Vol. 4, No. 1, April 2023, hal. 1-6

Jurnal Computer Science and Information Technology (CoSciTech)

p-ISSN: 2723-567X
e-ISSN: 2723-5661

http://ejurnal.umri.ac.id/index.php/coscitech/index

Klasifikasi multilabel komentar toxic pada sosial media twitter menggunakan convolutional neural network (CNN)

Regiolina Hayami [*1], Sofhia Mohnica[2], Soni[3]

Email: [1]regiolinahayami@umri.ac.id, [2]180401173@student.umri.ac.id, [3]soni@umri.ac.id

[12]Teknik Informatika, Ilmu Komputer, Universitas Muhammadiyah Riau

Diterima: 3 Desember 2023 | Direvisi: 19 April 2023 | Disetujui: 28 Mei 2023

Abstrak

Meningkatnya jumlah pengguna dari media sosial berarti jumlah konten akan meningkat. Apalagi pengguna media sosial yang membuat kontennya menarik cenderung ingin ditanggapi atau mendapat pengakuan dari pengguna lain, baik itu berupa suka maupun komentar. Tak jarang komentar berisi berisi kata-kata ancaman, cabul, penghinaan atau kebencian terhadap identitas atau disebut juga dengan komentar beracun. Meskipun ada peraturan yang mengatur semua aktivitas di media sosial, namun tetap saja tidak bekerja secara efektif karena ketidakmungkinan mengklasifikasikan komentar secara manual. Tujuan dari penelitian ini adalah membangun sebuah model klasifikiasi multilabel yang dapat mengklasifikasikan ke dalam kategori nya menggunakan algoritma *Convolutional Neural Network* serta *Word2Vec* yang digunakan sebagai pembobotan kata. Pada penelitian ini menghasilkan model klasifikasi dengan Nilai performa dari pengujian model mesin pembelajaran CNN dengan menggunakan optimizer adam menghasilkan akurasi sebesar 99%, presisi 100%, *recall* 99% dan *F1-Score* 99%.

Kata kunci: *convolutional neural network, klasifikasi, klasifikasi multilabel, komentar beracun, pembelajaran mesin, word2vec*

Multilabel classification of toxic comments on social media twitter using a convolutional neural network (CNN)

Abstract

The increasing number of users of social media means the amount of content will increase. Moreover, social media users who make their content interesting tend to want a response or recognition from other users, either in the form of likes or comments. Not infrequently comments contain words of threats, obscenity, insults or hatred of identity or also called toxic comments. Even though there are rules governing all activity on social media, they still don't work effectively due to the impossibility of manually classifying comments. The purpose of this study is to build a multilabel classification model that can classify into categories using the Convolutional Neural Network algorithm and Word2Vec which is used as word weighting. This study produces a classification model with performance values from testing the CNN machine learning model using the adam optimizer resulting in 99% accuracy, 100% precision, 99% recall and 99% F1-Score.

Keywords: *convolution neural netwrok, classification, multilabel classification, machine learning, toxic comment, word2vec*

1. PENDAHULUAN

Pengguna internet di Indonesia sudah mencapai jumlah 202,6 juta pengguna atau sekitar 73,7% dari total jumlah penduduk Indonesia[1]. Meningkatnya jumlah pengguna dari media sosial berarti jumlah konten akan meningkat. Apalagi pengguna media sosial yang membuat kontennya menarik cenderung ingin ditanggapi atau mendapat pengakuan dari pengguna lain, baik itu

berupa suka maupun komentar. Dengan begitu jumlah komentar akan semakin banyak. Terkadang komentar dan diskusi terbuka bisa memicu perdebatan, bisa karena perbedaan pendapat atau karena kesal dengan konten yang disajikan. Namun seringkali perdebatan yang terjadi muncul hal-hal yang tidak baik dan menggunakan cara-cara yang kotor untuk berdebat. Cara kotor dapat menyebabkan pertengkaran besar di media sosial, jadi gunakan komentar beracun untuk melakukan serangan.

