Pemodelan Machine Learning: Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Menggunakan Data Twitter

Syafrial Fachri Pane
Jurusan Teknik Informatika
Politeknik Pos Indonesia
Bandung, Indonesia
syafrial.fachri@poltekpos.ac.id

Jenly Ramdan
Jurusan Teknik Informatika
Politeknik Pos Indonesia
Bandung, Indonesia
Jenlyramdan3199@gmail.com

Abstract— In this pandemic era, the government is forced to implement policies that can reduce the daily positive rate of COVID-19. One of these policies is known as PPKM. It is unclear when the pandemic will end, causing data phenomena to be scattered on social media, one of which is Twitter. Therefore, in this study, we conducted an analysis of sentiment originating from tweets from Twitter social media users in the Jakarta area regarding the government's policy, namely PPKM in the face of the COVID-19 pandemic. In this research, we use Machine Learning approach, namely LSTM. This modeling produces a classification of positive and negative sentiments. The dataset used is 3000 tweets with a time period of September -November 2021. At the preprocessing stage the data that are ready to be used for modeling are 2176. The results of this study get an accuracy of 0.943. So the model that we propose, namely LSTM, has succeeded in classifying a satisfactory sentiment with a positive number of 92% and negative 8% of 2176 sentiments, so it can be concluded that the PPKM policy implemented by the Indonesian government in the DKI Jakarta area is said to be effective or positive.

Keywords—Modeling, Sentiment analysis, Tweets, Machine Learning, LSTM

I. PENDAHULUAN

COVID-19 adalah suatu penyakit menular yang disebabkan dari infeksi virus corona jenis baru yang bernama SARS-CoV-2. Organisasi WHO pertama kali mendeteksi virus ini pada tanggal 31 Desember 2019, menyusul laporan yang keluar tentang klaster kasus 'virus pneumonia' di Wuhan, Huabei, China.[1] Virus ini dapat menular dengan cepat dan menyebar ke beberapa wilayah di China dan sebagian besar negara di dunia, salah satunya Indonesia. Hal ini mengakibatkan beberapa negara menerapkan kebijakan-kebijakan baru untuk menaggulangi laju penyebaran COVID-19.[2]

Adapun kebijakan-kebijakan yang dilakukan oleh negara-negara guna menanggulangi laju penyebaran COVID-19 adalah dengan menerapkan peraturan kebijakan yang bersifat membatasi mobilitas masyarakat seperti lockdown nasional yang dijalankan oleh pemerintahan China, India dan Sydney di negaranya masing-masing[3-5]. Begitupun dengan pemerintahan Indonesia yang mana kebijakan ini dikenal dengan nama Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM).[6]

Pemberlakukan PPKM masih berlangsung di Indonesia, dengan ketidakjelasan kapan pandemi dan PPKM berakhir, mengakibatkan banyaknya terjadi fenomena-fenomena data. Tentunya mengakibatkan sentimen-sentimen yang beredar di sejumlah platform media sosial salah satunya adalah twitter. Ada beranekaragam sentimen-sentimen yang berasal dari komentar opini masyarakat terkait tanggapan terhadap kebijakan PPKM, baik itu opini positif maupun opini negatif. Komentar-komentar opini ini akan digunakan untuk analisis sentimen.

Oleh karena itu, opini tersebut sangat menarik apabila dilakukan kajian secara mendalam. Kajian komentar opini ini merupakan data yang relevan sehingga

dapat dijadikan sebagai bahan rekomendasi untuk pemerintah dalam mempertimbangan kebijakan PPKM. Pada penelitian sebelumnya, digunakan dataset twitter yang terdiri dari empat parameter yaitu tweet_id, tweet, skor_sentimen dan bulan. Kemudian parameter diuji menggunakan model *Deep Learning* dan LSTM sehingga menghasilkan multilabel yang didominisasi oleh label *joke*, *annoyed*, dan optimis yang kemudian label dapat diklasifikasikan menjadi sentimen negatif dan positif [7].

