

Perbandingan RoBERTa dan *Lexicon-Based* untuk *Pseudo-Labeling* Sentimen Komentar YouTube dengan *Machine Learning*

Valentino Juanito^{*1}, Manuele Jocelyn Blasius², Hafiz Irsyad³, Abdul Rahman⁴

^{1,2,3}Program Studi Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang

⁴Program Studi Teknik Elektro, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang

E-mail: ^{*1}valentinojuanito_2327250020@mhs.mdp.ac.id,

²manuelejocelynblasius_2327250087@mhs.mdp.ac.id, ³hafizirsyad@mdp.ac.id,

⁴arahman@mdp.ac.id

Abstrak

Penelitian ini bertujuan membandingkan pendekatan RoBERTa dan *Lexicon-Based* dalam menghasilkan pseudo-label sentimen komentar YouTube serta mengevaluasi kinerja algoritma machine learning dalam mengklasifikasikan label yang dihasilkan. Data penelitian berasal dari komentar pada video “Presiden Prabowo Menjawab (PART 2): Efisiensi, Pertaruhan MBG & Reformasi Kepolisian” dari Mata Najwa. Setelah proses penyaringan dan preprocessing, diperoleh 2.850 komentar yang dianalisis. Penelitian menerapkan dua jalur preprocessing, yaitu pemrosesan minimal pada RoBERTa untuk mempertahankan konteks kalimat serta pemrosesan tambahan pada *Lexicon-Based* melalui tokenisasi dan penghapusan stopword. Label sentimen yang dihasilkan oleh kedua pendekatan digunakan untuk melatih dan mengevaluasi algoritma Support Vector Machine (SVM), Naïve Bayes, dan Neural Network menggunakan fitur Term Frequency–Inverse Document Frequency (TF-IDF). Hasil evaluasi menunjukkan bahwa pada dataset berlabel RoBERTa, Neural Network memperoleh nilai accuracy sebesar 0,7333 dan F1-score sebesar 0,6024. Sementara itu, pada dataset berlabel *Lexicon-Based*, SVM menghasilkan accuracy sebesar 0,6439 dan F1-score sebesar 0,6323. Dengan demikian, Neural Network menghasilkan accuracy tertinggi pada dataset berlabel RoBERTa, sedangkan SVM menghasilkan F1-score tertinggi pada dataset berlabel *Lexicon-Based*. Perbedaan hasil tersebut menunjukkan adanya variasi kinerja klasifikasi pada kedua skema pelabelan otomatis. Namun, karena evaluasi didasarkan pada label hasil pseudo-labeling, nilai metrik tersebut menggambarkan kemampuan model dalam mereplikasi label otomatis dan belum secara langsung membuktikan ketepatan klasifikasi terhadap sentimen sebenarnya. Validasi menggunakan anotasi manusia diperlukan untuk menilai kualitas label dan memperkuat kesimpulan penelitian.

Kata kunci— Analisis Sentimen; *Pseudo-Labeling*; RoBERTa; *Lexicon-Based*; *Machine Learning*

Abstract

This study aims to compare RoBERTa and lexicon-based approaches for generating sentiment pseudo-labels for YouTube comments and to evaluate the performance of machine learning algorithms in classifying the resulting labels. The dataset consists of comments posted on the Mata Najwa video entitled “Presiden Prabowo Menjawab (PART 2): Efisiensi, Pertaruhan MBG & Reformasi Kepolisian.” Following data filtering and preprocessing, a total

of 2,850 comments were retained for analysis. Two preprocessing pipelines were implemented: minimal preprocessing for RoBERTa to preserve sentence-level context, and additional preprocessing for the lexicon-based approach involving tokenization and stopwords removal. The sentiment labels generated by both approaches were subsequently used to train and evaluate three machine learning algorithms: Support Vector Machine (SVM), Naïve Bayes, and Neural Network, using Term Frequency–Inverse Document Frequency (TF-IDF) features. The evaluation results indicate that, on the RoBERTa-labeled dataset, the Neural Network achieved an accuracy of 0.7333 and an F1-score of 0.6024. Meanwhile, on the lexicon-based-labeled dataset, SVM achieved an accuracy of 0.6439 and an F1-score of 0.6323. Thus, the Neural Network achieved the highest accuracy on the RoBERTa-labeled dataset, whereas SVM attained the highest F1-score on the lexicon-based-labeled dataset. These findings indicate variations in classification performance between the two automatic labeling schemes. However, because the evaluation was based on pseudo-labels, the reported metrics reflect the models' ability to reproduce automatically generated labels and do not directly establish the accuracy of sentiment classification against actual sentiment annotations. Human annotation-based validation is therefore necessary to assess label quality and strengthen the study's conclusions.

Keywords— Sentiment Analysis; Pseudo-Labeling; RoBERTa; Lexicon-Based; Machine Learning

1. PENDAHULUAN

Perkembangan teknologi digital menjadikan media sosial sebagai salah satu sarana bagi masyarakat untuk menyampaikan opini mengenai isu publik, termasuk kebijakan pemerintah dan dinamika politik. YouTube merupakan platform yang relevan untuk mengkaji fenomena tersebut karena menyediakan ruang bagi pengguna untuk menyampaikan tanggapan, kritik, dukungan, dan penilaian terhadap suatu isu melalui kolom komentar. Dalam konteks politik Indonesia, komentar YouTube telah dimanfaatkan sebagai sumber data untuk analisis sentimen berbasis teks guna mengidentifikasi kecenderungan opini pengguna [1][2]. Respons masyarakat terhadap kebijakan Presiden Republik Indonesia Prabowo Subianto, seperti efisiensi anggaran, program Makanan Bergizi Gratis (MBG), swasembada pangan, dan kebijakan strategis lainnya, juga menjadi objek kajian analisis sentimen di media sosial [3]. Potensi tersebut menunjukkan bahwa komentar YouTube dapat dimanfaatkan sebagai korpus teks untuk mengkaji pola sentimen pada konteks tertentu. Namun, hasil analisis terhadap komentar pada suatu video tidak dapat secara langsung dianggap mewakili keseluruhan opini masyarakat.

Analisis sentimen terhadap komentar media sosial menghadapi tantangan berupa penggunaan bahasa informal, singkatan, kesalahan penulisan (*typo*), emoji, simbol, dan struktur kalimat yang tidak selalu mengikuti kaidah bahasa formal. Karakteristik tersebut memerlukan tahapan *preprocessing* agar teks dapat dipersiapkan sesuai dengan kebutuhan metode analisis yang digunakan. Penelitian [1] menerapkan tahapan pembersihan teks, normalisasi, dan pemrosesan lanjutan dalam analisis komentar politik YouTube, sedangkan penelitian [4] menyoroti tantangan analisis sentimen yang berkaitan dengan variasi bahasa informal dan distribusi kelas sentimen. Meskipun *preprocessing* berperan penting dalam menyiapkan data, penerapannya perlu mempertimbangkan karakteristik metode yang digunakan. Pemrosesan yang menghilangkan kata atau struktur tertentu berpotensi memengaruhi informasi sentimen, khususnya pada metode yang bergantung pada konteks kalimat maupun pencocokan kata secara leksikal.

Salah satu pendekatan untuk memberikan label pada data yang belum memiliki anotasi

manual adalah *pseudo-labeling*, yaitu pemberian label secara otomatis menggunakan model atau aturan tertentu. Dalam penelitian ini, pendekatan tersebut diterapkan melalui RoBERTa dan *Lexicon-Based*. RoBERTa merupakan model berbasis *transformer* yang memanfaatkan representasi kontekstual untuk memproses teks. Penelitian [5] membahas penerapan BERT dan RoBERTa pada analisis sentimen berbahasa Indonesia serta melaporkan bahwa IndoBERT memberikan performa lebih tinggi daripada RoBERTa dalam konteks penelitian yang dikaji. Sementara itu, penelitian [6] mengkaji penerapan RoBERTa melalui *fine-tuning* untuk klasifikasi sentimen berbahasa Indonesia, termasuk pada kondisi distribusi kelas yang tidak seimbang. Berbeda dari pendekatan berbasis *transformer*, *Lexicon-Based* menentukan polaritas sentimen melalui pencocokan kata dengan kamus sentimen [7][8][9]. Pendekatan ini relatif sederhana dan transparan, tetapi hasilnya bergantung pada cakupan kosakata dalam kamus serta kemampuannya menangani konteks, negasi, dan penggunaan bahasa yang tidak literal. Perbedaan mekanisme tersebut menjadi dasar untuk mengkaji karakteristik label yang dihasilkan oleh kedua pendekatan.

Selain metode pelabelan, algoritma klasifikasi yang digunakan untuk mempelajari pola teks juga dapat memengaruhi hasil evaluasi. Penelitian [1][2][10][11][12] membahas perbandingan algoritma klasifikasi sentimen, termasuk SVM dan Naïve Bayes, pada berbagai jenis data teks. Sejumlah penelitian tersebut menunjukkan bahwa SVM dapat menjadi pendekatan yang kompetitif dalam klasifikasi teks, meskipun kinerja algoritma bergantung pada karakteristik dataset dan rancangan eksperimen. Di sisi lain, penelitian [13] membandingkan SVM dan RoBERTa dalam analisis sentimen terhadap liputan media dan menunjukkan relevansi perbandingan antara pendekatan berbasis fitur teks dan model *transformer*. Temuan dari penelitian-penelitian tersebut menunjukkan bahwa kinerja klasifikasi perlu dievaluasi secara empiris pada konteks data yang digunakan. Oleh karena itu, penelitian ini menggunakan SVM, Naïve Bayes, dan *Neural Network* sebagai algoritma pembanding untuk mengamati kemampuan masing-masing model dalam mengklasifikasikan data berdasarkan label yang dihasilkan oleh RoBERTa dan *Lexicon-Based*.

