

# PREDICTIVE ANALYSIS OF STOCK PRICE USING CONVOLUTIONAL NEURAL NETWORK LONG SHORT TERM MEMORY METHOD

Marcelina  
Departemen Matematika, FAST  
Universitas Pelita Harapan  
Jakarta, Indonesia  
mm80011@student.uph.edu

Kie Ivanky Saputra  
Departemen Matematika, FAST  
Universitas Pelita Harapan  
Jakarta, Indonesia  
kie.saputra@uph.edu

Dion Krisnadi  
Departemen Teknik Informatika, FIK  
Universitas Pelita Harapan  
Jakarta, Indonesia  
dion.krisnadi@uph.edu

**Abstrak**—Technological advances make it possible to predict stock prices using machine learning methods. Deep learning methods that will be used are Convolutional Neural Network Long Short Term Memory and Convolutional Neural Network Bidirectional Long Short Term Memory. The purpose of this research is to determine the impact of the number of layers on the model and use evaluation metrics to determine which model is better at predicting stock prices. The models are CNN LSTM and CNN BiLSTM models featuring 50 and 7 historical data. The results showed that the addition of the number of layers has a correlation with the predicted performance of the stock's closing price. The CNN BiLSTM model featuring 50 historical data is the best model in predicting stock prices with MSE, MAE, and RMSE values respectively are 0.12481, 0.29218, and 0.3520707 in the fifth layer of ITMG stocks, followed by the CNN LSTM model featuring 50 historical data with MSE, MAE, and RMSE values respectively are 0.02965, 0.13582, and 0.171661 on the fifth layer of ADRO stock, CNN BiLSTM featuring 7 historical data with MSE, MAE, and RMSE values respectively are 0.04988, 0.20037, and 0.2232269 on the third layer of ASII stocks, and CNN LSTM featuring 7 historical data with MSE, MAE, and RMSE values respectively are 0.03705, 0.15974, and 0.1924773 on the fifth layer of ADRO shares.

**Kata Kunci**—vaksin COVID-19, analisis sentimen, *Support Vector Machine*, pemodelan topik, *Latent Dirichlet Allocation*

## I. PENDAHULUAN

Pada Desember 2019, muncul untuk pertama kalinya penyakit COVID-19 di Kota Wuhan, Tiongkok. Virus penyebab COVID-19 diperkirakan berasal dari kelelawar dan bertransmisi sangat cepat antar manusia [1]. Di Indonesia, jumlah kasus terkonfirmasi pada 10 Juni 2021 adalah 1.877.050 kasus dengan 52.162 kematian, yang menempatkan Indonesia di posisi pertama jumlah kasus COVID-19 terkonfirmasi positif tertinggi di Asia Tenggara [2].

Dalam rangka mengatasi pandemi COVID-19, berbagai penelitian terhadap vaksin telah diadakan. Lebih dari 200 vaksin COVID-19 sudah berada dalam tahap pengembangan dalam kurun waktu kurang dari setahun [3]. Namun, pemilihan vaksin yang tepat masih menjadi suatu masalah yang perlu dijawab, mengingat protokol studi yang berbeda

antar vaksin, seperti studi populasi, risiko infeksi COVID-19 selama studi, lama paparan, perbedaan definisi populasi dalam analisis, dan juga titik akhir kemanjuran vaksin. Vaksin dengan tingkat efikasi pada populasi tertentu belum tentu memiliki tingkat efikasi yang sama pada populasi dengan latar belakang COVID-19 yang berbeda [4].

Sejumlah negara di dunia telah mewajibkan vaksinasi COVID-19, salah satunya adalah Indonesia [5]. Namun, vaksinasi COVID-19 di Indonesia menuai pro dan kontra dalam masyarakat. Hasil survei penerimaan vaksin COVID-19 di Indonesia yang diadakan oleh WHO bekerja sama dengan Kementerian Kesehatan, ITAGI, dan UNICEF terhadap 112.888 responden memperlihatkan bahwa sekitar 65% responden bersedia menerima vaksin COVID-19, delapan persen menolak, dan 27% menyatakan tidak yakin dengan vaksin COVID-19. Selain itu, 16.686 responden merasa khawatir dengan keamanan dan efektifitas vaksin, tidak percaya terhadap vaksin, dan mempersoalkan kehalalan vaksin [6].

Kekhawatiran masyarakat terhadap vaksin COVID-19 dapat menghambat rencana pemerintah dalam mencapai *herd immunity*. Oleh karena itu, diperlukan pemahaman mengenai persepsi masyarakat terhadap vaksin COVID-19. Penelitian ini akan melakukan analisis persepsi masyarakat Indonesia terhadap vaksin COVID-19 menggunakan metode *text mining*. Metode *text mining* yang akan dilakukan adalah analisis sentimen dan pemodelan topik.

Analisis sentimen pada *tweets* akan menggunakan model *machine learning Support Vector Machine* (SVM). Pelabelan sentimen untuk persiapan data latih akan dilakukan secara manual oleh penulis dan dengan metode leksikon. Kemudian, akan dilakukan pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA). Hasil yang didapatkan dari analisis sentimen diharapkan dapat menunjukkan bagaimana performa model dari data latih hasil pelabelan manual dan leksikon, serta bagaimana sikap masyarakat terhadap vaksin COVID-19. Adapun hasil dari pemodelan topik diharapkan dapat memperlihatkan topik permasalahan utama masyarakat terhadap

vaksin COVID-19.

## II. METODOLOGI

### A. Text Mining

*Text mining* merupakan suatu proses untuk mendapatkan informasi dari data yang tidak terstruktur melalui identifikasi dan eksplorasi pola [7]. Tahapan dalam *text mining* mirip dengan *data mining*, yaitu meliputi *pre-processing* data, algoritma pencarian pola, dan presentasi hasil dengan visualisasi. Data yang berada dalam tahap *pre-processing* akan melewati beberapa proses, yaitu pembersihan, *case folding*, *tokenization*, normalisasi, penghapusan *stopwords*, dan *stemming* [8]. Berikut adalah penjelasan dari proses-proses tersebut.

#### 1) Pembersihan Data

Pada tahap pembersihan data, data-data yang berduplikat akan dihapus. Selain itu, karakter-karakter yang tidak diperlukan, seperti tanda baca, juga dihapus.

#### 2) Case Folding

Pada tahap *case folding*, semua huruf akan diubah menjadi huruf kecil (*lowercase*) untuk mengurangi redundansi data.

#### 3) Normalisasi

Pada tahap ini, kata singkatan, *typo*, dan bahasa gaul akan diubah menjadi kata yang baku agar dapat meningkatkan performa proses komputasi.

#### 4) Penghapusan Stopwords

Pada tahap ini, kata yang umum digunakan seperti "saya", "anda", "adalah", dan sebagainya akan dihapus. Tujuan dari penghapusan kata-kata tersebut adalah mempercepat sistem dalam mengolah data.

#### 5) Stemming

Pada tahap ini, setiap kata akan diubah ke bentuk dasarnya. Imbuhan awalan dan akhiran akan dihapus, sehingga hanya tersisa kata dasarnya.

#### 6) Tokenization

Pada tahap *tokenization*, teks akan dipisah berdasarkan kata-kata penyusunnya.

Dalam *text mining*, terdapat beberapa istilah yang sering digunakan, yaitu kata, dokumen, dan *corpus*. Kata adalah suatu satuan dari data diskrit, yang didefinisikan sebagai bagian dari kosakata. Dokumen adalah rangkaian kata-kata. *Corpus* adalah koleksi dari dokumen-dokumen [9]. *Text mining* dapat digunakan untuk klasifikasi teks, seperti analisis sentimen, dan klasterisasi teks, seperti pemodelan topik.

