Sentiment Analysis Comments Covid-19 Variant Omicron on Social Media Instagram with Bidirectional Encoder from Transformers (BERT)

Sentimen Analisis Komentar Covid-19 Varian Omicron pada Media Sosial Instagram dengan Bidirectional Encoder from Transformers (BERT)

Dody Pradipta[1], Kusrini[2], Hanif Al Fatta[3]
[1,2,3] Teknik Informatika, Magister Teknik Informatika, Universitas Amikom Yogyakarta
E-mail: *[1]pradiptadody@gmail.com, [2] kusrini@amikom.ac.id, [3]hanif.a@amikom.ac.id

Abstract **–** *Technology and information in recent years has made the internet a communication tool that is in great demand. This is the background for changes in communication to be more up-to-date and completely digital. Social media is one of the changes in digital communication, social media is a digital platform that facilitates users to communicate with each other, exchange information, etc. Comments can be used to obtain opinions from users who are on the social media platform which is intended to obtain core input from users or consumers efficiently. One of the social media that is widely used today is Instagram, many of its users use Instagram, to express opinions (comments) about the Covid-19 pandemic that is happening. Community comments can later be classified into positive, negative and neutral sentiments using the Bidirectional Encoder from Transformers (BERT) method. The results of the sentiment analysis can see how people perceive the Omicron variant of the Covid-19 pandemic. The scrapping process obtained 1,052 data, which has been classified as 663 data negative comments, 388 data neutral comments and 1 positive comment data. The results of the test obtained an accuracy of 0.632 (63%).*

Keywords *— BERT, covid-19, media social, sentiment analysis*

Abstrak – Internet merupakan alat komunikasi yang banyak diminati karena pesatnya perkembangan teknologi dan informasi beberapa tahun belakangan ini. Ini adalah konteks untuk memodernisasi dan sepenuhnya mendigitalkan komunikasi. Salah satu perubahan dalam komunikasi digital adalah media sosial, *platform* digital yang memungkinkan orang berbicara satu sama lain, berbagi informasi, dan lainnya. Pada *platform* media sosial ini, yang dirancang untuk mendapatkan masukan inti dari pengguna atau konsumen secara efisien, komentar dapat digunakan untuk mengumpulkan opini dari pengguna. *Instagram* adalah salah satu *platform* media sosial paling populer saat ini, dan banyak penggunanya menggunakannya untuk menyuarakan pendapat (komentar) mereka tentang pandemi *Covid-19*. Menggunakan metode *Bidirectional Encoder from Transformers* (BERT), komentar masyarakat nantinya dapat diklasifikasikan menjadi sentimen positif, negatif, dan netral. Analisis sentimen mengungkapkan bagaimana perasaan orang tentang varian *Omicron* pandemi *Covid-19*. Prosedur scraping menghasilkan 1.052 data yang terdiri dari 663 komentar negatif, 388 komentar netral, dan 1 komentar positif. Hasil tes memiliki akurasi sebesar 0,632 (63%).

Kata Kunci — analisis sentimen, BERT, *covid-19*, sosial media

1. PENDAHULUAN

Covid-19 adalah penyakit menular yang disebabkan oleh virus *Corona* (juga dikenal sebagai *SARS-CoV-2*) [1]. Pada Desember 2019, pertama kali ditemukan di Wuhan, ibu kota provinsi Hubei

 doi: https://doi.org/10.32503/jtecs.v3i1.3219

China, dan sejak itu menyebar ke seluruh dunia. Karena penyebaran globalnya yang cepat, Organisasi Kesehatan Dunia (WHO) menyatakan *Covid-19* sebagai pandemi pada Maret 2020. Karena penyebarannya yang luar biasa, virus *Corona* tidak dianggap sebagai pandemi karena tidak terbatas pada wilayah tertentu.

