

# KLASIFIKASI SENTIMEN TENTANG PEMINDAHAN IBU KOTA NEGARA INDONESIA DENGAN CONVOLUTIONAL NEURAL NETWORK MENGGUNAKAN GLOVE DAN FASTTEXT

Desya Ristya Putri<sup>1</sup>, Eva Yulia Puspaningrum<sup>2</sup>, Hendra Maulana<sup>3</sup>

<sup>1,2,3</sup>Informatika, Universitas Pembangunan Nasional Veteran Jawa Timur; Jl. Rungkut Madya No.1, Surabaya

Received: 11 Juli 2024

Accepted: 31 Juli 2024

Published: 7 Agustus 2024

## Keywords:

GloVe, FastText, CNN, klasifikasi

## Correspondent Email:

evapuspaningrum.if@upnjatim.ac.id

**Abstrak.** Media sosial kini menjadi tempat untuk berkomunikasi jarak jauh yang marak digunakan. Di dalam media sosial, terdapat berbagai macam opini pengguna yang sering kali terjadi kesalahan penafsiran oleh pembaca. Kadang kala, informasi yang tersebar juga merupakan hoaks sehingga dapat mempersulit pemahaman aktual sentimen yang sesungguhnya ingin disampaikan. Beberapa waktu terakhir, topik pembicaraan mengenai pemindahan ibu kota negara Indonesia sangat banyak tersebar di media sosial. Dilakukannya penelitian ini memiliki tujuan untuk mendapatkan perbandingan hasil antara dua metode ekstraksi fitur yang digunakan. Penelitian ini menerapkan ekstraksi fitur GloVe dan FastText dengan besaran nilai ukuran vektor sebesar 100. Klasifikasi dalam penelitian ini dilakukan dengan algoritma Convolutional Neural Network yang menerapkan beberapa variasi skenario uji, yaitu dengan mengubah nilai *batch size* dan *epoch*. Penelitian dilakukan dengan 44957 data komentar YouTube yang besar perbandingannya adalah 70:30 untuk data pelatihan dan data pengujian. Hasil dari dilakukannya percobaan menunjukkan bahwa metode ekstraksi fitur GloVe menghasilkan akurasi yang lebih baik dibandingkan dengan FastText. Hasil akhir didapatkan bahwa implementasi Convolutional Neural Network dengan GloVe menghasilkan rata-rata nilai *precision* sebesar 74.3%, *recall* sebesar 73.6%, *f1-score* sebesar 73.6%, serta *accuracy* sebesar 76.1%.

**Abstract.** Social media is now a widely used place for long-distance communication. In social media, there are various kinds of user opinions which often cause errors in interpretation by readers. Sometimes, the information that is spread is also a hoax, so it can make it difficult to actually understand the sentiment that is actually being conveyed. In recent times, the topic of discussion regarding moving the Indonesian capital has been widely spread on social media. The aim of this research was to obtain a comparison of the results between the two feature extraction methods used. This research applies GloVe and FastText feature extraction with a vector size value of 100. Classification in this research is carried out using the Convolutional Neural Network algorithm which applies several variations of test scenarios, namely by changing the batch size and epoch values. The research was conducted with 44957 YouTube comment data with a ratio of 70:30 for training data and testing data. The results of the experiment show that the GloVe feature extraction method produces better accuracy compared to FastText. The final results showed that the implementation of a Convolutional Neural Network

*with GloVe produced an average precision value of 74.3%, recall of 73.6%, f1-score of 73.6%, and accuracy of 76.1%.*

## 1. PENDAHULUAN

Saat ini, perkembangan teknologi informasi telah mencapai tingkat yang mengagumkan. Interaksi sosial antar manusia tidak lagi terbatas pada pertemuan langsung, melainkan dapat dilakukan melalui media sosial. Media sosial memfasilitasi pertukaran informasi dalam berbagai bentuk, termasuk teks, gambar, audio, dan video. Namun, terkadang informasi yang disampaikan dalam bentuk teks tidak selalu mudah dipahami secara keseluruhan oleh pembaca. Hal ini dapat menyebabkan kebingungan atau bahkan miskomunikasi. Pada akhirnya, dapat memicu penyebaran informasi yang salah atau hoaks. Untuk mengatasi masalah ini, banyak penelitian menggunakan teknik pembelajaran mesin untuk menganalisis teks dengan lebih baik. Analisis sentimen adalah metode untuk memahami dan memproses data teks yang terkait dengan komputasi opini, perasaan, serta kondisi emosional [1].

Satu diantara proses penting dalam analisis teks adalah ekstraksi fitur. Ekstraksi fitur bertujuan untuk menemukan ciri-ciri khas dari suatu teks. Salah satu metode ekstraksi fitur yang umum digunakan adalah word embedding. Dalam word embedding, teks diubah menjadi vektor numerik, dan metode seperti Word2Vec, GloVe, FastText, dan TF-IDF sering digunakan untuk melakukan hal ini.

Selain sebagai sarana interaksi sosial, media sosial juga menjadi tempat diskusi berbagai topik, termasuk politik. Salah satu topik hangat belakangan ini di Indonesia adalah rencana pemindahan ibu kota negara ke Kalimantan Timur. Diskusi tentang hal ini sering memicu debat di media sosial. YouTube menjadi salah satu platform utama dimana berbagai pendapat diungkapkan dalam bentuk video dan komentar tulisan. Untuk menganalisis opini pengguna terkait rencana ini, Convolutional Neural Network digunakan untuk mengklasifikasikan opini menjadi positif, negatif, atau netral.

