

## Implementasi Metode CNN pada Klasifikasi Sentimen terhadap Pelaksanaan Piala Dunia U-17

(Implementation of the CNN Method in Classifying Sentiments Regarding the Implementation of the U-17 World Cup)

Sihol Paabanan Simanjuntak\*, Sunneng Sandino Berutu, Gogor C. Setyawan

Program Studi Informatika, Fakultas Sains dan Komputer, Universitas Kristen Immanuel, Yogyakarta, Indonesia

\*Corresponding author: sihol.p19@student.ukrimuniversity.ac.id

(Received 7 February 2024; Accepted 28 June 2024; Available online 29 June 2024)

### ABSTRACT

The World Cup is the world's largest tournament and is organized by the Federation International de Football Association (FIFA). Football is one famous sport in Indonesia; after the information that the U17 World Cup will be held in Indonesia, of course, many people talk about this, and it become a trending topic on Twitter. One way to analyze and process texts is through sentiment analysis. The purpose of this sentiment analysis is to identify positive, neutral, and negative sentences that originate from Twitter comments. The study also aims to evaluate the effectiveness of the Convolutional Neural Network (CNN) model in classifying sentimental sentences into positive, neutral and negative categories. The research implemented the CNN method. Tweet data was obtained through Google Collab. Sentiment analysis using CNN algorithms involves the results of two different data labelling methods. The results of the sentiment analysis with the CNN model using TextBlob labelling, showed 367 positive tweets, 253 neutral tweets and 48 negative tweets, with an accuracy of 89.67%, precision of 90.18%, recall of 89,67% and F1 score of 89.00%. The sentiment analysis results with the CNN using Vader Sentiment labelling, showing 80 positive tweets, 582 neutral tweets and 6 negative tweets, with an accuracy of 97.01%, precision of 96.21%, recall of 97.01% and F1 score of 96.46%. Results of sentiment analysis using the CNN model and 2 data labelling methods resulted in the following accuracy: the accuracy Model CNN with the textBlob data and the Vader Sentiment labelling were 89% and 97%, respectively.

### ABSTRAK

Piala Dunia adalah ajang kompetisi terbesar yang diprakarsai Federation International de Football Association (FIFA). Salah satu cabang olahraga yang terkenal yaitu sepak bola. Setelah adanya informasi bahwa Piala Dunia U17 akan dilaksanakan di Indonesia tentunya banyak yang membicarakan hal ini dan menjadi trending topik di twitter. Salah satu cara untuk menganalisis dan mengolah teks adalah melalui analisis sentimen. Tujuan dari analisis sentimen ini mengidentifikasi kalimat-kalimat positif, netral, dan negatif yang berasal dari komentar twitter. Penelitian ini juga bertujuan untuk mengevaluasi keefektifan model Convolutional Neural Network (CNN) dalam mengklasifikasikan kalimat-kalimat sentimen ke dalam kategori positif, netral dan negatif. Metode CNN diterapkan dalam penelitian ini. Tweet diperoleh melalui Google Collab. Analisis sentimen dengan menggunakan algoritma CNN melibatkan hasil dari dua metode pelabelan data yang berbeda. Hasil analisis sentimen dengan Model CNN menggunakan pelabelan TextBlob, menunjukkan 367 tweet positif, 253 tweet netral dan 48 tweet negatif, dengan akurasi 89,67%, presisi 90,18%, recall 89,67% dan F1 score 89,00%. Hasil analisis sentimen dengan Model CNN menggunakan pelabelan VaderSentiment, menunjukkan 80 tweet positif, 582 tweet netral, dan 6 tweet negatif. dengan akurasi 97,01%, presisi 96,21%, recall 97,01% dan F-1 Score 96,46%. Hasil penelitian analisis sentimen menggunakan model CNN dan 2 metode pelabelan data menghasilkan accuracy sebagai berikut; Akurasi Model CNN dengan pelabelan data TextBlob dan VaderSentiment masing-masing sebesar 89% dan 97%.

### Keywords:

FIFA;  
U17 World Cup;  
Twitter;  
Sentiment  
Analysis;  
Convolutional  
Neural Network

### Kata kunci:

FIFA;  
Piala Dunia U17;  
Twitter;  
Analisis  
Sentimen;  
Convolutional  
Neural Network