Virus ini mengembangkan varian baru sebagai hasil dari sejumlah mutasi. WHO membagi virus *SARS-CoV2* menjadi dua kelompok: varian minat (VOI) dan kekhawatiran (VOC) yang diidentifikasi oleh Organisasi Kesehatan Dunia (WHO) pada tahun 2021. Kategori VOI mencakup mutasi baru dengan *fenotipe implisit* yang telah terdeteksi di banyak negara dan meningkatkan kemungkinan penularan lokal [2]. WHO saat ini mengklasifikasikan *Omicron* sebagai VOC, meskipun kondisi tertentu dapat menyebabkan kategori VOI naik ke status VOC. Dasar klasifikasi ini adalah ditemukannya sejumlah besar mutasi pada varian ini, yang menunjukkan peningkatan risiko infeksi ulang.

Hingga lima kali lebih cepat dari varian sebelumnya, termasuk varian *Delta*, varian *omicron* mentransmisikan informasi [3]. Masyarakat mengalami kecemasan akibat angka penularan yang sangat tinggi; Namun, disiplin diri dalam protokol kesehatan dan vaksinasi dapat mencegah penularan. Pembatasan Sosial Berskala Besar (PSBB) adalah salah satunya, seperti penerapan protokol kesehatan [4]. Pembatasan sosial publik untuk perkantoran, kampus, sarana hiburan, sarana olah raga, dan lain-lain dipengaruhi oleh pemberlakuan PSBB.

Untuk melakukan analisis sentimen *review Amazon Fine Food*, peneliti [5] memanfaatkan teknik SVM, *Logistic Regression*, dan *Naive Bayes* dengan berbagai bobot. Dengan tingkat akurasi sebesar 78,11 persen, pendekatan Regresi Logistik merupakan yang paling akurat. Pada penelitian selanjutnya [6], analisis sentimen pada *dataset Amazon Fine Food Review* dilakukan dengan menggunakan teknik klasifikasi seperti *Naive Bayes*, *K-Nearest Neighbor*, *Logistic Regression*, *Decision Tree*, dan *Random Forest* serta bobot BOW sebesar 90%.

Terlepas dari kenyataan bahwa pemeriksaan sebelumnya hanya menggunakan model LSTM dan CNN satu arah, penelitian [7] menggunakan CNN dan *Word Embedding* sebagai penyorotan vektor kata dan eksplorasi [8] menggunakan LSTM dan *Word Embedding* yang telah disiapkan untuk mendapatkan deskripsi survei kata-kata masa lalu dan masa depan. untuk penemuan gambar, tetapi proses *downsampling* CNN menghasilkan lebih sedikit data untuk teks. Pemanfaatan Penyematan Kata, yang memanfaatkan vektor kata dengan dimensi lebih kecil dan semata-mata ditentukan oleh sejauh mana kata-kata itu berdekatan. Penelitian terbaru menggunakan BERT terlatih, yang peringatannya berasal dari wikipedia dan *Book Corpus* yang ekstensif, dan akurasinya mungkin telah meningkat.

BERT adalah model pembelajaran mendalam yang menghasilkan hasil mutakhir pada berbagai tugas Pemrosesan Bahasa Alami (NLP). BERT dapat menghasilkan model bahasa dua arah yang mendalam dengan memahami konteks komentar. Di BERT, ada enam lapisan transformer di setiap *encoder* dan *decoder*. *Google* telah membuka sumber model BERT yang sebelumnya dinonaktifkan, yang dapat dimulai dari sebuah kata dengan gambar Penyematan Kata yang disematkan, yang merupakan keuntungan dari BERT. Penulis memutuskan untuk menamai penelitiannya "Analisis Sentimen Komentar Varian *Omicron Covid-19* dengan *Bidirectional Encoder* dari Media Sosial *Instagram Transformers* (BERT)" karena fenomena tersebut di atas.