Terdapat penelitian yang membahas klasifikasi sentimen dengan fokus pada kebijakan publik di Indonesia, menggunakan data yang diperoleh melalui API Twitter dan

metode GloVe untuk ekspansi fiturnya. Metodologi yang diterapkan meliputi tahapan pengambilan data, preprocessing, representasi data, ekspansi fitur, dan pemodelan analisis sentimen. Dua algoritma klasifikasi yang digunakan dalam penelitian ini adalah Support Vector Machine (SVM) dan Naïve Bayes. Hasil akurasi dari kedua algoritma tersebut dibandingkan, kemudian diperoleh kesimpulan bahwa akurasi Naïve Bayes adalah 0.7786, sedangkan akurasi Support Vector Machine (SVM) adalah 0.8323 [2].

Analisis sentimen terhadap data teks mengenai program vaksinasi Covid-19 dilakukan menggunakan algoritma Support Vector Machine (SVM) dan metode word embedding FastText yang dikumpulkan dari media sosial Twitter. Data tersebut diproses melalui tahap penyeimbangan dataset, seleksi fitur, dan penyesuaian parameter. Hasil penelitian menunjukkan bahwa tingkat akurasi yang dicapai adalah sebesar 68% [3].

Meskipun banyak penelitian sebelumnya telah menggunakan Convolutional Neural Network untuk menganalisis sentimen, masih sedikit penelitian yang membandingkan penggunaan ekstraksi fitur menggunakan metode GloVe dan FastText secara menyeluruh. Oleh karena itu, penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan menggunakan data komentar YouTube tentang pemindahan ibu kota negara Indonesia.

## 2. TINJAUAN PUSTAKA

Penelitian yang dilakukan oleh [4] memiliki tujuan untuk mengevaluasi tingkat kepuasan pelanggan dengan melakukan analisis opini yang mereka tuliskan pada kolom komentar akun bisnis penyedia layanan di media sosial Instagram. Data dikumpulkan menggunakan teknik crawling, menghasilkan 1980 data. Setelah data terkumpul, dilakukan preprocessing. Selanjutnya, klasifikasi sentimen dilakukan menggunakan Convolutional Neural Network (CNN) dengan menerapkan ekstraksi fitur GloVe. Hasil penelitian menunjukkan nilai precision sebesar 95.80%, recall sebesar 88.12%, dan f1-score sebesar 91.62%. [4].

Penelitian lain yang menggunakan algoritma pembelajaran mesin CNN telah dilakukan oleh [5]. Dalam penelitian ini, data yang digunakan terdiri dari 3583 komentar yang diambil dari Twitter. Teks tersebut diubah menjadi vektor menggunakan metode word embedding FastText. Kombinasi FastText dengan CNN untuk klasifikasi sentimen menghasilkan akurasi sebesar 84,01% [5].

### **2.1 Analisis Sentimen**

Analisis sentimen adalah proses mengekstraksi, memahami, dan mengolah data berbentuk teks yang tidak teratur secara otomatis dengan tujuan memperoleh informasi sentimen yang terkandung dalam sebuah kalimat pendapat [6].

### **2.2 Text Mining**

Text mining merupakan metode untuk meraih informasi berkualitas tinggi dari teks dengan menganalisis pola statistiknya. Dalam praktik text mining, kata-kata diberi bobot untuk menunjukkan pentingnya dalam suatu dokumen [7]. Proses text mining berfokus pada ekstraksi beragam data dari sejumlah besar teks, diikuti dengan pencarian pola yang relevan dengan tujuan yang diharapkan [8].

### **2.3 Preprocessing**

Setelah dilakukan labelling, tahap selanjutnya adalah preprocessing. Preprocessing ialah tahap dalam analisis data teks yang berguna untuk membuat data bersih sehingga siap digunakan dalam tahap pemrosesan data selanjutnya [9]. Tahap preprocessing dilakukan dalam empat tahap sebagai berikut.

#### **2.3.1 Case Folding**

Case folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil atau besar (biasanya huruf kecil) untuk konsistensi dalam analisis teks. Seluruh data dalam dokumen tidak semuanya konsisten penggunaan huruf kapitalnya. Guna case folding adalah untuk melakukan konversi semua teks dalam dataset menjadi distandarisasi ke huruf kecil [10]. Ini membantu dalam memperlakukan kata yang sama dengan cara yang sama, terlepas dari bagaimana mereka ditulis dalam dokumen.

#### **2.3.2 Tokenization**

Tokenization adalah proses memecah teks menjadi bagian-bagian kecil yang disebut token. Token bisa berupa kata, frasa, atau karakter terpisah, tergantung pada kebutuhan analisis. Misalnya, kalimat "Saya suka makan nasi goreng" dapat dipecah menjadi token individu seperti "Saya", "suka", "makan", "nasi", "goreng".

#### **2.3.3 Stopwords Removal**

Stopwords adalah kata-kata umum yang sering muncul dalam teks tetapi tidak memiliki nilai informatif yang signifikan dalam analisis, seperti "dan", "atau", "di", "dari", dan sebagainya. Penghapusan stopwords dari teks membantu dalam fokus pada kata-kata kunci yang lebih penting untuk pemrosesan lebih lanjut.

#### **2.3.4 Stemming**

Stemming adalah proses reduksi kata-kata ke bentuk dasarnya dengan menghilangkan afiks (imbuhan seperti akar kata). Tujuannya adalah untuk mengatasi variasi kata yang sama yang berasal dari akar kata yang sama. Misalnya, "berlari", "berlari", dan "lari" semuanya dapat disesuaikan menjadi bentuk dasar "lari". Ini membantu dalam menghitung frekuensi kata yang konsisten.