### B. Representasi Dokumen

1) *Model bags-of-words*: Pada model *bag-of-words*, semua kata yang digunakan dalam dokumen akan digunakan sebagai fitur, sehingga dimensi dari fitur yang ada sama dengan banyaknya jumlah kata berbeda pada semua dokumen. Kemudian, setiap dokumen dalam *corpus* akan direpresentasikan sebagai vektor yang menyimpan frekuensi munculnya setiap kata pada dokumen tersebut. Dengan menggunakan model *bag-of-words*, struktur semantik tersembunyi dari teks akan hilang karena tidak menyimpan urutan kata dalam dokumen. Model LDA akan menggunakan fitur *bags-of-words* dalam

perancangan modelnya, dengan menggunakan data *tweets* hasil *tokenization*.

2) *Term Frequency – Inverse Document Frequency (TF-IDF)*: Metode *Term Frequency – Inverse Document Frequency* (TF-IDF) adalah metode statistik yang digunakan untuk mengetahui tingkat kepentingan kata dalam dokumen [10]. Jika sebuah kata atau frasa sering ditemukan di sebuah artikel, tetapi jarang ditemukan di artikel lain, maka kata tersebut dianggap memiliki kemampuan membagi kelas dengan baik dan cocok digunakan dalam proses klasifikasi. Metode TF-IDF terdiri dari dua bagian utama, yaitu frekuensi kata (*term frequency*) dan inversi dari frekuensi dokumen (*inverse document frequency*).

Frekuensi kata adalah banyaknya sebuah kata muncul dalam dokumen. Frekuensi kata untuk kata ke- $i$  pada dokumen ke- $j$  dinyatakan dengan persamaan

$$tf_{i,j} = n_{i,j}. \quad (1)$$

Dari persamaan tersebut,  $n_{i,j}$  merupakan jumlah munculnya kata ke- $i$  ( $t_i$ ) pada dokumen ke- $j$ . Inversi dari frekuensi dokumen untuk kata ke- $i$  dinyatakan dengan persamaan

$$idf_i = \log \left( \frac{1 + \|D\|}{1 + \|\{j : t_i \in d_j\}\|} \right) + 1. \quad (2)$$

Pada persamaan (2),  $\|D\|$  adalah jumlah semua dokumen. Himpunan  $\{j : t_i \in d_j\}$  merupakan himpunan dokumen ke- $j$  ( $d_j$ ) yang memuat kata  $t_i$  sehingga  $\|\{j : t_i \in d_j\}\|$  merupakan jumlah dokumen yang memiliki kata  $t_i$ . Misalkan nilai TF-IDF dinotasikan sebagai  $W$ , bentuk persamaan TF-IDF yang paling umum [7] untuk kata  $t_i$  pada dokumen  $d_j$  adalah

$$W_{i,j} = tf_{i,j} \times idf_i. \quad (3)$$

Perancangan model SVM akan menggunakan fitur TF-IDF, dengan menggunakan data *tweets* yang telah melewati proses *stemming*. Perhitungan TF-IDF akan menggunakan *library* scikit-learn [11]. *Library* scikit-learn merupakan *library* yang sering digunakan untuk melakukan *data pre-processing*, perancangan model pembelajaran *supervised* dan *unsupervised*, serta evaluasi model. Pada *library* scikit-learn, nilai TF-IDF yang menjadi fitur adalah nilai TF-IDF yang telah dinormalisasi L2. Misalkan terdapat vektor berdimensi- $n$  yang dinyatakan sebagai

$$\vec{x} = (x_1, x_2, x_3, \dots, x_n),$$

maka persamaan normalisasi L2 untuk data  $x_i$  yang disimbolkan dengan  $x_i'$  adalah

$$x_i' = \frac{x_i}{\|\vec{x}\|_2} = \frac{x_i}{\sqrt{\sum_{i=1}^n |x_i|^2}}. \quad (4)$$

Normalisasi L2 menghitung jarak koordinat suatu vektor dari titik asal ruang vektor. Vektor yang telah dinormalisasi L2 memiliki panjang vektor sebesar 1. Hal ini dilakukan karena dalam kasus klasifikasi sentimen, panjang vektor (banyaknya kata) tidak penting, namun makna atau nilai TF-IDF kata tersebutlah yang penting.

### C. Analisis Sentimen

Analisis Sentimen adalah suatu bentuk analisis untuk mengetahui pendapat seseorang terhadap suatu entitas [12]. Data yang dianalisis adalah dokumen *corpus* dengan berbagai format. Dokumen dalam *corpus* akan diubah menjadi teks, dan dilakukan *pre-processing* dengan berbagai teknik seperti *stemming*, *tokenization*, *part of speech* (POS) *tagging*, ekstrasi entitas dan ekstrasi hubungan. Hasil dari analisis sentimen dapat digunakan untuk memantau reputasi dan saran terhadap produk mereka di media sosial sebagai acuan dalam pengambilan keputusan.

Analisis sentimen dapat dilakukan dengan pendekatan pengetahuan atau leksikon (*lexicon-based*) dan pendekatan *machine learning* [13]. Pada pendekatan berdasarkan leksikon, sentimen pada dokumen ditentukan dengan menghitung orientasi semantik dokument tersebut. Dengan menggunakan kamus yang berisikan daftar leksikon, akan dihitung nilai sentimen untuk setiap kata. Lalu nilai sentimen tersebut akan dimasukkan ke fungsi agregasi, yang akan menghitung orientasi semantik akhir dokumen tersebut. Pendekatan ini bersifat lebih fleksibel karena dapat diaplikasikan pada berbagai topik dan mudah digunakan. Namun, kelemahan pendekatan leksikon adalah sangat bergantung pada kamus leksikon yang dimiliki. Selain itu, sebuah kata terkadang juga dapat memiliki makna sentimen yang berbeda pada topik-topik tertentu.

Pendekatan *machine learning* menggunakan teknik klasifikasi, yaitu mempelajari karakteristik dokumen untuk sentimen tertentu. Dokumen akan dibagi menjadi *training data* dan *testing data*. Model analisis sentimen akan dibangun berdasarkan *training data*. Kemudian, model akan memprediksi sentimen pada *testing data*. Keunggulan pendekatan *machine learning* adalah dapat secara otomatis mempelajari fitur-fitur untuk klasifikasi. Namun, pendekatan ini sangat bergantung pada *training data* yang diberikan.

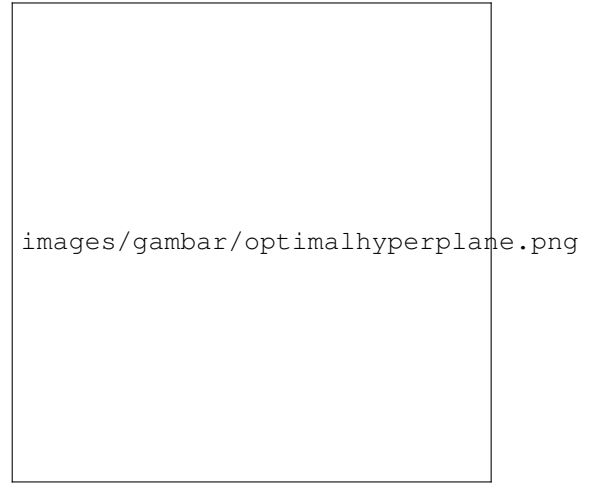
Penelitian dari Y. Choi and H Lee menunjukkan bahwa properti dari data latih dapat mempengaruhi performa klasifikasi sentimen [14]. Secara umum, pendekatan *machine learning* memberikan performa lebih baik daripada pendekatan leksikon. Namun, untuk *corpus* yang terdiri dari dokumen berukuran kecil, performa pendekatan leksikon hampir mendekati performa pendekatan *machine learning*. Untuk pendekatan *machine learning*, kriteria data latih yang baik adalah berukuran setidaknya 3% data dari keseluruhan data untuk setiap sentimen, memiliki 100-200 kata untuk setiap dokumen, dan tingkat subjektivitas dokumen sebesar 0,5-1,0. Untuk data yang memiliki tingkat subjektivitas yang tinggi seperti ulasan pembeli dan *blog*, *machine learning* dapat bekerja dengan baik dengan data berukuran kecil, dengan ukuran data latih minimum adalah 3% dari keseluruhan data.