Berdasarkan penelitian terdahulu yang dirujuk, kajian mengenai analisis sentimen telah mencakup perbandingan algoritma klasifikasi [1][2][10][11][13], perbandingan pendekatan pelabelan [4], serta penerapan analisis sentimen pada isu politik dan kebijakan publik [2][3]. Namun, dalam lingkup literatur yang ditinjau pada penelitian ini, masih diperlukan analisis komparatif yang secara khusus mengkaji perbedaan distribusi label yang dihasilkan oleh RoBERTa dan *Lexicon-Based*, kemudian mengevaluasi kinerja beberapa algoritma *machine learning* pada masing-masing dataset berlabel tersebut. Fokus ini penting karena kedua pendekatan pelabelan memiliki mekanisme yang berbeda dan dapat menghasilkan distribusi kelas yang berbeda pula. Selain itu, perbedaan distribusi label dapat memengaruhi interpretasi metrik klasifikasi, sehingga nilai evaluasi tidak cukup ditafsirkan hanya berdasarkan akurasi. Dengan demikian, perbandingan perlu mempertimbangkan distribusi kelas serta metrik evaluasi yang digunakan, tanpa langsung mengasumsikan bahwa label yang lebih seimbang menunjukkan kualitas pelabelan yang lebih baik.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan membandingkan skema *pseudo-labeling* berbasis RoBERTa dan *Lexicon-Based* pada komentar YouTube berbahasa Indonesia mengenai isu kebijakan publik, serta mengevaluasi kinerja SVM, Naïve Bayes, dan *Neural Network* pada masing-masing dataset hasil pelabelan. Perbandingan dilakukan dengan menerapkan jalur *preprocessing* yang disesuaikan dengan karakteristik kedua pendekatan, kemudian merepresentasikan teks menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk proses klasifikasi. Kinerja model dievaluasi menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*. Kontribusi penelitian ini terletak pada analisis komparatif terhadap distribusi pseudo-label dan kinerja klasifikasi pada dua skema pelabelan

otomatis dalam konteks komentar YouTube berbahasa Indonesia. Hasil penelitian diharapkan memberikan gambaran empiris mengenai perbedaan karakteristik kedua skema tersebut dan implikasinya terhadap klasifikasi sentimen. Karena label yang digunakan diperoleh secara otomatis, evaluasi dalam penelitian ini terutama menunjukkan kemampuan model dalam mereplikasi pseudo-label; penilaian terhadap ketepatan sentimen sebenarnya memerlukan validasi menggunakan anotasi manusia.

2. METODE PENELITIAN

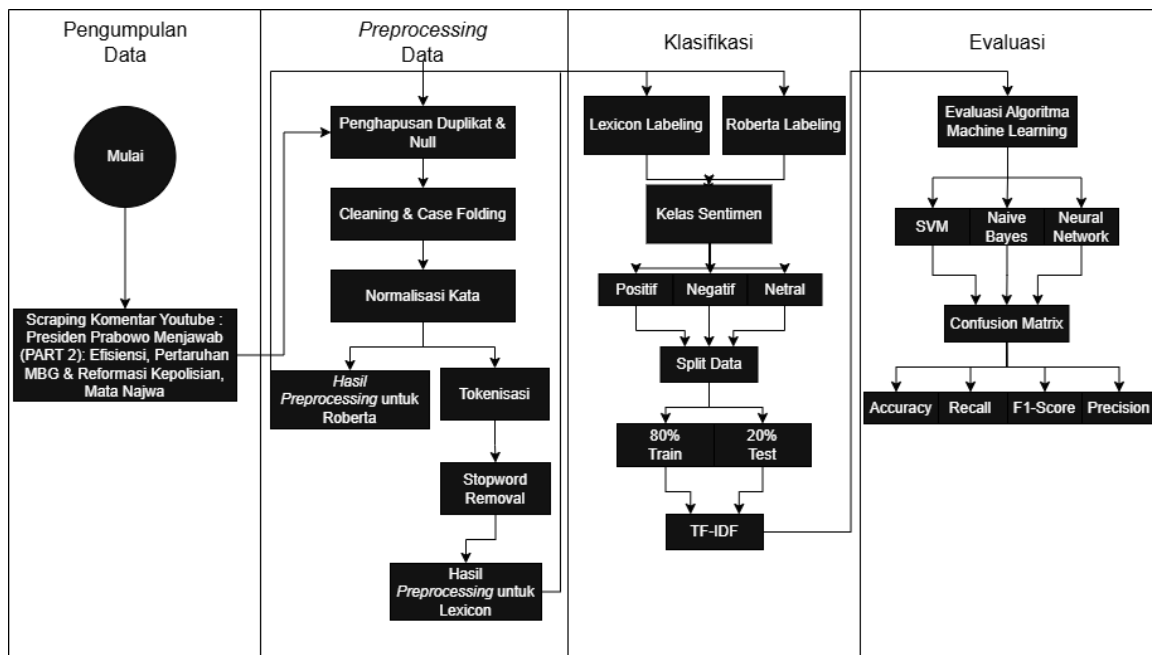
Penelitian ini menerapkan metode eksperimen komparatif untuk membandingkan dua skema *pseudo-labeling*, yaitu RoBERTa Labeling dan *Lexicon-Based Labeling*, dalam menghasilkan label sentimen pada komentar YouTube. Label yang dihasilkan oleh masing-masing pendekatan selanjutnya digunakan untuk melatih dan mengevaluasi beberapa algoritma *machine learning*. Perbandingan dilakukan dengan mengamati distribusi label sentimen yang dihasilkan serta kinerja algoritma klasifikasi dalam mereplikasi label tersebut pada data uji.

Penelitian ini menggunakan pendekatan komparatif dengan dua skema pelabelan otomatis yang diterapkan pada sumber data komentar yang sama. Skema pertama menggunakan model RoBERTa yang telah dilatih untuk klasifikasi sentimen bahasa Indonesia, sedangkan skema kedua menggunakan kamus sentimen berbahasa Indonesia melalui pendekatan berbasis leksikon. Hasil dari kedua skema tersebut diproses secara terpisah menggunakan algoritma *Support Vector Machine* (SVM), *Naïve Bayes*, dan *Neural Network* dengan representasi fitur *Term Frequency-Inverse Document Frequency* (TF-IDF). Pemisahan proses ini bertujuan untuk mengidentifikasi perbedaan kinerja klasifikasi berdasarkan sumber label yang digunakan.

Perlu ditegaskan bahwa label yang digunakan dalam eksperimen ini merupakan label otomatis (*pseudo-label*), bukan label yang diperoleh melalui anotasi manual oleh anotator manusia. Oleh karena itu, metrik evaluasi pada tahap klasifikasi digunakan untuk mengukur kemampuan algoritma dalam mereplikasi label hasil *pseudo-labeling*. Metrik tersebut tidak secara langsung menunjukkan tingkat ketepatan klasifikasi terhadap sentimen sebenarnya dari komentar pengguna. Keterbatasan ini menjadi pertimbangan dalam interpretasi hasil perbandingan kedua skema pelabelan.

2.1 Rancangan Penelitian

Rancangan penelitian ini disusun untuk menggambarkan alur kerja analisis sentimen komentar YouTube terhadap video “Presiden Prabowo Menjawab (PART 2): Efisiensi, Pertaruhan MBG & Reformasi Kepolisian | Mata Najwa”. Penelitian menggunakan pendekatan eksperimen komparatif untuk membandingkan skema RoBERTa Labeling dan *Lexicon-Based Labeling* dalam menghasilkan label sentimen otomatis yang selanjutnya diuji menggunakan algoritma *machine learning*. Secara umum, alur penelitian terbagi ke dalam empat tahapan utama, yaitu pengumpulan data, *preprocessing* data, klasifikasi, serta evaluasi model. Rincian alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Rancangan Penelitian

Penelitian ini menerapkan dua jalur *preprocessing* yang berbeda sesuai dengan karakteristik kebutuhan representasi input dari masing-masing metode pelabelan. Jalur RoBERTa menggunakan skema *preprocessing* minimal agar hubungan antarkata dan informasi kontekstual dalam kalimat tetap terjaga. Sementara itu, jalur *Lexicon-Based* menerapkan tahapan tambahan berupa tokenisasi dan *stopword removal* untuk mendukung proses pencocokan kata dengan kamus sentimen. Perbedaan ini disesuaikan dengan mekanisme kerja kedua pendekatan, tetapi sekaligus menjadi salah satu faktor yang perlu diperhatikan ketika menginterpretasikan hasil perbandingan.

Secara operasional, penelitian dimulai dengan mengumpulkan komentar dari video yang menjadi objek penelitian. Data mentah kemudian melalui tahap pembersihan dan standarisasi teks sebelum dipisahkan ke dalam dua jalur *preprocessing*. Pada jalur pertama, teks digunakan untuk memperoleh label sentimen melalui model RoBERTa. Pada jalur kedua, teks diproses lebih lanjut untuk memperoleh label berdasarkan pencocokan kata dengan kamus sentimen. Masing-masing dataset berlabel kemudian dibagi menjadi data latih dan data uji, direpresentasikan menggunakan TF-IDF, serta digunakan dalam pelatihan tiga algoritma *machine learning*. Tahap akhir dilakukan dengan mengevaluasi prediksi model menggunakan *confusion matrix*, akurasi, presisi, *recall*, dan F1-score.

Rancangan tersebut memungkinkan penelitian membandingkan hasil klasifikasi berdasarkan dua sumber label otomatis. Namun, karena skema pelabelan dan jalur *preprocessing* diterapkan secara berbeda, hasil eksperimen perlu ditafsirkan sebagai perbandingan dua skema pelabelan beserta konfigurasi pemrosesannya, bukan sebagai pengujian yang mengisolasi pengaruh metode pelabelan semata. Untuk mengukur ketepatan label terhadap sentimen sebenarnya, diperlukan data pembanding yang telah dianotasi oleh manusia.