### **2.4 Labelling**

VADER (Valence Aware Dictionary and sEntiment Reasoner) adalah alat yang menggabungkan leksikon dan aturan untuk analisis sentimen, khususnya selaras dengan sentimen di media sosial. Kategori penilaian sentimen dalam VADER terdiri dari empat bagian: pos, neg, neu, dan compound. Nilai pos, neg, dan neu menunjukkan proporsi rasio teks dalam masing-masing kategori. Pos mewakili teks positif, neg mewakili teks negatif, dan neu mewakili teks netral. Kategori ini digunakan untuk menganalisis konteks teks. Nilai compound adalah hasil penjumlahan dari semua skor valensi setiap kata dalam leksikon, yang diatur oleh aturan VADER. Aturan ini memperhitungkan urutan kata, penguat kata, tanda baca, serta penggunaan konjungsi yang berlawanan makna. Nilai ini kemudian dinormalisasi ke dalam rentang -1 (sangat negatif) hingga +1 (sangat positif). Nilai

compound inilah yang digunakan untuk menentukan label sentimen [11].

### 2.5 GloVe

GloVe, yang merupakan singkatan dari Global Vector, mengadopsi dua pendekatan dalam tekniknya, yakni pendekatan berbasis hitungan dan prediksi langsung. Pendekatan berbasis hitungan dalam GloVe melibatkan pembelajaran hubungan antarkata dengan mengukur frekuensi kemunculan kata bersamaan dalam korpus teks. Dalam metode global matrix factorization yang digunakan oleh GloVe, sebuah matriks merepresentasikan keberadaan atau ketiadaan kata dalam dokumen [12]. Persamaan dari model GloVe adalah sebagai berikut.

$$w_i^T + \vec{w}_k + b_i + \vec{b}_k = \log(X_{ik}) \tag{1}$$

Pada persamaan di atas,  $w$  merupakan word vector sedangkan  $\vec{w}$  merupakan context word vector. Nilai  $b$  adalah scalar biases dari  $i$ -word dan  $k$ -word context. Kemudian, nilai  $X$  adalah word co-occurrence matrix sedangkan  $X_{ik}$  adalah jumlah kemunculan  $i$  dalam  $k$ -word context.

Dalam persamaan tersebut, setiap kata diberi bobot yang sama. Ini menyebabkan sulitnya memisahkan sinyal dari noise karena keduanya dianggap memiliki tingkat penting yang sama. Oleh karena itu, diperlukan suatu fungsi pembobotan  $f(x)$  yang dapat menyesuaikan bobot setiap kata dengan konten informasi yang terkandung. Fungsi  $f(x)$  yang dimaksud adalah sebagai berikut.

$$J = \sum_{i,k=1}^V f(X_{ik})(w_i^T \vec{w}_k + b_i + b_k - \log(X_{ik}))^2 \tag{2}$$

### 2.6 FastText

FastText merupakan teknik embedding kata yang dapat melatih model dengan kecepatan tinggi pada kumpulan data yang besar dan mampu menggambarkan kata-kata yang tidak ada dalam data pelatihan. Ketika sebuah kata tidak terdapat dalam model pelatihan, vektor nilainya dapat dihasilkan dengan membaginya menjadi  $n$ -grams [12].

### 2.7 Convolutional Neural Network

Convolutional Neural Network (CNN) merupakan suatu teknik klasifikasi yang memanfaatkan lapisan khusus untuk

memproses input dengan menggunakan filter. CNN memiliki dua fase utama, yaitu pembelajaran fitur dan klasifikasi. Fase pembelajaran fitur terdiri dari lapisan konvolusi, fungsi aktivasi ReLu, dan lapisan penggabungan (pooling). Sementara fase klasifikasi, terdiri dari mengubah dimensi data (flatten), lapisan terhubung penuh (fully-connected layer), dan proses prediksi. Tiap komponen CNN melibatkan dua langkah utama, yaitu penerusan (feed-forward) dan propagasi balik (backpropagation) [13].

### 2.8 Confusion Matrix

Confusion Matrix adalah sebuah metode yang digunakan untuk mengevaluasi kinerja model klasifikasi dalam machine learning yang menghasilkan output berupa dua atau lebih kelas.