Pada penelitian ini, data *tweets* akan dibagi dengan proporsi 80% dan 20%, dan setiap data akan diberi label sentimen secara manual maupun dengan pendekatan leksikon. Pelabelan dengan metode leksikon menggunakan leksikon dari InSet (*Indonesia Sentiment Lexicon*) [15] dan leksikon Sen-

tiStrengthID<sup>1</sup>. Lalu, penulis menyesuaikan nilai sentimen pada beberapa kata dengan konteks data *tweets* vaksin COVID-19. Penulis juga menambahkan beberapa kata yang tidak ada di leksikon InSet maupun SentiStrengthID, untuk memperkaya leksikon yang digunakan. Model SVM akan dilatih dan diuji dengan berbagai kombinasi data, untuk menentukan karakteristik *training data* yang baik dalam membuat model analisis sentimen.

### D. Support Vector Machine

*Support Vector Machine* (SVM) adalah metode klasifikasi yang dikembangkan pada tahun 1990-an dan telah banyak digunakan hingga saat ini. [16]. SVM melakukan klasifikasi dengan menentukan *hyperplane* atau garis pembagi yang dapat memisahkan suatu kelas dengan kelas yang lain. Sebuah vektor



Gambar 1: *Hyperplane* pemisah optimal  
Sumber : *The Nature of Statistical Learning Theory* [17]

dikatakan terpisahkan oleh *hyperplane* optimal jika vektor-vektor tersebut dapat dipisahkan tanpa galat dan jarak antara dua vektor terdekat yang berasal dari kelas yang berbeda dengan *hyperplane* adalah maksimum [17]. Misalkan terdapat  $N$  data latih yang terdiri dari fitur  $\vec{x}$  dan kelas  $y$ , fitur tersebut berdimensi ruang  $m$ . Nilai kelas pada data tersebut terdiri dari  $+1$  atau  $-1$ . Data latih tersebut dapat dirumuskan sebagai

$$\{\vec{x}_1, y_1\}, \dots, \{\vec{x}_N, y_N\}, \vec{x}_i \in R^m, y \in \{+1, -1\},$$

yang dapat dipisahkan oleh *hyperplane* yang memenuhi persamaan

$$(\vec{w} \cdot \vec{x}) - b = 0, \quad (5)$$

dengan  $\vec{w}$  adalah vektor yang tegak lurus terhadap *hyperplane* dan  $b$  adalah skalar bias. *Hyperplane* akan membagi data menjadi dua kelas. Perlu diketahui bahwa

$$(\vec{w} \cdot \vec{x}) - b \geq 1 \quad \text{jika } y_i = +1, \quad (6)$$

dan

$$(\vec{w} \cdot \vec{x}) - b \leq -1 \quad \text{jika } y_i = -1. \quad (7)$$

<sup>1</sup><https://github.com/agusmakmun/SentiStrengthID>

Pencarian *hyperplane* optimal akan diselesaikan dengan menggunakan fungsi Lagrange dan *quadratic programming*. Hasil dari optimisasi tersebut menunjukkan bahwa setiap titik yang merupakan *support vector* akan memiliki  $\alpha_i > 0$  dan setiap titik yang bukan merupakan *support vector* akan memiliki  $\alpha_i = 0$ , dengan  $\alpha_i$  adalah *Lagrange multiplier*. Dengan mengetahui fakta tersebut maka persamaan (5) menjadi

$$\sum_{i=1}^N \alpha_i (\vec{x} \cdot \vec{x}_i) + b_0 = 0. \quad (8)$$

Untuk kasus di mana data-data tidak dapat dipisahkan secara linier, maka dapat menggunakan konsep *soft margin*. Pada konsep ini, akan terdapat variabel tak negatif  $\xi_i$ , yang merupakan besar galat pada klasifikasi, dan konstanta  $C$ , yang menentukan seberapa besar pelanggaran data yang dapat dilakukan terhadap *hyperplane*. Permasalahan optimisasi yang baru adalah

$$\Phi(\vec{w}) = \frac{1}{2}(\vec{w} \cdot \vec{w}) + C \sum_{i=1}^N \xi_i, \quad (9)$$

dengan batasan

$$\xi_i \geq 0, \quad (10)$$

dan

$$y_i[(\vec{w} \cdot \vec{x}_i) - b] \geq 1 - \xi_i, \quad i = 1, \dots, N. \quad (11)$$

Hasil kali dalam antara vektor  $\vec{x}_i$  dengan  $\vec{x}_j$  pada persamaan (8) dapat digeneralisasikan dengan bentuk

$$K(\vec{x}_i, \vec{x}_j), \quad (12)$$

dengan  $K$  adalah fungsi yang disebut kernel. Pada pemetaan linear, fungsi kernel linear adalah

$$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j. \quad (13)$$

Ketika penggunaan SVM linear tidak memberikan performa yang baik, maka perlu dilakukan pemetaan  $\vec{x}$  dengan menggunakan fungsi pemetaan tak linier  $\Phi(\vec{x})$  ke dimensi yang lebih tinggi untuk mendapatkan *hyperplane* optimal. Pemetaan taklinier ini dinyatakan dengan rumus

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j). \quad (14)$$

Fungsi kernel yang sering digunakan adalah *radial basis function* (RBF), yang didefinisikan dengan

$$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{\sigma^2}\right), \quad (15)$$

dan fungsi polinomial, yang didefinisikan dengan

$$K(\vec{x}_i, \vec{x}_j) = \left((\vec{x}_i^T \cdot \vec{x}_j) + 1\right)^d. \quad (16)$$

Model SVM dapat digunakan untuk melakukan klasifikasi data dengan lebih dari dua kelas dengan menggunakan pendekatan *one-versus-one* dan *one-versus-all*. Misalkan terdapat  $K > 2$  kelas, dengan pendekatan *one-versus-one*, akan dikonstruksi  $\binom{K}{2}$  SVM. Pada setiap SVM yang dibangun, akan dilakukan perbandingan sepasang kelas, klasifikasi terakhir

dilakukan dengan menempatkan data observasi ke kelas yang paling sering didapatkan. Pada pendekatan *one-versus-all* akan dilakukan dengan membuat  $K - 1$  model SVM. Pada setiap SVM yang dibuat, akan dilakukan perbandingan antara salah satu kelas dengan  $K - 1$  sisa kelas yang ada. Penelitian ini akan menggunakan menggunakan pendekatan *one-versus-one* dalam merancang model SVM. Model tersebut akan dirancang dengan *library* scikit-learn [11].

#### E. Latent Dirichlet Allocation

*Latent Dirichlet Allocation* (LDA) adalah jenis pemodelan topik yang paling sederhana [18]. Pemodelan topik adalah sebuah model statistik yang digunakan untuk menganalisis kata-kata pada teks asli dan mencari tema dalam teks tersebut. LDA adalah model probabilistik generatif dari *corpus*. LDA menggunakan asumsi "bag of words", yaitu urutan kata pada dokumen tidak penting. Urutan dokumen dalam *corpus* juga bukan permasalahan utama dalam LDA. Selain itu, LDA mengasumsikan bahwa setiap dokumen terdiri dari topik-topik tersembunyi dengan proporsi tertentu. Setiap topik merupakan distribusi dari kata-kata [9]. LDA mengasumsikan jumlah topik sudah diketahui.