2.2 Pengumpulan Data

Pengumpulan data dilakukan menggunakan teknik *scraping* komentar YouTube pada video “Presiden Prabowo Menjawab (PART 2): Efisiensi, Pertaruhan MBG & Reformasi Kepolisian | Mata Najwa”. Video tersebut dipilih karena memuat pembahasan mengenai isu kebijakan publik yang berpotensi memunculkan beragam respons masyarakat, seperti efisiensi anggaran, program

Makanan Bergizi Gratis (MBG), dan reformasi kepolisian. Komentar YouTube digunakan sebagai sumber data karena memuat respons tekstual pengguna terhadap isi video yang dibahas.

Pemilihan satu video sebagai sumber data memberikan batasan konteks yang relatif terfokus untuk membandingkan kedua skema pelabelan. Meskipun demikian, komentar pada satu video tidak dapat secara otomatis dianggap mewakili seluruh opini masyarakat Indonesia terhadap isu politik atau kebijakan publik. Oleh karena itu, ruang lingkup temuan penelitian dibatasi pada data komentar yang dikumpulkan dari video tersebut.

Data yang dikumpulkan berfokus pada atribut isi teks komentar pengguna, sedangkan metadata lain, seperti nama pengguna, waktu komentar, jumlah *like*, dan identitas akun, tidak digunakan dalam proses klasifikasi. Pembatasan atribut tersebut dilakukan agar analisis berfokus pada informasi tekstual yang menjadi masukan bagi metode pelabelan dan algoritma klasifikasi. Data hasil *scraping* kemudian disimpan dalam format *Comma-Separated Values* (CSV) agar dapat dikelola dan diproses lebih lanjut menggunakan Python.

Proses pengumpulan data menghasilkan 2.917 komentar mentah. Dataset tersebut digunakan sebagai data awal sebelum melalui tahapan pembersihan. Data mentah belum dapat langsung digunakan dalam proses *pseudo-labeling* dan klasifikasi karena teks media sosial dapat memuat data kosong, komentar duplikat, URL, simbol, singkatan, kata tidak baku, serta variasi penulisan lainnya. Oleh sebab itu, diperlukan tahap pembersihan dan standarisasi teks untuk menyiapkan data sebelum diproses lebih lanjut.

Dalam penelitian ini, jumlah komentar yang diperoleh dari proses pengumpulan data dibedakan dari jumlah komentar yang dipertahankan setelah pembersihan dan *preprocessing*. Sebanyak 2.917 komentar merupakan jumlah data mentah, sedangkan jumlah data setelah pembersihan duplikasi dan data kosong adalah 2.896 komentar. Setelah proses *preprocessing*, jumlah data yang digunakan dalam analisis adalah 2.850 komentar. Perbedaan jumlah tersebut menunjukkan adanya data yang tidak dipertahankan selama tahapan pembersihan dan penyiapan teks.

Untuk meningkatkan keterulangan penelitian, informasi teknis pengumpulan data perlu dilaporkan secara konsisten dengan pelaksanaan eksperimen, meliputi tanggal pengambilan data, perangkat atau pustaka yang digunakan untuk melakukan *scraping*, cara pengambilan komentar, serta kriteria penghapusan data. Informasi tersebut perlu diisi berdasarkan catatan eksperimen yang sebenarnya karena tidak seluruh rinciannya tersedia dalam uraian metode awal. Pelaporan informasi ini diperlukan agar pembaca dapat memahami batasan dataset dan menilai kemungkinan pengulangan proses pengumpulan data.

2.3 Preprocessing Data

Tahap *preprocessing* dilakukan untuk membersihkan dan menyiapkan data komentar sebelum digunakan pada proses *pseudo-labeling*. Penelitian ini menerapkan dua jalur *preprocessing* yang berbeda sesuai dengan karakteristik input yang dibutuhkan oleh pendekatan RoBERTa dan *Lexicon-Based*. Perbedaan tersebut didasarkan pada mekanisme pemrosesan teks masing-masing metode dalam menentukan sentimen. Pada pendekatan RoBERTa, teks dipertahankan agar informasi kontekstual tetap tersedia bagi model, sedangkan pendekatan *Lexicon-Based* memerlukan pemrosesan yang mendukung pencocokan kata dengan kamus sentimen.

Tahap awal *preprocessing* diterapkan pada seluruh data komentar. Proses ini diawali dengan penghapusan data duplikat dan data kosong agar komentar yang berulang atau tidak memiliki isi teks tidak ikut diproses pada tahap berikutnya. Penghapusan duplikasi bertujuan mengurangi pengulangan teks yang sama, sedangkan penghapusan data kosong dilakukan karena data tanpa isi tekstual tidak dapat digunakan untuk menentukan polaritas sentimen melalui pendekatan yang diterapkan dalam penelitian ini.

Selanjutnya, dilakukan *cleaning* untuk menghapus elemen yang tidak relevan dengan analisis teks, seperti URL, *mention*, *hashtag*, simbol tertentu, karakter khusus, dan spasi berlebih [9]. Tahap ini bertujuan mengurangi unsur yang dapat mengganggu penyeragaman teks. Namun, penerapan *cleaning* perlu dilakukan secara hati-hati karena simbol, tanda baca, atau bentuk penulisan tertentu dapat membawa informasi pragmatis dalam komentar media sosial. Oleh sebab itu, elemen yang dihapus perlu mengikuti aturan pembersihan yang diterapkan dalam kode eksperimen, bukan diasumsikan seluruhnya tidak memiliki makna.

Setelah proses *cleaning*, dilakukan *case folding* dengan mengubah seluruh huruf menjadi huruf kecil agar bentuk penulisan kata menjadi lebih seragam [9]. Penyeragaman ini membantu mengurangi perbedaan representasi kata yang hanya disebabkan oleh penggunaan huruf kapital. Tahap berikutnya adalah normalisasi kata, yaitu proses mengubah kata tidak baku, singkatan, atau bahasa informal menjadi bentuk yang lebih baku [6]. Pada tahap ini juga dilakukan koreksi *typo* ringan dengan mempertimbangkan kemiripan kata terhadap daftar kata baku. Normalisasi diharapkan dapat mengurangi variasi penulisan yang berpotensi menghambat pemrosesan teks, khususnya pada pencocokan kata dengan kamus sentimen.

Meskipun demikian, normalisasi kata dan koreksi *typo* dapat memengaruhi makna komentar apabila kata informal, singkatan, atau bentuk penulisan tertentu diubah secara tidak tepat. Oleh karena itu, penerapannya perlu memperhatikan konteks dan makna kata, terutama pada komentar politik yang dapat memuat ironi, sarkasme, ungkapan tidak langsung, atau istilah khusus. Aturan normalisasi dan daftar kata yang digunakan sebaiknya didokumentasikan berdasarkan implementasi aktual agar proses tersebut dapat ditelusuri dan diulang.

Hasil *preprocessing* umum kemudian dipisahkan ke dalam dua jalur sesuai kebutuhan masing-masing pendekatan *pseudo-labeling*. Pada jalur RoBERTa, data hasil *preprocessing* umum digunakan secara langsung tanpa penerapan tokenisasi manual maupun *stopword removal*. Pendekatan ini dipilih karena RoBERTa merupakan model berbasis *transformer* yang memanfaatkan hubungan kontekstual antarkata untuk memahami makna kalimat. Penghapusan kata secara agresif berpotensi menghilangkan informasi penting, seperti kata negasi atau kata penghubung yang dapat memengaruhi interpretasi sentimen. Oleh karena itu, teks dipertahankan tanpa penghapusan *stopword* secara manual agar konteks kalimat tetap tersedia ketika dilakukan pelabelan sentimen.

Tidak diterapkannya tokenisasi manual pada jalur RoBERTa tidak berarti model memproses teks tanpa tokenisasi. Model *transformer* tetap menggunakan mekanisme tokenisasi internal yang sesuai dengan arsitektur dan *tokenizer*-nya. Dengan demikian, yang dibedakan dalam penelitian ini adalah tidak dilakukannya tokenisasi manual pada tahap *preprocessing* sebelum teks dimasukkan ke dalam *pipeline* RoBERTa.

Sebaliknya, pada jalur *Lexicon-Based* dilakukan tahapan tambahan berupa tokenisasi dan *stopword removal*. Tokenisasi digunakan untuk memecah kalimat menjadi unit kata sehingga setiap kata dapat dicocokkan dengan entri dalam kamus sentimen. Setelah itu, dilakukan penghapusan *stopword* untuk mengurangi kata-kata umum yang dianggap tidak memberikan kontribusi langsung terhadap polaritas sentimen. Tahap ini dimaksudkan untuk mendukung proses pencocokan kata dengan kamus yang digunakan pada tahap pelabelan.

Penghapusan *stopword* perlu memperhatikan kata yang berfungsi sebagai negasi atau pengubah makna. Kata seperti “tidak”, “bukan”, atau “belum”, misalnya, dapat mengubah polaritas suatu pernyataan apabila dihilangkan. Oleh karena itu, daftar *stopword* yang digunakan perlu diperiksa agar kata yang berperan penting dalam penentuan sentimen tidak ikut terhapus. Ketentuan ini penting terutama pada pendekatan berbasis leksikon karena penentuan label sangat bergantung pada kata-kata yang berhasil dicocokkan dengan kamus.

Dengan demikian, penelitian ini menerapkan dua jalur *preprocessing* untuk menyesuaikan representasi teks dengan karakteristik masing-masing pendekatan *pseudo-labeling*. Jalur RoBERTa berfokus pada pemeliharaan konteks kalimat, sedangkan jalur *Lexicon-Based* berfokus

pada pencocokan unit kata dengan kamus sentimen. Karena jalur pemrosesan tersebut berbeda, pengaruh metode pelabelan tidak dapat sepenuhnya dipisahkan dari pengaruh *preprocessing*. Hal ini perlu diperhatikan ketika membandingkan distribusi label dan hasil klasifikasi dari kedua skema.