2. METODE PENELITIAN

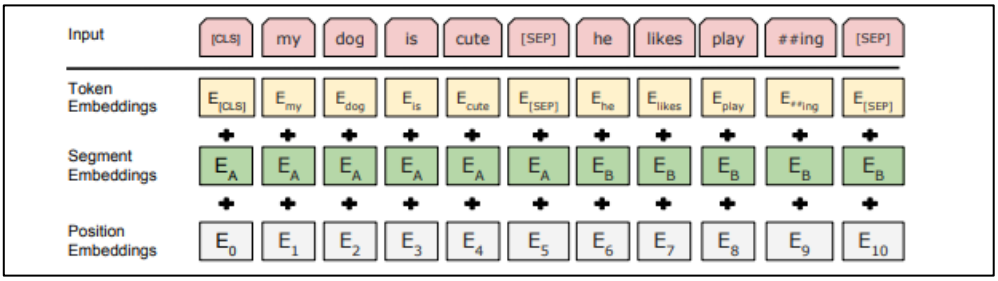
2.1. Analisis Sentimen

Bidang penelitian *Text Mining* disebut Analisis Sentimen, juga dikenal sebagai *Opinion Mining* yang merupakan sebagai pembaharuan terhadap sebuah opini [9]. Menentukan perspektif pembicara atau penulis tentang sejumlah topik atau polaritas keseluruhan konteks, diperlukan penggalian opini, pendapat, penilaian, atau evaluasi, serta keadaan afektif atau komunikasi emosional, menjadi sikap yang diambil [1]. Produk konsumen, layanan kesehatan, acara sosial dan politik, dan analisis tren menggunakan analisis sentimen positif atau negatif hanyalah beberapa contoh dari banyak penerapan

dari analisis sentimen. Tujuan dari penelitian ini adalah untuk menyelidiki opini publik mengenai varian *omicron* yang diklasifikasikan menggunakan BERT dari *Covid-19*.

2.2. Bidirectional Encoder from Transformers (BERT)

Deep Learning model yang dikembangkan peneliti *Google AI Language* mengembangkan model representasi bahasa terlatih yang dikenal sebagai BERT. BERT memanfaatkan *Transformers*, sebuah sistem yang melihat bagaimana kata-kata dalam sebuah teks berhubungan dengan konteksnya [10].



Gambar 1. Arsitektur BERT

Gambar 1 merupakan arsitektur BERT menunjukkan cara tokenisasi setiap kata menjadi *Word Embedding* dalam struktur vektor yang merupakan kontribusi dari lapisan *Encoder* dari *Transformers* BERT. Vektor ini memungkinkan model transformer untuk memahami posisi setiap kata, yang mungkin memiliki makna kontekstual yang berbeda meskipun memiliki bentuk yang sama [11].

2.3. Dataset

Pengumpulan *dataset* didapatkan dengan cara *scrapping* dengan *Webharvy*, dimana komentar yang digunakan adalah postingan dari akun indozone.id dengan kata kunci "*Covid*, *Covid-19*, *Omicron*, Pandemi".

Tabel 1. *Dataset*

No	*Dataset*
1	Herannya sebelum tuh virus muncul pasti sudah diprediksi terlebih dahulu 😮😮
2	klo mirip flu biasa knp mesti takut
3	Banyak makan, banyak tedor, banyak ketawa, jangan banyak baca berita. Aman wesss... Wkwkwk...

Tabel 1 merupakan beberapa hasil scrapping dengan *Webharvy*, berdasarkan hasil *scrapping* didapatkan sejumlah 1.052 data komentar pada Instagram, yang selanjutnya akan dilabeli positif, negatif serta netral.