## 1. Pendahuluan

Piala Dunia adalah ajang kompetisi terbesar yang diprakarsai oleh *Federation International de Football Association* (FIFA). Salah satu cabang olahraga terkenal yaitu sepak bola, dari anak-anak sampai dewasa. Setelah adanya informasi bahwa kegiatan ini akan dilaksanakan di Indonesia tentunya banyak yang membicarakan hal ini dan menjadi trending topik di twitter. Secara resmi, FIFA telah menunjuk Indonesia sebagai penyelenggara Piala Dunia U-17 sejak Juni 2023 dalam sidang Council di Zurich, Swiss. Hal ini berdampak baik bagi Indonesia sebagai batu loncatan agar Indonesia lebih dikenal oleh negara luar, baik dari dunia sepak bola Indonesia dan berbagai bidang lainnya. Keputusan FIFA menunjuk Indonesia sebagai tuan rumah menuai banyaknya opini masyarakat maupun pemerintah. Indonesia menjadi negara Asia Tenggara pertama yang menjadi tuan rumah Piala Dunia U17. Selain itu turnamen tersebut menjadi *event* ternama FIFA pertama yang akan digelar Indonesia, setelah sebelumnya Piala Dunia U20 dibatalkan. Dalam konteks ini, diperlukan analisis sentimen dengan tujuan mendapatkan hasil atau pandangan masyarakat terhadap keputusan FIFA dengan metode *Convolutional Neural network* telah terbukti efektif untuk tugas ini dan mampu mencapai kinerja yang baik dalam mengklasifikasikan kalimat (Listyarini & Anggoro, 2021). Penelitian yang bertujuan pada platform Twitter untuk mendeteksi sarkasme yang menunjukkan model CNN memiliki kemampuan yang kuat dalam mengekstraksi fitur dan cocok untuk analisis opini (Rhanoui et al., 2019). Penelitian lain juga menunjukkan bahwa model CNN yang menggunakan lapisan *convolutional* dan *max pooling* mampu mengekstrak fitur dengan tingkat keakuratan yang lebih tinggi (Rehman et al., 2019).

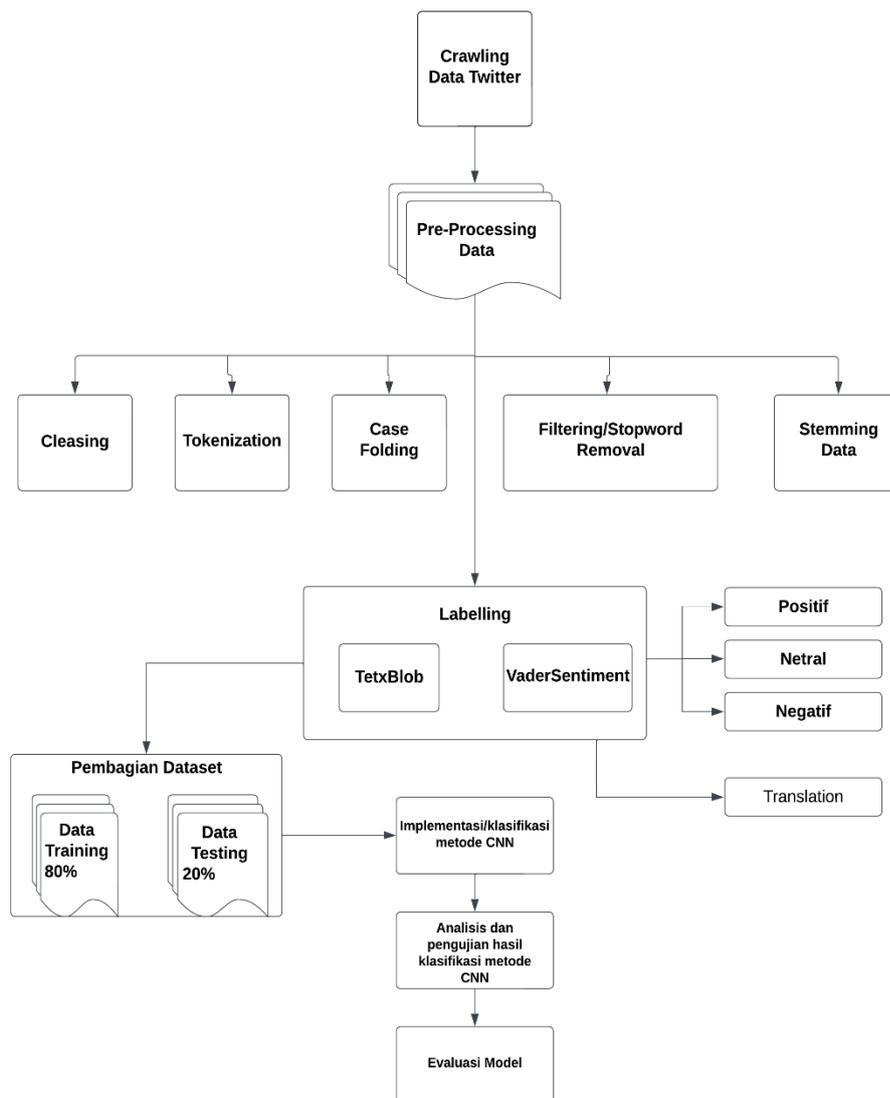
Untuk itu penulis melakukan analisis sentimen terhadap pelaksanaan Piala Dunia U-17 dengan data yang diambil dari Twitter menggunakan *Google Colab* dengan bahasa pemrograman *Python*, dengan menggunakan metode *Convolution Neural Network* (CNN). Penulis memilih metode ini karena memiliki kemampuan yang tinggi dalam mengekstraksi fitur. Hasil penelitian ini mencakup akurasi dan dampak dari jumlah *epoch* yang digunakan dalam model, serta hasil label prediksi pada pengujian kemampuan terhadap input baru.

