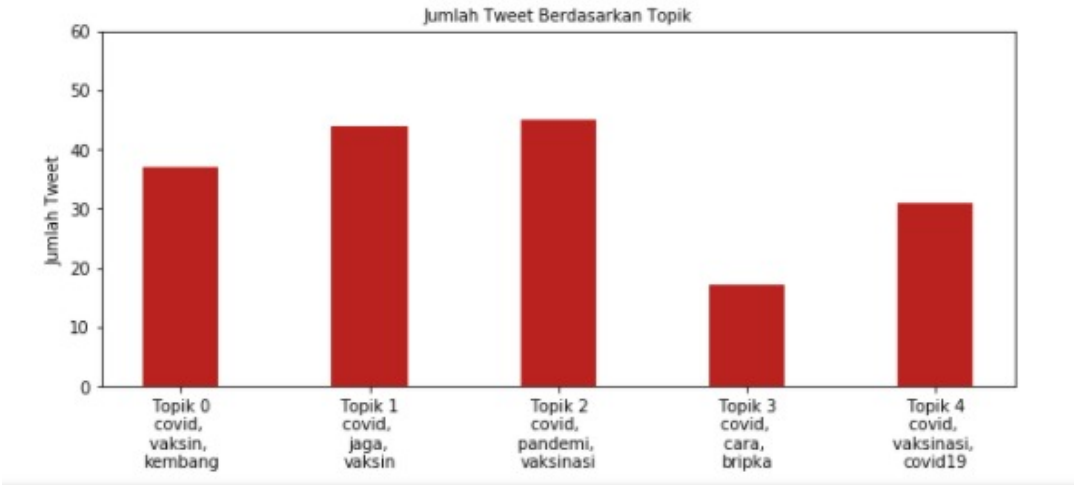
4.4 Hasil Pemodelan Topik

Data yang telah dilakukan preprocessing kemudian masuk dalam proses pemodelan topik menggunakan LDA . Dengan melakukan pemetaan pada setiap *tweet* dengan menggunakan username akun bot, data yang ada kemudian diubah menjadi *term dictionary*. Kemudian dari data tersebut, dilakukan pemodelan topik dengan menggunakan modul gensim untuk menghasilkan *document-term matrix*. Untuk menghasilkan LDA model maka menggunakan *package LdaModel*

dan kemudian menggunakan *Coherence Model* untuk menghitung banyaknya topik yang saling koheren. Semakin tinggi nilai kohernsi topik, semakin bagus hasilnya. Untuk menetapkan nilai optimal dari banyaknya topik dapat dilakukan dengan menggunakan perhitungan *perplexity*. Di mana model LDA yang baik itu dilihat dari nilai *perplexity* yang rendah.

4.5 Data Visualisasi

Setelah melakukan LDA pada data lalu mengambil hasilnya untuk divisualisasikan. Jumlah topik yang optimal dapat dilihat dengan memperhatikan nilai *perplexity*. Semakin rendah nilainya akan semakin bagus. Jika melihat dari nilai *perplexity* untuk menentukan jumlah topik yang baik maka dipilihlah dengan banyak lima topik, di mana menunjukkan nilai *perplexity* paling rendah yaitu (-8.018809653605734). Selain itu pada lima topik juga memiliki nilai kherensi topik tertinggi yaitu (0.5223357979928904). Kemudian untuk hasil topik yang koheren ditunjukkan pada gambar tiga.



Gambar 7 Hasil dengan 5 topik

Di mana dapat dijabarkan secara lebih jelas detail melalui Tabel 1 yang menunjukkan jenis topik beserta jumlahnya. Pada analisis data, dihasilkan daftar topik yang dominan dari *Tweet* tagar #Covid-19 akun bot dengan menggunakan metode LDA.