Analisis sentimen melibatkan penggunaan metode *Natural Languange Processing* (NLP), metode NLP ini digunakan untuk mengolah data dan mengekstrak data dari media sosial twitter. [11] Yang mana hasil preprocessing selanjutnya akan diklasifikasikan dengan menggunakan pendekatan model *Machine Learning* yaitu LSTM. Penulis berpendapat bahwa model *Machine Learning* cocok untuk melakukan tugas yang akan dibuat visualisasi. Hasil dari penelitian ini adalah didapatkannya klasifikasi model LSTM yang menghasilkan nilai akurasi tertinggi bila dibandingkan dengan metode lainnya.[2]

Oleh karena itu dari ulasan diatas penulis tertarik untuk melakukan penelitian ini. Adapun perbedaan pada penelitian ini adalah penulis akan lebih fokus pada analisis karakteristik data dan sentimen di media sosial twitter menggunakan kata kunci yang berhubungan dengan COVID–19 dan PPKM menggunakan model *Machine Learning* yaitu LSTM dengan studi kasus di Jakarta, Indonesia.

II. METODOLOGI PENELITIAN

Metode penelitian ini pembahasan tentang langkah yang dilakukan penulis pada penelitian. Dalam penelitian ini penulis menggunakan metode LSTM untuk klasifikasi sentimen twitter, dimana hal ini meliputi beberapa Langkah-langkah dimulai dari studi literatur, collecting data, labeling, preprocessing, word embedding, modeling, dan evaluasi model. Berikut adalah gambaran dari metode penelitian:



Gambar 1. Metodologi penelitian

A. Collecting Data

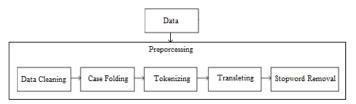
Collecting data tweet dilakukan secara crawling dengan memanfaatkan api twitter dan library tweepy, dimana diharuskan untuk memiliki akun twitter developer dan token. Setelah itu dilakukan crawling data sebanyak 3100 tweet menggunakan keyword ppkm, pemberlakuan pembatasan kegiatan masyarakat dan hastag ppkm yang memiliki geocode daerah Jakarta.

Tabel 1. Tabel indeks kata

No	Tweet
1	#BersamaRakyat #bersama #rakyat #jokowi #presiden #presidenjokowi #covid_19 #covid #vaksin #vaksincovid19 #vaksinasicovid19 #ppkm #ppkmterbaru #ppkmupdate https://t.co/IEnucGFuuJ","Indonesia Tutup Pintu Masuk Bagi Negara Ini !!
2	#BeritaTerkini #beritaViral #berita #NEWS #NewsUpdate #Viral #viralTiktok #viraltwitter #viralkan #BeritaJogja #BeritaJakarta #Omricon #OmriconVariant #COVID19 #covidvariant #Covid #Covid_19 #COVID—19 #ppkm #vaksinCOVID19 https://t.co/1VWEXaONoJ","NEGARA MANA DILARANG INDONESIA? #BeritaTerkini #beritaViral #berita #NEWS #NewsUpdate #Viral #viralTiktok #viraltwitter #viralkan #BeritaJogja #BeritaJakarta #Omricon #OmriconVariant #COVID19 #covidvariant #Covid #Covid_19 #COVID—19 #ppkm #vaksinCOVID19 https://t.co/1VWEXaONoJ","['BeritaTerkini', 'beritaViral', 'berita', 'NEWS', 'NewsUpdate', 'Viral', 'viralTiktok']"
	#BeritaTerkini #beritaViral #berita #NEWS #NewsUpdate #Viral #viralTiktok #viraltwitter #viralkan #BeritaJogja #BeritaJakarta #Omricon #OmriconVariant #COVID19 #covidvariant #Covid #Covid_19 #COVID—19 #ppkm #vaksinCOVID19 https://t.co/8HKdzESBXx","APA GEJALA TIDAK BIASA OMICRON? #BeritaTerkini #beritaViral #berita #NEWS #NewsUpdate #Viral #viralTiktok #viraltwitter #viralkan #BeritaJogja #BeritaJakarta #Omricon #OmriconVariant #COVID19 #covidvariant #Covid #Covid_19 #COVID—19 #ppkm #vaksinCOVID19 https://t.co/8HKdzESBXx","['BeritaTerkini', 'beritaViral', 'berita', 'NEWS', 'NewsUpdate', 'Viral',
3	'viralTiktok']"