## 2. Metodologi

Permasalahan dalam klasifikasi sentimen melibatkan identifikasi sentimen dalam sebuah teks, apakah teks, apakah teks itu positif, negatif atau netral. Gambaran umum dari penelitian ini adalah sebagai berikut (Gambar 1). Tahap awal melibatkan pengumpulan data twitter, yang kemudian diikuti oleh proses pre-processing data seperti pembersihan, tokenisasi, penyamarataan huruf, penyaringan, pemangkasan kata dan penerjemahan. Selanjutnya, dataset dipisahkan menjadi dua bagian, yaitu 80% untuk data pelatihan dan 20% untuk data pengujian. Kemudian, model dibangun dengan algoritma CNN. Terakhir, model dilatih dan diuji. Hasil pengujian dievaluasi dengan metode confusion matrix.

**Table 1.** Hasil Crawling Data Twitter

Tanggal	Tweet
19-08-2023	PSSI mematangkan persiapan tim utk Piala Dunia U17 agar mendapat hasil terbaik #KitaGaruda
08-11-2023	Polri menjamin pelaksanaan Piala Dunia U-17 di Indonesia dapat berjalan dengan aman. #PialaDuniaU17
08-11-2023	Menyuksesan gelaran FIFA World Cup U17, PLN memastikan kesiapan kelistrikan di seluruh venue stadion utama yaitu Jakarta International Stadium (JIS), Si Jalak Harupat Bandung, Manahan Solo, dan Gelora Bung Tomo Surabaya. #PialaDuniaU17



Gambar 1. Alur Penelitian

## 2.1. Crawling Data

Data diperoleh dengan menggunakan bahasa pemrograman Python. *Crawling data* menggunakan *harvest* di dalam *google collab* yang bisa diakses secara online. Hasil *crawling Twitter* berjumlah 7920 data, seperti Tabel 1.

## 2.2. Pre-processing data

*Pre-processing* data adalah serangkaian langkah atau teknik yang dilakukan untuk membersihkan, mengubah, atau mengorganisasi data mentah (*raw data*) sebelum data tersebut diolah atau dianalisis lebih lanjut. Tujuan utama dari *pre-processing* data adalah untuk meningkatkan kualitas data, menghilangkan *noise* atau gangguan, dan membuat data siap digunakan dalam analisis atau pemodelan.

### 2.2.1 Cleansing

Dalam melakukan *cleansing* penting untuk memperhatikan beberapa hal termasuk identifikasi data yang tidak valid, identifikasi data yang tidak akurat, identifikasi data yang tidak lengkap, identifikasi data yang duplikat normalisasi, validasi data dan beberapa karakter yang dihapus seperti tanda baca, *username*, *url*, *mention*, *hashtag*, *retweet* serta *symbol* atau karakter numerik seperti : (, ~ & ? ! > < # % { } [ ( [ 0 - 9 ] + ; ; " ) [ 1 1 2 3 ] , seperti ditampilkan dalam Table 1.

**Table 2.** Hasil Cleansing

Text Asli	Sesudah Cleansing
PSSI mematangkan persiapan tim untuk #Piala Dunia.	PSSI mematangkan persiapan tim untuk Piala Dunia

### 2.2.2 Tokenisasi

Tokenisasi adalah proses mengurai sebuah teks atau kalimat menjadi bagian kecil, yang disebut sebagai "token." Token adalah unit terkecil yang dapat berupa kata, frasa, atau simbol yang memiliki makna dalam bahasa tertentu. Tujuan utama dari tokenisasi adalah untuk memproses teks atau kalimat menjadi bentuk yang lebih mudah diolah dalam analisis teks, pemrosesan bahasa alami, atau tugas-tugas lainnya. Contoh hasil tokenisasi seperti Tabel 3.