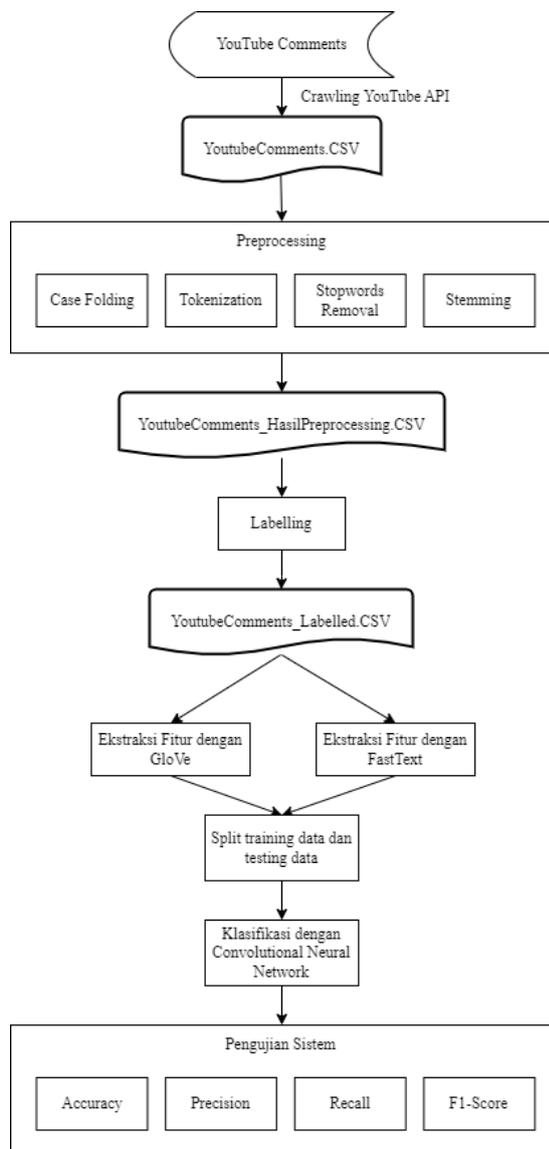
**Tabel 1.** Nilai Confusion Matrix

	<b>Aktual Positif</b>	<b>Aktual Negatif</b>
<b>Prediksi Positif</b>	True Positive (TP)	False Positive (FP)
<b>Prediksi Negatif</b>	False Negative (FN)	True Negative (TN)

Berikut merupakan ketentuan dari setiap nilai confusion matrix yang ada pada tabel di atas.

- TP (True Positive) adalah jumlah data yang benar-benar termasuk kelas True dan berhasil diprediksi dengan tepat sebagai kelas True.
- TN (True Negative) adalah jumlah data yang benar-benar termasuk kelas False dan berhasil diprediksi dengan tepat sebagai kelas False.
- FP (False Positive) adalah jumlah data yang benar-benar termasuk kelas True tetapi salah diprediksi sebagai kelas False.
- FN (False Negative) adalah jumlah data yang benar-benar termasuk kelas False tetapi salah diprediksi sebagai kelas True.

### 3. METODE PENELITIAN



Gambar 1. Desain Sistem

Penelitian ini dilakukan dengan alur yang terdiri dari proses pengumpulan data, pemrosesan data, dan pengujian hasil.

#### 3.1 Crawling Data

Data diperoleh melalui teknik crawling dari YouTube dengan memanfaatkan YouTube API. Video yang diambil adalah lima video teratas dengan kata kunci "Pemindahan Ibu Kota Negara Indonesia," menggunakan filter tipe Video dan filter urutan berdasarkan jumlah penonton. Video-video ini diurutkan mulai dari yang paling banyak ditonton. Jumlah total data yang digunakan adalah 44.957, dengan

persebaran jumlah data di masing-masing video ditunjukkan pada tabel berikut.

Tabel 2. Jumlah Data yang Digunakan

Judul Video	Jumlah Data Komentar
Yang Tidak Dikatakan Jokowi soal Ibu Kota Baru   Buka Mata	13,147
Resmi! Presiden Jokowi Umumkan Lokasi Ibu Kota Baru	18,080
Ganjar Tanya Anies soal Pemindahan Ibu Kota #debatcapres #aniesbaswedan #ganjarpranowo	5,933
Chan Foong Hin - Pemindahan ibu kota Indonesia, bagaimana Sabah & Sarawak mendapat manfaat?	5,614
Rencana Memindahkan Ibu Kota, Malah Jadi Kota Hantu Yang Sepi Penduduk Rencana Memindahkan Ibu Kota, Malah Jadi Kota Hantu Yang Sepi Penduduk	2,183
<b>TOTAL</b>	<b>44,957</b>

#### 3.2 Preprocessing

Secara singkat, proses preprocessing dimulai dengan menghapus karakter-karakter yang tidak diperlukan. Langkah selanjutnya adalah memecah teks menjadi daftar kata. Kemudian, dilakukan penghapusan kata-kata yang tidak memberikan informasi penting. Pada tahap terakhir, imbuhan dari setiap kata dihilangkan. Secara berturut-turut, preprocessing melibatkan empat tahap: case folding, tokenisasi, penghapusan kata berhenti (stopwords), dan stemming.

#### 3.3 Labelling

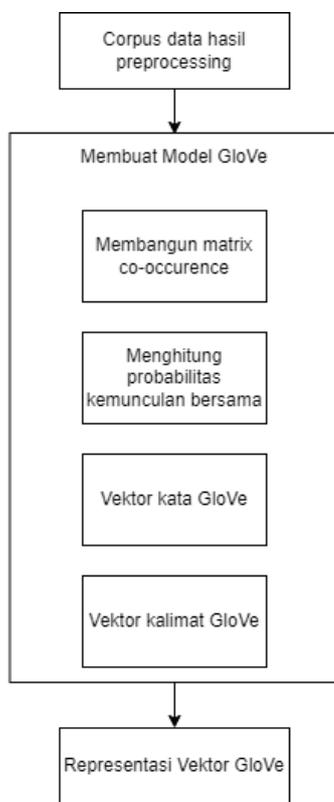
Metode yang digunakan untuk pelabelan adalah VADER Lexicon. Mengingat dataset peneliti berbahasa Indonesia, penerjemahan dilakukan terlebih dahulu sebelum VADER Lexicon diterapkan. Setelah proses penerjemahan selesai, pelabelan dengan VADER Lexicon dilakukan untuk mendapatkan nilai compound. Nilai compound ini menjadi dasar dalam menentukan label data. Ketentuan pelabelan adalah sebagai berikut.

- a. Teks dengan nilai compound kurang dari 0 diberi label negatif.
- b. Teks dengan nilai compound sama dengan 0 diberi label netral.
- c. Teks dengan nilai compound lebih dari 0 diberi label positif.

**3.4 Ekstraksi Fitur**

Tahap ini digunakan untuk mendapatkan ciri khas dari setiap kata dalam dataset. Setiap kata akan diubah menjadi bentuk vektor. Hasil akhir dari tahap ini adalah berupa representasi vektor.