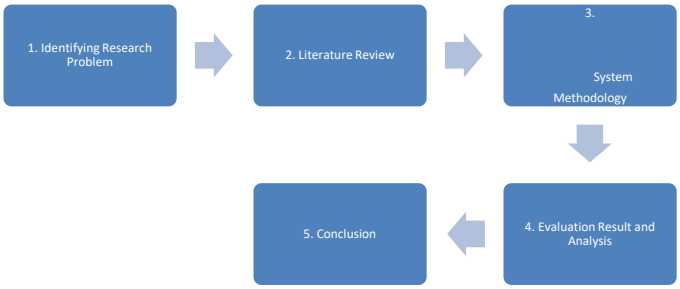
Toxic Comment dapat berisi kata-kata ancaman, cabul, penghinaan atau kebencian terhadap identitas, sehingga akan menimbulkan pelecehan di media sosial, atau biasa disebut pelecehan online. Akibat dari tindakan pelecehan tersebut, sebagian orang akan berhenti memberikan pendapat atau berusaha menghindari perdebatan di media sosial yang berujung pada diskusi yang tidak sehat dan tidak adil [2]. Contoh dari Toxic Comment sepertimenyebarkan kebencian, pencemaran nama baik, pornografi, radikalisme, SARA (Suku, Ras, dan Agama), dan lain sebagainya [3].

Berdasarkan pemaparan fakta dan permasalahan di atas, diperlukan adanya penelitian dalam klasifikasi *toxic comment*, penelitian mengenai klasifikasi toxic comment sebelumnya pernah dilakukan oleh Kemudian [4] mengambil data dari Kaggle dengan membandingkan hasil performa dari metode klasifikasi Long Short Term Memory (LSTM) dan Naive Bayes. Dimana hasil menunjukkan bahwa penggunaan metode LSTM lebih unggul daripada metode Naive Bayes, yaitu sebesar 64% dan 73%. Kemudian [5] dimana pada penelitian ini digunakan dataset berjumlah 1.500 yang diambil dari media sosial Facebook dan dari dataset tersebut dibagi menjadi 2 kelas yaitu kelas toxic, dan non-toxic, pada penelitian tersebut hanya untuk menentukan komentar termasuk kategori toxic comment atau bukan, penelitian ini menggunakan Naive Bayes dengan transformasi TF-IDF dan menghasilkan akurasi sebesar sebesar 75%, precision sebesar 63%, recall sebesar 67%, dan F-measure sebesar 64%.

Berdasarkan penelitian sebelumnya yang telah dilakukan, penelitian ini akan dibangun klasifikasi multi-label, hal ini dilakukan agar analisis komentar tidak hanya membedakan antara *toxic comment* dan tidak *toxic comment*, tetapi akan di kategorikan ke dalam beberapa kategori yang terdapat pada dataset yaitu toxic, severe_toxic, obscene, threat, insult, identity_hate. Penelitian ini menggunakan algoritlma Convolutional Neural Network dan Word2Vec sebagai pembobotan kata.

2. METODE PENELITIAN

Metode penelitian yang digunakan pada penelitian ini dibagi menjadi beberapa tahap agar proses yang dilakukan bisa berjalan dengan baik. Tahapan-tahapan penelitian digambarkan sebagai berikut :



Gambar 1 Kerangka Konseptual Penelitian

2.1. Identifikasi Masalah

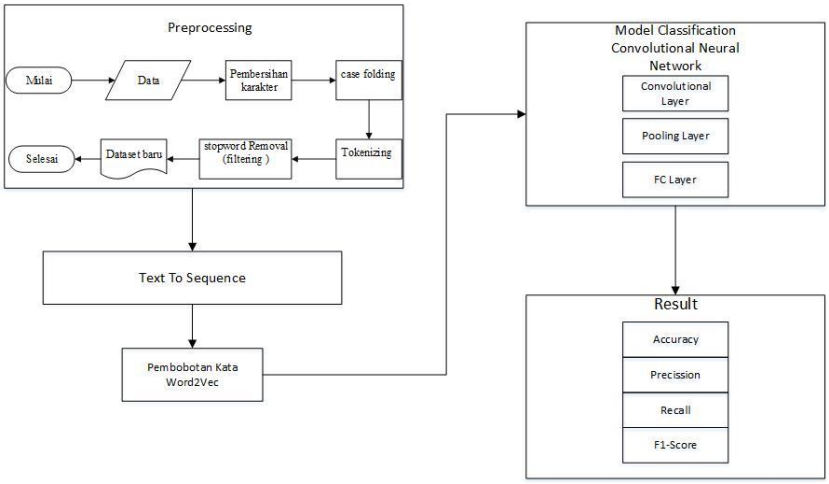
Identifikasi Masalah merupakan langkah awal yang dilakukan untuk memperoleh dan menentukan topik penelitian yang akan diteliti lebih lanjut. Pada tahapan ini dimulai dengan melihat berbagai fenomena, kejadian dan informasi yang didapatkan dengan berbagai cara. Dalam hal ini penulis melihat terdapat suatu dataset mengenai toxic comment terbaru yang masih bisa digunakan untuk pengujian klasifikasi toxic comment dengan tujuan untuk mengetahui kecocokan data dengan model klasifikasi dan juga performa machine learning yang akan dibuat.