2.4 Pseudo-Labeling (Klasifikasi)

Tahap *pseudo-labeling* dilakukan untuk memberikan label sentimen secara otomatis pada data komentar yang telah melalui *preprocessing*. Pendekatan ini digunakan karena dataset komentar YouTube yang diperoleh belum memiliki label sentimen hasil anotasi manual. Dalam penelitian ini, *pseudo-labeling* dilakukan menggunakan dua pendekatan, yaitu RoBERTa Labeling dan *Lexicon-Based Labeling*. Kedua pendekatan menghasilkan tiga kelas sentimen, yaitu positif, negatif, dan netral (*neutral*). Hasil pelabelan dari masing-masing pendekatan kemudian digunakan sebagai dataset berlabel untuk proses pelatihan model *machine learning*.

Penggunaan dua pendekatan pelabelan bertujuan untuk mengamati bagaimana perbedaan mekanisme penentuan sentimen menghasilkan label yang kemudian dipelajari oleh algoritma klasifikasi. RoBERTa menentukan label melalui model klasifikasi sentimen berbasis *transformer*, sedangkan *Lexicon-Based* menentukan label berdasarkan kemunculan kata positif dan negatif dalam kamus sentimen. Kedua pendekatan tersebut memiliki karakteristik yang berbeda sehingga distribusi label yang dihasilkan tidak harus sama.

Perbedaan distribusi label merupakan bagian penting dari analisis, tetapi tidak dapat digunakan sebagai satu-satunya dasar untuk menyimpulkan bahwa salah satu metode menghasilkan label yang lebih akurat. Distribusi label yang lebih seimbang tidak secara otomatis menunjukkan kualitas label yang lebih baik, sebagaimana distribusi yang didominasi satu kelas tidak secara langsung membuktikan bahwa label tersebut salah. Penilaian terhadap kualitas label memerlukan pembandingan yang independen, misalnya anotasi manusia pada sebagian data.

2.4.1 RoBERTa Labeling

RoBERTa Labeling dilakukan dengan memanfaatkan model *transformer* yang telah dilatih untuk tugas klasifikasi sentimen bahasa Indonesia. Model yang digunakan adalah *willwo/indonesian-RoBERTa-base-sentiment-classifier* melalui *pipeline* analisis sentimen. Pendekatan ini dipilih karena RoBERTa memproses teks secara kontekstual dan mempertimbangkan hubungan antarkata dalam suatu kalimat sehingga berpotensi sesuai untuk menangani komentar media sosial dengan struktur bahasa informal dan variasi penulisan [5][6].

Pada tahap ini, data yang digunakan adalah hasil *preprocessing* khusus RoBERTa, yaitu teks yang telah melalui *cleaning*, *case folding*, dan normalisasi kata tanpa tokenisasi manual maupun *stopword removal*. Hal ini dilakukan agar konteks kalimat tetap terjaga selama proses pelabelan. Setiap komentar dimasukkan ke dalam *pipeline* analisis sentimen, kemudian model menghasilkan prediksi label berdasarkan keluaran klasifikasi untuk kelas positif, negatif, atau netral (*neutral*).

Dalam implementasinya, model klasifikasi sentimen menghasilkan keluaran yang berkaitan dengan kelas sentimen dan skor prediksi. Penetapan label dalam penelitian ini mengikuti label dengan probabilitas atau skor prediksi tertinggi sebagaimana dihasilkan oleh *pipeline* yang digunakan. Agar prosedur ini dapat direplikasi, versi pustaka, konfigurasi *pipeline*, pemetaan keluaran model ke dalam tiga kelas yang digunakan pada penelitian, serta aturan penanganan keluaran perlu dilaporkan berdasarkan implementasi aktual. Jika model menghasilkan skor keyakinan, perlu dijelaskan apakah skor tersebut hanya digunakan untuk menentukan label tertinggi atau juga digunakan sebagai ambang batas penerimaan label. Informasi mengenai penggunaan ambang batas tidak boleh diasumsikan apabila tidak diterapkan dalam eksperimen.

Hasil RoBERTa Labeling menghasilkan dataset berlabel yang digunakan secara terpisah pada tahap pelatihan dan evaluasi algoritma *machine learning*. Dengan pendekatan ini, penelitian dapat mengamati kemampuan SVM, Naïve Bayes, dan *Neural Network* dalam mereplikasi label

yang dihasilkan oleh model RoBERTa. Namun, karena label RoBERTa tidak dibandingkan dengan anotasi manusia dalam rancangan eksperimen yang dilaporkan, hasil klasifikasi hanya dapat menunjukkan kesesuaian prediksi algoritma terhadap label RoBERTa, bukan ketepatan sentimen sebenarnya secara langsung.

2.4.2 *Lexicon-Based Labeling*

Lexicon-Based Labeling dilakukan menggunakan kamus sentimen berbahasa Indonesia yang terdiri atas daftar kata positif dan negatif. Pada penelitian ini, data yang digunakan merupakan hasil *preprocessing* khusus *Lexicon-Based*, yaitu teks yang telah melalui tokenisasi, pembersihan karakter nonalfabet sesuai aturan yang diterapkan, dan *stopword removal*. Tahapan tersebut dilakukan agar kata-kata dalam komentar dapat dicocokkan dengan entri yang terdapat pada kamus sentimen.

Proses pelabelan dilakukan dengan menghitung jumlah kata positif dan negatif yang muncul pada setiap komentar. Setiap kata yang cocok dengan kamus positif atau negatif diakumulasikan untuk memperoleh skor sentimen. Skor sentimen dihitung berdasarkan selisih antara jumlah kata positif dan jumlah kata negatif. Secara konseptual, skor sentimen dapat dituliskan sebagai berikut:

$$S(c) = N_+(c) - N_-(c)$$

Keterangan:

$S(c)$ = skor sentimen komentar.

$N_+(c)$ = jumlah kata positif yang ditemukan dalam komentar.

$N_-(c)$ = jumlah kata negatif yang ditemukan dalam komentar.

Tanda minus (−) berarti jumlah kata negatif dikurangkan dari jumlah kata positif.

Berdasarkan skor tersebut, aturan penentuan label sentimen ditetapkan sebagai berikut. Jika skor sentimen lebih besar dari nol, komentar diberi label positif. Jika skor sentimen lebih kecil dari nol, komentar diberi label negatif. Jika skor sentimen sama dengan nol, komentar diberi label netral (*neutral*). Dengan aturan ini, *Lexicon-Based Labeling* menghasilkan dua informasi utama, yaitu skor sentimen dan label sentimen.

Aturan tersebut bersifat deterministik berdasarkan kemunculan kata dalam kamus. Dengan demikian, komentar yang tidak memiliki kata yang cocok dengan kamus positif maupun negatif akan memperoleh skor nol dan masuk ke kelas netral. Kelas netral dalam pendekatan ini tidak selalu berarti komentar benar-benar netral secara semantik; label tersebut juga dapat muncul karena komentar tidak mengandung kata yang terdaftar dalam kamus atau karena kata yang menunjukkan sentimen tidak berhasil dicocokkan. Hal ini merupakan salah satu keterbatasan pendekatan berbasis leksikon.

Selain itu, pencocokan kata berdasarkan kamus memiliki keterbatasan dalam menangani negasi, ironi, sarkasme, konteks kalimat, dan kata yang memiliki makna berbeda bergantung pada penggunaannya. Misalnya, keberadaan kata positif tidak selalu menunjukkan sentimen positif apabila kata tersebut berada dalam kalimat bernada negatif atau sarkastis. Oleh karena itu, sumber kamus, cakupan kosakata, cara pencocokan kata, dan daftar *stopword* perlu dijelaskan agar pembaca dapat menilai karakteristik serta keterbatasan label yang dihasilkan.

Pendekatan *Lexicon-Based* dipilih sebagai pembanding karena bersifat transparan, relatif sederhana, dan tidak memerlukan pelatihan model klasifikasi yang kompleks. Pendekatan berbasis leksikon juga telah digunakan dalam penelitian analisis sentimen berbahasa Indonesia, terutama pada data media sosial dan komentar politik [1][9]. Meskipun demikian, penggunaan pendekatan ini sebagai pembanding tidak dengan sendirinya menjamin bahwa label yang dihasilkan menjadi label acuan yang benar. Dalam penelitian ini, label *Lexicon-Based* digunakan sebagai salah satu skema *pseudo-labeling* yang dibandingkan dengan RoBERTa, bukan sebagai *ground truth* yang telah tervalidasi oleh anotator manusia.

Untuk memperkuat keterulangan eksperimen, informasi mengenai nama dan sumber kamus sentimen, versi atau tanggal akses kamus, aturan pencocokan kata, penanganan kata berulang, serta penanganan negasi perlu dicantumkan sesuai implementasi aktual. Jika informasi tersebut tidak tersedia dalam kode atau catatan penelitian, informasi itu perlu dilengkapi sebelum naskah dinyatakan final.

2.5 Pembagian Data dan Ekstraksi Fitur

Setelah proses *pseudo-labeling* selesai dilakukan, dataset hasil RoBERTa Labeling dan *Lexicon-Based Labeling* diproses secara terpisah pada tahap pembagian data dan ekstraksi fitur. Pemisahan ini dilakukan agar performa model *machine learning* dapat dibandingkan berdasarkan sumber label yang berbeda. Setiap dataset terdiri atas teks hasil *preprocessing* dan label sentimen yang terbagi ke dalam tiga kelas, yaitu positif, negatif, dan netral (*neutral*).

Pembagian data dilakukan menggunakan metode *train-test split* dengan rasio 80:20. Sebanyak 80% data digunakan sebagai data latih, sedangkan 20% sisanya digunakan sebagai data uji. Berdasarkan jumlah 2.850 komentar yang dilaporkan, pembagian tersebut menghasilkan 2.280 data latih dan 570 data uji untuk setiap dataset, sepanjang seluruh komentar digunakan dalam pembagian tanpa pengurangan tambahan pada tahap ini. Pembagian dilakukan secara terpisah pada dataset RoBERTa dan *Lexicon-Based* sehingga setiap skema pelabelan memiliki data latih dan data uji masing-masing.