**Table 3.** Hasil tokenisasi

Text Asli	Sesudah Tokenisasi
PSSI mematangkan persiapan tim untuk Piala Dunia.	'PSSI','mematangkan','persiapan','tim','untuk','Piala','Dunia'

### 2.2.3 Case Folding

*Case folding* adalah proses pemrosesan teks yang mengubah semua karakter dalam teks menjadi huruf kecil/ besar, tergantung pada konvensi yang dipilih

**Table 4.** Hasil *Case Folding*

Text Asli	Sesudah Case Folding
PSSI mematangkan persiapan tim untuk Piala Dunia..	psSI mematangkan persiapan tim untuk piala dunia

### 2.2.4 Filtering/Stopword Removal

*Filtering* atau *stopword removal* adalah salah satu tahap dalam pemrosesan yang bertujuan untuk menghilangkan kata-kata yang disebut "*stopwords*" dari sebuah teks. *Filtering* bertujuan untuk menghilangkan kata yang tidak memiliki arti penting, yang mungkin mempengaruhi kecepatan dan kinerja proses analisis. Dalam penelitian ini *Filtering* menggunakan library *pythonsastrawi*. Beberapa contoh *stopword* dalam *python sastrawi* adalah 'yang', 'untuk', 'pada', 'dan', 'ke', 'dari', dan sebagainya.

### 2.2.5 Stemming

*Stemming* data adalah proses untuk menghilangkan awalan atau akhiran sehingga menyisakan akar kata. Tujuan utama dari *stemming* adalah untuk menghapus infleksi kata seperti akhiran (*suffixes*) atau awalan (*prefixes*) sehingga kata-kata yang memiliki akar yang sama akan dianggap sama, meskipun mereka mungkin memiliki bentuk yang berbeda. Contoh hasil *stemming* seperti Tabel 5.

**Table 5.** Hasil *Stemming*

Text Asli	Sesudah <i>Stemming</i>
PSSI mematangkan persiapan tim untuk Piala Dunia..	psSI matang siap tim untuk piala dunia

### 2.2.6 Translation

Terjemahan adalah proses mengubah teks sambil mempertahankan arti, konteks, dan niat dari konten aslinya.

### 2.3. Labelling Data

Penandaan data, atau yang dikenal sebagai pelabelan data merupakan langkah dimana label atau kategori diberikan kepada data awal guna menyediakan konteks dan informasi yang signifikan dalam konteks pembelajaran mesin. Tujuan dari proses ini adalah membantu mesin memahami dataset dan mengidentifikasi pola atau karakteristik khusus yang dapat dimanfaatkan untuk melakukan prediksi atau pengambilan keputusan.

#### 2.3.1 *TextBlob*

*TextBlob* merupakan sebuah perpustakaan *Python* yang digunakan untuk pemrosesan teks dan analisis bahasa alami (*Natural Language Processing/NLP*). Pustaka ini dapat digunakan untuk melakukan pelabelan data dengan menerapkan fungsi analisis sentimen yang disediakan. Fungsi sentimen dalam *TextBlob* memberikan skor sentimen pada suatu teks berdasarkan analisis yang dilakukan oleh algoritma yang telah dilatih sebelumnya. *TextBlob* menyediakan beragam fitur dan fungsi yang mempermudah manipulasi dan analisis teks kedalam kategori ‘negatif’, ‘positif’, dan ‘netral’ berdasarkan sentimen yang terkandung, *TextBlob* umumnya digunakan untuk analisis sentimen umum pada teks, memberikan skor polaritas sentimen dalam rentang dari -1 hingga 1. Nilai negatif menunjukkan nilai sentiment negatif, nilai positif menunjukkan nilai sentiment positif sementara nilai nol menunjukkan nilai netral

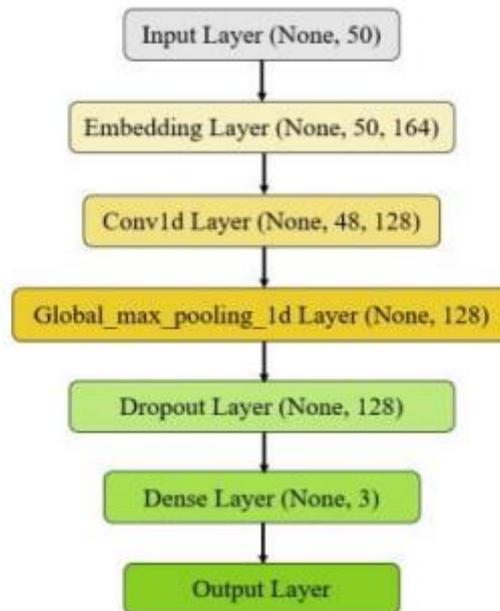
#### 2.3.2 *VaderSentiment*

*Vader (Valence Aware Dictionary and Sentiment Reasoner)* Sentimen analisis adalah sebuah algoritma analisis sentimen yang dikembangkan khusus untuk memahami sentimen dalam teks yang berasal dari media sosial. Algoritma ini dapat mengekstraksi dan mengukur tingkat sentimen positif, negatif dan netral dari suatu teks. Dibandingkan metode analisis sentimen tradisional yang mengandalkan kamus kata-kata berlabel sentimen, *VADER Sentiment Analysis* menggunakan kamus sentimen yang telah dikurasi secara spesifik untuk menangani sentimen umum dalam teks media sosial, yang sering kali mengandung kosakata informal, kata-kata berimbuhan, emotikon, dan tanda baca.

