E-ISSN 2963-9174

Indonesian Journal of Computer Science Research

Journal page: https://subset.id/index.php/IJCSR

Implementasi Regresi Logistik untuk Klasifikasi *Cyberbullying* pada Komentar Instagram Berbahasa Indonesia

Implementation of Logistic Regression for Cyberbullying Classification on Indonesian Instagram Comments

Pita Penengah[1,a)] Erna Daniati [2,b)] M. Najibulloh Muzaki[3,c)]

[1, 2, 3)] Program Studi Sistem Informasi , Fakultas Teknik dan Ilmu Komputer,
Universitas Nusantara PGRI Kediri, Jl. KH. Achmad Dahlan No.76, Kediri 64112, Indonesia

Author Emails
[a)] *Corresponding author:* Erna Daniati. Email: *ernadaniati@unpkediri.ac.id*
Received: 2 Juli 2026; Accepted: 27 Juli 2026

KEYWORDS

Kata kunci:
Bag of Words
Cyberbullying
Instagram
Logistic Regression
Natural Language Processing
Text Classification

Keywords:
Bag of Words
Cyberbullying
Instagram
Logistic Regression
Natural Language Processing
Text Classification

ABSTRAK

Abstrak. Perkembangan media sosial, khususnya Instagram, telah meningkatkan intensitas komunikasi antar pengguna sehingga mempermudah penyebaran informasi dan interaksi secara daring. Namun, tingginya aktivitas tersebut juga memunculkan permasalahan *cyberbullying* berupa komentar yang mengandung hinaan, ejekan, ujaran kebencian, maupun pelecehan verbal yang berpotensi memberikan dampak negatif terhadap kesehatan psikologis korban. Urgensi penelitian ini terletak pada perlunya sistem deteksi otomatis yang mampu mengidentifikasi komentar bermuatan *cyberbullying* secara cepat dan akurat guna membantu menciptakan lingkungan media sosial yang lebih aman. Kebaruan (novelty) penelitian ini adalah penerapan algoritma Regresi Logistik yang dipadukan dengan tahapan *Natural Language Processing* (NLP) dan representasi fitur *Bag of Words* berbasis *CountVectorizer* untuk mendeteksi komentar Instagram berbahasa Indonesia. Penelitian ini bertujuan membangun serta mengevaluasi model klasifikasi yang mampu membedakan komentar *cyberbullying* dan *non-cyberbullying*. Metode yang digunakan mengacu pada kerangka kerja CRISP-DM yang meliputi *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Dataset yang digunakan berjumlah 650 komentar Instagram berbahasa Indonesia yang telah melalui proses pelabelan ke dalam dua kelas, yaitu *cyberbullying* dan non-*cyberbullying*. Tahap *data preparation* meliputi *cleaning*, *case folding*, *tokenization*, *stopword removal*, dan *stemming* menggunakan Sastrawi, kemudian teks dikonversi menjadi fitur numerik menggunakan *CountVectorizer*. Hasil evaluasi menunjukkan bahwa model Regresi Logistik mampu mengklasifikasikan komentar dengan baik, ditunjukkan oleh nilai accuracy sebesar 83%, precision sebesar 0,83, recall sebesar 0,83, dan f1-score sebesar 0,83. Hasil tersebut menunjukkan bahwa kombinasi *Bag of Words* dan Regresi Logistik efektif digunakan untuk mendeteksi *cyberbullying* pada komentar Instagram berbahasa Indonesia.

Abstract. *The rapid growth of social media, particularly Instagram, has significantly increased user interaction and information sharing. However, this growth has also led to the emergence of cyberbullying in the form of insulting, offensive, hateful, and abusive comments that may negatively affect victims' psychological well-being. The urgency of this study lies in the need for an automated detection system capable of identifying cyberbullying content accurately and efficiently to support a safer social media environment. The novelty of this research is the implementation of the Logistic Regression algorithm combined with Natural Language Processing (NLP) techniques and Bag of Words feature representation using CountVectorizer for detecting cyberbullying in Indonesian-language Instagram comments. This study aims to develop and evaluate a classification model capable of distinguishing between cyberbullying and non-cyberbullying comments. The research follows the CRISP-DM methodology, which consists of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The dataset comprises 650 Indonesian Instagram comments that were manually labeled into two categories: cyberbullying and non-cyberbullying. The data preparation stage includes cleaning, case folding, tokenization, stopword removal, and stemming using the Sastrawi library. Subsequently, the text data are transformed into numerical features using CountVectorizer. The evaluation results demonstrate that the Logistic Regression model performs well in classifying comments, achieving an accuracy of 83%, precision of 0.83, recall of 0.83, and F1-score of 0.83. These findings indicate that the combination of the Bag of Words approach and Logistic Regression is effective for detecting cyberbullying in Indonesian-language Instagram comments.*

