BAB 2
LANDASAN TEORI

2.1 Vaksin COVID-19

Sejak urutan genetik (*genetic sequence*) dari virus SARS-CoV-2 dipublikasikan pada 11 Januari 2020, berbagai instansi telah berlomba mengembangkan vaksin terhadap penyakit yang diakibatkan COVID-19 [17]. Yayasan Coalition for Epidemic Preparedness Innovations (CEPI) bekerja sama bersama pihak otoritas kesehatan global dan berbagai pengembang vaksin untuk mengawasi perkembangan vaksin COVID-19. Sejak 8 April 2020, pengembangan vaksin memiliki total 115 kandidat vaksin, yang diantaranya 78 telah dikonfirmasi telah aktif berfungsi, dimana salah satunya adalah vaksin Moderna. Namun sikap oposisi dan enggan dari masyarakat terhadap vaksin juga telah bertumbuh besar. Dari survei yang diangkat pada penelitian terkait respon masyarakat oleh Machingaidze & Wiysonge [18], rasa khawatir terhadap efek samping yang ditimbulkan oleh vaksin merupakan alasan utama di balik keraguan masyarakat. Dari survei yang sama juga diketahui bahwa kebanyakan orang menganggap tenaga kesehatan merupakan sumber panduan paling terpercaya terkait vaksin COVID-19.

2.2 Twitter

Twitter merupakan media sosial yang menyediakan layanan *microblogging*. Tidak seperti blog pada umumnya, Twitter sebagai *microblog* memungkinkan pengguna untuk menuliskan konten singkat, cukup dengan panjang maksimal berjumlah 280 karakter [19]. Konten singkat itu dinamakan dengan *tweet* dan bersifat publik. Selama penggunanya tidak membuatnya *private*, pengguna Twitter lain dapat melihat bahkan meresponnya dengan *tweet* lainnya [20]. Hal inilah yang membuat Twitter sangat diminati, terbukti dengan jumlah pengguna yang mencapai 315 juta jiwa di seluruh dunia pada tahun 2020 [21], dan mencapai 15,7 juta pengguna di Indonesia sejak bulan Juli 2021 [22]. Dalam buku yang berjudul Twitter: Social Communication in the Twitter Age oleh Dhiraj Murthy, disebutkan bahwa yang menjadi faktor penarik utama bagi Twitter merupakan sifat kemudahannya dalam mengakses dan penggunaannya [23].

Dikarenakan populernya penggunaan media sosial tersebut, Twitter men-

geluarkan layanan berupa Application Programming Interface (API) untum memudahkan pengguna dalam mengambil data *tweet* untuk dianalisa [20]. Untuk menggunakan Twitter API, pengguna hanya perlu mengajukan permintaan untuk akun *developer* pada Twitter. Jika disetujui, pengguna akan diberikan 4 kunci, yaitu *consumer key, consumer secret, access token*, dan *access secret*, yang berfungsi dalam proses autentikasi untuk pengumpulan data menggunakan API tersebut. Dengan Twitter API, pengguna dapat mengambil sejumlah *tweet* sesuai dengan kata kunci yang diinginkan dalam 1 minggu terakhir terhitung sejak pengambilan dilakukan.

2.3 *Topic Modelling*

Topic modelling merupakan suatu teknik yang menerapkan asosiasi antar kata dalam suatu teks untuk menghasilkan topik yang bersifat laten atau tersebunyi. Kemudian akan dihasilkan pengelompokan berdasarkan kata-kata yang bermunculan secara bersamaan yang berkaitan dengan topic laten yang telah diperoleh [24]. *Topic modelling* termasuk dalam *unsupervised probabilistic model*, dan dapat digunakan untuk memahami kumpulan data yang tidak terstruktur. Contoh yang paling mudah merupakan segala jenis *user-generated content*, seperti blog, *review* terhadap suatu tempat atau produk, dan postingan sosial media [25].