Dalam pemodelan probabilistik generatif, kita menganggap data muncul dari sebuah proses generatif yang memiliki variabel tersembunyi [18]. Proses generatif tersebut mendefinisikan sebuah distribusi probabilitas gabungan yang terdiri dari variabel yang diamati dan variabel acak tersembunyi. Analisis data dilakukann dengan menggunakan distribusi gabungan untuk menghitung *conditional distribution* dari variabel tersembunyi dengan diberikan variabel yang diamati. *Conditional distribution* ini juga dikenal sebagai *posterior distribution*.

LDA menganggap variabel yang diamati adalah kata-kata pada dokumen, variabel yang tidak diketahui adalah struktur topik, dan proses generatifnya seperti yang telah dijelaskan sebelumnya. Secara formal, proses generatif LDA bersesuaian dengan distribusi probabilitas gabungan dari variabel tersembunyi dan variabel yang sudah diketahui, dengan persamaan

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \cdot \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) \cdot p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right). \quad (17)$$

Notasi matematika pada persamaan (17) memiliki makna sebagai berikut.

- $K$  adalah jumlah topik.
- $D$  adalah jumlah dokumen pada *corpus*.
- $N$  adalah jumlah kata pada setiap dokumen.
- $\beta_{1:K}$  adalah topik-topik. Setiap  $\beta_k$  terdiri dari distribusi kosakata, yaitu probabilitas suatu kata muncul di topik  $k$ .
- $\theta_d$  adalah distribusi topik pada dokumen ke- $d$ .
- $z_{d,n}$  adalah pemberian topik untuk kata ke- $n$  pada dokumen  $d$ .

- $w_{d,n}$  adalah kata ke- $n$  yang diamati pada dokumen  $d$  dan merupakan elemen dari kosa-kata yang tetap.

Persamaan *posterior* dirumuskan sebagai

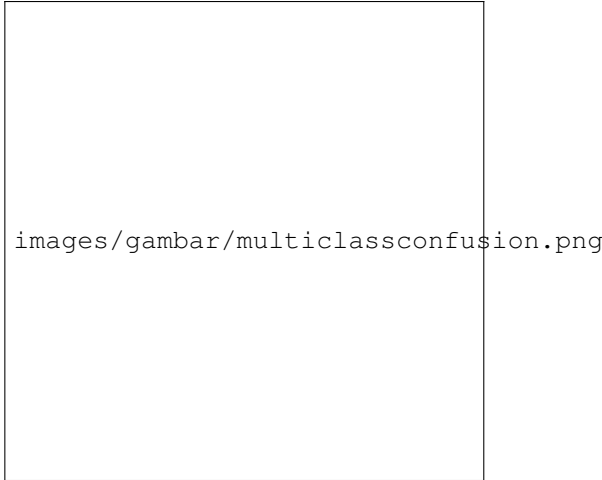
$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \quad (18)$$

Pembilang pada persamaan *posterior* adalah distribusi gabungan dari semua variabel acak. Penyebut pada persamaan tersebut adalah *marginal probability* dari observasi, yaitu peluang melihat *corpus* yang diamati dari model topik apapun. Model LDA akan dibangun dengan *library* Gensim [19]. *Library* Gensim merupakan *library* yang digunakan untuk melakukan pemodelan topik secara *unsupervised* dan tugas *text mining* lainnya.

#### F. Evaluasi

1) *Evaluasi Model Klasifikasi*: Dalam melakukan evaluasi model klasifikasi, terdapat beberapa metode pengukuran umum yang dilakukan dengan menggunakan *confusion matrix*. *Confusion matrix* adalah ilustrasi yang menggambarkan perbandingan antara hasil prediksi dan nilai asli. Berikut adalah istilah-istilah yang digunakan pada pengukuran tersebut [20].

- 1) *True Positive* (TP) = jumlah data yang merupakan bagian suatu kelas dan diprediksi dengan benar.
- 2) *False Negative* (FN) = jumlah data yang merupakan bagian suatu kelas, tapi tidak diprediksi dengan benar.
- 3) *False Positive* (FP) = jumlah data yang bukan merupakan dari suatu kelas, tapi diprediksi sebagai bagian dari kelas tersebut.
- 4) *True Negative* (TN) = jumlah data yang bukan merupakan bagian suatu kelas dan diprediksi dengan benar.



Gambar 2: *Confusion matrix* untuk data *multiclass*

Indikator pengukuran yang digunakan adalah *recall* atau sensitivitas, *precision*, dan *F1-score* dari suatu model. *Recall* atau sensitivitas mengukur seberapa baik model memprediksi sesuai dengan kelas aslinya. *Recall* dinyatakan dengan persamaan

$$R_i = \frac{TP_i}{TP_i + FN_i}. \quad (19)$$

*Precision* mengukur proporsi data yang diprediksi pada kelasnya dan benar-benar merupakan bagian dari kelas tersebut. *Precision* dirumuskan sebagai

$$P_i = \frac{TP_i}{TP_i + FP_i}. \quad (20)$$

*F1-score* adalah rata-rata harmonik dari *precision* dan *recall*. *F1-score* untuk setiap kelas ke- $i$  dinyatakan dengan persamaan

$$F1 - score_i = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i}. \quad (21)$$

Nilai rata-rata *F1-score* pada suatu model dapat diperoleh melalui perhitungan *unweighted mean* atau *weighted mean* dari *F1-score* pada setiap kelas. Dengan *unweighted mean*, nilai *F1-score* pada setiap kelas akan dijumlahkan, kemudian dibagi dengan banyaknya kelas. Pada perhitungan *weighted mean*, proporsi banyaknya data asli untuk setiap kelas juga turut diperhitungkan dalam menghitung rata-rata. *F1-score* sering digunakan sebagai metrik evaluasi untuk data yang memiliki distribusi kelas tidak seimbang, karena *F1-score* memberikan suatu nilai tunggal yang menggabungkan *precision* dan *recall* dengan bobot yang sama. Hasil dari model SVM akan diukur dengan menggunakan *F1-score* dengan perhitungan *unweighted mean*.

2) *Evaluasi Pemodelan Topik*: Dalam melakukan pemodelan topik, masalah yang sering dihadapi adalah menentukan jumlah topik dari *corpus*. Perhitungan koherensi topik yang disebut  $C_V$  dinilai memberikan performa yang cukup baik dalam mengevaluasi topik [21]. Perhitungan nilai  $C_V$  terbagi menjadi empat tahap. Berikut adalah tahapan perhitungan  $C_V$ .

- 1) *Segmentasi pasangan kata*

Pada setiap topik, akan diambil  $N$  kata teratas yang diurutkan berdasarkan nilai peluang kata tersebut secara menurun. Misalkan  $W$  adalah himpunan  $N$  kata yang memiliki peluang tertinggi pada setiap topik, maka  $W$  dirumuskan dengan

$$W = \{W_1, \dots, W_N\}. \quad (22)$$

Kemudian, terdapat  $S_i$  yang merupakan pasangan segmentasi untuk setiap kata  $W' \in W$  yang dipasangkan ke kata lain  $W^* \in W$ . Himpunan  $S$  adalah himpunan pasangan yang didefinisikan dengan

$$S = \{(W', W^*) \mid W' = \{w_i\}; w_i \in W; W^* = W\}. \quad (23)$$

- 2) *Peluang suatu kata atau setiap pasangan kata*

Peluang suatu kata ( $p(w_i)$ ) atau peluang gabungan dari dua kata ( $p(w_i, w_j)$ ) dapat diestimasi dengan menghitung *boolean documentation*, yaitu jumlah dokumen yang memuat kata ( $w_i$ ) atau ( $w_i, w_j$ ) dibagi dengan jumlah total dokumen. Perhitungan *boolean documentation* pada  $C_V$  menggunakan teknik *sliding window*.