B. Preprocessing

Preprocessing adalah suatu proses dimana dataset yang di collect dari media sosial twitter akan dibersihkan dari elemen-elemen yang tidak dibutuhkan sehingga nantinya akan mendapatkan data yang memiliki kualitas dan sesuai apa yang di inginan peneliti. Proses preprocessing terbagi menjadi kedalam beberapa tahapan. yaitu data cleaning, case folding, transleting, tokenizing, stopword removal dan stemming.



Gambar 2. Preprocessing

1) Data cleaning

Pada proses ini dilakukan pembersihan dataset tweet seperti *link*, gambar, tanda baca, *retweet*, simbol-simbol, dan *emoticon* yang terdapat pada dataset.

Tabel 2. Tabel data cleaning

No	Tweet
1	Indonesia Tutup Pintu Masuk Bagi Negara Ini
2	NEGARA MANA DILARANG INDONESIA
3	APA GEJALA TIDAK BIASA OMICRON

2) Case Folding

Proses case folding adalah tahapan dimana dataset yang sudah dibersihkan kemudian diubah menjadi font kecil atau lower.

Tabel 3. Tabel case folding

No	Tweet
1	indonesia tutup pintu masuk bagi negara ini
2	negara mana dilarang Indonesia
3	apa gejala tidak biasa omicron

3) *Transleting*

Proses menterjemahkan tweet bahasa Indonesia kedalam bahasa inggris. Hal ini dilakukan karena *library TextBlob* hanya dapat menghitung *polarity* dan sentimen dataset berbahasa inggris.

Tabel 4. Tabel translating

No	Tweet asli	Tweet terjemahan
1	indonesia tutup pintu masuk bagi negara ini	indonesia close the entrance for this country
2	negara mana dilarang Indonesia	which country is prohibited by indonesia
3	apa gejala tidak biasa omicron	what are the unusual symptoms omicron

4) Tokenizing

Proses memberikan token dengan cara membagi tweet menjadi beberapa potongan kecil kata yang disebut dengan nama token.

Tabel 5. Tabel tokenizing

No	Tweet	Tokenizing
1	indonesia close the entrance for this country	indonesia, close, the, entrance, for, this, country
2	which country is prohibited by indonesia	which, country, is, prohibited, by, indonesia
3	what are the unusual symptoms omicron	what, are, the, unusual, symptoms, omicron

5) Stopword removal

Setelah kata-kata menjadi token berikutnya akan dilakukan proses eliminasi kata-kata yang tidak memiliki arti penting. Sebelum dilakukan *stopword removal* dilakukan terlebih dahulu membuat *stoplist*. Nantinya kata-kata yang masuk kedalam daftar *stoplist* akan dihapus.

Tabel 5. Tabel stopword removal

No	Tweet	Stopword
1	indonesia close the entrance for this country	indonesia, close, entrance, country
2	which country is prohibited by Indonesia	country, prohibited, indonesia
3	what are the unusual symptoms omicron	unusual, symptoms, omicron

C. Labeling

Labeling adalah suatu proses untuk memberikan label sentiment pada dataset. Data yang ditampung kemudian dilabeli menggunakan *library TextBlob* dengan memanfaatkan *polarity*. 3100 tweet dijadikan satu *variabel* dengan tipe *data List*. Kemudian dari masing-masing tweet tersebut ditetapkan polaritasnya apakah berlabel positif atau negatif. Dari *variabel List*, data ditampilkan menggunakan *Pandas DataFrame* dengan *field*(tweet, sentimen, label). Kemudian data dari *Pandas DataFrame* di-*export* menjadi file CSV.