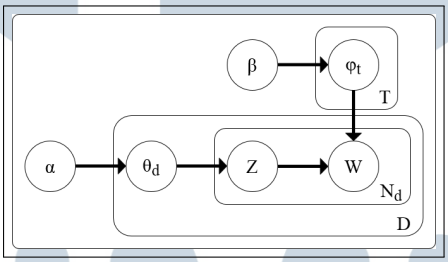
2.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation merupakan salah satu algoritma *topic modelling* yang adalah bagian dari *unsupervised algorithm*, digunakan untuk mencari hubungan semantik antar kata dalam sebuah grup dengan bantuan indikator [9]. Latent Dirichlet Allocation melihat dokumen sebagai campuran acak dari topik laten (tersebunyi), dimana setiap topik merupakan distribusi antar kata dalam dokumen. Secara garis besar, Latent Dirichlet Allocation mengasumsikan bahwa sebuah dokumen merupakan sebuah kantong yang berisi banyak kata tanpa dipengaruhi oleh sintaks maupun letaknya dalam dokumen [24]. Setiap distribusi topik yang dihasilkan berisi semua kata yang terdaftar, namun setiap kata memiliki distribusi kemungkinan yang berbeda satu dengan yang lain, tergantung dari seberapa tinggi tingkat kemunculan kata tersebut pada dokumen [26]. Menurut Jelodar [27], LDA merupakan sebuah model probabilistik generatif terhadap suatu dokumen, dan setiap topik laten yang dihasilkan oleh model LDA juga merupakan distribusi probabilistic

terhadap kata-kata dan distribusi kata terhadap topik. Proses generatif LDA dapat dijelaskan sebagai berikut [27]:

1. Pilih sebuah distribusi multinomial φ_t untuk topik t(t$\in\{1, ..., T\}$) dari sebuah distribusi Dirichlet dengan parameter β.
2. Pilih sebuah distribusi multinomial θ_d untuk dokumen d(d$\in\{1, ..., D\}$) dari sebuah distribusi Dirichlet dengan parameter α.
3. Untuk setiap kata w_n di dalam dokumen d:
   (a) Pilih sebuah topik Z dari θ_d.
   (b) Pilih sebuah kata W dari φ_t terhadap topik Z.

Probabilitas dari data yang diamati (D) kemudian dihitung dan dihasilkan sebuah *corpus*. *Corpus* tersebut akan dibandingkan dengan data yang asli untuk menghasilkan probabilitas masing-masing topik pada data.



Gambar 2.1. *Plate notation* model LDA
Sumber: [28]

Plate notation dari model LDA terdapat pada Gambar 2.1, dengan keterangan sebagai berikut:

1. θ_d merupakan distribusi multinomial terhadap topik untuk dokumen ke-d.
2. φ_t merupakan distribusi multinomial terhadap kata untuk topik t.
3. α merupakan parameter Dirichlet untuk θ.
4. β merupakan parameter Dirichlet untuk φ.
5. D merupakan jumlah dokumen.
6. N_d merupakan jumlah kata di dokumen ke-d.

7. T merupakan jumlah topik.
8. Z merupakan topik dari kata ke-n di dokumen ke-d.
9. W merupakan kata ke-N di dokumen ke-d.

2.5 *Topic Coherence*

Topic coherence merupakan metode pendekatan kualitatif yang menilai seberapa baik topik dihasilkan dari suatu teks. Metode ini berdasar pada teori dimana kata-kata dengan makna yang sama atau mirip cenderung muncul di dalam konteks yang sama. Maka dari itu, topik dianggap koheren jika didominasi oleh kata-kata yang berhubungan, dan menghasilkan nilai *coherence* yang tinggi [26]. Terdapat banyak metode *topic coherence* yang telah diusulkan pada penelitian sebelumnya, seperti pada penelitian oleh Röder dkk. [29]. Pada penelitian tersebut, diusulkan 7 metode *topic coherence* dimana metode C_v memperoleh hasil dengan tingkat korelasi tertinggi terhadap penilaian manual oleh manusia. Metode *topic coherence* C_v menghitung *co-occurence counts* (jumlah kemunculan 2 kata atau lebih secara bersamaan) dengan menggunakan perhitungan Normalized Pointwise Mutual Information (NPMI) dari setiap kata populer (kata dengan tingkat kemunculan tertinggi pada suatu dokumen, dinotasikan dengan w_i) terhadap kata populer lainnya (dinotasikan dengan w_j) [30].