### 2.4. Pembagian (*Split*) Dataset

Split data adalah cara membagi data set agar menjadi dua bagian terpisah yaitu subset utama yang digunakan dalam *machine learning* dan analisis data. Data *training* adalah subset dari dataset yang digunakan untuk melatih model *machine learning*. Pada tahap ini, model belajar pola dan relasi dalam data untuk memahami dan memprediksi hasil tertentu. Tujuan dari data *training* digunakan untuk mengejar model. *Data Testing* adalah data yang tidak digunakan selama proses pelatihan. Setelah model selesai dilatih dengan data *training*, data *testing* digunakan untuk menguji sejauh mana model dapat melakukan prediksi yang akurat pada data yang belum pernah dilihat

sebelumnya. Dalam penelitian ini, pembagian dataset mengikuti rasio 80/20, di mana 80% digunakan untuk data training dan 20% untuk data testing.



**Gambar 2.** Arsitektur CNN

## 2.5. Membangun Model

Dalam penelitian ini model dibangun berbasis algoritma CNN dengan arsitektur yang ditunjukkan pada Gambar 2. Arsitektur terdiri dari beberapa layer seperti *input layer*, *embedding layer*, *conv1D layer*, *global max pooling layer*, *dropoutlayer*, *dense layer* dan *output layer*. Dengan beberapa layer ini model dapat mempelajari pola dan hubungan antar kata dalam teks dan menghasilkan prediksi kelas atau label yang akurat untuk tugas klasifikasi teks. Proses ini melibatkan proses pelatihan dengan menggunakan data teks yang telah dilabeli (labeled text dataset).

## 2.6. Evaluasi Model

Evaluasi akan dipergunakan untuk mengetahui seberapa bagus metode yang digunakan. Caranya adalah dengan melakukan perhitungan nilai performa berdasarkan *accuracy*, *recall*, *precision*, dan *f1-score* (Berutu et al., 2023). Pada penelitian ini, model *Confusion Matrix* menjadi pilihan. Tabel 6 memperlihatkan deskripsi sederhana mengenai *confusion matrix*

Table 6. *Confusion Matrix*

Kelas	Aktual Positif	Aktual Negatif	Aktual Netral
Aktual Positif	<i>True Positive (TP)</i>	<i>False Negative1 (FNg1)</i>	<i>False Netral1 (FNt1)</i>
Aktual Negatif	<i>False Positive1 (FP1)</i>	<i>True Negative (TNg)</i>	<i>False Netral2 (FNt2)</i>
Aktual Netral	<i>False Positive2 (FP2)</i>	<i>False Negative2 (FNg2)</i>	<i>True Netral (TNt)</i>

Keterangan :

1. *True Negative* = data asli yang berada di kelas negatif dan model memprediksi negative
2. *True Positive* = data asli yang berada di kelas positif dan model memprediksi positif
3. *False Negative* = data asli yang berada di kelas positif, namun model memprediksi negatif
4. *False Positive* = data asli yang berada di kelas negatif, namun model memprediksi positif

Evaluasi model mendapatkan nilai akurasi, *recall*, *precision*, dan F1-Score dari empat kemungkinan *confusion matrix*.

- a. Akurasi = Total keseluruhan dari nilai benar model melakukan klasifikasi. Rumus akurasi dapat dilihat dibawah ini:

$$Accuracy = \frac{TP + TNg + TNt}{TP + FN1 + \dots + FN2 + TNt} \times 100\% \quad (1)$$