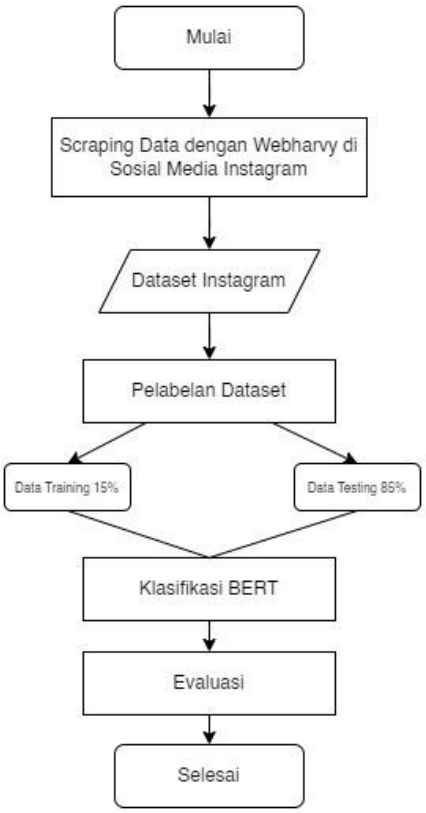
2.4. Jenis, Sifat dan Pendekatan Penelitian

Penelitian eksperimental, yang menggunakan serangkaian tindakan untuk mendemonstrasikan suatu konsep adalah jenis penelitian yang digunakan peneliti [12]. Melalui *scraping* data *Instagram*, penulis menerapkan penelitian kuantitatif deskriptif artinya tentang penelitian yang didefinisikan berlandaskan filsafat positivisme yang digunakan untuk mempelajari populasi atau sampel tertentu dengan mengumpulkan data dengan instrumen tertentu [13]. Metodologi kuantitatif dilibatkan oleh peneliti untuk mengukur penilaian dalam strategi BERT dalam mengklasifikasikan komentar.

2.5. Desain Penelitian

Metode dalam penelitian ini dilakukan melalui beberapa tahap. Secara garis besar, alur penelitian dapat digambarkan pada diagram Gambar dibawah:

 doi: https://doi.org/10.32503/jtecs.v3i1.3219



Gambar 2. Desain Penelitian

Gambar 2 merupakan desain kerja dari penelitian yang dilakukan, dimulai dengan proses *scrapping* sehingga didapatkan sebuah dataset, kemudian dataset tersebut akan dilabeli ,selanjutnya akan diklasifikan dengan *Bidirectional Encoder from Transformers* (BERT) dengan model yang dibangun adalah 15% untuk data *train* dan 85% untuk data *test*. Setelah proses klasifikasi telah dilakukan maka akan dilakukan evaluasi untuk memperbaiki model yang akan dikembangkan.

3. HASIL DAN PEMBAHASAN

Sebelum diklasifikasikan ada beberapa tahap yang harus dilakukan yaitu tokenisasi dan *padding*, BERT memanfaatkan tokenizer model dari BERT yang sudah ada, yang memiliki dua buah spesial token yaitu [cls] dan [sep]. Kedua token tersebut digunakan untuk tokenisasi terhadap 2 kalimat atau lebih. Namun penambahan ini bersifat optional dalam penerapannya. Contoh proses tokenisasi dapat dilihat pada Gambar 3 dibawah ini.

```
print("Original: ", sentences[78])

print("Tokenized: ", tokenizer.tokenize(sentences[78]))

print("Token IDS: ", tokenizer.convert_tokens_to_ids(tokenizer.tokenize(sentences[78])))
```

```
Original:  BODO AMAT NGENTTOTTT
Tokenized:  ['bodo', 'amat', 'ng', '##ent', '##tott', '##t']
Token IDS:  [71945, 58314, 10822, 11604, 20272, 10123]
```

Gambar 3. Tokenisasi

Gambar 3 diatas ialah proses tokenisasi untuk membagi teks yang berupa kalimat menjadi token-token tertentu, digunakan untuk mengamankan dan mengurangi kepekaan data dengan mengganti data asli dengan nilai yang tidak terkait dengan panjang dan format yang sama.

```
[ ]  input_ids = []

     for sent in sentences:
       encoded_sent = tokenizer.encode(
           sent,
           add_special_tokens = True
       )
       input_ids.append(encoded_sent)

     print("Original: ", sentences[0])
     print("Token IDs: ", input_ids[0])
```

```
Original:  Brarti sakit flu gitu kah?
Token IDs:  [101, 47237, 26538, 43380, 19341, 10136, 21464, 10998, 10237, 10243, 136, 102]
```