**3.4.1 GloVe**



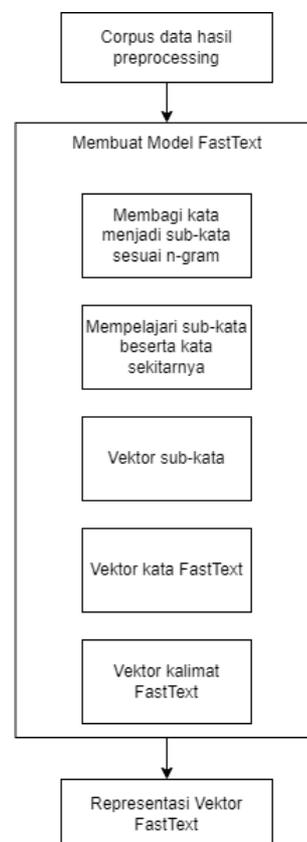
**Gambar 2.** Alur GloVe

Pada tahap ini, data yang telah dibersihkan melalui preprocessing dan diberi label melalui proses labelling akan diproses lebih lanjut dengan tahapan ekstraksi fitur menggunakan Glove. Tahap ini bertujuan untuk mengubah setiap kata menjadi vektor representasi berdasarkan kemunculannya dalam korpus. Vektor-vektor ini kemudian akan digunakan sebagai input pada proses pelatihan convolutional neural network.

Metode Word Embedding GloVe (Global Vectors for Word Representation) merupakan pendekatan yang menghasilkan vektor representasi kata dengan memanfaatkan

informasi co-occurrence atau kemunculan bersama antar kata dalam suatu korpus teks. Proses ini dimulai dengan membangun matriks co-occurrence yang mencatat frekuensi kemunculan bersama antar kata dalam konteks yang ditentukan, menghasilkan vektor representasi kata. Vektor ini kemudian digunakan sebagai input dalam proses klasifikasi. Pada penelitian ini, klasifikasi dilakukan menggunakan algoritma Convolutional Neural Network.

**3.4.2 FastText**



**Gambar 3.** Alur FastText

FastText adalah sebuah algoritma yang digunakan untuk menghasilkan representasi vektor dari kata-kata dalam teks, yang juga dikenal sebagai embedding. Algoritma ini mempertimbangkan konteks dari kata tersebut, sehingga kata-kata yang sering muncul bersama akan memiliki representasi vektor yang mirip.

Berbeda dengan GloVe, FastText menggunakan pendekatan yang mengembangkan konsep Word2Vec dengan memperhitungkan sub-kata atau karakter dalam proses pembentukan vektor kata. Proses ini

dimulai dengan membagi kata menjadi n-gram karakter yang saling tumpang tindih, yang kemudian direpresentasikan sebagai vektor. Setiap kata direpresentasikan sebagai penjumlahan dari vektor sub-kata atau karakternya, sehingga memungkinkan untuk menangkap makna kata berdasarkan struktur internalnya. Berikut adalah langkah-langkah dalam ekstraksi fitur menggunakan FastText.

### 3.5 Klasifikasi

Setelah tahap ekstraksi fitur selesai, langkah berikutnya adalah pengklasifikasian data. Proses klasifikasi ini menggunakan algoritma Convolutional Neural Network (CNN). Untuk klasifikasi, CNN dikombinasikan dengan setiap metode ekstraksi fitur yang dipakai.

Arsitektur yang digunakan dalam klasifikasi ini adalah DoubleMax CNN. Kelebihan utama arsitektur DoubleMax adalah adanya dua pooling layer. Urutan layer dalam arsitektur DoubleMax CNN adalah input layer, embedding layer, convolutional layer 1D (pertama), pooling layer (pertama), convolutional layer 1D (kedua), pooling layer (kedua), flatten layer, fully-connected/dense layer, drop-out layer, dan output layer.

### 3.6 Pengujian

Pengujian dilakukan dengan mengubah berbagai nilai parameter yang ada. Perubahan ini dilakukan dalam proses pelatihan model CNN, khususnya dengan memodifikasi nilai epoch dan batch\_size. Terdapat 30 skenario pengujian, yang terdiri dari 15 skenario untuk GloVe-CNN dan 15 skenario untuk FastText-CNN. Variasi skenario terlihat pada nilai batch size, yaitu 2, 16, 32, 64, dan 128. Selain itu, setiap nilai batch size diuji dengan epoch 5, 10, dan 15.

## 4. HASIL DAN PEMBAHASAN

### 4.1 Preprocessing

Dalam preprocessing dilakukan pembersihan data sehingga data hasil preprocessing merupakan data yang mengandung informasi penting dari sebuah teks pada dataset.

#### 4.1.1 Case Folding

Case folding, yaitu mengubah teks menjadi lowercase dan menghilangkan angka, whitespace, serta tanda baca.

**Tabel 3.** Data Setelah Diterapkan Case Folding

Sebelum	Sesudah
Tidak setuju dengan rencana pemindahan ibu kota karena biayanya terlalu besar dan berpotensi merusak lingkungan.	tidak setuju dengan rencana pemindahan ibu kota karena biayanya terlalu besar dan berpotensi merusak lingkungan

#### 4.1.2 Tokenization

Tokenization, yaitu setiap kata pada teks masing-masing dipisahkan menjadi satu per satu.

**Tabel 4.** Data Setelah Diterapkan Tokenization

Sebelum	Sesudah
tidak setuju dengan rencana pemindahan ibu kota karena biayanya terlalu besar dan berpotensi merusak lingkungan	["tidak", "setuju", "dengan", "rencana", "pemindahan", "ibu", "kota", "karena", "biayanya", "terlalu", "besar", "dan", "berpotensi", "merusak", "lingkungan"]

#### 4.1.3 Stopwords Removal

Stopwords removal, yaitu menghilangkan kata yang tidak memiliki makna.