1. Pendahuluan

Perkembangan teknologi digital telah mengubah pola komunikasi masyarakat secara signifikan, terutama melalui pemanfaatan media sosial sebagai sarana berbagi informasi dan berinteraksi secara daring. Salah satu platform yang memiliki jumlah pengguna aktif tinggi adalah Instagram, yang memungkinkan pengguna mengunggah foto, video, serta memberikan komentar terhadap unggahan pengguna lain secara bebas [1]. Di balik kemudahan tersebut, muncul berbagai dampak negatif, salah satunya adalah *cyberbullying*, yaitu tindakan perundungan yang dilakukan melalui media digital dalam bentuk hinaan, ejekan, body shaming, ujaran kebencian, maupun komentar yang bersifat menyerang individu lain. Fenomena ini menjadi perhatian karena dapat memberikan dampak serius terhadap kondisi psikologis korban, seperti menurunnya rasa percaya diri, stres, kecemasan, depresi, hingga gangguan kesehatan mental dalam jangka panjang [2]. Meningkatnya jumlah komentar yang dipublikasikan setiap hari menyebabkan proses moderasi secara manual menjadi tidak efektif, sehingga diperlukan suatu sistem deteksi otomatis yang mampu mengidentifikasi komentar bermuatan cyberbullying secara cepat, akurat, dan konsisten. Pengembangan sistem tersebut menjadi semakin penting untuk mendukung terciptanya lingkungan digital yang lebih aman, sehat, dan nyaman bagi seluruh pengguna media sosial.

Berbagai penelitian sebelumnya telah menerapkan teknik Machine Learning dan Natural Language Processing (NLP) untuk mendeteksi *cyberbullying* pada media sosial dengan memanfaatkan algoritma klasifikasi seperti Naïve Bayes, Support Vector Machine, Decision Tree, dan Random Forest. Meskipun demikian, hasil yang diperoleh masih menunjukkan adanya keterbatasan dalam menangani karakteristik komentar berbahasa Indonesia yang cenderung menggunakan bahasa tidak baku, singkatan, kata gaul, serta variasi penulisan yang beragam. Selain itu, belum banyak penelitian yang secara khusus mengombinasikan tahapan prapemrosesan teks menggunakan NLP, representasi fitur Bag of Words berbasis *CountVectorizer*, dan algoritma Regresi Logistik pada dataset komentar Instagram berbahasa Indonesia. Berdasarkan kesenjangan tersebut, penelitian ini menawarkan pendekatan klasifikasi menggunakan Regresi Logistik yang didukung tahapan cleaning, case folding, tokenization, stopword removal, dan stemming menggunakan Sastrawi, kemudian dilakukan ekstraksi fitur menggunakan *Bag of Words.* Tujuan penelitian ini adalah membangun dan mengevaluasi model klasifikasi yang mampu membedakan komentar cyberbullying dan non-cyberbullying pada Instagram berbahasa Indonesia. Hasil penelitian diharapkan dapat memberikan alternatif metode yang efektif dalam mendeteksi cyberbullying secara otomatis serta menjadi referensi bagi pengembangan sistem moderasi konten pada media sosial di masa mendatang [3], [4].

1. Tinjauan Pustaka

2.1. Cyberbullying

Cyberbullying merupakan tindakan agresif yang dilakukan melalui media digital dengan tujuan menghina, merendahkan, mengintimidasi, atau menyerang individu lain secara verbal maupun nonverbal. Perkembangan media sosial menyebabkan perilaku *cyberbullying* semakin mudah terjadi karena pengguna dapat memberikan komentar secara bebas tanpa batas ruang dan waktu. Bentuk *cyberbullying* pada media sosial umumnya berupa hinaan, ejekan, body shaming, ujaran kebencian, maupun ancaman yang dilakukan secara berulang [5]. Dampak *cyberbullying* dapat memengaruhi kondisi psikologis korban seperti stres, kecemasan, depresi, dan menurunnya rasa percaya diri, sehingga diperlukan sistem yang mampu membantu mendeteksi komentar negatif secara otomatis.

2.2. Media Sosial Instagram

Instagram merupakan salah satu platform media sosial yang memiliki jumlah pengguna aktif sangat besar dan banyak digunakan untuk berbagi foto, video, serta berinteraksi melalui komentar [6]. Tingginya aktivitas pengguna menyebabkan Instagram menjadi salah satu media yang rentan terhadap penyebaran komentar negatif dan *cyberbullying*. Komentar pada Instagram umumnya bersifat tidak terstruktur karena mengandung bahasa gaul, singkatan, emotikon, dan berbagai variasi bahasa lainnya sehingga proses analisis teks menjadi lebih kompleks. Oleh karena itu, diperlukan pendekatan pengolahan bahasa alami untuk memahami pola bahasa yang digunakan dalam komentar media sosial.