Untuk meningkatkan keterulangan dan mengurangi potensi variasi hasil akibat pembagian data yang berbeda, nilai *random seed* yang digunakan dalam *train-test split* perlu dilaporkan. Selain itu, proporsi kelas dalam data latih dan data uji perlu diperiksa. Apabila pembagian menggunakan parameter *stratify* berdasarkan label sentimen, hal tersebut perlu dinyatakan secara eksplisit. Jika parameter tersebut tidak digunakan dalam eksperimen awal, naskah tidak boleh menyatakan bahwa pembagian dilakukan secara terstratifikasi. Penerapan pembagian terstratifikasi dapat dipertimbangkan pada eksperimen ulang agar proporsi kelas lebih terjaga, terutama ketika distribusi label tidak seimbang.

Setelah pembagian data dilakukan, teks direpresentasikan menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk mengubah data tekstual menjadi vektor numerik yang dapat diproses oleh algoritma *machine learning*. TF-IDF memberikan bobot pada kata berdasarkan frekuensi kemunculannya dalam dokumen dan tingkat kekhasan kata tersebut terhadap kumpulan dokumen. Representasi ini memungkinkan algoritma klasifikasi mempelajari pola hubungan antara fitur tekstual dan label sentimen yang digunakan.

Dalam proses evaluasi yang benar, pembentukan representasi TF-IDF harus memperhatikan pemisahan data latih dan data uji. Parameter representasi, termasuk kosakata dan bobot IDF, perlu dipelajari hanya dari data latih. Data uji kemudian ditransformasikan menggunakan representasi yang telah dipelajari dari data latih, tanpa melakukan *fit* ulang pada data uji. Prosedur ini diperlukan untuk mencegah kebocoran informasi (*data leakage*) dari data uji ke proses pelatihan. Karena itu, implementasi sebaiknya menggunakan alur *fit* TF-IDF pada data latih dan *transform* pada data uji secara terpisah, atau menggunakan *pipeline* yang memastikan urutan proses tersebut.

Jumlah fitur TF-IDF yang dilaporkan dalam eksperimen adalah 8.739 fitur. Nilai tersebut perlu dipastikan berasal dari hasil ekstraksi fitur pada konfigurasi yang benar-benar digunakan. Untuk meningkatkan keterulangan penelitian, parameter TF-IDF yang digunakan, seperti pengaturan rentang *n-gram*, batas frekuensi kata, dan parameter lain yang diaktifkan, perlu dicantumkan sesuai kode program. Parameter yang tidak digunakan tidak perlu ditambahkan hanya untuk melengkapi uraian metode.

Penting untuk diperhatikan bahwa kedua dataset memiliki sumber label dan jalur *preprocessing* yang berbeda. Oleh karena itu, hasil klasifikasi pada kedua dataset dibandingkan dengan tetap memperhatikan perbedaan distribusi kelas, representasi teks, dan sumber label.

Perbandingan akurasi saja dapat memberikan gambaran yang kurang memadai ketika proporsi kelas berbeda. Oleh sebab itu, interpretasi hasil perlu mempertimbangkan presisi, *recall*, F1-score, *confusion matrix*, serta metrik agregat yang digunakan dalam evaluasi.

2.6 Pelatihan dan Evaluasi Model Machine Learning

Tahap pelatihan dan evaluasi model *machine learning* dilakukan untuk mengukur kemampuan algoritma dalam mereplikasi label otomatis (*pseudo-labeling*) ke dalam tiga kelas, yaitu positif, negatif, dan netral (*neutral*). Pada penelitian ini, proses pelatihan dilakukan secara terpisah pada dua dataset hasil *pseudo-labeling*, yaitu dataset RoBERTa dan dataset *Lexicon-Based*. Pemisahan ini dilakukan agar kinerja algoritma dapat diamati berdasarkan sumber label yang berbeda.

Algoritma *machine learning* yang digunakan dalam penelitian ini terdiri atas *Support Vector Machine* (SVM), Naïve Bayes, dan *Neural Network*. SVM digunakan dengan *kernel* linear karena sesuai untuk data teks berdimensi tinggi yang direpresentasikan menggunakan TF-IDF serta telah banyak digunakan dalam klasifikasi sentimen berbasis teks tidak terstruktur [1][2][10][11]. Naïve Bayes digunakan sebagai pembanding karena merupakan pendekatan probabilistik yang banyak diterapkan pada klasifikasi teks dan sering digunakan sebagai model dasar (*baseline*) dalam penelitian analisis sentimen [1][10][11]. Sementara itu, *Neural Network* digunakan untuk mengamati kemampuan model berbasis jaringan saraf dalam mempelajari pola klasifikasi sentimen yang lebih kompleks.

Ketiga algoritma diterapkan pada masing-masing dataset berlabel agar perbandingan dapat dilakukan dalam dua konteks, yaitu klasifikasi terhadap label yang dihasilkan RoBERTa dan klasifikasi terhadap label yang dihasilkan *Lexicon-Based*. Dengan rancangan ini, penelitian dapat mengidentifikasi algoritma yang menghasilkan kinerja tertinggi pada masing-masing skema pelabelan. Namun, kinerja tersebut harus ditafsirkan sesuai dengan label yang menjadi target pelatihan dan evaluasi, bukan sebagai bukti bahwa salah satu algoritma lebih tepat dalam mengenali sentimen manusia secara umum.

Setiap model dilatih menggunakan data latih yang telah direpresentasikan dalam bentuk matriks TF-IDF. Setelah proses pelatihan selesai, model diuji menggunakan data uji yang tidak digunakan dalam proses pelatihan. Prediksi terhadap data uji kemudian dibandingkan dengan label *pseudo-label* yang tersedia pada data uji untuk memperoleh nilai evaluasi. Agar proses eksperimen dapat direplikasi, parameter utama setiap algoritma perlu dilaporkan berdasarkan konfigurasi aktual, termasuk pengaturan SVM dengan *kernel* linear, konfigurasi varian Naïve Bayes yang digunakan, arsitektur dan parameter pelatihan *Neural Network*, serta pengaturan lain yang berpengaruh terhadap hasil. Jika suatu algoritma menggunakan parameter bawaan pustaka, hal tersebut dapat dinyatakan secara eksplisit beserta pustaka dan versinya.

Evaluasi dilakukan menggunakan *confusion matrix* serta metrik akurasi, presisi, *recall*, dan F1-score. Akurasi digunakan untuk menunjukkan proporsi prediksi yang sesuai dengan label acuan terhadap seluruh data uji. Presisi menunjukkan proporsi prediksi pada suatu kelas yang sesuai dengan label acuan kelas tersebut, sedangkan *recall* menunjukkan proporsi data dari suatu kelas pada label acuan yang berhasil dikenali oleh model. F1-score merupakan rata-rata harmonik antara presisi dan *recall*, sehingga dapat memberikan gambaran keseimbangan kedua metrik tersebut.

Secara matematis, akurasi dapat dituliskan sebagai:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Pada persamaan tersebut, (TP) (*true positive*) menyatakan jumlah prediksi positif yang sesuai dengan label acuan, (TN) (*true negative*) menyatakan jumlah prediksi negatif yang sesuai dengan label acuan, (FP) (*false positive*) menyatakan jumlah prediksi positif yang tidak sesuai dengan label acuan, dan (FN) (*false negative*) menyatakan jumlah data positif pada label acuan yang tidak dikenali sebagai positif. Rumus tersebut merupakan ilustrasi untuk klasifikasi biner; karena penelitian ini menggunakan tiga kelas sentimen, perhitungan metrik dilakukan berdasarkan setiap kelas dan kemudian dapat diagregasikan sesuai metode yang digunakan.

Untuk klasifikasi multikelas, presisi, *recall*, dan F1-score dapat dilaporkan dalam bentuk *macro average*, *weighted average*, atau *micro average*. *Macro average* menghitung rata-rata antarkelas tanpa mempertimbangkan jumlah data setiap kelas, sedangkan *weighted average* memperhitungkan proporsi jumlah data pada setiap kelas. *Micro average* menghitung metrik berdasarkan agregasi kontribusi prediksi seluruh kelas. Jenis agregasi yang digunakan dalam pelaporan hasil harus dinyatakan secara eksplisit dan konsisten dengan kode evaluasi. Apabila penelitian menggunakan nilai F1-score agregat, nilai tersebut tidak boleh disebut sebagai F1-score *macro* atau *weighted* sebelum metode perhitungannya dipastikan.

Penggunaan *confusion matrix* juga diperlukan untuk mengamati pola kesalahan klasifikasi antarkelas, misalnya ketika komentar berlabel negatif diprediksi sebagai netral atau komentar berlabel positif diprediksi sebagai negatif. Analisis ini membantu menjelaskan apakah nilai akurasi yang diperoleh didukung oleh kemampuan mengenali ketiga kelas atau lebih banyak dipengaruhi oleh kelas yang memiliki proporsi lebih besar. Karena itu, penyajian metrik per kelas, apabila tersedia dari hasil eksperimen, dapat melengkapi metrik agregat dan memberikan gambaran yang lebih utuh mengenai kinerja model.

Selain metrik evaluasi, interpretasi hasil harus mempertimbangkan distribusi label pada masing-masing dataset. Jika suatu dataset didominasi satu kelas, model dapat memperoleh akurasi yang relatif tinggi meskipun kemampuannya mengenali kelas lain masih terbatas. Sebaliknya, nilai F1-score dapat memberikan perspektif tambahan mengenai keseimbangan presisi dan *recall*. Oleh sebab itu, perbandingan model tidak sebaiknya hanya didasarkan pada satu metrik, melainkan mempertimbangkan keseluruhan hasil evaluasi yang tersedia.

Meskipun demikian, label yang digunakan sebagai acuan evaluasi dalam penelitian ini berasal dari proses *pseudo-labeling*, bukan dari anotasi manusia yang independen. Dengan demikian, *confusion matrix*, akurasi, presisi, *recall*, dan F1-score terutama menggambarkan kesesuaian prediksi algoritma terhadap label otomatis yang digunakan dalam pelatihan dan pengujian. Nilai metrik tersebut belum cukup untuk menyatakan bahwa prediksi model sesuai dengan sentimen sebenarnya dari komentar YouTube.