- b. *Precision* = Seberapa sering prediksi benar saat model melakukan prediksi positif. Rumus *Precision* dapat dilihat dibawah ini:

$$\text{Positif} = \frac{TP}{FP1 + FP2 + TP}$$
$$\text{Negatif} = \frac{TNg}{FNg1 + FNg2 + TNg} \quad (2)$$

$$\text{Positif} = \frac{TNt}{FNt1 + FNt2 + tnT}$$

- c. *Recall* = Seberapa sering model melakukan prediksi positif saat kelas aktualnya positif. Rumus *Recall* dapat dilihat dibawah ini:

$$\text{Positif} = \frac{TP}{FPg1 + FNt1 + TP}$$
$$\text{Negatif} = \frac{TNg}{FP1 + FNt2 + TNg} \quad (3)$$

$$\text{Positif} = \frac{TNt}{FP2 + FNg2 + TNt}$$

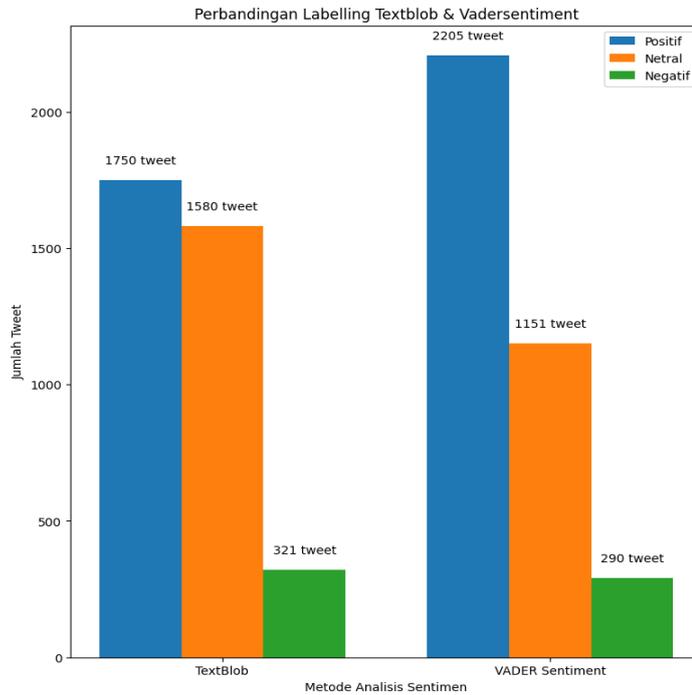
- d. *F1-Score* = Rata-Rata harmonik dari nilai *Precision* dan *Recall*. Rumus *F1-Score* dapat dilihat dibawah ini:

$$Recall = 2X \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

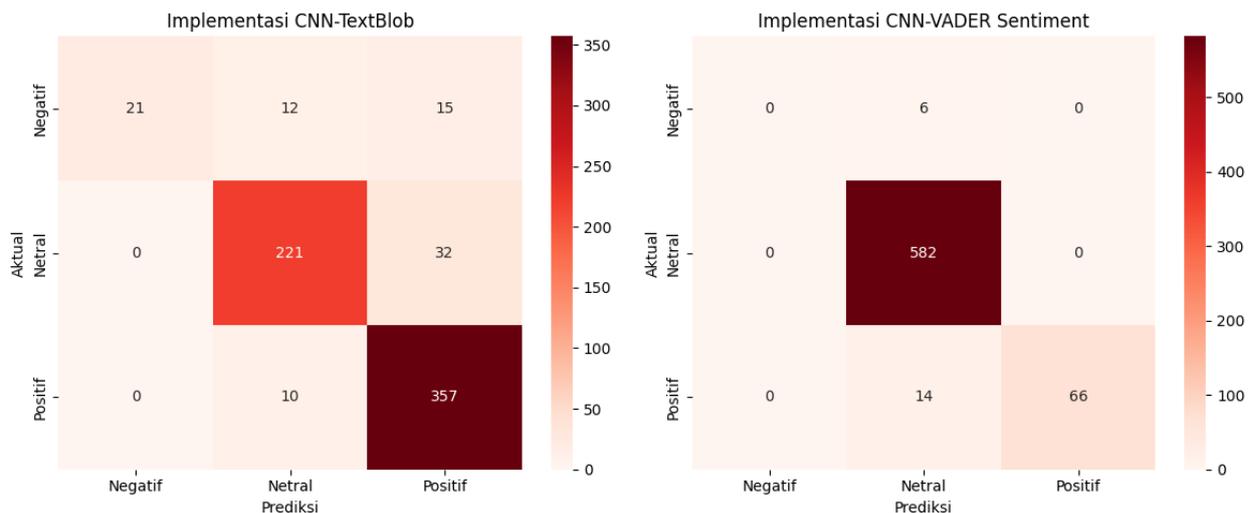
### 3. Hasil dan Pembahasan

Sebanyak 7920 data *tweet* setelah dilakukan *pre-processing* data menjadi 3646 *tweet*. Selanjutnya data akan dilakukan pelabelan dan hasil pelabelan dapat dilihat pada Gambar 3.