**Tabel 5.** Data Setelah Diterapkan Stopwords Removal

Sebelum	Sesudah
["tidak", "setuju", "dengan", "rencana", "pemindahan", "ibu", "kota", "karena", "biayanya", "terlalu", "besar", "dan", "berpotensi", "merusak", "lingkungan"]	["setuju", "rencana", "pemindahan", "ibu", "kota", "biayanya", "besar", "berpotensi", "merusak", "lingkungan"]

4.1.4 Stemming

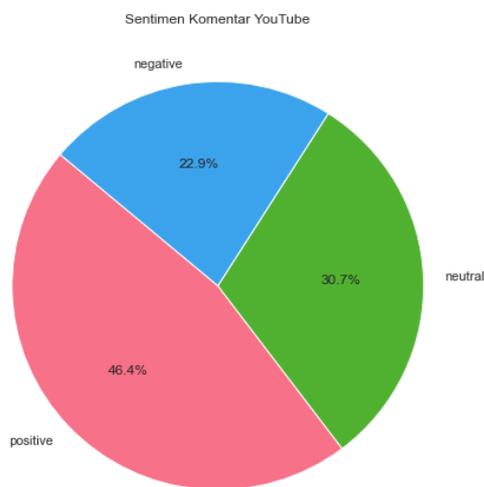
Stemming, yaitu mengubah tiap kata menjadi kata dasar dengan menghapus prefiks maupun sufiks-nya.

Tabel 6. Data Setelah Diterapkan Stemming

Sebelum	Sesudah
["setuju", "rencana", "pemindahan", "ibu", "kota", "biayanya", "besar", "berpotensi", "merusak", "lingkungan"]	["setuju", "rencana", "pindah", "ibu", "kota", "biaya", "besar", "potensi", "rusak", "lingkung"]

4.2 Labelling

Setelah mengumpulkan 44,957 data melalui proses akuisisi data dan membersihkannya melalui proses preprocessing, kami melakukan labelling menggunakan metode VADER Lexicon. Labelling dilakukan untuk mengelompokkan data berdasarkan sentimennya ke dalam tiga kelas, yaitu positif, negatif, dan netral. Persebaran hasil labelling dapat dilihat pada gambar berikut.



Gambar 4. Persebaran Data Berdasarkan Label

4.3 Ekstraksi Fitur

4.3.1 GloVe

Gambar berikut merupakan bentuk vektor kata hasil penerapan ekstraksi fitur dengan GloVe. Terdapat kata “negara” dan bentuk

vektor katanya yang memiliki panjang dimensi vektor sebesar 100.

```
Vektor untuk kata "negara":
[-3.80909579e-03  2.11623375e-03 -3.90914704e-03  1.34635349e-03
 5.81486804e-03 -3.59977446e-03  2.00489516e-03 -4.16158977e-03
-1.76732430e-03  1.13107556e-03  1.07229904e-03  1.94841350e-03
 3.36004455e-03 -4.43812673e-03  2.99036411e-03 -4.42330586e-05
 4.40952944e-03  2.80628951e-03  2.34522377e-03  8.55306952e-04
 4.40860598e-03  2.41887231e-03 -2.62644395e-03 -3.28794802e-04
 3.83270542e-03 -2.55725154e-03 -1.69162331e-03 -4.78837554e-03
-2.94884890e-03 -5.42909310e-03 -5.49504353e-04  4.20774102e-03
 1.33547063e-03  2.50826687e-03 -5.98479230e-04 -3.23677557e-03
 2.81735530e-03  2.05016773e-03 -9.71765126e-04  4.12267848e-03
-3.85204773e-03  2.62517594e-03  7.55295810e-04  3.66028671e-03
-3.09303827e-03 -1.28928014e-03 -3.60645254e-03  3.24517817e-03
-2.91806282e-03 -1.31536762e-03 -1.05562954e-03 -1.71654018e-03
 5.09974422e-03  1.42936818e-03 -4.51410583e-03  2.71274055e-03
-5.42022740e-03 -1.70377351e-03 -2.34127990e-03  5.27404865e-03
 4.20571934e-04  1.17881360e-03  4.04644812e-03  4.43231127e-03
-1.94105497e-03 -5.20502508e-05 -1.40260254e-03 -1.87519976e-03
 4.88128051e-04  3.46795958e-03 -4.67446979e-04  3.25908171e-03
 2.19558847e-03  2.92894196e-04 -2.17245132e-03 -4.20391646e-04
-1.84646816e-03  2.27744138e-03  4.47218488e-06 -2.37864787e-03
 3.90568141e-03  3.99900019e-03 -3.85204520e-03 -9.58084138e-04
-3.29134270e-03 -1.25768025e-03  2.68596279e-03 -3.31486655e-03
-2.42309071e-03  3.34054162e-03  2.63069248e-03  4.08594913e-03
 5.19442495e-03  4.36971421e-03 -2.70477176e-03  3.31638745e-04
 2.34214378e-03 -5.39142333e-04 -4.70543856e-03 -4.08813362e-03]
Index kata "negara": 63
```