- 3) *Perhitungan ukuran konfirmasi*

Ukuran konfirmasi merupakan perhitungan yang menentukan seberapa kuat suatu kata mendukung kata lain. Untuk setiap  $S_i = (W', W^*)$ , nilai ukuran konfirmasi  $\phi$  akan menghitung seberapa kuat  $W^*$  mendukung kata

$W'$  berdasarkan kesamaan kata  $W'$  dan  $W^*$  terhadap semua kata  $W$ . Perhitungan ini merupakan perhitungan secara tidak langsung. Himpunan  $W'$  dan  $W^*$  akan direpresentasikan secara berurutan sebagai vektor  $\vec{v}(W')$  dan vektor  $\vec{v}(W^*)$ , untuk mengetahui dukungan semantik kata-kata dalam  $W$ . Persamaan untuk representasi vektor  $\vec{v}(W')$  dirumuskan sebagai

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} \text{NPMI}(w_i, w_j)^\gamma \right\}_{j=1, \dots, |W|}, \quad (24)$$

dan persamaan representasi vektor  $\vec{v}(W^*)$  dirumuskan sebagai

$$\vec{v}(W^*) = \left\{ \sum_{w_i \in W^*} \text{NPMI}(w_i, w_j)^\gamma \right\}_{j=1, \dots, |W|}. \quad (25)$$

Nilai hubungan antara kata  $w_i$  dan  $w_j$  dihitung dengan *Normalized Pointwise Mutual Information* (NPMI), yang dinyatakan dengan rumus

$$\text{NPMI}(w_i, w_j)^\gamma = \left( \frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} \right)^\gamma. \quad (26)$$

Pada persamaan (26), variabel  $\epsilon$  adalah nilai yang ditambahkan untuk mencegah logaritma nol, dan variabel  $\gamma$  untuk memberi beban kepada nilai NPMI yang lebih tinggi. Nilai konfirmasi  $\phi$  pada pasangan  $S_i$  didapatkan dengan menghitung *cosine vector similarity* (ukuran kemiripan vektor) dengan menggunakan persamaan

$$\phi_{S_i(\vec{u}, \vec{w})} = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\| \cdot \|\vec{w}\|}, \quad (27)$$

dengan  $\vec{u} = \vec{v}(W')$  dan  $\vec{w} = \vec{v}(W^*)$

#### 4) Agregasi dari ukuran konfirmasi

Nilai akhir koherensi didapatkan dengan menghitung rata-rata aritmetika dari semua nilai ukuran konfirmasi  $\phi$ , yang dinyatakan sebagai

$$CV = \frac{1}{|S|} \sum_{i=1}^{|S|} \phi_i. \quad (28)$$

### III. ANALISIS DAN PEMBAHASAN

#### A. Data Tweets

Data yang dikumpulkan adalah *tweets* mengenai vaksin COVID-19 berbahasa Indonesia pada 1 Mei 2021 – 30 Juni 2021. Kata kunci yang digunakan adalah "vaksin covid", "vaksin corona", "vaksin sinovac", "vaksin astrazeneca", atau "vaksin sinopharm". Jumlah *tweets* yang dikumpulkan adalah 12.776 *tweets*. Namun, *tweets* berbahasa Malaysia juga ditemukan dalam data yang telah diperoleh. Penghapusan *tweets* berbahasa Malaysia menyisakan 8.689 *tweets*. Kemudian, *tweets* yang berduplikat juga dihapus, menyisakan 8.643 *tweets*. Gambar 3 menunjukkan sisa data yang dikumpulkan, yaitu *tweets* berbahasa Indonesia.

Setelah mengumpulkan dan menghapus data yang tidak relevan, data tersebut perlu melalui *pre-processing*. Tahapan

*pre-processing* ini mencakup pembersihan (penghapusan tanda baca, emoji, *hashtag*, *username*, URL, serta pengubahan huruf besar menjadi huruf kecil), pengubahan kata *slang* menjadi bahasa baku, penghapusan *stopwords*, *stemming*, dan pengubahan bentuk teks menjadi token. Data yang telah melalui *pre-preprocessing* kemudian akan dibagi menjadi dua bagian dengan proporsi 80% dan 20%. Untuk kemudahan penyebutan, data dengan proporsi 80% akan disebut sebagai data A, dan data dengan proporsi 20% akan disebut dengan data B.

Selanjutnya, setiap data akan diberi label sentimen, baik dengan manual maupun leksikon. Pada saat pelabelan sentimen dengan manual, penulis menemukan bahwa terdapat sejumlah *tweets* yang tidak relevan, yaitu *tweets* mengenai kuis berhadiah oleh akun @arusbaik\_id. *Tweets* tersebut pun dihapus dari data A dan data B. Jumlah data yang tersisa adalah 8.473 *tweets*, dengan 6.782 *tweets* pada data A, dan 1.691 *tweets* pada data B. Pelabelan leksikon dapat memberikan hasil yang sama dengan pelabelan manual, dengan persentase sebesar 47% pada data A dan 53% pada data B. Gambar 4 menunjukkan bagaimana data melalui proses *pre-processing* dan hasil label sentimennya. Tabel I menunjukkan bahwa kelas sentimen pelabelan manual pada data A dan data B tidak terdistribusi secara merata.

TABLE I: Jumlah sentimen pada data

	Data A		Data B	
Sentimen	Manual	Leksikon	Manual	Leksikon
Positif	984	1911	468	452
Netral	3788	1699	559	394
Negatif	2010	3172	664	845

#### B. Analisis Sentimen

Fitur yang digunakan adalah nilai TF-IDF yang telah dinormalisasi L2 pada setiap kata dalam dokumen. Data A terdiri dari 6782 dokumen dan memiliki 111.878 fitur, sedangkan data B terdiri dari 1691 dokumen dan memiliki 33.820 fitur.

Setelah membuat fitur dari data A dan data B, maka akan terdapat tiga model SVM yang dibuat, yaitu model dengan data latih berupa data A dengan pelabelan leksikon, data B dengan pelabelan manual, dan data A dengan pelabelan manual. Ketiga model tersebut akan diuji dengan data uji yang berbeda-beda, menghasilkan empat kombinasi pengujian. Hasil dari pengujian tercantum pada Tabel II.