Tabel 1 Persebaran pada setiap topik

No	Dominant_topic	Keywords	Jumlah
1	0.0	Covid, vaksin, kembang	37
2	0.1	Covid, jaga, vaksin	44
3	0.2	Covid, pandemi, vaksinasi	45
4	0.3	Covid, cara, bripka	17
5	0.4	Covid, vaksinasi, covid19	31

Dari Tabel satu dan Gambar tiga menunjukkan hasil persebaran topik dalam tiga kata di setiap topiknya. Topik nol menunjukkan hasil yang membahas mengenai perkembangan penyebaran Covid-19 yang ada di Indonesia. Topik nol ini mencul pada *Tweet* dengan jumlah 37 *Tweet*. Topik satu membahas mengenai himbauan untuk menjaga jarak agar Kesehatan tetap terjaga. Topik satu ini muncul pada 44 *Tweet* yang ada. Topik dua yang membahas tentang kondisi

dan dampak pandemi saat ini. Topik dua ini muncul paling banyak dengan jumlah 45 *Tweet*. Topik tiga menunjukkan hasil di mana membahas mengenai cara menghadapi Covid-19. Topik tiga kali ini muncul pada 17 *Tweet*. Terakhir yaitu topik empat yang membahas vaksinasi yang terjadi di beberapa wilayah di Indonesia. Topik ini memiliki jumlah 31 *Tweet*.

4. KESIMPULAN

Berdasarkan analisis dari metode LDA yang dilakukan pada 62 akun bot didapatkan lima topik teratas yang menjadi pembicaraan yang sedang ramai dibahas oleh para akun bot tersebut. Topik yang paling banyak dibahas adalah topik dua membahas tentang kondisi dan dampak pandemi saat ini. Selanjutnya ada topik satu menempati urutan yang kedua mengenai himbauan untuk menjaga jarak agar Kesehatan tetap terjaga. Untuk topik yang menempati posisi yang ketiga adalah topik nol membahas mengenai perkembangan penyebaran Covid-19 yang ada di Indonesia. Kemudian topik yang menempati posisi yang keempat adalah topik empat yang membahas vaksinasi yang terjadi di beberapa wilayah di Indonesia. Selanjutnya, topik yang menempati posisi nomor lima yaitu topik tigsa yang membahas mengenai cara menghadapi Covid-19. Berdasarkan data *crawling* yang didapat, ada bebebrapa akun yang memiliki *Tweet* yang identik dengan akun yang lainnya. Hal ini membuktikan bahwa, bot digunakan sebagai penyebaran informasi mengenai Covid-19 yang ads pada media sosial Twitter.

5. SARAN

Dari pengerjaan penelitian ini, terdapat beberapa saran untuk pengembangan penelitian ke depan :

1. Dari hasil pemodelan topik, terdapat kata-kata yang tergolong sebagai nama gelar dalam topik sehingga untuk memperoleh hasil yang lebih optimal, diperlukan suatu pendeteksian nama gelar sebagai penyaringan data sebelum melakukan pemodelan topik. Terdapat pula kata-kata yang memiliki arti sama pada topik yang berbeda sehingga diperlukan normalisasi sinonim sebelum dilakukan pemodelan topik.
2. Data yang digunakan hanya berjumlah 62 baris sehingga menyebabkan topik kurang bervariasi. Untuk itu diperlukan lebih banyak data agar topik yang dihasilkan lebih beragam.

UCAPAN TERIMA KASIH

Puji dan syukur peneliti panjatkan kepada Tuhan Yang Maha Esa atas segala rahmat dan kasih karunia-Nya yang memberikan Kesehatan dan kesempatan pada peneliti sehingga dapat menyelesaikan paper ini dengan baik. Dalam menyelesaikan paper ini peneliti megalami kendala error pada codingan, namun dengan bimbimngan dan dorongan dari berbagai pihak yang akhirnya penulisan ini dapat diselesaikan dengan sebagaimana mestinya. Pada kesempatan ini, peneliti menyampaikan terima kasih pada Bu Nur Aini Rakhmawati, S.Kom., M.Sc.Eng., Ph.D sebagai dosen pengampu mata kuliah Etika Profesi yang telah banyak memberikan bimbingan dan saran kepada peneliti sejak awal hingga terselesaikannya paper ini. Semoga dengan adanya paper ini, dapat bermanfaat bagi kita semua dan menjadi bahan masukan bagi pengembang dunia Pendidikan.