$$NPMI(w_i, w_j) = \sum_{j}^{N-1} \frac{log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-logP(w_i,w_j)} \tag{2.1}$$

Perhitungan NPMI terhadap kata dengan tingkat kemunculan tertinggi dapat dilihat pada Persamaan 2.1, dengan keterangan sebagai berikut [30]:

1. $P(w_i)$ merupakan probabilitas kemunculan kata w_i pada sebuah dokumen acak.
2. $P(w_j)$ merupakan probabilitas kemunculan kata w_j pada sebuah dokumen acak.
3. $P(w_i,w_j)$ merupakan probabilitas kemunculan kata w_i dan w_j pada sebuah dokumen acak secara bersamaan.
4. N merupakan jumlah kata dengan tingkat kemunculan tertinggi yang dibandingkan.

Topik yang koheren sangatlah penting bagi *topic modelling*, dikarenakan lebih bermakna dan akan mempermudah proses identifikasi label topik [31]. Dalam penelitian ini, metode *topic coherence* C_v akan digunakan sebagai evaluasi untuk menghitung seberapa baik topik yang dihasilkan oleh model Latent Dirichlet Allocation.

2.6 Analisis Sentimen

Analisis sentimen merujuk kepada mengambil opini ataupun emosi dari suatu data, yang kemudian akan dianalisis. Data yang dianalisis tersebut dapat berupa beragam media dari teks, suara, gambar hingga video. Teknik ini sangat penting dalam menilai sentimen publik terhadap suatu produk atau kejadian [32]. Terdapat beberapa jenis metode *Sentiment Analysis*, diantaranya adalah *Lexicon-Based*, *Learning-Based (Machine Learning), Hybrid,* dan *Graph-Based* [33]. Metode *Lexicon-Based* memanfaatkan daftar kata yang disertai dengan angka polaritasnya masing-masing untuk menentukan nilai opini secara keseluruhan dari suatu teks.

2.7 *Lexicon-based Sentiment Analysis*

Metode *lexicon based* merupakan salah satu jenis Sentiment Analysis, dimana digunakan *sentiment lexicon* dalam menentukan sentimen sebuah kata atau kalimat. *Sentiment lexicon* merupakan suatu kumpulan kata (kamus) dimana masing-masing kata telah dilabel sebagai positif, netral ataupun negatif. Label positif, netral, dan negatif juga dapat direpresentasikan sebagai kisaran nilai yang mewakili seberapa kuat sentimen pada kata tersebut [34]. Metode *lexicon based* bekerja dengan menghitung nilai sentimen per kata dalam sebuah kalimat atau dokumen, dimana kemudian masing-masing nilai tersebut dijumlahkan dan menjadi nilai representasi akhir sentimen kalimat atau dokumen tersebut [35].

Sentiment lexicon yang digunakan pada penelitian ini merupakan InSet yang dibuat oleh Koto & Rahmaningtyas [36]. *Sentiment lexicon* InSet terdiri atas 3.609 *lexicon* positif dan 6.609 *lexicon* negatif dan memiliki jangkauan bobot nilai sentimen (*polarity score*) dari -5 hingga 5, dengan -5 melambangkan sangat negatif dan 5 melambangkan sangat positif. *Sentiment lexicon* ini dikembangkan dari *microblog* yang ada di Indonesia.