Untuk menilai validitas label secara lebih langsung, penelitian lanjutan atau eksperimen tambahan perlu menggunakan sebagian komentar yang dianotasi oleh manusia berdasarkan pedoman anotasi yang jelas. Label hasil anotasi manusia dapat digunakan untuk membandingkan hasil RoBERTa dan *Lexicon-Based* serta menilai ketepatan klasifikasi secara lebih independen. Jika validasi tersebut belum dilakukan, kondisi ini perlu dinyatakan sebagai keterbatasan penelitian dan menjadi dasar untuk membatasi kesimpulan pada perbandingan kinerja klasifikasi terhadap dua skema *pseudo-labeling* yang diuji.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan dan Preprocessing Data

Hasil pengumpulan data menghasilkan 2.917 komentar mentah dari video YouTube yang menjadi objek penelitian. Data tersebut kemudian melalui proses pemeriksaan awal untuk menghapus komentar duplikat dan data kosong (*null*). Setelah proses penghapusan duplikat dan data kosong dilakukan, jumlah data berkurang menjadi 2.896 komentar. Selanjutnya, data diproses melalui tahap *preprocessing* yang meliputi *cleaning*, *case folding*, dan normalisasi kata. Khusus pada jalur *Lexicon-Based*, dilakukan tahapan tambahan berupa tokenisasi dan *stopword removal*. Setelah seluruh tahapan *preprocessing* selesai, dilakukan kembali pemeriksaan terhadap data kosong yang muncul akibat proses pembersihan teks. Hasil akhir *preprocessing* menghasilkan 2.850 komentar yang digunakan pada tahap *pseudo-labeling*.

Tabel 1. Statistik Dataset Penelitian

Keterangan	Jumlah
Dataset awal	2.917
Dataset setelah penghapusan duplikat dan <i>null</i>	2.896
Dataset setelah <i>preprocessing</i> final	2.850

Berdasarkan Tabel 1, terdapat pengurangan jumlah data dari 2.917 menjadi 2.896 komentar setelah proses penghapusan duplikasi dan data kosong. Dengan demikian, terdapat selisih sebanyak 21 komentar pada tahap pemeriksaan awal. Setelah tahap *preprocessing* final, jumlah data menjadi 2.850 komentar, atau berkurang sebanyak 46 komentar dari dataset setelah pemeriksaan awal. Pengurangan tersebut terjadi karena terdapat teks yang menjadi kosong setelah proses pembersihan, khususnya pada jalur *Lexicon-Based*. Agar perbandingan kedua pendekatan dilakukan menggunakan jumlah komentar yang sama, data yang tidak memiliki teks setelah *preprocessing* pada jalur *Lexicon-Based* juga dikeluarkan dari dataset yang digunakan dalam kedua pendekatan. Dengan demikian, masing-masing pendekatan menggunakan 2.850 komentar yang sama sebagai dasar proses *pseudo-labeling*.

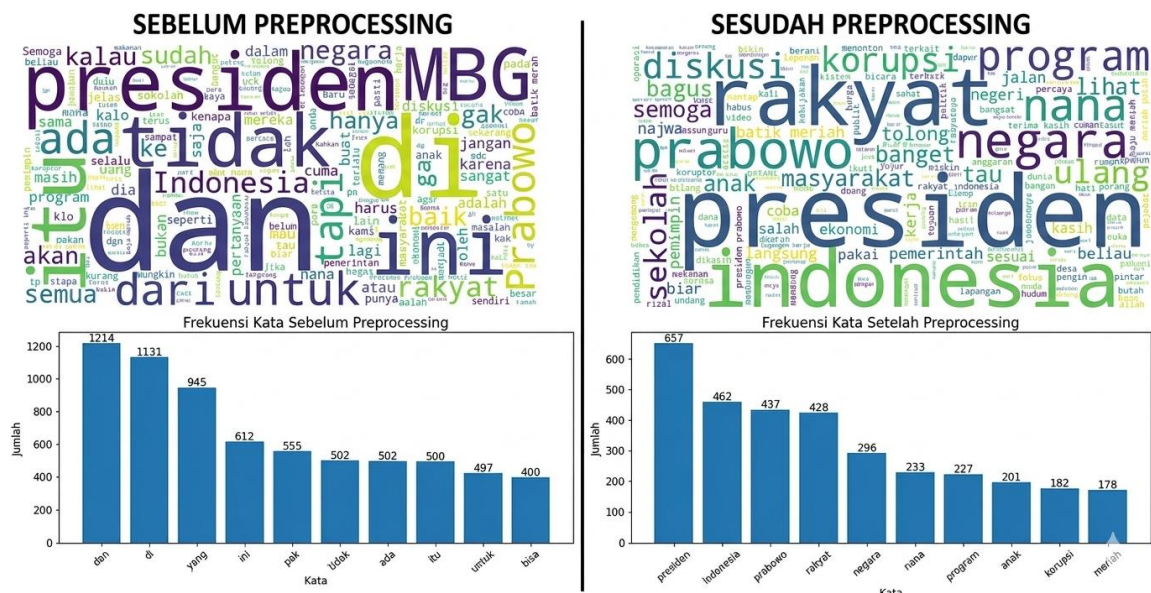
Penggunaan dataset yang sama dimaksudkan untuk menjaga kesetaraan jumlah dan sumber komentar dalam perbandingan kedua pendekatan. Namun, kesamaan jumlah data tidak berarti bahwa kedua pendekatan menghasilkan representasi teks dan label sentimen yang sama. Perbedaan *preprocessing* dan mekanisme penentuan label tetap dapat menghasilkan distribusi sentimen yang berbeda. Oleh karena itu, kesamaan jumlah data dalam penelitian ini berfungsi sebagai pengendalian terhadap sumber dan jumlah observasi, bukan sebagai bukti bahwa seluruh faktor eksperimen telah sepenuhnya dikendalikan. Proses *preprocessing* menghasilkan beberapa bentuk data teks, yaitu teks hasil *cleaning*, teks hasil *preprocessing* umum, teks untuk RoBERTa, serta teks untuk *Lexicon-Based*. Contoh hasil *preprocessing* ditampilkan pada Tabel 2.

Tabel 2. Sampel Hasil Preprocessing Data

Komentar asli	Hasil <i>cleaning</i>	<i>Preprocessing</i> untuk RoBERTa	<i>Preprocessing</i> untuk <i>Lexicon-Based</i>
Sekolah di saya juga pak garut Selatan pameungpeuk tepatnya kecamatan Cibalong desa mekar kp cibanen sekolah nyaris runtuh rapi masih di pake belum ada perbaikan	sekolah di saya juga pak garut selatan pameungpeuk tepatnya kecamatan cibalong desa mekar kp cibanen sekolah nyaris runtuh rapi masih di pake belum ada perbaikan	sekolah di saya juga pak garut selatan pameungpeuk tepatnya kecamatan cibalong desa mekar kp cibanen sekolah nyaris runtuh rapi masih di pakai belum ada perbaikan	sekolah garut selatan pameungpeuk tepatnya kecamatan cibalong desa mekar kp cibanen sekolah runtuh rapi pakai perbaikan

Berdasarkan Tabel 2, terlihat bahwa tahapan *preprocessing* menghasilkan bentuk teks yang berbeda sesuai dengan kebutuhan masing-masing pendekatan. Pada jalur RoBERTa, teks masih mempertahankan susunan kalimat agar informasi kontekstual tetap tersedia bagi model *transformer*. Sementara itu, pada jalur *Lexicon-Based*, teks menjadi lebih ringkas karena kata-kata yang termasuk dalam daftar *stopword* dihapus setelah proses tokenisasi. Perbedaan ini mencerminkan penyesuaian tahapan pemrosesan terhadap mekanisme kerja masing-masing metode. Namun, teks yang lebih ringkas tidak selalu berarti informasi sentimennya lebih lengkap, sebab penghapusan kata tertentu juga berpotensi menghilangkan konteks yang diperlukan untuk memahami polaritas komentar.

Perbandingan teks sebelum dan sesudah *preprocessing* pada jalur *Lexicon-Based* juga ditampilkan melalui visualisasi pada Gambar 2.



Gambar 2. WordCloud dan Frekuensi Kata Sebelum dan Sesudah Preprocessing

Berdasarkan Gambar 2, visualisasi *WordCloud* menunjukkan adanya perubahan pola kata sebelum dan sesudah *preprocessing*. Sebelum *preprocessing*, kata yang dominan meliputi kata umum seperti “dan”, “di”, “yang”, “ini”, dan “pak”. Kata-kata tersebut memiliki frekuensi kemunculan tinggi, tetapi tidak seluruhnya secara langsung menunjukkan polaritas sentimen. Setelah *preprocessing*, kata yang dominan berubah menjadi “presiden”, “indonesia”, “prabowo”, “rakyat”, “negara”, dan “program”. Kata-kata tersebut lebih berkaitan dengan topik yang dibahas dalam komentar pada video yang dianalisis.

Perubahan tersebut juga terlihat pada grafik frekuensi kata. Sebelum *preprocessing*, kata “dan” muncul sebanyak 1.214 kali, “di” sebanyak 1.131 kali, dan “yang” sebanyak 945 kali. Setelah *preprocessing*, kata yang dominan berubah menjadi “presiden” sebanyak 657 kali, “indonesia” sebanyak 462 kali, “prabowo” sebanyak 437 kali, dan “rakyat” sebanyak 428 kali. Perubahan frekuensi tersebut menunjukkan bahwa proses pembersihan dan penghapusan *stopword* mengubah komposisi kata yang paling sering muncul dalam dataset. Hasil visualisasi ini mendukung penjelasan mengenai perubahan representasi teks setelah *preprocessing*, tetapi tidak dengan sendirinya membuktikan peningkatan akurasi pelabelan sentimen. Penilaian terhadap ketepatan label tetap memerlukan evaluasi menggunakan data acuan yang independen.