Berdasarkan hasil pelabelan menggunakan *TextBlob*, data yang paling banyak adalah *tweet* positif sebanyak 1745 *tweet*, di ikuti dengan *tweet* netral sebanyak 1580 *tweet* dan 321 *tweet* negatif. Dan berdasarkan hasil pelabelan menggunakan *VaderSentiment*, data yang paling banyak adalah *tweet* positif sebanyak 2205 diikuti dengan *tweet* netral sebanyak 1151 *tweet* dan *tweet* negatif sebanyak 290 *tweet*.



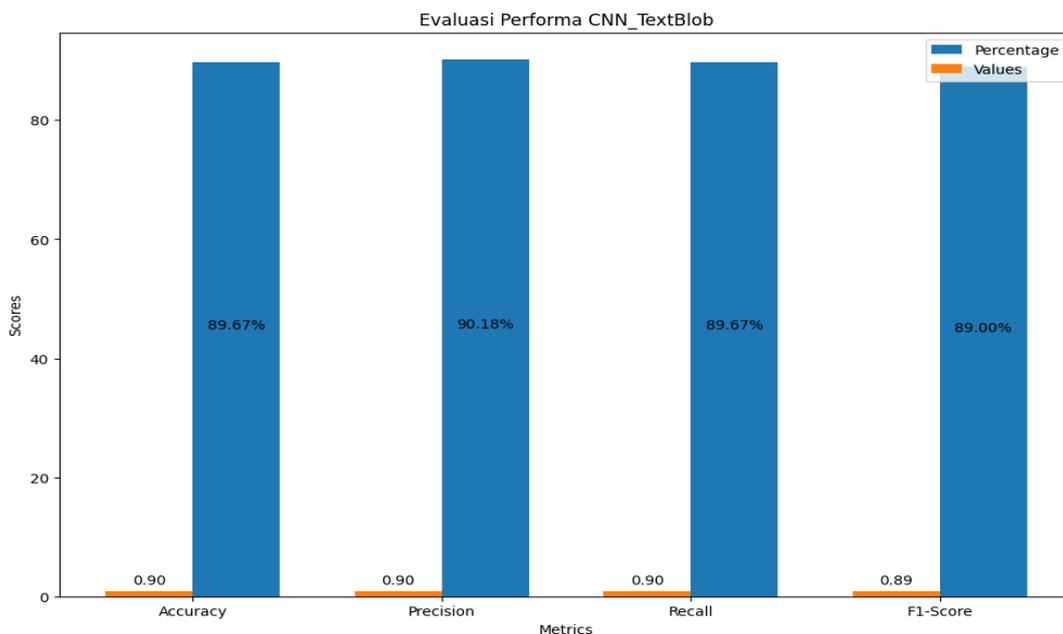
**Gambar 3.** Perbandingan Pelabelan *Textblob* & *Vadersentiment*



**Gambar 4.** Hasil Uji Model CNN menggunakan *TextBlob* & *Vadersentiment*

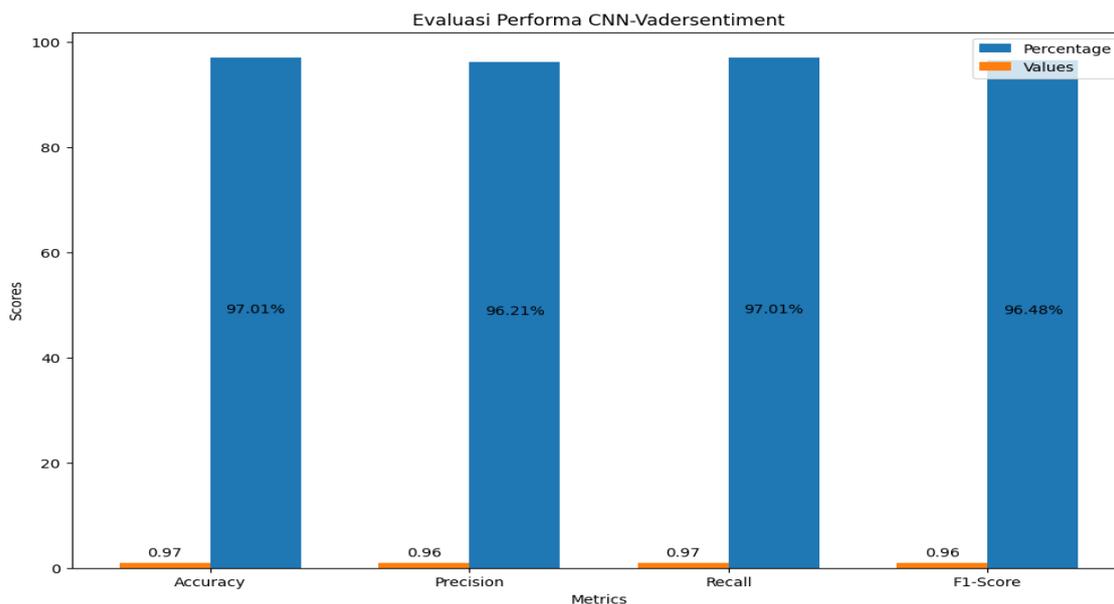
Selanjutnya implementasi dilakukan di setiap metode pelabelan menggunakan model *CNN*. Hasil pelabelan *TextBlob* yang diuji oleh model *CNN*, *tweet* yang prediksi negatif adalah 21 *tweet*, dan *tweet* yang diprediksi netral sebanyak 221 *tweet*, dan *tweet* yang diprediksi positif adalah 357 *tweet* (lihat Gambar 4). Hasil pelabelan *VaderSentiment* yang diuji oleh model *CNN* *tweet* yang paling banyak benar diprediksi adalah *tweet* netral sebanyak 582, diikuti dengan *tweet* positif sebanyak 66 *tweet* dan *tweet* negatif sebanyak 0.