DAFTAR PUSTAKA

[1] D. Inayah dan F. L. Purba, "Implementasi Social Network Analysis Dalam Penyebaran Informasi Virus Corona (Covid-19) Di Twitter," *Semin. Nas. Off. Stat.*, vol. 2020, no. 1, hal. 292–299, 2021, doi: 10.34123/semnasoffstat.v2020i1.

[2] J. Clement, "twitter: number of monthly active users 2010-2019," *statista.com*, 2019, [Daring]. Tersedia pada: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[3] J. Sinuhaji, "Peneliti Sebut Persen dari 200 Juta Cuitan Virus Corona di Twitter Adalah Bot," *pikiran-rakyat.com*, 2020. https://www.pikiran-rakyat.com/teknologi/prpeneliti-sebutpersen-dari-200-juta-cuitan-virus-corona-di-twitter-adalah-bot (diakses Apr 06, 2021).

[4] F. F. Rachman dan S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Heal. Inf. Manag. J.*, vol. 8, no. 2, hal. 100–109, 2020, [Daring]. Tersedia pada: https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/

[5] A. A. Amrullah, A. Tantoni, N. Hamdani, R. T. R. L. Bau, M. R. Ahsan, dan E. Utami, "Reviewatas Analisis Sentimen Pada Twitter Sebagai Representasi Opini Publik Terhadap Bakal Calon Pemimpin," *Pros. Semin. Nas. Multi Disiplin Ilmu Call Pap. Unisbank*, vol. 2, no. 1, hal. 978–979, 2016.

[6] I. Putra dan I. Dana, "Pengaruh Profitabilitas, Leverage, Likuiditas Dan Ukuran Perusahaan Terhadap Return Saham Perusahaan Farmasi Di Bei," *E-Jurnal Manaj. Univ. Udayana*, vol. , no. 11, hal. 249101, 2016.

[7] P. G. Pratama dan N. A. Rakhmawati, "Social bot detection on 2019 indonesiapresident candidates supporters tweets," *Procedia Comput. Sci. 161*, hal. 813–820, 2019.

[8] "Unsupervised Machine Learning: What is, Algorithms, Example," *Guru99*. https://www.guru99.com/unsupervised-machine-learning.html (diakses Jun 27, 2021).

[9] V. Sharma, "Unsupervised Learning an Angle for Unlabelled Data World," *TechTarget*, 2018. https://www.datasciencecentral.com/profiles/blogs/unsupervised-learning-an-angle-for-unlabelled-data-world (diakses Jun 27, 2021).

[10] M. Steyvers dan T. Griffiths, "Probalistic Topic Models," *Latent Semant. Anal. A Road To Mean.*, vol. 3, no. 3, hal. 993–1022, 2010, [Daring]. Tersedia pada: http://www.sciencedirect.com/science/article/pii/S0140366413001047ceas.cc/2004/167.pdfdoi.acm.org/101806338eprints.soton.ac.ukieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7033160

[11] Y. Sahria dan D. H. Fudholi, "Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA," *J. Rekayasa Sist. dan Teknol. Inf.*, vol. 4, no. 2, hal. 336–344, 2020, [Daring]. Tersedia pada: http://jurnal.iaii.or.id.

[12] W. M. Darling, "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling," *Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*,

hal. 1–10, 2011.

[13] R. F. Rizki, "Mau Deteksi Akun Bot di Twitter? Gunakan Botometer," *Teknologi.id*, 2020. https://teknologi.id/insight/mau-deteksi-akun-bot-di-twitter-gunakan-botometer (diakses Mar 30, 2021).

[14] "Topic modeling," *cfss.uchicago.edu*, 2021. https://cfss.uchicago.edu/notes/topic-modeling/#:~:text=Perplexity is a statistical measure,of words in your documents (diakses Apr 06, 2021).

[15] F. Rashif, G. I. P. Nirvana, dan M. A. N. Febriansyach, "Dataset dari Akun Bot Twitter," 2021. https://zenodo.org/record/4679107#.YHKM8OgzaUk.

.