Tabel 6. Tabel labeling

No	Tweet	Polarity	Subject	Sentimen	Neg	Pos	Comp
1	indonesia close the entrance for this country	0.534300	0.765402	Positif	0.0	0.098	0.6278
2	which country is prohibited by indonesia	0.633200	0.554020	Positif	0.0	0.074	0.7546
3	what are the unusual symptoms omicron	0.500300	0.654622	Positif	0.0	0.088	0.7732

D. Word Embedding

Setelah melalui tahapan *preprocessing*, selanjutnya akan dilakukan perubahan kata-kata yang memiliki kaitan secara semantik kedalam bentuk angka dan dibuat menjadi word embedding. Nantinya word embedding ini akan dijadikan input untuk metode klasifikasi sentimen LSTM. Adapun berikut tahapan yang terdapat dalam proses word embedding:

a) Indeks Kata

Melakukan pengindeksan dari data yang sudah di preprocessing dan sudah menjadi *list* dengan cara merubah kata-kata unik menjadi angka yang mewakilkan kata-kata sebelum memasuki tahapan word embedding.

Tabel 7. Tabel indeks kata

Kata	Indeks
hope	1
Ends	2
covid	3

b) *Padding*

Setelah mendapatkan indeks untuk setiap kata, langkah selanjutnya adalah melakukan padding untuk setiap frase. Padding berfungsi untuk menyamakan panjang list dari setiap kalimat sehingga dapat dimasukkan pada saat proses klasifikasi. Dalam pembuatan padding pertama harus melakukan deklarasi panjang maksimum vektor. Dengan melakukan padding, daftar kalimat yang panjangnya tidak memenuhi panjang nilai yang ditentukan akan memiliki tambahan nilai 0 di awal.

Tabel 8. Tabel *padding*

Hasil preprocessing	Indeks	padding
hope pandemic ends soon	[1623]	[0000001623]

c) Word Embedding

Setelah penindeksan dan *padding*, selanjutnya list kalimat dapat dilakukan konversi ke dalam vektor untuk menemukan kesamaan semantic.

Tabel 9. Tabel word embedding
Hasil word embedding
[[-0.00736422 -0.03235253 -0.022900940.03847875 -0.047875831
-0.03987709]
[0.01309475 0.07812738 0.006908 0.0312356 0.02665273
-0.05730297]
[-0.08347892 0.03180597 0.03123798 0.02611795 -0.000234211
-0.03423871]

E. Modeling

Modeling merupakan proses dalam membuat model LSTM. Proses modeling terbagi kedalam beberapa tahap, yaitu splitting data, vektorisasi teks, dan arsitektur jaringan LSTM.

1) Splitting data

Data dibagi menjadi *data training* dan *data test*. Pembagian data dilakukan dengan rasio 80:20 dengan cara *test_size*=0.2 yang berarti data yang digunakan sebagai *data testing* adalah sebesar 20% dari keseluruhan dataset yang berarti sebanyak 1740 data digunakan untuk *data training* dan 436 data untuk *data test*. Selanjutnya, data tersebut dibagi secara acak dengan *random_state* = 42.

```
train_df, test_df = train_test_split(df, test_size = 0.2, random_state = 42)
print("Training data size: ", train_df.shape)
print("Test data size: ", test_df.shape)

Training data size: (1740, 13)
Test data size: (436, 13)
```