Gambar 5. Vektor Kata GloVe

4.3.2 FastText

Gambar berikut merupakan bentuk vektor kata hasil penerapan ekstraksi fitur dengan FastText. Terdapat kata “negara” dan bentuk vektor katanya yang memiliki panjang dimensi vektor sebesar 100.

```
Vector representation of "negara":
[-0.03224507  0.0053715  -0.03220132  0.02998173 -0.04258324 -0.07263137
 0.09851802 -0.02233984  0.012253  0.09058601 -0.04281488  0.01884676
 0.06551841  0.00839031  0.01442692 -0.00047767 -0.02110284  0.02085458
 0.09146469 -0.03936826 -0.03494108  0.02238316  0.00851843 -0.04276636
 0.10249083  0.03663471 -0.00455725 -0.0118829  0.04058761  0.03908813
-0.05263791 -0.01985779 -0.10201575 -0.00258318  0.00772959  0.00486328
 0.0754292 -0.00666846 -0.00845567 -0.00579217  0.0721885 -0.02326209
 0.02391917 -0.01091618 -0.10692307  0.02399169 -0.0477818 -0.04994968
-0.0327309  0.01463424  0.01414562  0.0524498 -0.05750851 -0.07658795
 0.0779916 -0.00099052  0.02689641 -0.00459589 -0.01863438 -0.02453976
-0.08861142  0.03232953 -0.048535 -0.02076997 -0.02287904  0.00891832
 0.01316473  0.0045049  0.05656601  0.02998755 -0.02510779 -0.03646505
-0.00454733  0.04256386 -0.01893015 -0.00136956  0.04847341  0.05803168
-0.04185209 -0.07276151 -0.06724676  0.03929774  0.03506952 -0.09691916
-0.09358246 -0.09801447  0.03915358  0.03560373  0.05053356 -0.04101868
 0.02499353 -0.00349924 -0.00776613  0.0378278  0.040309 -0.03152878
 0.05024701  0.07380073 -0.01325603  0.03463152]
```

Gambar 6. Vektor Kata FastText

4.4 Hasil Pengujian

Pada tabel berikut disajikan hasil akurasi dari seluruh skenario uji yang dilakukan.

Tabel 7. Hasil Akurasi Seluruh Skenario Uji

Metode	Batch Size	Epoch	Accuracy
GloVe-CNN	2	5	0.5389
		10	0.5779
		15	0.5431
	16	5	0.7619
		10	0.7461
		15	0.7431
32	5	0.7542	

	64	10	0.7423		
		15	0.7392		
		5	0.7552		
		10	0.7408		
		15	0.7348		
		5	0.7515		
	128	10	0.7313		
		15	0.7409		
		FastText-CNN	2	5	0.4676
				10	0.6026
				15	0.5583
		16	5	0.7539	
10	0.7463				
15	0.7332				
32	5	0.7551			
	10	0.7441			
	15	0.7383			
64	5	0.7556			
	10	0.7438			
	15	0.7439			
128	5	0.7530			
	10	0.7453			
	15	0.7368			

Hasil dari semua skenario uji dapat ditemukan dalam tabel di atas. Untuk model GloVe-CNN, akurasi tertinggi yang tercapai adalah 0,7619 dengan pengaturan batch size = 16 dan epoch = 5. Sementara itu, untuk model FastText-CNN, akurasi tertinggi yang dicapai adalah 0,7556 dengan pengaturan batch size = 64 dan epoch = 5.

**Tabel 8.** Rata-Rata Hasil Skenario Terbaik

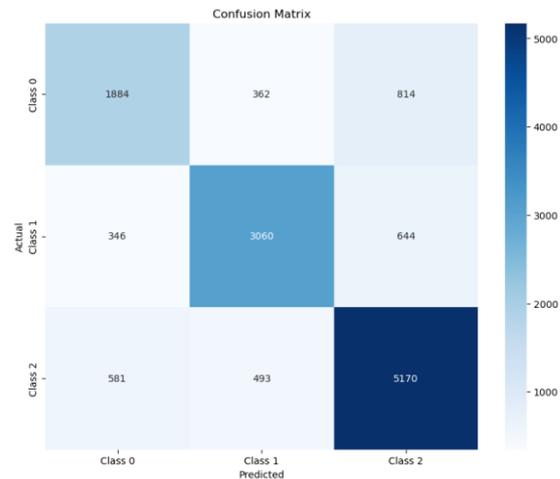
	GloVe-CNN	FastText-CNN
<b>Precision</b>	0.743	0.736
<b>Recall</b>	0.736	0.746
<b>F1-Score</b>	0.736	0.74
<b>Accuracy</b>	76.1%	75.5%

Dalam analisis yang telah dilakukan dalam penelitian ini, terlihat perbandingan antara tingkat akurasi Word Embedding GloVe dan Word Embedding FastText. Hasilnya menunjukkan bahwa Word Embedding GloVe mencatat tingkat akurasi yang lebih tinggi, mencapai 76.1%, sementara Word Embedding FastText hanya mencapai 75.5%.

Pada implementasi menggunakan algoritma GloVe-CNN, diperoleh nilai precision sebesar 0.743, recall sebesar 0.736, dan f1-score sebesar 0.736, dengan tingkat akurasi mencapai 76.1%.

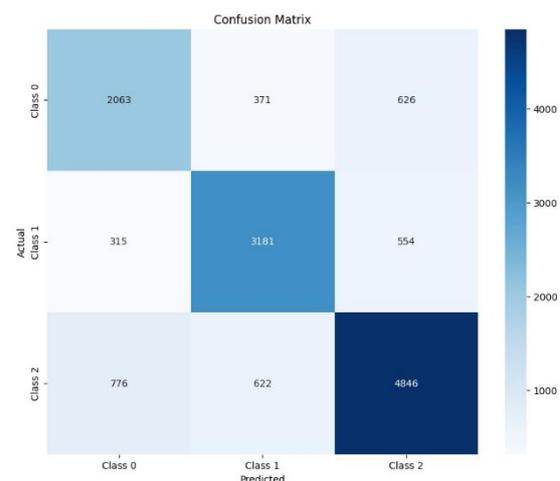
Sementara itu, untuk FastText-CNN, nilai precision, recall, dan f1-score berturut-turut adalah 0.736, 0.746, dan 0.74, dengan tingkat akurasi mencapai 75.5%.