2.3. Natural Language Processing (NLP)

Natural Language Processing (NLP) merupakan cabang kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia. NLP memungkinkan komputer memahami, memproses, dan menganalisis data teks sehingga dapat digunakan dalam berbagai bidang seperti analisis sentimen, klasifikasi teks, *chatbot*, dan deteksi *cyberbullying* [7]. Dalam penelitian ini, NLP digunakan untuk membantu sistem mengenali pola kata pada komentar Instagram sehingga komentar dapat diklasifikasikan ke dalam kategori *cyberbullying* dan non-*cyberbullying* secara otomatis.

2.4. Preprocessing Teks

Preprocessing teks merupakan tahapan awal dalam pengolahan data teks yang bertujuan membersihkan dan menyiapkan data agar lebih mudah diproses oleh algoritma *Machine Learning* [8]. Komentar pada media sosial umumnya mengandung banyak noise seperti simbol, URL, emotikon, angka, dan penggunaan bahasa tidak baku sehingga perlu dilakukan *preprocessing* untuk meningkatkan kualitas data. Tahapan *preprocessing* sangat penting karena dapat membantu model *Machine Learning* mengenali pola kata dengan lebih baik dan meningkatkan performa klasifikasi.

2.5. Bag of Words (BoW)

Bag of Words (BoW) merupakan metode representasi teks yang digunakan untuk mengubah data teks menjadi bentuk numerik berdasarkan frekuensi kemunculan kata. Metode ini bekerja dengan menghitung jumlah kemunculan kata pada setiap dokumen tanpa memperhatikan urutan kata dalam kalimat. Hasil representasi tersebut berupa vektor numerik yang digunakan sebagai fitur input dalam proses klasifikasi menggunakan algoritma *Machine Learning*. Dalam penelitian ini, metode *Bag of Words* diimplementasikan menggunakan *CountVectorizer* dari library *Scikit-learn*. Setiap komentar Instagram diubah menjadi representasi numerik berdasarkan frekuensi kata yang muncul pada komentar tersebut sehingga model dapat mengenali pola kata yang sering digunakan dalam komentar *cyberbullying* maupun non-*cyberbullying* [9].

2.6. Machine Learning

Machine Learning merupakan cabang kecerdasan buatan yang memungkinkan sistem mempelajari pola dari data dan membuat prediksi secara otomatis tanpa diprogram secara eksplisit. *Machine Learning* banyak digunakan dalam berbagai bidang seperti analisis sentimen, klasifikasi teks, deteksi spam, dan pengenalan pola bahasa. Dalam penelitian ini, *Machine Learning* digunakan untuk melakukan klasifikasi komentar Instagram menjadi kategori *cyberbullying* dan non-*cyberbullying* berdasarkan pola kata yang dipelajari dari dataset [5]. Dengan memanfaatkan *Machine Learning*, proses deteksi *cyberbullying* dapat dilakukan secara otomatis dan lebih efisien dibandingkan proses moderasi manual.

2.7. Klasifikasi (Classification)

Klasifikasi merupakan salah satu teknik dalam *Machine Learning* yang digunakan untuk mengelompokkan data ke dalam kategori tertentu berdasarkan pola yang terdapat pada data. Pada penelitian ini, klasifikasi digunakan

untuk menentukan apakah suatu komentar termasuk kategori *cyberbullying* atau non-*cyberbullying*. Proses klasifikasi dilakukan menggunakan algoritma Regresi Logistik dengan memanfaatkan fitur hasil ekstraksi teks menggunakan metode *Bag of Words*. Hasil klasifikasi digunakan untuk membantu mendeteksi komentar negatif secara otomatis pada media sosial Instagram [10].

2.8. Regresi Logistik

Regresi Logistik merupakan salah satu algoritma *Machine Learning* yang digunakan untuk menyelesaikan permasalahan klasifikasi biner dengan memprediksi probabilitas suatu data termasuk ke dalam kelas tertentu melalui fungsi sigmoid [11]. Algoritma ini banyak diterapkan pada klasifikasi teks karena memiliki proses komputasi yang relatif ringan, mudah diimplementasikan, serta mampu menghasilkan performa klasifikasi yang baik. Pada penelitian ini, Regresi Logistik dimanfaatkan untuk mengklasifikasikan komentar Instagram ke dalam kategori *cyberbullying* dan *non-cyberbullying* berdasarkan fitur yang diperoleh dari proses ekstraksi menggunakan metode *Bag of Words* [12]. Probabilitas suatu data dihitung menggunakan fungsi sigmoid yang dinyatakan pada Persamaan (1).