2.2. Studi Literatur

Studi Literatur pada penelitian ini dilakukan dengan cara mengumpulkan serta mempelajari literatur-literatur yang berhubungan dengan toxic comment, algoritma convolutional neural network, dan word2vec. Sumber literatur yang digunakan berupa paper dan jurnal penelitian sebelumnya yang sesuai dengan topik penelitian ini.

2.3. Metodologi Sistem

Pada tahapan ini dilakukan untuk mengetahui bahan dan alur kerja dalam pelaksanaan penelitian. Adapun bahan yang digunakan dalam penelitian ini adalah *dataset* toxic comment yang telah disediakan oleh Kaggle dalam Toxic Comment Classification Challenge. Dataset berjumlah 159571 komentar. Pada *dataset* tersebut terdapat 6 kategori yaitu toxic, severe_toxic, obscene, threat, insult, identity_hate. Berikut ini merupakan alur kerja penelitian:



Gambar 2 Alur Kerja Penelitian

2.4. Analisis dan Evaluasi Hasil

Hasil dari implementasi dari algoritma CNN selanjutnya dilakukan proses evaluasi. Evaluasi dilihat tiga aspek yaitu precision, recall, dan accuracy. Tiga aspek tersebut yang akan digunakan untuk melihat keakurasian atau performa yang dihasilkan dari metode yang digunakan. Sehingga dapat menarik kesimpulan dari hasil penerapan algoritma tersebut.

2.5. Kesimpulan

Pada tahapan akhir ini dapat kesimpulan yang diperoleh adalah memuat bagaimana proses penerapan model *Multi-Label Classification* dan hasil dari performa dari algoritma yang diterapkan

3. HASIL DAN PEMBAHASAN

Bab ini menguraikan tentang tahapan implementasi lebih lanjut, dimulai dari tahapan preprocessing, pembobotan kata, pemodelan, pengujian.

3.1. Import Dataset

Pada penelitian ini terdapat 2 dataset yang akan digunakan, yaitu dataset Train yang digunakan untuk melatih model CNN dan dataset Test yang digunakan untuk validasi atau pengujian dari model CNN yang telah dibuat. Dataset train memuat komentar sebanyak 159571 sedangkan dataset test memuat komentar sebanyak 63978. Dengan jumlah masing masing label dengan toxic sebanyak 15249, severe_toxic 1595, obscene 8449, threat 478, insult 7877, dan identity hate 1405.

3.2. Preprocessing

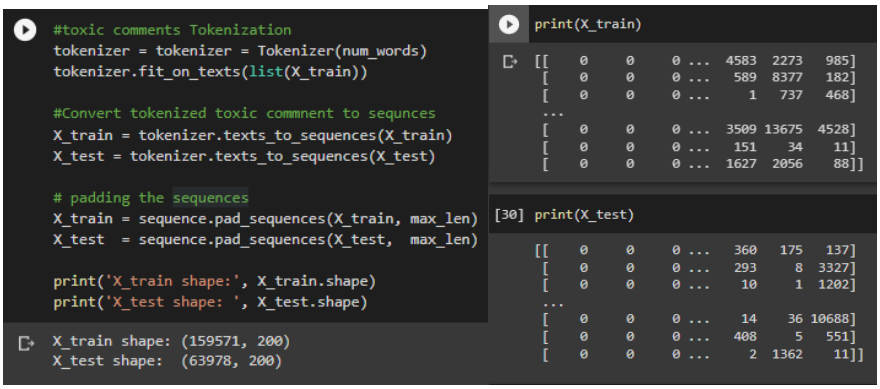
Preprocessing dilakukan pada kolom Tweet berisikan baris data berupa kalimat yang terdiri atas beberapa kata-kata. Hal tersebut disebabkan oleh data yang belum terstruktur dengan baik serta mengurangi jumlah noise atau gangguan [6], [7] . Agar data dapat digunakan, maka harus dilakukan beberapa tahapan preprocessing sebagai berikut [8]:

1. Pembersihan Karakter
   Proses ini bertujuan untuk adalah proses menghilangkan karakter pada dokumen text yang bukan alfabet seperti simbol, angka dan tanda baca.
2. Case Folding
   Proses ini merupakan penyamaan *case* dalam sebuah dokumen, tahap ini dilakukan untuk mengubah dokumen teks yang tidak konsisten dalam penggunaan huruf kapital. Pada tahap ini teks pada data akan diubah menjadi bentuk lowercase.
3. Tokenizing
   Tokenizing adalah proses memecah aliran teks menjadi kata, frasa, simbol, atau elemen bermakna lainnya yang disebut *token*. Ini adalah langkah wajib sebelum semua jenis pemrosesan. Proses ini akan membagi teks menjadi kalimat dan kalimat menjadi *token* tipografi.
4. Stopword Removal
   Proses penghapusan kata penghubung. Hal tersebut membuat dataset sulit untuk diolah lebih lanjut. Oleh karena itu pada tahap ini dilakukan penghilangan kata-kata tersebut untuk memperingkas dataset.

3.3. Text To Sequence

Pada tahap ini dataset akan diubah menjadi bentuk sequence atau bilangan tertentu yang kemudain akan dijadikan input pada proses Word2Vec. Pada tahap ini digunakan atribut tokenizer.texts_to_sequnces untuk mengubah teks menjadi sequence dan sequence.pad_sequences untuk membentuk padding. Padding adalah proses yang dilakukan untuk menyamakan panjang

sequence pada tiap data. Padding yang digunakan adalah post padding, yaitu menambahkan vektor kosong pada akhir kata hingga ukuran kalimat menjadi sama dengan kalimat terpanjang pada keseluruhan data.



Gambar 3 Text To Sequence

3.4 Pembobotan Kata Word2Vec

Word2Vec adalah salah satu model yang digunakan untuk vector representation of words atau word embeddings. Word2Vec mempelajari sekumpulan kalimat dengan mempertimbangkan kata-kata yang berdekatan untuk mengekstraksi makna kata [9]. Word2Vec memiliki dua jenis model arsitektur untuk merepresentasikan vektor kata, yaitu continuous bag of words (CBOW) dan skip-gram [10].

Pada proses pembobotan kata dilakukan training word2vec dengan menggunakan library gensim.Size merupakan parameter yang menentukan jumlah hidden layer dan dimensionalitas dari vektor. min_count adalah jumlah minimal kemunculan kata yang yang akan masuk didalam training, window adalah jumlah kata sebelum dan sesudah yang bisa dilihat. Adapun parameter pembuatan word2vec adalah sebagai berikut. Dan pada penelitian ini digunakan Word2Vec model skipgram maka ditandai dengan sg=1

Tabel 1 Tabel Parameter Word2Vec

No	Parameter	Nilai
1	Dimention Size	128
2	Min_count	1
3	Window	5
4	workers	4
5	sg	1

Setelah proses pembuatan model selesai, maka sistem menghasilkan vektor-vektor dari setiap kata dari data korpus. Selanjutnya penyimpanan hasil word2vec, Model word2vec hasil training akan disimpan ke dalam bentuk file txt. Yang nantinya akan disematan training Word2Vec dalam pemodelan CNN, Model word embeddings dibutuhkan sebagai input training model CNN. Model tersebut akan dibaca dengan kelas Embeddings dan disimpan nilai vektornya sebagai objek.

3.5 Pemodelan Convolution Neural Network

Pemodelan CNN dibangun dengan arsitektur layer dimulai menggunakan fungsi Sequential dimana terdapat satu input tensor dan satu output tensor. Input tensor merepresentasikan matriks sebagai inputan data yang dapat dimasukkan nilai dari hasil pembobotan word2vec sebelumnya. Output tensor merepresentasikan hasil keluaran yang dalam penelitian adalah klasifikasi. Dropout layer dibuat untuk mengurangi terjadinya data overfitting, Setelah menerima output dari lapisan embedding. selanjutnya, Lapisan konvolusi akan mengekstrak fitur dari data input. Dalam pemodelan CNN menggunakan filter 64 dan 2 unit. Convolution layer akan membentuk vektor feature map sebanyak filter yang digunakan dalam proses konvolusi. Tiap filter akan digunakan untuk semua window. Selanjutnya lapisan pooling, pada lapisan ini melakukan fungsi untuk mengurangi dimensi input. Pooling dilakukan dengan Max pooling dengan mengambil nilai maksimum dari feture maps. Lapisan flatten untuk mereduksi dimensi menjadi vektor satu dimensi. Terkhir yaitu lapisan fully connector layer lapisan ini menghasilkan vektor output dengan menggunakan fungsi aktivasi softmax layer sesuai dengan jumlah type yang digunakan. Pada Gambar 4 dapat dilihat tampilan hasil pemodelan CNN.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 200, 128)          25421312