Gambar 3. Arsitektur jaringan

2) Vektorisasi teks

Pada tahap ini dilakukan *vektorisasi* teks dengan menggunakan kelas *tokenizer* dari modul Keras. *Vektorisasi* teks memberikan setiap kata unik nomor *integer*. Yang artinya setiap kata akan diidentifikasi dengan bilangan bulat. Dengan ketentuan *top words* sebesar 10000, teks *review* sebesar 200 dan *train_df* menggunakan parameter label.

```
top_words = 10000
tokenizer = Tokenizer(num_words=top_words)
tokenizer.fit_on_texts(train_df['transalted'])
list_tokenized_train = tokenizer.texts_to_sequences(train_df['transalted'])
max_review_length = 200
X_train = pad_sequences(list_tokenized_train, maxlen=max_review_length)
y_train = train_df['label']
```

Gambar 4. Arsitektur jaringan

3) Arsitektur Jaringan

Arsitektur ini memiliki 3 layer. Layer pertama adalah Embedding Layer yang menggunakan vector dengan besaran panjang dimensi sebesar 32 dan kata sebanyak 200. Layer kedua adalah LSTM layer dengan besaran 100 neuron. Layer terakhir adalah Dense Layer dengan besaran 1 neuron dan fungsi aktivasi menggunakan aktivasi sigmoid. Dikarenakan penelitian ini merupakan binary classification yaitu klasifikasi sentimen positif atau negatif, maka digunakanlah loss function binary_crossentropy (keras) dan optimasi 'Adam'. Parameter lainnya adalah batch size sebesar 64, epoch sebesar 20 dan validation split sebesar 0.2 atau 20%. Bedasarkan uraian parameter yang dijelaskan maka berikut adalah arsitektur jaringan pada model yang terbentuk:

Model: "sequential"

Layer (type)	Output	Shape	Param #
embedding (Embedding)	(None,	200, 32)	320032
lstm (LSTM)	(None,	100)	53200
dense (Dense)	(None,	1)	101
Total params: 373,333 Trainable params: 373,333 Non-trainable params: 0			=======

Gambar 5. Arsitektur jaringan

F. Evaluasi Model

Model kemudian dilakukan evaluasi dengan cara pengukuran kinerja menggunakan *confusion matrix*. adapun rumus dari *confusion matrix* yang digunakan untuk menghitung performa klasifikasi adalah mencari nilai akurasi, presisi, f1 dan *recall*. Berikut rumus nya:

$$Akurasi = \frac{(TP+TN)}{(TP+FN+FP+TN)} * 100\%$$
 (1)

$$Presisi = \frac{TP}{(TP+FP)} * 100\%$$
 (2)

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision}$$
 (3)

$$Recall = \frac{TP}{(TP+FN)} * 100\% \tag{4}$$

III. PEMBAHASAN DAN HASIL

Hasil pada tabel dibawah menunjukkan bahwa metode LSTM yang diusulkan dalam hal sentimen analisis menggunakan data tweet twitter dapat menghasilkan akurasi, presisi, dan daya ingat yang tinggi. Ini terjadi karena kernel LSTM sangat cocok digunakan pada masalah klasifikasi data berupa teks. LSTM merupakan metode yang terbukti mampu mengklasifikasikan data teks dengan hasil akurasi yang tinggi.

Tabel 10. Tabel evaluasi model

Nama	LSTM
Accuracy score	0.917
F1 score	0.956
Precision score	1.0
Recall score	0.917

IV. KESIMPULAN

Berdasarkan dari hasil pengujian yang dilakukan, maka dapat disimpulkan bahwasanya metode LSTM berhasil melakukan klasifikasi sentiment data tweet dengan nilai akurasi sebesar 0,917. Pada tahapan *preprocessing* menggunakan 3336 data tranning dan 835 data testing. Model dapat melakukan klasifikasi analisis sentimen yang dikelompokan menjadi 2 jenis kelas sentimen, yaitu

sentimen negatif dan sentimen positif dengan sentimen positif sebesar 92% dan negatif 8% dari 2176 sentimen.