$$P(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{1}$$

dengan (x) merupakan fitur masukan hasil ekstraksi *Bag of Words*, (β_0) adalah bias (*intercept*), (β_1) merupakan parameter model, dan (P(y)) menyatakan probabilitas suatu data termasuk ke dalam kelas tertentu. Nilai probabilitas yang dihasilkan berada pada rentang 0 hingga 1 dan digunakan sebagai dasar dalam menentukan apakah suatu komentar diklasifikasikan sebagai *cyberbullying* atau *non-cyberbullying.*

2.9. Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa model klasifikasi yang dibangun menggunakan algoritma Regresi Logistik dalam mengklasifikasikan komentar *cyberbullying* dan *non-cyberbullying*. Pada penelitian ini, model dievaluasi menggunakan *Confusion Matrix*, yaitu metode evaluasi yang membandingkan hasil prediksi model dengan kelas sebenarnya. *Confusion Matrix* terdiri atas empat komponen, yaitu *True Positive (TP)* yang menyatakan jumlah data *cyberbullying* yang berhasil diprediksi dengan benar sebagai *cyberbullying*, *True Negative (TN)* yang menyatakan jumlah data *non-cyberbullying* yang berhasil diprediksi dengan benar sebagai *non-cyberbullying, False Positive (FP)* yang menunjukkan jumlah data *non-cyberbullying* yang salah diprediksi sebagai *cyberbullying*, dan *False Negative (FN)* yang menunjukkan jumlah data *cyberbullying* yang salah diprediksi sebagai *non-cyberbullying.* Berdasarkan nilai TP, TN, FP, dan FN tersebut, dihitung metrik evaluasi berupa *accuracy, precision, recall*, dan *F1-score* untuk mengukur kinerja model. *Accuracy* menunjukkan tingkat ketepatan prediksi secara keseluruhan, *precision* menunjukkan ketepatan model dalam memprediksi komentar *cyberbullying*, *recall* mengukur kemampuan model mendeteksi komentar *cyberbullying* yang sebenarnya, sedangkan *F1-score* menunjukkan keseimbangan antara nilai *precision* dan *recall*. Hasil evaluasi digunakan untuk mengetahui tingkat performa model Regresi Logistik dalam mendeteksi *cyberbullying* pada komentar Instagram berbahasa Indonesia [8].

2. Metode penelitian

Penelitian ini menggunakan metodologi *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*) sebagai kerangka kerja dalam pengembangan model deteksi *cyberbullying* pada komentar Instagram berbahasa Indonesia. Metodologi ini dipilih karena menyediakan tahapan yang sistematis, terstruktur, dan banyak diterapkan dalam penelitian berbasis *data mining* maupun *Machine Learning*, sehingga mampu mendukung proses pengembangan model secara menyeluruh mulai dari identifikasi permasalahan hingga evaluasi hasil [13]. Tahapan dalam CRISP-DM meliputi *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Pada tahap *business understanding*, penelitian diawali dengan mengidentifikasi permasalahan *cyberbullying* pada media sosial Instagram, yaitu tingginya jumlah komentar negatif berupa hinaan, ejekan, *body shaming*, dan ujaran kebencian yang sulit dimoderasi secara manual akibat banyaknya komentar yang dipublikasikan setiap hari [10]. Berdasarkan permasalahan tersebut, penelitian ini bertujuan membangun model klasifikasi otomatis yang mampu membedakan komentar *cyberbullying* dan *non-cyberbullying* menggunakan pendekatan *Machine Learning*. Selanjutnya, dilakukan proses pemahaman dataset, prapemrosesan teks menggunakan teknik *Natural Language Processing* (NLP), ekstraksi fitur dengan metode *Bag of Words*, pembangunan model menggunakan algoritma Regresi Logistik, serta evaluasi performa model menggunakan *Confusion Matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*. Tahapan-tahapan tersebut dilakukan secara berurutan

agar model yang dihasilkan memiliki kemampuan klasifikasi yang baik dan dapat digunakan sebagai dasar dalam pengembangan sistem deteksi *cyberbullying* secara otomatis.