3.2 Hasil Pseudo-Labeling

Kedua pendekatan *pseudo-labeling* menghasilkan tiga kelas sentimen, yaitu positif (*positive*), negatif (*negative*), dan netral (*neutral*). Jumlah data yang digunakan pada masing-masing

pendekatan sama, yaitu 2.850 komentar. Hasil pelabelan otomatis tersebut disajikan pada Tabel 3.

Tabel 3. Distribusi Sentimen Hasil Pseudo-Labeling

Pendekatan <i>Labeling</i>	<i>Positive</i>	<i>Negative</i>	<i>Neutral</i>
RoBERTa	698 (24,49%)	1.826 (64,07%)	326 (11,44%)
<i>Lexicon-Based</i>	1.159 (40,67%)	799 (28,04%)	892 (31,30%)

Berdasarkan Tabel 3, RoBERTa *Labeling* menghasilkan 1.826 komentar berlabel negatif atau sebesar 64,07% dari keseluruhan data. Kelas positif berjumlah 698 komentar atau 24,49%, sedangkan kelas netral berjumlah 326 komentar atau 11,44%. Distribusi ini memperlihatkan dominasi label negatif pada hasil pelabelan menggunakan RoBERTa. Dengan demikian, distribusi label yang dihasilkan model tersebut tidak seimbang karena sebagian besar komentar ditempatkan pada kelas negatif.

Sementara itu, *Lexicon-Based Labeling* menghasilkan 1.159 komentar berlabel positif atau 40,67%, 799 komentar berlabel negatif atau 28,04%, dan 892 komentar berlabel netral atau 31,30%. Dibandingkan dengan distribusi RoBERTa, distribusi *Lexicon-Based* lebih merata di antara ketiga kelas sentimen. Perbedaan ini memperlihatkan bahwa dua pendekatan pelabelan otomatis dapat menghasilkan komposisi label yang berbeda meskipun diterapkan pada sumber komentar dan jumlah data yang sama.

Perbedaan distribusi tersebut dapat berkaitan dengan karakteristik mekanisme pelabelan masing-masing pendekatan. RoBERTa menentukan label berdasarkan prediksi model *transformer* yang memanfaatkan representasi kontekstual teks, sedangkan *Lexicon-Based* menentukan label berdasarkan selisih jumlah kata yang cocok dengan kamus positif dan negatif. Pada pendekatan berbasis leksikon, komentar yang tidak memiliki kata yang cocok dengan kamus atau memiliki jumlah kata positif dan negatif yang sama akan memperoleh skor nol dan diberi label netral. Oleh karena itu, jumlah label netral yang dihasilkan *Lexicon-Based* dapat dipengaruhi oleh cakupan kamus, aturan pencocokan kata, serta proses *preprocessing* yang diterapkan.

Dominasi label negatif pada hasil RoBERTa tidak dapat langsung diartikan sebagai bukti bahwa mayoritas komentar benar-benar bernada negatif. Demikian pula, distribusi label *Lexicon-Based* yang lebih merata tidak otomatis membuktikan bahwa metode tersebut lebih akurat. Kedua distribusi tersebut merupakan hasil keluaran sistem pelabelan otomatis dan belum dibandingkan dengan anotasi manusia yang independen. Dengan demikian, interpretasi mengenai sentimen aktual pengguna perlu dibatasi karena distribusi label pada Tabel 3 terutama menunjukkan kecenderungan masing-masing pendekatan dalam memberikan label terhadap data yang dianalisis.

Hasil *pseudo-labeling* ini menjadi dasar bagi tahap pelatihan model *machine learning*. Distribusi kelas yang tidak seimbang pada dataset RoBERTa berpotensi memengaruhi proses pembelajaran karena model dapat lebih sering menemukan contoh dari kelas negatif. Kondisi tersebut dapat memengaruhi kemampuan model dalam mengenali kelas positif dan netral. Sebaliknya, distribusi label yang lebih merata pada dataset *Lexicon-Based* menyediakan jumlah contoh yang relatif lebih seimbang untuk ketiga kelas. Meskipun demikian, distribusi kelas bukan satu-satunya faktor yang menentukan hasil klasifikasi. Kualitas label otomatis, karakteristik teks, representasi fitur, serta konfigurasi algoritma juga dapat berkontribusi terhadap perbedaan kinerja model. Oleh sebab itu, pengaruh distribusi label perlu dianalisis bersama metrik evaluasi dan *confusion matrix*, bukan dijadikan satu-satunya penjelasan atas hasil eksperimen.

3.3 Hasil Ekstraksi Fitur, Pembagian Data, dan Evaluasi Model

Dataset hasil *pseudo-labeling* diproses secara terpisah pada tahap ekstraksi fitur, pembagian data, dan evaluasi model. Pembagian data dengan rasio 80:20 menghasilkan 2.280 data latih dan 570 data uji pada masing-masing pendekatan. Berdasarkan metode yang dilaporkan, pembagian

dilakukan secara *stratified* untuk mempertahankan proporsi kelas pada data latih dan data uji. Penerapan pembagian tersebut dimaksudkan agar evaluasi tidak hanya bergantung pada distribusi kelas yang terbentuk secara kebetulan akibat pembagian data.

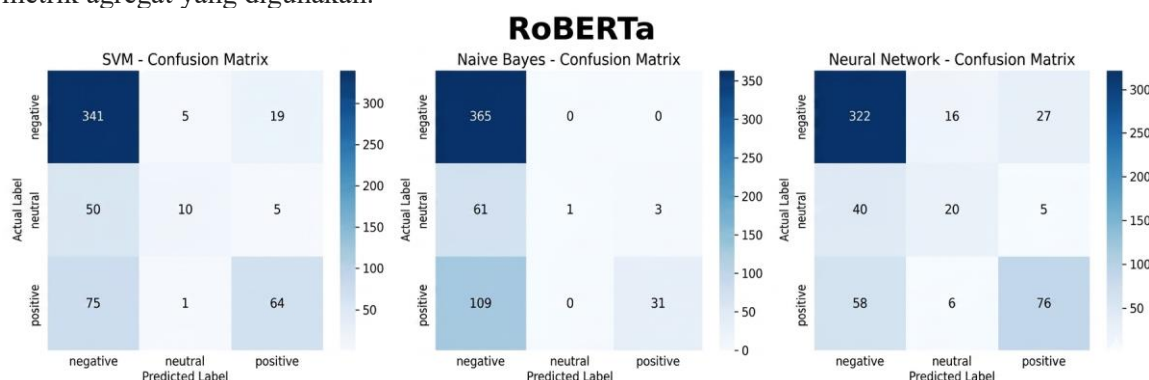
Ekstraksi fitur menggunakan TF-IDF menghasilkan matriks data latih berukuran 2.280×8.739 dan matriks data uji berukuran 570×8.739 . Dengan demikian, setiap komentar direpresentasikan dalam ruang fitur yang terdiri atas 8.739 fitur. Matriks tersebut selanjutnya digunakan sebagai masukan untuk melatih tiga model *machine learning*, yaitu SVM, Naïve Bayes, dan *Neural Network*. Agar hasil ini dapat direplikasi, proses pembentukan fitur perlu dipastikan menggunakan data latih untuk mempelajari kosakata dan bobot IDF, kemudian menerapkan representasi yang sama pada data uji. Dengan prosedur tersebut, informasi dari data uji tidak digunakan dalam pembentukan fitur selama pelatihan.

Tabel 4. Performa Model Machine Learning pada RoBERTa Labeling

<i>Machine Learning</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	F1-Score
<i>Neural Network</i>	0,7333	0,6489	0,5776	0,6024
SVM	0,7281	0,6947	0,5151	0,5430
Naïve Bayes	0,6965	0,8647	0,4123	0,3992

Berdasarkan Tabel 4, *Neural Network* memperoleh nilai *accuracy* tertinggi sebesar 0,7333, diikuti SVM sebesar 0,7281 dan Naïve Bayes sebesar 0,6965. Berdasarkan F1-score yang dilaporkan, *Neural Network* juga memperoleh nilai tertinggi sebesar 0,6024, diikuti SVM sebesar 0,5430 dan Naïve Bayes sebesar 0,3992. Hasil ini menunjukkan bahwa pada dataset berlabel RoBERTa, *Neural Network* memberikan kinerja terbaik di antara ketiga algoritma berdasarkan nilai *accuracy* dan F1-score yang disajikan.

Naïve Bayes menghasilkan nilai *precision* sebesar 0,8647, lebih tinggi dibandingkan *Neural Network* sebesar 0,6489 dan SVM sebesar 0,6947. Namun, nilai *recall* Naïve Bayes hanya sebesar 0,4123, sedangkan *Neural Network* memperoleh 0,5776 dan SVM memperoleh 0,5151. Perbedaan nilai *precision* dan *recall* tersebut menunjukkan bahwa kinerja Naïve Bayes tidak seimbang berdasarkan metrik yang dilaporkan. Nilai *precision* yang tinggi tidak dengan sendirinya menunjukkan bahwa model mampu mengenali seluruh data pada kelas yang relevan. Interpretasi lebih lanjut perlu mempertimbangkan hasil *confusion matrix* serta cara penghitungan metrik agregat yang digunakan.



Gambar 4. Confusion Matrix Model pada RoBERTa Labeling

Confusion matrix pada Gambar 4 digunakan untuk mengamati kesesuaian prediksi dengan label *pseudo-label* pada data uji. Berdasarkan uraian hasil yang tersedia, ketidakseimbangan distribusi kelas pada dataset RoBERTa disertai kecenderungan Naïve Bayes salah mengklasifikasikan sejumlah data positif dan netral ke dalam kelas negatif. Pola tersebut dapat mengurangi kemampuan model dalam mengenali kelas minoritas. Meskipun demikian, kesimpulan mengenai kecenderungan prediksi harus merujuk pada jumlah prediksi per kelas dan urutan label yang digunakan dalam *confusion matrix*, bukan hanya pada distribusi label keseluruhan.