V. SARAN

Model LSTM ini tidak dapat melakukan klasifikasi sentimen emoji. Oleh karena itu sangat menarik jika model dikembangkan dengan menambahkan suatu metode atau algoritma untuk meningkatkan nilai akurasi model dengan melakukan klasifikasi sentimen emoji.

REFERENCES

- [1] WHO, "Pneumonia of unknown cause China," 2020. https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229 (accessed Apr. 13, 2022).
- [2] Fajria A (2020, 17 Maret). 11 Negara Ini Lakukan Lockdown Demi Tekan Corona, Indonesia Kok Belum?. Diakses dari http://www.myblog.com/117893
- [3] Marry C (2020, 25 Maret). Coronavirus: India enters 'total lockdown' after spike in cases. Diakses dari https://www.bbc.com/news/world-asia-india-52024239
- [4] Rollin A (2020, 23 Maret). Australia starts lockdown measures as coronavirus cases jump. Diakses dari https://www.straitstimes.com/asia/australianz/australia-starts-lockdown-measures-as-coronavirus-cases-jump
- [5] Lin (2020, 13 Maret). China's coronavirus lockdown strategy: brutal but effective. Diakses dari https://www.theguardian.com/world/2020/mar/19/chinas-coronavirus-lockdown-strategy-brutal-but-effective
- [6] Michael R (2021, 1 July). Covid-19: Indonesia announces lockdown in Java and Bali as cases surge. Diakses dari https://www.bbc.com/news/world-asia-57647693
- [7] R. Chandra and A. Krishna, "COVID-19 sentiment analysis via deep learning during the rise of novel cases," *PLoS ONE*, vol. 16, no. 8 August, Aug. 2021, doi: 10.1371/journal.pone.0255615.
- [8] Y. Astari and S. Wahib Rozaqi, "Analisis Sentimen Multi-Class pada Sosial Media menggunakan metode Long Short-Term Memory (LSTM)," 2021.
- [9] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model," *IEEE Access*, vol. 8, pp. 138162–138169, 2020, doi: 10.1109/ACCESS.2020.3012595.
- [10] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 802–808, Aug. 2021, doi: 10.29207/resti.v5i4.3308.
- [11] A. Tiwari, R. Gupta, and R. Chandra, "Delhi air quality prediction using LSTM deep learning models with a focus on COVID-19 lockdown," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.10551
- [12] F. Aldi Nugraha, N. Hanum Harani, R. Habibi, and R. Nuraini Siti Fatonah, "Sentiment Analysis on Social Distancing and Physical Distancing on Twitter Social Media using Recurrent Neural Network (RNN) Algorithm," *Jurnal Online Informatika*, vol. 5, no. 2, pp. 195–204, 2020, doi: 10.15575/join.
- [13] R. Chandra, A. Jain, and D. S. Chauhan, "Deep learning via LSTM models for COVID-19 infection forecasting in India," Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.11881
- [14] C. K. N. Paputungan and A. Jacobus, "Sentiment Analysis of Social Media Users Using Long-Short Term Memory Method Analisis Sentimen Pengguna Sosial Media Menggunakan Metode Long Short Term Memory," *Jurnal Teknik Elektro dan Komputer*, vol. 10, no. 2.

- [15] R. Lamsal, "Design and analysis of a large-scale COVID-19 tweets dataset," *Applied Intelligence*, vol. 51, no. 5, pp. 2790–2804, May 2021, doi: 10.1007/s10489-020-02029-z.
- [16] A. C. M. V. Srinivas, Ch. Satyanarayana, Ch. Divakar, and K. P. Sirisha, "Sentiment Analysis using Neural Network and LSTM," *IOP Conference Series: Materials Science and Engineering*, vol. 1074, no. 1, p. 012007, Feb. 2021, doi: 10.1088/1757-899x/1074/1/012007.