Selanjutnya performa model dievaluasi pada pelabelan data *textblob*. Data hasil *labelling Textblob* yang diuji oleh model *CNN* menghasilkan *Precision* dengan hasil tertinggi dengan nilai 90% dan *Accuracy* 89,67%, *Recall* 89,67% dan *F1-Score* 89% dan hasilnya dapat dilihat pada Gambar 5.



**Gambar 5.** Evaluasi CNN-Textblob

Selanjutnya performa model dievaluasi pada pelabelan data *Vadersentiment*, Data hasil labelling *VaderSentiment* yang diuji oleh model *CNN* menghasilkan *Accuracy* dan *Reccal* dengan hasil tertinggi dengan nilai 97,01%, *FI-Score* 96,48% dan *Precision* 96,21% dan hasilnya dapat dilihat pada Gambar 6.



**Gambar 6.** Evaluasi CNN-Vadersentiment

#### 4. Kesimpulan

Berdasarkan hasil pelabelan dengan *Textblob*, tweet sentiment positif memperoleh persentase tertinggi sebesar 47,9% di ikuti tweet sentimen netral sebesar 43.3% dan sentimen negatif sebesar 8,8%. Sementara itu hasil pelabelan dengan *Vadersentiment* diperoleh persentase tertinggi pada tweet sentiment positif sebesar 60,5% diikuti dengan tweet netral sebesar 31,6% dan tweet negatif sebesar 8,0%. Selanjutnya hasil pengujian model dengan dataset pelabelan *TextBlob*, menunjukkan persentase nilai *accuracy* sebesar 89,67% ,*precision* sebesar 90.18%, *recall* sebesar

89,67% dan *f1-score* sebesar 89.00%. Sedangkan dengan dataset VaderSentiment, menunjukkan nilai *accuracy* sebesar 97,01%, *precision* sebesar 96,21%, *recall* sebesar 97,01% dan *f1-score* sebesar 96,46%.

Untuk memperoleh informasi lebih detail terkait hasil sentimen maka perlu dilakukan penelitian lebih lanjut dalam pengukuran sentimen berbasis aspek

## Referensi

- Listyarni, S. N., & Anggoro, D. A. 2021. analisis sentimen pilkada di tengah pandemi Covid-19 Menggunakan Convolution Neural Network (CNN). *Jurnal Pendidikan Dan Teknologi Indonesia*, 1(7), 261–268. <https://doi.org/10.52436/1.jpti.60>
- Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. 2019. A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, 78(18), 26597–26613. <https://doi.org/10.1007/s11042-019-07788-7>
- Rhanoui, M., Mikram, M., Yousfi, S., & Barzali, S. 2019. A CNN-BiLSTM Model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction*, 1(3), 832–847. <https://doi.org/10.3390/make1030048>
- Widayat, W. 2021. Analisis sentimen movie review menggunakan Word2Vec dan metode LSTM Deep Learning. *Jurnal Media Informatika Budidarma*, 5(3), 1018. <https://doi.org/10.30865/mib.v5i3.3111>
- Reza, M., Pulungan, U., Ratnawati, D. E., & Rahayudi, B. 2022. Analisis sentimen ulasan aplikasi pedulilindungi dengan Metode Random Forest. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 6(9), 4378–4385. <http://j-ptiik.ub.ac.id>
- Irawan, F. A., & Rochmah, D. A. 2022. Penerapan Algoritma CNN untuk mengetahui sentimen masyarakat terhadap kebijakan vaksin Covid-19. *Jurnal Informatika*, 9(2), 148–158. <https://doi.org/10.31294/inf.v9i2.13257>
- Berutu, S. S., Budiati, H., & Gulo, F. 2023. Data preprocessing approach for machine learning-based sentiment classification. *Jurnal Infotel*, 5(4), 317–325.