 spatial_dropout1d (SpatialD  (None, 200, 128)         0
 ropout1D)

 conv1d (Conv1D)             (None, 200, 100)          51300

 batch_normalization (BatchN  (None, 200, 100)         400
 ormalization)

 global_max_pooling1d (Globa  (None, 100)              0
 lMaxPooling1D)

 dropout (Dropout)           (None, 100)               0

 dense (Dense)               (None, 50)                5050

 dense_1 (Dense)             (None, 6)                 306

=================================================================
Total params: 25,478,368
Trainable params: 56,856
Non-trainable params: 25,421,512
_________________________________________________________________
```

Gambar 4 Tampilan Hasil Pemodelan CNN

3.6 Pengujian

Pada pengujian model yang diujikan menggunakan optimizer adam dengan loss binary_crossentropy, Untuk percobaan pengujian menggunakan epoch 25 dan bath size 256, untuk menemukan perfoma model CNN. Adaktif momen estimation (Adam) adalah algoritme pengoptimalan yang dapat digunakan sebagai ganti dari prosedur stochastic gradient descent klasik untuk memperbarui weight network secara iteratif berdasarkan data training.

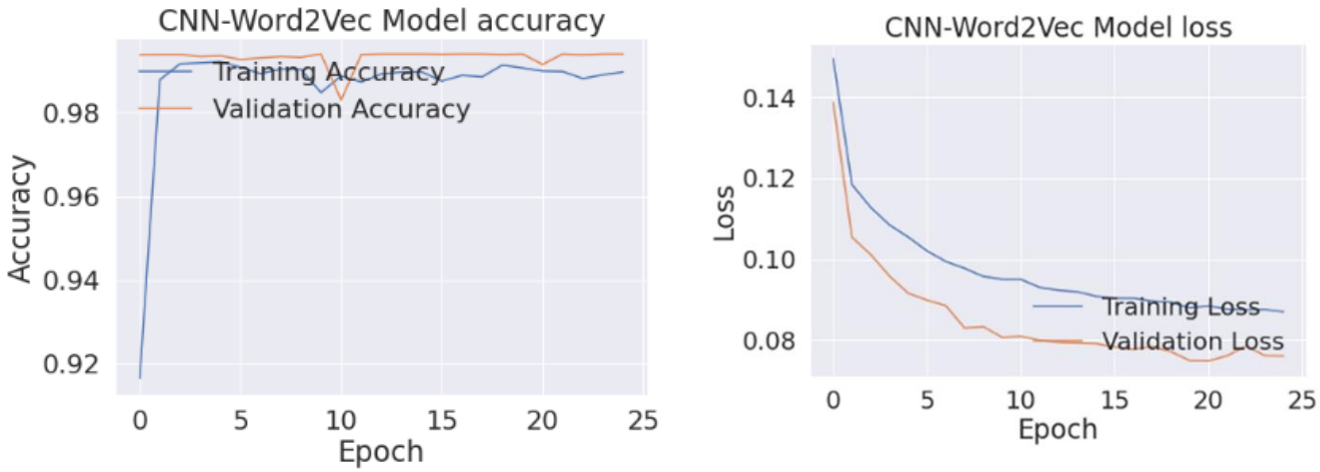
```
CNN_Word2Vec_test_score = CNN_Word2Vec_model.evaluate(X_test, y_test, batch_size=batch_size2, verbose=1)
print('Test Loss:', CNN_Word2Vec_test_score[0]*100)
print('Test Accuracy:', CNN_Word2Vec_test_score[1]*100)