Gambar 4. Tokenisasi Ids

Gambar 4 Token Ids ialah pemberian token-token tertentu, ketika model BERT dilatih setiap kata akan dilabeli dengan token unik. Oleh karena itu perlu adanya konversi token kedalam kalimat input menjadi ID yang unik.

```
[ ]  from keras_preprocessing.sequence import pad_sequences

     MAX_LEN = 170

     print("Padding/truncating all sentences to %d values" % MAX_LEN)
     print('Padding token: "{:}", ID: {:}'.format(tokenizer.pad_token, tokenizer.pad_token_id))

     input_ids = pad_sequences(input_ids, maxlen=MAX_LEN, dtype='long', value=0, truncating='post', padding='post')

     print("Done")
```

```
Padding/truncating all sentences to 170 values
Padding token: "[PAD]", ID: 0
Done
```

Gambar 5. Proses *Padding*

Gambar 5 merupakan proses tokenisasi, data yang diperoleh dari opini pengguna Instagram yang memiliki bentuk tidak terstruktur, masih memiliki suku kata yang beragam. Proses ini berfungsi untuk mengubah setiap *sequence* agar memiliki suku kata yang sama. Sehingga semua *sequence* memiliki panjang yang disamakan agar terstruktur. Penelitian kali ini, peneliti menyamakan seluruh *sequence* memiliki panjang maksimal 170 kata.

Hasil *scrapping* pada Instagram dengan *keywords* "*Covid-19*, *Omicron*, Corona dan Pandemi" didapatkan sejumlah 1.052 data.

Tabel 2. Hasil Pelabelan

Sumber Data	Jumlah Data			Total Data
	Komentar Positif	Komentar Negatif	Komentar Netral	
Instgram	1	663	388	1052

Tabel 2 Sebanyak 1052 data yang didapatkan dari *scrapping*, selanjutnya diklasifikasikan menjadi data positif, negatif dan netral. Hasil dari pelabelan menghasilkan 663 komentar negatif, 388 komentar netral dan 1 komentar positif. Model yang dibangun menggunakan BERT dengan proporsi data *training* 85% dan data *testing* 15%.

 doi: https://doi.org/10.32503/jtecs.v3i1.3219

```
[ ]  from sklearn.model_selection import train_test_split

     train_input, test_input, train_labels, test_labels = train_test_split(input_ids,
                                                                           labels,
                                                                           random_state=2017,
                                                                           test_size=0.1)
     train_mask, test_mask, _, _ = train_test_split(attention_mask,
                                                    labels,
                                                    random_state=2017,
                                                    test_size=0.1)

     train_input, validation_input, train_labels, validation_labels = train_test_split(train_input,
                                                                                       train_labels,
                                                                                       random_state=2018,
                                                                                       test_size=0.15)
     train_mask, validation_mask, _, _ = train_test_split(train_mask,
                                                          train_mask,
                                                          random_state=2018,
                                                          test_size=0.15)
```

Gambar 6. Training Model

Gambar 6 merupakan serangkaian proses untuk train model yang tengah dibangun, dengan mengunakan data *train* sebesar 15% dan data *test* sebesar 85%. Menggunakan *library sklearn* digunakan untuk proses *split dataset*.

```
[ ]  from sklearn.metrics import accuracy_score

     acc = accuracy_score(flat_true_labels, flat_prediction)

     print("Akurasi: %.3f" %acc)

     Akurasi: 0.632
```

Gambar 7. Hasil Evaluasi Klasifikasi

Gambar 7 merupakan hasil evaluasi klasifikasi dengan menggunakan *library sklearn* digunakan untuk inisialisasi untuk mencari akurasi dari model yang tengah dibangun. Hasil pengujian dengan klasifikasi BERT dengan proporsi data *train* 85% dan data *test* 15% dapat diketahui pada Gambar 7. Nilai akurasi klasifikasi BERT yang diperoleh pada Instagram adalah 63%.