3.1. Business understanding

Tahap *business understanding* dilakukan untuk memahami permasalahan, tujuan, serta kebutuhan penelitian sebagai dasar dalam pengembangan model deteksi *cyberbullying*. Pada tahap ini dilakukan identifikasi terhadap fenomena *cyberbullying* yang sering muncul pada kolom komentar Instagram berbahasa Indonesia, seperti hinaan, ejekan, *body shaming*, ujaran kebencian, maupun bentuk perundungan verbal lainnya yang dapat memberikan dampak negatif terhadap kondisi psikologis korban, antara lain stres, kecemasan, depresi, menurunnya rasa percaya diri, hingga terganggunya kesehatan mental [13]. Tingginya jumlah komentar yang dipublikasikan setiap hari menyebabkan proses identifikasi dan moderasi secara manual menjadi kurang efektif, sehingga diperlukan suatu sistem yang mampu melakukan deteksi secara otomatis. Berdasarkan permasalahan tersebut, penelitian ini difokuskan pada pembangunan model klasifikasi menggunakan algoritma Regresi Logistik untuk membedakan komentar *cyberbullying* dan *non-cyberbullying*. Model yang dikembangkan diharapkan mampu mengidentifikasi komentar yang mengandung unsur *cyberbullying* secara cepat, akurat, dan konsisten sehingga dapat membantu proses moderasi konten pada media sosial serta mendukung terciptanya lingkungan digital yang lebih aman, sehat, dan nyaman bagi para pengguna Instagram.

3.2. Data Understanding

Tahap *data understanding* dilakukan untuk memahami karakteristik, kualitas, dan struktur dataset yang digunakan sebagai dasar dalam proses pembangunan model klasifikasi. Dataset penelitian berupa 650 komentar Instagram berbahasa Indonesia yang dikumpulkan dari beberapa unggahan publik pada akun Instagram yang memiliki tingkat interaksi tinggi. Proses pengumpulan data dilakukan dengan mengambil komentar dari unggahan yang dapat diakses secara publik, kemudian setiap komentar melalui proses pelabelan (*labeling*) secara manual ke dalam dua kategori, yaitu *cyberbullying* dan *non-cyberbullying,* berdasarkan kandungan bahasa dan konteks komentar. Seluruh data disimpan dalam format *CSV (Comma Separated Values)* dan diolah menggunakan bahasa pemrograman *Python* melalui *Google Colab* atau *Jupyter* Notebook. Pada tahap ini juga dilakukan eksplorasi data (*exploratory data analysis*) untuk memperoleh pemahaman mengenai kondisi awal dataset, yang meliputi pemeriksaan jumlah data, distribusi kelas, struktur atribut, analisis frekuensi kemunculan kata, serta verifikasi terhadap keberadaan *missing value* dan data duplikat [13]. Hasil eksplorasi menunjukkan bahwa distribusi data antara kelas *cyberbullying* dan *non-cyberbullying* relatif seimbang sehingga tidak menimbulkan ketimpangan kelas (*class imbalance*) yang signifikan. Selain itu, tidak ditemukan *missing value* maupun data duplikat, sehingga dataset dinilai memiliki kualitas yang baik dan layak digunakan sebagai data pelatihan serta pengujian model Regresi Logistik pada penelitian ini.

3.3. Data preparation

Tahap *data preparation* merupakan tahapan penting dalam penelitian berbasis *Natural Language Processing* (NLP). Pada tahap ini dilakukan proses pembersihan dan transformasi data agar data teks dapat diproses oleh algoritma *Machine Learning* dengan lebih optimal [13]. Komentar pada media sosial umumnya bersifat tidak terstruktur dan mengandung banyak noise seperti simbol, emotikon, URL, singkatan, serta penggunaan bahasa tidak baku. Oleh karena itu, diperlukan tahapan *preprocessing* untuk meningkatkan kualitas data.

GAMBAR 1. Tahap *data preparation*

Pada Gambar 1 ditunjukkan Tahap *data preparation* yang merupakan salah satu tahapan paling penting dalam penelitian berbasis *Natural Language Processing* (NLP) karena bertujuan untuk membersihkan dan menyiapkan data teks agar dapat diproses dengan baik oleh algoritma *Machine Learning*. Data komentar pada media sosial umumnya bersifat tidak terstruktur dan mengandung banyak noise seperti simbol, emotikon, URL, singkatan, maupun penggunaan bahasa tidak baku. Oleh karena itu, diperlukan beberapa proses *preprocessing* agar kualitas data meningkat dan model dapat melakukan klasifikasi dengan lebih optimal. Proses pertama yang dilakukan adalah seleksi data, yaitu memilih atribut yang relevan dengan kebutuhan penelitian. Pada tahap ini, data yang digunakan difokuskan pada kolom komentar sebagai fitur input dan kolom label sebagai target klasifikasi. Data lain yang tidak memiliki hubungan dengan proses klasifikasi dihapus agar pengolahan data menjadi lebih efisien dan terarah. Setelah proses seleksi data selesai dilakukan, tahap berikutnya adalah *cleaning*. Tahap *cleaning* bertujuan membersihkan data teks dari elemen-elemen yang tidak diperlukan seperti URL, mention, hashtag, angka, tanda baca, emotikon, dan simbol tertentu yang dapat mengganggu proses analisis. Proses ini dilakukan agar teks menjadi lebih bersih dan mudah diproses oleh sistem.