TABLE II: Hasil pengujian model SVM prediksi sentimen

Pengujian	Data Latih	Data Uji	F1-score
1	Data A (leksikon)	Data B (manual)	0,4945
2	Data B (manual)	Data A (leksikon)	0,4775
3	Data A (manual)	Data B (leksikon)	0,4704
4	Data A (manual)	Data B (manual)	0,5805

Pengujian 1 dengan pengujian 2 dan pengujian 1 dengan pengujian 3 tidak dapat dibandingkan. Dari hasil pengujian 1 dan pengujian 4, terlihat bahwa selisih dari F1-score yang didapat sebesar 0,0860 atau 8,6%. Selisih yang cukup signifikan ini menunjukkan bahwa model dengan data latih dari pelabelan leksikon tidak dapat memberikan keakuratan hasil prediksi sebaik model dengan data latih pelabelan manual.



images/gambar/sampletweet.png

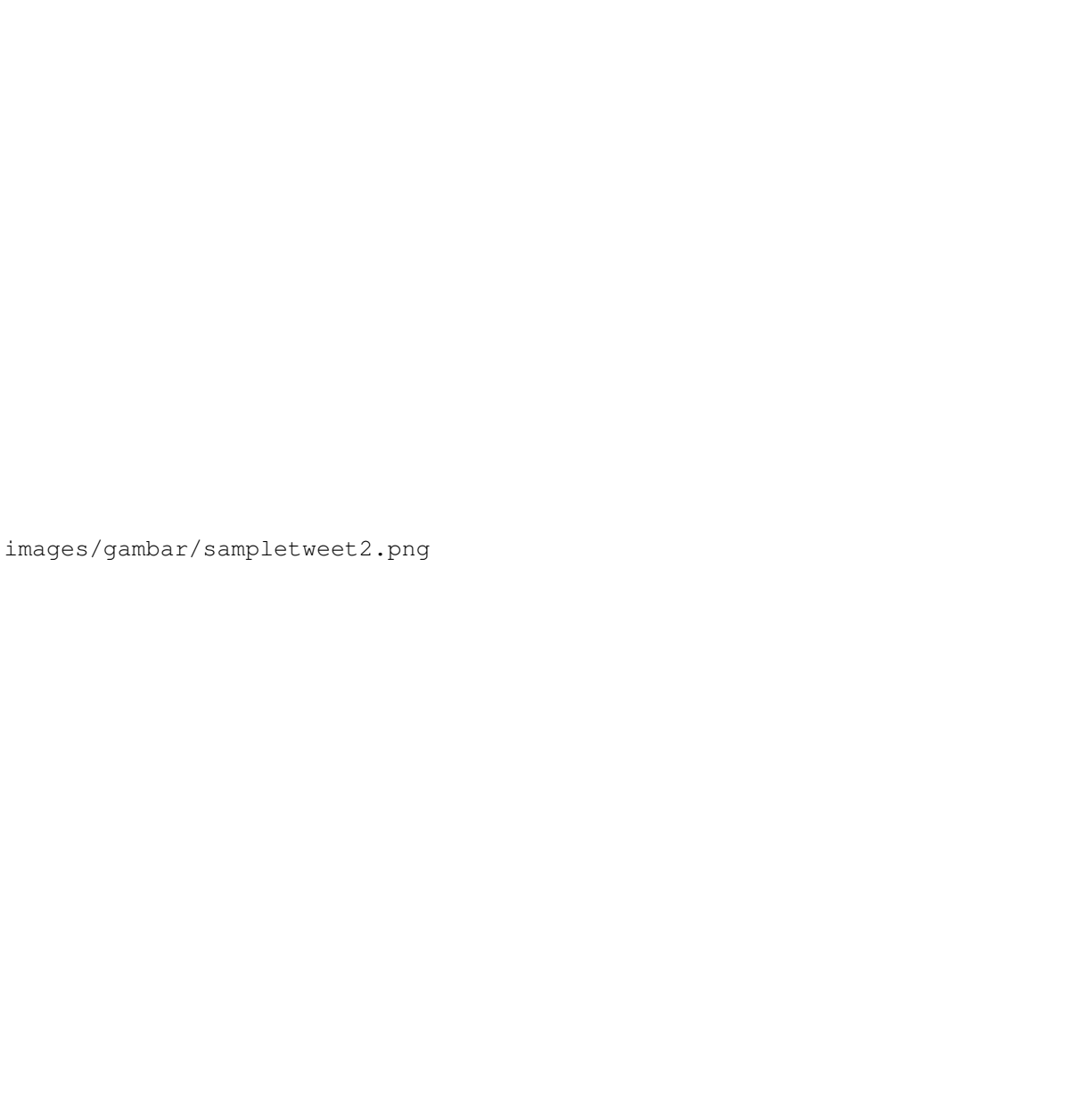
Gambar 3: Contoh data yang dikumpulkan

Dari hasil pengujian 2 dan 3, terlihat bahwa nilai *F1-score* pengujian 2 tidak jauh berbeda dengan nilai *F1-score* pengujian 3. Perbedaan nilai sebesar 0,7% ini menunjukkan bahwa model dengan jumlah latih yang kecil dapat memprediksi sama baiknya dengan model dengan jumlah data latih yang besar. Temuan ini sesuai dengan hasil dari penelitian sebelumnya yang menyimpulkan bahwa untuk dokumen *blog*, ukuran data latih yang kecil tetap dapat memberikan performa baik pada model *machine learning* [14].

Hasil dari pengujian 2 lebih rendah 10,3% dibandingkan pengujian 4. Hal ini dikarenakan pengujian 4 menggunakan data latih yang lebih banyak dibandingkan pengujian 2 dan data uji yang digunakan merupakan hasil pelabelan yang benar (metode manual).

Pada pengujian 3 dan 4, terlihat bahwa pengujian 4 memberikan hasil *F1-score* lebih tinggi daripada pengujian 3. Selisih nilai *F1-score* mencapai 11%. Hal ini disebabkan oleh penggunaan data uji yang kurang tepat pada pengujian 3, yaitu menggunakan metode leksikon. Dari hasil pengujian 1, 2, 3, dan 4, terlihat bahwa model 4 merupakan model terbaik, yaitu model yang menggunakan 80% dari data yang dikumpulkan sebagai data latih, 20% dari data yang dikumpulkan sebagai data uji, dan menggunakan teknik pelabelan manual. Model ini akan digunakan untuk memprediksi sentimen masyarakat terhadap vaksin COVID-19 pada data *tweets* dari Mei hingga Juni 2021.

Selama bulan Mei hingga Juni 2021, mayoritas *tweets* yang dibuat oleh masyarakat di *Twitter* bersentimen netral, diikuti



images/gambar/sampletweet2.png

Gambar 4: Contoh hasil *pre-processing* dan label sentimen

oleh sentimen negatif di urutan kedua, dan sentimen positif di urutan terakhir. Hal ini dapat dilihat pada Gambar 5 dan Gambar 6 yang memperlihatkan perbedaan jumlah sentimen seluruh data Mei-Juni 2021, baik *data training* maupun *data testing*, antara hasil pelabelan manual dengan model. Gambar 5 menunjukkan bahwa hasil pelabelan manual dengan prediksi menggunakan model terbaik dari pengujian 4 tidak jauh berbeda. Hal ini dikarenakan 80% dari data *tweets* hasil pelabelan manual merupakan data latih dari model. Gambar 6 menunjukkan bahwa jumlah *tweets* untuk setiap sentimen meningkat seiring berjalannya waktu. Terlihat bahwa hasil prediksi memiliki pola yang mirip dengan pelabelan manual. Hasil prediksi model memiliki jumlah *tweets* bersentimen netral lebih banyak. Peningkatan jumlah *tweets* mengenai

vaksin COVID-19 mengikuti tren jumlah kasus terkonfirmasi positif dan kasus kematian di Indonesia.