250/250 [==============================] - 29s 117ms/step - loss: 0.0883 - accuracy: 0.9976 - mean_pred: 0.1667
Test Loss: 8.829766511917114
Test Accuracy: 99.76085424423218
```

Gambar 5 Hasil Pengujian Nilai Loss dan Nilai Accuracy

Pemodelan CNN menggunakan Word2Vec dengan optimizer Adam memperoleh test loss : 8,8% dan akurasi yang diperoleh dan test accuracy : 99,7%.

Berikut ditampilkan grafik nilai akurasi pada data latih dan data uji menggunakan optimizer adam pada pemodelan CNN menggunakan Word2Vec:



Gambar 6 Grafik Penujian Akurasi dan Loss

3.6 Evaluasi dan Analisa Hasil

Dalam penelitian ini dilakukan beberapa pemodelan untuk menghasilkan akurasi terbaik, dalam penelitian ini terdapat 4 percobaan yaitu dataset split 70% train 30% test, 80% train 20% test, 90% train 10% test dan dataset train dan test terpisah.

Tabel 2 Perbandingan Akurasi Model CNN

No	Parameter	Nilai
1	Train 70% test 30%	98,48%
2	Train 80% test 20%	99,23%
3	Train 90% test 10%	99,33%
4	Dataset train dan test terpisah	99,76%

Pada Tabel 2 dapat dilihat bahwa akurasi dari model yang dibentuk hanya tidak mengalami perbedaan yang signifikan dengan train 70% test 30% menghasilkan akurasi terendah dengan 98,48% dan dataset terpisah dengan akurasi tertinggi yaitu 99,76%.

4. KESIMPULAN

Dari hasil penelitian klasifikasi multilabel toxic comment menggunakan convolutional neural network dapat diambil kesimpulan sebagai berikut:

1. Metode Convolutional Neural Network dapat diterapkan untuk klasifikasi multilabel toxic comment pada sosial media twitter dengan menggunakan Word2Vec sebagai pembobotan kata
2. Nilai performa yang didapat dari pengujian model mesin pembelajaran CNN dengan menggunakan optimizer adam menghasilkan akurasi tertinggi sebesar 99,76%, presisi 100%, recall 99% dan F1-Score 99%. Banyaknya data, Jumlah Filter, dan pengunaan optimizer yang berbeda mempengaruhi tingkat *accuracy* dan *loss*

DAFTAR PUSTAKA

[1] A. Perwira, J. Dwitama, and S. Hidayat, "Identifikasi Ujaran Kebencian Multilabel Pada Teks Twitter Berbahasa Indonesia Menggunakan Convolution Neural Network," vol. 3, pp. 117–127, 2021, doi: 10.30865/json.v3i2.3610.

[2] R. Y. Rumagit, "Multilabel Classification for Toxic Comments in Indonesian," *Engineering, MAthematics and Computer Science (EMACS) Journal*, vol. 2, no. 1, pp. 29–34, 2020, doi: 10.21512/emacsjournal.v2i1.6256.

[3] N. D. Kusumawati, S. Al Faraby, and M. D. P, "Analisis Sentimen Komentar Beracun pad Media Sosial Menggunakan Word2Vec dan Support Vectore Machine (SVM)," vol. 8, no. 5, pp. 10038–10050, 2021.

[4] S. Zaheri, J. Leath, D. Stroud, S. Zaheri, J. Leath, and D. Stroud, "SMU Data Science Review Toxic Comment Classification Toxic Comment Classification," vol. 3, no. 1, 2020.

[5] R. P. Sidiq, B. A. Dermawan, and Y. Umaidah, "Sentimen Analisis Komentar Toxic pada Grup Facebook Game Online Menggunakan Klasifikasi Naïve Bayes," *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 3, p. 356, 2020, doi: 10.32493/informatika.v5i3.6571.

[6] A. N. Muhammad, S. Bukhori, and P. Pandunata, "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes-Support Vector Machine (NBSVM) Classifier," *Proceedings - 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering, ICOMITEE 2019*, vol. 1, pp. 199–205, 2019, doi: 10.1109/ICOMITEE.2019.8920923.

[7] Ash Shiddicky and Surya Agustian, "Analisis Sentimen Masyarakat Terhadap Kebijakan Vaksinasi Covid-19 pada Media Sosial Twitter menggunakan Metode Logistic Regression," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 2, pp. 99–106, Aug. 2022, doi: 10.37859/coscitech.v3i2.3836.

[8] R. Hayami, Soni, and I. Gunawan, "Klasifikasi Jamur Menggunakan Algoritma Naïve Bayes," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 1, pp. 28–33, May 2022, doi: 10.37859/coscitech.v3i1.3685.

[9] N. A. Shafirra and I. Irhamah, "Klasifikasi Sentimen Ulasan Film Indonesia dengan Konversi Speech-to-Text (STT) Menggunakan Metode Convolutional Neural Network (CNN)," *Jurnal Sains dan Seni ITS*, vol. 9, no. 1, 2020, doi: 10.12962/j23373520.v9i1.51825.

[10] A. Nurdin, B. A. S. Aji, A. Bustamin, and Z. Abidin, "Perbandingan Kinerja Word Embedding Word2Vec , Glove ," *Jurnal TEKNOKOMPAK*, vol. 14, no. 2, pp. 74–79, 2020.