Selanjutnya dilakukan proses *case folding*, yaitu mengubah seluruh huruf pada teks menjadi huruf kecil (*lowercase*). Tahapan ini bertujuan menjaga konsistensi data sehingga sistem tidak menganggap kata yang sama sebagai fitur yang berbeda hanya karena penggunaan huruf kapital. Setelah itu dilakukan proses *tokenization*, yaitu memecah kalimat menjadi unit-unit kata atau token agar setiap kata dapat dianalisis secara terpisah oleh sistem *Machine Learning*. Tahapan *tokenization* sangat penting karena algoritma klasifikasi bekerja berdasarkan pola kemunculan kata pada data teks. Tahap berikutnya adalah *stopword removal*, yaitu proses menghapus kata-kata umum yang tidak memiliki pengaruh besar terhadap proses klasifikasi. Penghapusan stopword bertujuan agar model lebih fokus pada kata-kata penting yang berkaitan dengan *cyberbullying*. Setelah *stopword removal* selesai dilakukan, proses dilanjutkan dengan *stemming*. Tahap *stemming* bertujuan mengubah kata menjadi bentuk dasar untuk mengurangi variasi kata yang memiliki arti sama. Dengan *stemming*, sistem dapat mengenali hubungan antar kata dengan lebih baik sehingga membantu model dalam mempelajari pola klasifikasi. Setelah seluruh proses *preprocessing* selesai dilakukan, data teks kemudian diubah menjadi representasi numerik menggunakan metode *Bag of Words (BoW)* dengan bantuan *CountVectorizer*. Metode ini bekerja dengan menghitung frekuensi kemunculan kata pada setiap komentar sehingga menghasilkan representasi data dalam bentuk vektor numerik. Hasil transformasi tersebut digunakan sebagai fitur input dalam proses pelatihan model menggunakan algoritma Regresi Logistik. Tahap terakhir dalam *data preparation* adalah pembagian dataset menjadi data training dan data testing dengan perbandingan 80:20. Data training digunakan untuk melatih model, sedangkan data testing digunakan untuk menguji kemampuan model dalam melakukan klasifikasi terhadap data baru yang belum pernah dipelajari sebelumnya.

3. Hasil dan Pembahasan

Tahap hasil dan pembahasan dilakukan untuk menganalisis performa model Regresi Logistik dalam mendeteksi *cyberbullying* pada komentar Instagram berbahasa Indonesia. Penelitian ini menggunakan pendekatan *Natural Language Processing* (NLP) dengan metode *Bag of Words* berbasis *CountVectorizer* sebagai teknik ekstraksi fitur. Dataset yang digunakan terdiri dari 650 komentar Instagram berbahasa Indonesia yang telah diberi label ke dalam dua kategori, yaitu *cyberbullying* dan *non-cyberbullying* [14]. Sebelum proses klasifikasi dilakukan, seluruh data melalui tahapan *preprocessing* yang meliputi *cleaning*, *case folding*, *tokenization*, *stopword removal*, *stemming*, transformasi data menggunakan *Bag of Words*, serta pembagian dataset menjadi data *training* dan *testing*. Tahapan *preprocessing* berperan penting dalam meningkatkan kualitas data dengan mengurangi *noise*, menyeragamkan bentuk kata, serta menyederhanakan struktur teks sehingga pola bahasa pada komentar lebih mudah dipelajari oleh model. Selanjutnya, data teks diubah menjadi representasi numerik menggunakan *CountVectorizer* yang menghasilkan vektor berdasarkan frekuensi kemunculan kata pada setiap komentar. Representasi numerik tersebut digunakan sebagai fitur masukan bagi algoritma Regresi Logistik untuk mempelajari karakteristik komentar *cyberbullying* dan *non-cyberbullying*. Pada tahap *modeling*, dataset dibagi menjadi 80% *data training* dan 20% *data testing*, di mana data *training* digunakan untuk melatih model, sedangkan data *testing* digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data yang belum pernah dipelajari sebelumnya [15].

Hasil pengujian menunjukkan bahwa model Regresi Logistik mampu mengklasifikasikan komentar Instagram berbahasa Indonesia dengan performa yang baik. Evaluasi dilakukan menggunakan *Confusion Matrix* serta metrik *accuracy, precision, recall,* dan *F1-score* untuk mengukur kemampuan model dalam melakukan klasifikasi. Berdasarkan hasil evaluasi, model memperoleh nilai *accurac***y** sebesar 83%, *precision* sebesar 0,83, *recall* sebesar 0,83, dan *F1-score* sebesar 0,83. Nilai tersebut menunjukkan bahwa model memiliki tingkat ketepatan yang baik dalam membedakan komentar *cyberbullying* dan *non-cyberbullying*, serta mampu mendeteksi sebagian besar komentar *cyberbullying* dengan tingkat kesalahan yang relatif rendah. Hasil ini menunjukkan bahwa kombinasi

tahapan *preprocessing*, metode *Bag of Words*, dan algoritma Regresi Logistik mampu menghasilkan model klasifikasi yang efektif dalam mendeteksi *cyberbullying* pada komentar Instagram berbahasa Indonesia serta memberikan potensi untuk diterapkan sebagai dasar pengembangan sistem moderasi komentar otomatis pada media sosial.