Melalui Gambar 8, terlihat bahwa *tweets* bersentimen netral didominasi oleh kata "efek", "informasi", "dan" "dosis". Masyarakat bertanya-tanya mengenai informasi seputar vaksin COVID-19 seperti efek samping, lokasi vaksinasi, dan dosis vaksin. Pada *tweets* bersentimen negatif, kata "habis" dan "efek" memiliki frekuensi penyebutan yang tinggi. Kata "habis" dalam *tweets* bermakna "setelah melakukan sesuatu". Lalu, kata-kata yang mendominasi lainnya merupakan kata-kata yang berkaitan dengan efek pasca vaksinasi, seperti "demam" dan "pegal". Tak hanya itu, kata "takut" juga memiliki frekuensi besar dalam *tweets*. Rangkaian kata-kata yang mendominasi ini menunjukkan bahwa *tweets* negatif berisikan pen-





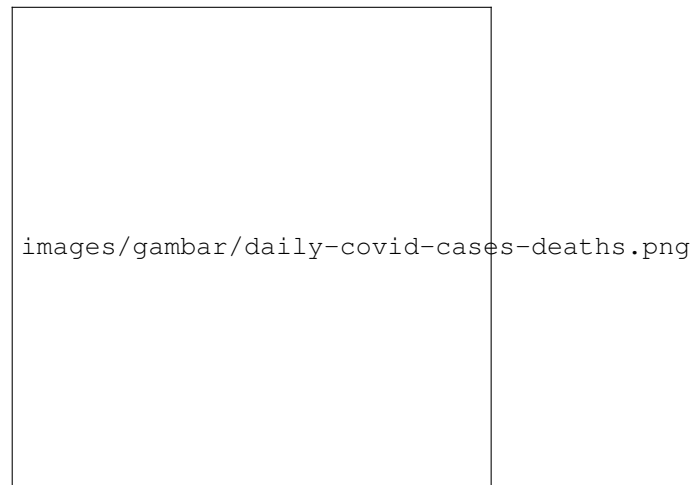
Gambar 5: Perbandingan jumlah sentimen *tweets* Mei hingga Juni 2021 hasil pelabelan manual dan prediksi model (mencakup data latih dan data uji)



Gambar 6: Perbandingan jumlah sentimen *tweets* Mei hingga Juni 2021 hasil pelabelan manual dan prediksi model per hari (mencakup data latih dan data uji)

galaman masyarakat setelah mendapatkan vaksinasi maupun rasa takut akan kegiatan vaksinasi itu sendiri.

*Tweets* bersentimen positif didominasi kata "alhamdulillah", "efek", "semoga", dan "aman". Masyarakat menunjukkan rasa syukur setelah vaksin COVID-19, terutama ketika efek samping yang diberikan tidak parah atau mengganggu kesehatan mereka. Masyarakat juga kerap menunjukkan harapan agar vaksin tersebut tidak memberikan efek samping yang berlebihan dan dapat efektif mencegah penularan virus COVID-19.



Gambar 7: Total kasus terkonfirmasi positif dan kasus kematian akibat COVID-19  
Sumber : *Coronavirus Pandemic (COVID-19)* [22]

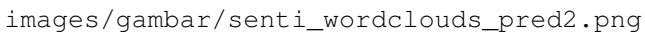
### C. Pemodelan Topik

Pemodelan topik dilakukan dengan seluruh data *tweets* yang telah dikumpulkan dan telah melewati tahap *pre-processing* hingga *tokenization*. Kemudian, data tersebut akan diubah ke dalam model *bag-of-words*. Model *bag-of-words* akan menjadi fitur yang digunakan dalam perancangan model LDA. Pada fitur ini, frekuensi kata dalam setiap dokumen akan menjadi bobot kepentingan pada setiap kata.

Untuk mendapatkan model dengan jumlah topik terbaik, maka dirancang 16 model LDA, dengan topik berjumlah 5 hingga 20. Model terbaik dipilih berdasarkan nilai *topic coherence*  $C_V$  pada setiap model. Gambar 9 menunjukkan bahwa model dengan jumlah topik 11 adalah model terbaik, dengan nilai  $C_V$  sebesar 0,44.

Tabel III memperlihatkan interpretasi dari topik yang dihasilkan model LDA. Dari 11 topik, terdapat 1 topik yang sulit diinterpretasi. Pada 10 topik yang tersisa, terlihat bahwa topik permasalahan mengenai vaksin COVID-19 di *Twitter* cukup beragam, mulai dari persetujuan penggunaan vaksin oleh WHO, uji klinis vaksin, informasi pendaftaran vaksin, hingga efek samping yang ditimbulkan oleh vaksin COVID-19. Topik-topik tersebut dapat menjadi rangkuman pendapat masyarakat terhadap vaksin COVID-19.

Sebagai vaksin yang melewati proses pembuatan yang sangat cepat (kurang dari setahun), masyarakat mengkhawatirkan efikasi dan keamanan vaksin COVID-19. Pemberian izin penggunaan darurat (*emergency use of listing*) oleh WHO merupakan salah aspek yang meningkatkan kepercayaan masyarakat akan keamanan dan efektifitas vaksin. Namun pada bulan Mei 2021, vaksin Sinovac adalah satu-satunya vaksin COVID-19 yang digunakan di Indonesia dan belum mendapatkan izin penggunaan darurat dari WHO. Hal ini juga menyebabkan penolakan Arab Saudi terhadap jemaat umrah yang menggunakan vaksin Sinovac. Masyarakat kecewa karena terancam ditolak untuk menunaikan ibadah haji di negara tersebut. Masyarakat juga kerap bertanya-tanya mengenai



Gambar 8: *WordCloud* sentimen *tweets* terhadap vaksin COVID-19 pada Mei sampai Juni 2021 berdasarkan prediksi model

hasil uji klinis vaksin COVID-19. seperti tingkat efektivitas masing-masing vaksin.

Berbagai hoaks mengenai vaksin COVID-19 juga tersebar di masyarakat. Dari segi agama, terdapat berita bahwa vaksin COVID-19 mengandung babi, sehingga haram untuk digunakan bagi pemeluk agama Islam. Dari segi keamanan privasi, beredar hoaks bahwa vaksin Sinovac memiliki *chip* didalamnya. Selain itu, pada pertengahan Mei 2021 terdapat hoaks bahwa vaksin AstraZeneca menyebabkan kematian. Vaksin AstraZeneca *batch* CTMAV547 pun sempat diberhentikan sementara oleh BPOM guna melakukan uji toksisitas dan sterilitas.

Informasi mengenai lokasi vaksinasi, tata cara pendaftaran vaksin, hingga syarat sebelum melakukan vaksinasi merupakan

topik yang sangat ramai dibicarakan oleh masyarakat. Banyak masyarakat yang bertanya-tanya mengenai lokasi vaksinasi di daerah mereka. Tampaknya, informasi mengenai lokasi vaksinasi belum tersebar secara menyeluruh dalam masyarakat.

Selain membahas mengenai vaksin Sinovac dan AstraZeneca dari program vaksinasi pemerintah, masyarakat juga kerap membandingkan vaksin tersebut dengan vaksin Sinopharm yang digunakan dalam program Vaksin Gotong Royong mulai pertengahan Mei 2021. Efek samping dari vaksin merupakan hal yang sangat dipertanyakan oleh masyarakat. Masyarakat juga kerap menceritakan pengalamannya setelah mendapatkan vaksin COVID-19 di *Twitter*. *Tweets* yang beredar juga kerap kali diikuti oleh harapan agar vaksin COVID-19 dapat memberikan efek yang baik kepada