4. KESIMPULAN

Berdasarkan hasil analisa dan pengujian terhadap analisis sentimen pada opini publik terhadap *Covid-19 Omicron* Media Sosial Instagram denggan menggunakan metode BERT, maka dapat ditarik kesimpulan bahwa metode BERT dapat digunakan untuk mengetahui sentimen terhadap *Covid-19* varian *Omicron* pada sosial media Instagram. Analisis sentimen pada *Covid-19* varian *Omicron* menghasilkan nilai akurasi 63% dengan proporsi *dataset* 85% data *training* dan 15 data *testing*. Token [cls] dan [sep] pada BERT dapat mempengaruhi klasifikasi analisis sentimen.

Untuk pengembangan analisis sentimen pada opini publik terhadap *Covid-19* varian *Omicron* dapat dilakukan beberapa cara agar diperoleh hasil yang maksimal yaitu dengan cara mendetailkan *preprocessing* data, sehingga memungkinkan data akan berjalan efektif dan efisien, karena data yang telah melalui preprocessing data, merupakan data yang telah melalui tahap pembersihan dan siap digunakan. Selain itu melakukan distribusi *dataset* sehingga memiliki jumlah data latih dan data uji proporsional yang sesuai untuk menunjang dalam proses klasifikasi.

DAFTAR PUSTAKA

[1] Saraswati. N.S. 2011. "Text Mining dengan Metode Naïve Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis". Program Studi Teknologi Informasi Fakultas Teknik UGM Yogyakarta.

[2] Amalia H. 2021. "Omicron Penyebab COVID-19 sebagai Variant of Concern". Jurnal Biomedika dan Kesehatan. Vol 4(4).

[3] Dyer O. Covid-19: "Omicron is causing more infection but fewer hospital admissions than delta, South African data show". BMJ 2021; 375:n3104. Doi: 10.1136/bmj.n3104.

[4] Samsir., Ambiyar., Verawardina. U., Edi. F., Watrianthos. R., 2021. "Analisis Sentimen Pembelajaran Daring pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes". Jurnal Media Informatika Budidarma. Vol.5(1) Hal 157-163.

[5] Anees, Aiman Abdullah et al. 2019. "Performance Analysis of Multiple Classifiers Using Different Term Weighting Schemes for Sentiment Analysis." 2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019 (Iciccs): 637–41.

[6] Prabhavathi, C. 2019. "Machine Learning Model for Classifying L _ Text Using Nlp (Amazon Product Reviews)." 6(04): 161–78.

[7] Chemchem, Amine, François Alin, and Michael Krajecki. 2019. "Improving the Cognitive Agent Intelligence by Deep Knowledge Classification." International Journal of Computational Intelligence and Applications 18(1): 1–25.

[8] Zhou, Zhenxiang, and Lan Xu. 2016. "Amazon Food Review Classification Using Deep Learning and Recommender System." Stanford University: 1–7.

[9] H. Irsyad dan A. Taqwiym, "Community Analysis Sentiment Against Palestinian People with Naive Bayes Classification," JTECS : Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem dan Komputer, vol. 1, no. 2, 2021, doi: 10.32503/jtecs.v1i2.1623.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. 2017. Attention is All you need. Advances in Neural Information Processing Systems 2017-Decem(Nips), 5999-6009.

[11] Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1(Mlm): 4171–86.

[12] Prasetyo, Eko. 2014. Data Mining Mengolah Data Menjadi Informasi menggunakan Matlab. Yogyakarta: ANDI

[13] Sugiyono. 2015. Metode Penelitian Kuantitatif, Kualitatif, dan R&D. 22 ed. Bandung: Alfabeta.

doi: https://doi.org/10.32503/jtecs.v3i1.3219