Tabel 1. Hasil evaluasi model

Metrik	Nilai
Accuracy	83%
Precision	0,83
Recall	0,83
F1-Score	0,83

Selain menggunakan metrik evaluasi, hasil klasifikasi juga dianalisis menggunakan *Confusion matrix* untuk mengetahui jumlah prediksi yang benar maupun salah pada masing-masing kategori. Berdasarkan *Confusion matrix*, model mampu mengklasifikasikan sebagian besar komentar dengan benar, meskipun masih terdapat beberapa kesalahan klasifikasi pada komentar yang bersifat ambigu atau menggunakan bahasa tidak baku. Hal ini menunjukkan bahwa model masih memiliki keterbatasan dalam memahami konteks kalimat secara mendalam.

GAMBAR 2. *Confusion matrix*

Berdasarkan GAMBAR 2, hasil *Confusion Matrix* menunjukkan bahwa model Regresi Logistik memiliki kemampuan yang cukup baik dalam mendeteksi komentar *cyberbullying* pada komentar Instagram berbahasa Indonesia. Hal ini ditunjukkan oleh nilai *precision* yang tinggi, yang menunjukkan bahwa sebagian besar komentar yang diprediksi sebagai *cyberbullying* memang termasuk ke dalam kategori tersebut, serta nilai *recall* yang baik, yang mengindikasikan bahwa model mampu mendeteksi sebagian besar komentar *cyberbullying* yang terdapat pada dataset. Performa model tersebut dipengaruhi oleh tahapan *preprocessing* yang berperan penting dalam meningkatkan kualitas data, di mana proses *cleaning* mampu mengurangi *noise* pada teks, sedangkan *tokenization*, *stopword removal*, dan *stemming* membantu menyederhanakan bentuk kata sehingga pola bahasa pada komentar Instagram lebih mudah dikenali oleh model. Selain itu, metode *Bag of Words* berbasis *CountVectorizer* terbukti efektif dalam merepresentasikan data teks menjadi fitur numerik yang dapat diproses oleh algoritma *Machine Learning*. Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan, yaitu model belum mampu memahami konteks semantik secara mendalam sehingga beberapa komentar yang mengandung sindiran halus, ironi, atau bahasa gaul masih berpotensi mengalami kesalahan klasifikasi. Di samping itu, jumlah dataset yang digunakan masih terbatas sehingga belum sepenuhnya merepresentasikan keragaman bahasa yang digunakan oleh pengguna media sosial. Secara keseluruhan, hasil penelitian menunjukkan bahwa algoritma Regresi Logistik yang dipadukan dengan pendekatan *Natural Language Processing* dan metode *Bag of Words* efektif digunakan untuk mendeteksi *cyberbullying* pada komentar Instagram berbahasa Indonesia, sehingga model yang dihasilkan berpotensi menjadi dasar dalam pengembangan sistem moderasi komentar otomatis untuk mendukung terciptanya lingkungan media sosial yang lebih aman, sehat, dan nyaman bagi para pengguna.

4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma Regresi Logistik yang dipadukan dengan pendekatan Natural Language Processing (NLP) mampu mendeteksi cyberbullying pada komentar Instagram berbahasa Indonesia dengan performa yang baik. Model klasifikasi yang dibangun berhasil mengelompokkan komentar ke dalam kategori cyberbullying dan non-cyberbullying secara otomatis melalui tahapan preprocessing yang meliputi cleaning, case folding, tokenization, stopword removal, dan stemming. Tahapan tersebut terbukti meningkatkan kualitas data dengan mengurangi noise, menyeragamkan bentuk kata,