TABLE III: Interpretasi hasil pemodelan topik LDA

Topik	Tema	Kata-Kata yang Mewakikan
1	Tidak dapat diinterpretasi	"mudah", "depok", "lancar", "ujung", "kategori", "ih", "sasar", "diagnosis", "johnson", "jasa", "nyuntik", "abal", "dada", "barang", "paksin"
2	Persetujuan penggunaan vaksin oleh WHO dan penolakan vaksin Sinovac oleh Arab Saudi	"vaksin", "sinovac", "who", "juni", "guna", "sinopharm", "saudi", "indonesia", "tanggal", "arab", "darurat", "efikasi", "astrazeneca", "tuju", "negara"
3	Hoaks mengenai vaksin COVID-19 dan jual beli vaksin ilegal	"temu", "bayi", "kontra", "vaksin", "desa", "covid", "halal", "cacar", "beli", "adil", "tanda", "kutu", "jual", "abang", "judul"
4	Prasyarat sebelum vaksin	"emg", "alam", "cek", "normal", "telah", "tensi", "booster", "kmren", "us", "iyaa", "beres", "penasaran", "darah", "smpe", "lebih"
5	Informasi pendaftaran vaksin	"vaksin", "covid", "tidak", "informasi", "sinovac", "iya", "terima", "usia", "tahun", "anak", "daftar", "kasih", "astrazeneca", "negara", "sehat"
6	Pengalaman pasca vaksin	"vaksin", "demam", "tidak", "pegal", "habis", "efek", "astrazeneca", "sinovac", "jam", "ngantuk", "suntik", "badan", "lapar", "sakit", "pusing"
7	Program Vaksinasi Gotong Royong	"vaksin", "dosis", "vaksinasi", "covid", "sinopharm", "juta", "indonesia", "program", "terima", "gotong", "royong", "sinovac", "darah", "astrazeneca", "perintah"
8	Uji klinis vaksin COVID-19	"vaksin", "sinovac", "pfizer", "varian", "covid", "efektif", "tahap", "uji", "cegah", "corona", "moderna", "sehat", "klinis", "was", "astrazeneca"
9	Pemberhentian penggunaan vaksin AstraZeneca	"vaksin", "covid", "mati", "aman", "astrazeneca", "hasil", "virus", "sebab", "tidak", "batch", "positif", "henti", "bentuk", "isi", "antibodi"
10	Efek samping vaksin COVID-19	"vaksin", "tidak", "sinovac", "efek", "iya", "covid", "astrazeneca", "sih", "samping", "suntik", "habis", "takut", "tertawa", "emang", "banget"
11	Konsultasi dengan dokter	"vaksin", "sinovac", "astrazeneca", "tidak", "covid", "iya", "pakai", "teman", "dok", "kemarin", "boleh", "gejala", "sakit", "dokter", "habis"

Gambar 9: Nilai  $C_V$  pada model LDA

masyarakat.

#### IV. KESIMPULAN

Dari penelitian yang dilakukan, didapatkan bahwa hasil dari pengujian model SVM dengan berbagai kombinasi data latih dan data uji menunjukkan bahwa model SVM yang menggunakan data latih pelabelan leksikon tidak dapat memprediksi data uji sebaik model dengan data latih pelabelan manual, sehingga dapat disimpulkan bahwa metode pelabelan sentimen dengan leksikon tidak cocok untuk menyiapkan data latih untuk model analisis sentimen dengan SVM pada data *tweets* vaksin COVID-19.

Model SVM yang digunakan untuk analisis sentimen adalah model yang menggunakan data latih dan data uji pelabelan manual, dengan proporsi 80% dan 20%, secara berurutan. Hasil dari model SVM menunjukkan bahwa *tweets* mengenai

vaksin COVID-19 selama bulan Mei dan Juni 2021 didominasi oleh sentimen netral. Jumlah *tweets* bertambah seiring bertambahnya jumlah kasus terkonfirmasi positif dan kematian akibat COVID-19. Kata-kata yang mendominasi *tweets* bersentimen netral adalah "efek", "informasi", dan "dosis". Kata-kata yang mendominasi *tweets* bersentimen negatif adalah "efek", "habis", "demam", dan "pegal", sedangkan kata-kata yang mendominasi *tweets* bersentimen positif adalah "alhamdulillah", "dosis", dan "semoga".

Pemodelan topik dengan LDA menunjukkan bahwa model LDA terbaik adalah model dengan 11 topik. Dari 11 topik yang dihasilkan, terdapat 1 topik yang sulit diinterpretasi. Topik-topik permasalahan yang beredar di *tweets* mengenai vaksin COVID-19 adalah persetujuan penggunaan vaksin oleh WHO, penolakan vaksin Sinovac oleh Arab Saudi, hoaks dan jual beli vaksin ilegal, pemberhentian sementara penggunaan vaksin AstraZeneca *batch* CMAV547, hasil uji klinis vaksin COVID-19, informasi dan tata cara pendaftaran vaksin, syarat vaksinasi, konsultasi dengan dokter, vaksin Gotong Royong, hingga pengalaman dan efek samping pasca vaksinasi.

Pemerintah Indonesia diharapkan dapat meningkatkan penyebaran informasi mengenai vaksin COVID-19 dan klarifikasi mengenai hoaks agar dapat mengurangi keraguan masyarakat akan vaksin COVID-19. Selain itu, penelitian selanjutnya diharapkan dapat mempertimbangkan untuk menggunakan model dan algoritma yang lebih bervariasi dalam analisis sentimen maupun pemodelan topik, sehingga dapat memberikan hasil yang lebih akurat.

#### DAFTAR PUSTAKA

- [1] M. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "Covid-19 infection: Origin, transmission, and characteristics of human coronaviruses," *Journal of Advanced Research*, vol. 24, pp. 91–98, 03 2020.
- [2] W. H. Organization, "Who coronavirus (covid-19) dashboard," 2020, dapat diakses di <https://covid19.who.int/>. [Diakses pada 10 Juni 2021]. [Online]. Available: <https://covid19.who.int/>

- [3] E. Parker, M. Shrotri, and B. Kampmann, "Keeping track of the sars-cov-2 vaccine pipeline," *Nature reviews. Immunology*, vol. 20, 09 2020.
- [4] P. Olliaro, E. Torreale, and M. Vaillant, "Covid-19 vaccine efficacy and effectiveness—the elephant (not) in the room," *The Lancet Microbe*, 04 2021.
- [5] K. C. Media, "Berita terkini hari ini, kabar akurat terpercaya - kompas.com," dapat diakses di <https://www.kompas.com/>. [Diakses pada 10 Juni 2021]. [Online]. Available: <https://www.kompas.com/>
- [6] K. Kesehatan, ITAGI, and U. dan WHO, "Survei penerimaan vaksin covid-19 di indonesia," Tech. Rep., 2020.
- [7] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.
- [8] F. F. Rachmand and S. Pramana, "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, 2020.
- [9] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 05 2003.
- [10] C.-z. Liu, Y.-x. Sheng, Z.-q. Wei, and Y.-Q. Yang, "Research of text classification based on improved tf-idf algorithm," 08 2018, pp. 218–222.
- [11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [12] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, 2013.
- [13] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.
- [14] Y. Choi and H. Lee, "Data properties and the performance of sentiment classification for electronic commerce applications," *Information Systems Frontiers*, vol. 19, pp. 1–20, 10 2017.
- [15] F. Koto and G. Rahmangtyas, "Inset lexicon: Evaluation of a word list for indonesian sentiment analysis in microblogs," 12 2017.
- [16] G. James, D. Witten, T. Hastie, and R. Tibshirani, *Introduction to Statistical Learning*, 1st ed. Springer, 2013.
- [17] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Springer, 2000.
- [18] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models: A focus on graphical model design and applications to document and image analysis," *IEEE signal processing magazine*, vol. 27, pp. 55–65, 11 2010.
- [19] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.
- [20] M. Ali, D.-H. Son, S.-H. Kang, and S.-R. Nam, "An accurate ct saturation classification using a deep learning approach based on unsupervised feature extraction and supervised fine-tuning strategy," *Energies*, vol. 10, p. 1830, 2017.
- [21] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 399–408. [Online]. Available: <https://doi.org/10.1145/2684822.2685324>
- [22] R. Hannah, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser, "Coronavirus pandemic (covid-19)," *Our World in Data*, 2020, dapat diakses di <https://ourworldindata.org/coronavirus>. [Diakses pada 1 Desember 2021]. [Online]. Available: <https://ourworldindata.org/coronavirus>