#### 4.4.1 Confusion Matrix



**Gambar 7.** Confusion Matrix GloVe-CNN

Gambar di atas menunjukkan hasil dari confusion matrix untuk skenario CNN-GloVe menunjukkan nilai True Negative, True Neutral, dan True Positive masing-masing adalah 1884, 3060, dan 5170. Nilai True Positive yang signifikan mungkin disebabkan oleh komposisi dataset yang dominan berlabel positif.



**Gambar 8.** Confusion Matrix FastText-CNN

Gambar di atas menunjukkan hasil dari confusion matrix untuk skenario CNN-FastText

menunjukkan nilai True Negative, True Neutral, dan True Positive masing-masing adalah 2063, 3181, dan 4846. Nilai True Positive yang signifikan mungkin dikarenakan komposisi dataset yang mayoritas berlabel positif.

## 5. KESIMPULAN

### 5.1 Kesimpulan

- a. Dari hasil penelitian yang dilakukan oleh peneliti, berikut merupakan kesimpulan yang dapat ditarik, yakni dilihat dari tingkat akurasi, presisi, dan recall yang didapatkan melalui perhitungan nilai confusion matrix, penggunaan ekstraksi fitur GloVe lebih baik dibandingkan dengan FastText.
- b. Performa GloVe dianggap lebih baik dikarenakan model GloVe-CNN menghasilkan nilai akurasi sebesar 76.1%, presisi sebesar 74.3%, recall 73.6%, dan f1-score 73.6% sedangkan model FastText-CNN menghasilkan nilai akurasi sebesar 75.5%, presisi sebesar 73.6%, recall 74.6%, dan f1-score 74%.
- c. Meskipun metode GloVe-CNN lebih baik dibandingkan dengan metode FastText-CNN, tetapi tidak menunjukkan perbedaan yang signifikan berdasarkan hasil akurasi yang didapatkan.

### 5.2 Saran

- a. Membandingkan hasilnya dengan data yang telah diberi label secara manual oleh ahli.
- b. Diharapkan penelitian di masa depan dapat mempertimbangkan berbagai aspek tambahan sehingga kelebihan dan kekurangan program dapat lebih mudah diidentifikasi.

## DAFTAR PUSTAKA

- [1] S. Budi, "Text Mining Untuk Analisis Sentimen Review Film Menggunakan Algoritma K-Means," *Techno.Com: Jurnal Teknologi Informasi*, vol. 16, pp. 1-8, 2017.
- [2] M. D. D. Sreya and E. B. Setiawan, "Penggunaan Metode GloVe untuk Ekspansi Fitur pada Analisis Sentimen Twitter dengan Naïve Bayes dan Support Vector Machine," *e-Proceeding of Engineering*, vol. 9, pp. 2008-2015, 2022.
- [3] M. M. Kusairi and S. Agustian, "SVM Method with FastText Representation Featurefor Classification of Twitter Sentiments Regarding the Covid-19 Vaccination Program," *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, vol. 13, pp. 140-150, 2022.
- [4] M. A. Rahman, H. Budianto and E. I. Setiawan, "Aspect Based Sentimen Analysis Opini Publik Pada Instagram dengan Convolutional Neural Network," *INSYST: Journal of Intelligent System and Computation*, vol. 1, pp. 50-57, 2019.
- [5] F. Alfariqi, W. Maharani and J. H. Husen, "Klasifikasi Sentimen pada Twitter dalam Membantu Pemilihan Kandidat Karyawan dengan Menggunakan Convolutional Neural Network dan Fasttext Embeddings," *e-Proceeding of Engineering*, vol. 7, pp. 8052-8062, 2020.
- [6] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (Svm)," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, vol. 8, pp. 147-156, 2021.
- [7] A. Deolika, Kusri and E. T. Luthf, "Analisis Pembobotan Kata Pada Klasifikasi Text Mining," *JurTI: Jurnal Teknologi Informasi*, vol. 3, pp. 179-184, 2019.
- [8] F. Fathonah and A. Herliana, "Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid -19 Menggunakan Metode Naïve Bayes," *Jurnal Sains dan Informatika*, vol. 7, pp. 155-164, 2021.
- [9] A. Rhamadanti, A. Rifa'i, F. Dikananda and K. Anam, "Analisis Sentimen pada Ulasan Access By Kereta Api Indonesia dengan K-Nearest Neighbor," *JITET (Jurnal Informatika dan Teknik Elektro Teraoan)*, vol. 12, pp. 775-783, 2024.
- [10] D. Alita and A. Rahman, "Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier," *Jurnal Komputasi*, vol. 8, pp. 50-58, 2020.
- [11] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.
- [12] E. M. Dharma, F. L. Gaol, H. L. H. S. Warnars and B. Soewito, "The Accuracy Comparison Among Word2vec, Glove, And Fasttext Towards Convolution Neural Network (Cnn)

- Text Classification," *Jatit: Journal of Theoretical and Applied Information Technology*, vol. 100, pp. 349-359, 2022.
- [13] A. Yusuf, R. C. Wihandika and C. Dewi, "Klasifikasi Emosi Berdasarkan Ciri Wajah Menggunakan Convolutional Neural Network," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, pp. 10595-10604, 2019.