serta menyederhanakan teks sehingga pola bahasa yang berkaitan dengan cyberbullying lebih mudah dikenali oleh model. Selanjutnya, metode Bag of Words berbasis CountVectorizer berhasil mengubah data teks menjadi representasi numerik yang sesuai untuk proses klasifikasi menggunakan algoritma Regresi Logistik. Berdasarkan hasil evaluasi, model memperoleh nilai accuracy sebesar 83%, precision sebesar 0,83, recall sebesar 0,83, dan F1-score sebesar 0,83, yang menunjukkan bahwa model memiliki kemampuan yang cukup baik dan stabil dalam mengidentifikasi komentar cyberbullying, sebagaimana ditunjukkan oleh hasil Confusion Matrix. Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan, di antaranya kemampuan model yang belum optimal dalam memahami konteks semantik, ironi, sindiran, maupun penggunaan bahasa gaul yang sering dijumpai pada media sosial, serta jumlah dataset yang masih terbatas sehingga belum sepenuhnya merepresentasikan keragaman bahasa pengguna Instagram di Indonesia. Oleh karena itu, penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan beragam serta membandingkan performa Regresi Logistik dengan algoritma lain atau pendekatan berbasis deep learning untuk meningkatkan akurasi deteksi. Secara keseluruhan, penelitian ini menunjukkan bahwa kombinasi Natural Language Processing, Bag of Words, dan algoritma Regresi Logistik merupakan pendekatan yang efektif untuk mendeteksi cyberbullying pada komentar Instagram berbahasa Indonesia serta berpotensi menjadi dasar dalam pengembangan sistem moderasi komentar otomatis guna menciptakan lingkungan media sosial yang lebih aman, sehat, dan nyaman bagi pengguna.

Daftar pustaka

[1] W. N. Aji, Y. Tara, and A. D. Hartanto, "KEBENCIAN DALAM BAHASA INDONESIA Abstrak Keywords : Pendahuluan Tinjauan Pustaka Metode Penelitian," vol. 3, no. 1, pp. 6–9, 2021.

[2] K. Lukman and S. Novianto, "Komparasi Algoritma Naïve Bayes dan SVM untuk Identifikasi Cyberbullying Selebriti di Media Sosial Twitter," pp. 970–981, 2025, doi: 10.33364/algoritma/v.22-1.2196.

[3] E. Daniati et al., "ANALISIS SENTIMEN DENGAN PENDEKATAN ENSEMBLE LEARNING DAN WORD EMBEDDING PADA TWITTER Abstraksi Keywords : Pendahuluan Tinjauan Pustaka," vol. 4, no. 2, 2023.

[4] D. Yuliana, A. Ningrum, E. Daniati, M. N. Muzaki, and S. Informasi, "Perbandingan Model BERT dan RNN-LSTM pada Analisis Sentimen Aplikasi BRI Mobile," vol. 4, no. 2, pp. 75–85, 2025.

[5] Y. Setiawan, N. U. Maulidevi, and K. Surendro, "DETEKSI CYBERBULLYING DENGAN MESIN PEMBELAJARAN KLASIFIKASI (SUPERVISED LEARNING) : PELUANG DAN TANTANGAN CYBERBULLYING DETECTION USING SUPERVISED LEARNING :," vol. 9, no. 7, pp. 1577–1582, 2022, doi: 10.25126/jtiik.202296747.

[6] N. Lestari, M. Iqbal, and D. Larasati, "Bullying Verbal dalam Komentar di Media Sosial Instagram," pp. 99–108, 2025.

[7] E. Daniati, A. Prasetya, W. Sakti, G. Irianto, and L. Hernandez, "Few-Shot-BERT-RNN Narrative Structure Analysis for Andersen ' s Stories," vol. 9, no. July, pp. 1773–1782, 2025.

[8] S. Khairunnisa and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi," vol. 5, no. April, pp. 406–414, 2021, doi: 10.30865/mib.v5i2.2835.

[9] W. A. Luqyana, I. Cholissodin, and R. S. Perdana, "Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," vol. 2, no. 11, pp. 4704–4713, 2018.

[10] D. Nugraha, P. Astuti, P. S. Informatika, U. N. Mandiri, P. S. Informatika, and U. N. Mandiri, "MEDIA INSTAGRAM MENGGUNAKAN METODE," vol. 8, no. 2, pp. 153–164, 2023.

[11] N. Abdulloh, "Deteksi Cyberbullying pada Cuitan Media Sosial Twitter".

[12] M. Fahmuddin, M. K. Aidid, and M. J. Taslim, "IMPLEMENTASI ANALISIS REGRESI LOGISTIK DENGAN METODE MACHINE LEARNING UNTUK MENGKLASIFIKASI BERITA DI INDONESIA , dimana :," vol. 5, no. 03, pp. 155–162, 2023, doi: 10.35580/variansiunm116.

[13] D. Ruswanti, D. Susilo, and Riani, "Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C.45," vol. 30, no. 1, pp. 111–121, 2024, doi: 10.36309/goi.v30i1.266.

[14] A. Machmud, B. Wibisono, and N. Suryani, "Analisis Sentimen Cyberbullying Pada Komentar X Menggunakan Metode Naïve Bayes," vol. 5, no. 1, 2025.

[15] M. Faruqziddan, E. Daniati, and M. N. Muzaki, "Pendekatan BERT Dalam Analisis Sentimen Terhadap Kominfo Di Media Sosial X," vol. 4, no. 2, pp. 107–116, 2025.