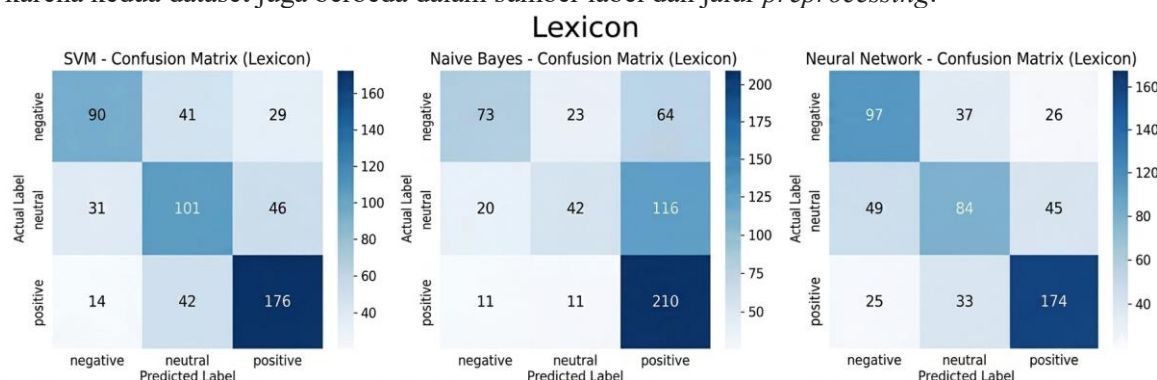
Neural Network menunjukkan jumlah prediksi benar pada diagonal *confusion matrix* sebanyak 322 data negatif, 20 data netral, dan 76 data positif. Jumlah ketiga nilai tersebut adalah 418 data. Jumlah ini konsisten dengan nilai *accuracy* sebesar 0,7333 apabila dihitung terhadap 570 data uji, karena 418 dibagi 570 menghasilkan sekitar 0,7333. Dengan demikian, hasil pada diagonal *confusion matrix* tersebut mendukung nilai *accuracy* yang dilaporkan untuk *Neural Network*. Namun, jumlah prediksi benar pada diagonal saja belum cukup untuk menjelaskan keseluruhan pola kesalahan model. Untuk mengetahui jenis kesalahan antarkelas secara lebih terperinci, seluruh nilai pada *confusion matrix* perlu diperiksa.

Tabel 5. Performa Model Machine Learning pada Lexicon-Based Labeling

Machine Learning	Accuracy	Precision	Recall	F1-Score
SVM	0,6439	0,6389	0,6295	0,6323
Neural Network	0,6228	0,6076	0,6094	0,6072
Naïve Bayes	0,5702	0,5977	0,5325	0,5197

Berdasarkan Tabel 5, SVM memperoleh nilai *accuracy* sebesar 0,6439 dan F1-score sebesar 0,6323. *Neural Network* menempati posisi berikutnya dengan nilai *accuracy* sebesar 0,6228 dan F1-score sebesar 0,6072, sedangkan Naïve Bayes memperoleh nilai *accuracy* sebesar 0,5702 dan F1-score sebesar 0,5197. Berdasarkan nilai yang disajikan, SVM menghasilkan kinerja tertinggi di antara ketiga algoritma pada dataset berlabel *Lexicon-Based*.

Nilai *precision*, *recall*, dan F1-score SVM masing-masing sebesar 0,6389, 0,6295, dan 0,6323. Ketiga nilai tersebut relatif berdekatan sehingga menunjukkan keseimbangan yang lebih baik di antara metrik yang dilaporkan dibandingkan Naïve Bayes. Sementara itu, *Neural Network* memperoleh nilai *precision* sebesar 0,6076 dan *recall* sebesar 0,6094, dengan F1-score sebesar 0,6072. Naïve Bayes memperoleh nilai *precision* sebesar 0,5977, *recall* sebesar 0,5325, dan F1-score sebesar 0,5197. Hasil tersebut menunjukkan bahwa SVM memperoleh nilai F1-score tertinggi pada dataset *Lexicon-Based*, sedangkan Naïve Bayes memperoleh nilai terendah di antara ketiga algoritma yang diuji. Dibandingkan dengan hasil pada dataset RoBERTa, kinerja algoritma pada dataset *Lexicon-Based* memperlihatkan pola yang berbeda. Pada dataset RoBERTa, *Neural Network* memperoleh F1-score tertinggi, sedangkan pada dataset *Lexicon-Based*, SVM memperoleh F1-score tertinggi. Perbedaan ini menunjukkan bahwa urutan kinerja algoritma dapat berubah sesuai dengan dataset berlabel yang digunakan. Namun, perubahan tersebut tidak dapat secara langsung dikaitkan hanya dengan keseimbangan distribusi kelas karena kedua dataset juga berbeda dalam sumber label dan jalur *preprocessing*.



Gambar 5. Confusion Matrix Model pada Lexicon-Based Labeling

Berdasarkan Gambar 5, *confusion matrix* pada dataset *Lexicon-Based* memperlihatkan distribusi prediksi pada tiga kelas sentimen. Pada model SVM, jumlah prediksi benar pada diagonal *confusion matrix* adalah 90 data negatif, 101 data netral, dan 176 data positif. Jumlah tersebut mencapai 367 data, sesuai dengan nilai *accuracy* sebesar 0,6439 pada 570 data uji. Hasil ini mendukung nilai *accuracy* SVM yang dilaporkan pada Tabel 5.

Neural Network menunjukkan jumlah prediksi benar sebanyak 97 data negatif, 84 data netral, dan 174 data positif. Total prediksi benar tersebut adalah 355 data, sesuai dengan nilai *accuracy* sebesar 0,6228 pada 570 data uji. Sementara itu, Naïve Bayes cenderung lebih banyak memprediksi data sebagai positif. Berdasarkan uraian *confusion matrix*, terdapat 210 data positif yang berhasil dikenali, tetapi sejumlah data negatif dan netral juga diprediksi sebagai positif. Pola ini menunjukkan perlunya memeriksa prediksi salah pada masing-masing kelas untuk memahami keterbatasan Naïve Bayes dalam membedakan ketiga label sentimen.

Secara keseluruhan, hasil pada Tabel 4, Tabel 5, Gambar 4, dan Gambar 5 menunjukkan bahwa kinerja algoritma tidak sama pada kedua dataset berlabel. *Neural Network* memperoleh *accuracy* tertinggi pada dataset RoBERTa, sedangkan SVM memperoleh *accuracy* dan F1-score tertinggi pada dataset *Lexicon-Based*. Meskipun demikian, angka-angka tersebut menggambarkan kinerja model dalam memprediksi label otomatis yang tersedia pada data uji. Oleh karena itu, hasil evaluasi belum dapat digunakan sebagai bukti langsung bahwa prediksi tersebut sesuai dengan sentimen sebenarnya dari setiap komentar.

3.4 Pembahasan Komparatif

Berdasarkan hasil evaluasi, pendekatan RoBERTa memperoleh kinerja terbaik pada algoritma *Neural Network* dengan nilai *accuracy* sebesar 0,7333 dan F1-score sebesar 0,6024. Sementara itu, pendekatan *Lexicon-Based* memperoleh kinerja terbaik pada algoritma SVM dengan nilai *accuracy* sebesar 0,6439 dan F1-score sebesar 0,6323. Dengan demikian, hasil penelitian menunjukkan bahwa *Neural Network* pada dataset RoBERTa menghasilkan nilai *accuracy* tertinggi di antara kombinasi model dan dataset yang diuji, sedangkan kombinasi *Lexicon-Based* dan SVM menghasilkan nilai F1-score tertinggi berdasarkan hasil yang dilaporkan.

Perbedaan antara nilai *accuracy* dan F1-score tersebut penting untuk diperhatikan. *Accuracy* mengukur proporsi seluruh prediksi yang sesuai dengan label acuan, sedangkan F1-score menggabungkan *precision* dan *recall* sesuai metode agregasi yang digunakan. Oleh karena itu, nilai *accuracy* yang lebih tinggi tidak selalu diikuti F1-score yang lebih tinggi. Pada penelitian ini, *Neural Network* pada dataset RoBERTa menghasilkan *accuracy* 0,7333, sedangkan SVM pada dataset *Lexicon-Based* menghasilkan *accuracy* 0,6439. Namun, F1-score SVM pada dataset *Lexicon-Based* sebesar 0,6323 lebih tinggi daripada F1-score *Neural Network* pada dataset RoBERTa sebesar 0,6024. Perbandingan tersebut memperlihatkan bahwa kesimpulan mengenai model terbaik bergantung pada metrik evaluasi yang digunakan sebagai acuan.

Perbedaan hasil tersebut juga perlu dikaitkan dengan distribusi label yang dihasilkan masing-masing pendekatan *pseudo-labeling*. RoBERTa menghasilkan distribusi yang tidak seimbang, dengan kelas negatif mendominasi sebesar 64,07%, sedangkan kelas netral hanya sebesar 11,44%. Kondisi ini dapat memengaruhi proses pembelajaran karena jumlah contoh pada kelas negatif jauh lebih besar dibandingkan jumlah contoh pada kelas netral. Model berpotensi lebih sering mempelajari pola kelas mayoritas, sementara kemampuan mengenali kelas yang jumlahnya lebih sedikit perlu dinilai melalui metrik per kelas dan *confusion matrix*. Meskipun demikian, ketidakseimbangan kelas tidak dapat dianggap sebagai satu-satunya penyebab perbedaan kinerja karena karakteristik teks, kualitas label, dan konfigurasi model juga dapat berpengaruh.

Sebaliknya, *Lexicon-Based* menghasilkan distribusi label yang lebih merata, yaitu kelas positif sebesar 40,67%, netral sebesar 31,30%, dan negatif sebesar 28,04%. Distribusi tersebut menyediakan jumlah contoh yang relatif lebih seimbang untuk proses pelatihan ketiga kelas. Dalam eksperimen ini, SVM memperoleh F1-score sebesar 0,6323, lebih tinggi daripada *Neural Network* sebesar 0,6072 dan Naïve Bayes sebesar 0,5197. Hasil tersebut menunjukkan bahwa SVM memberikan kinerja terbaik berdasarkan F1-score yang dilaporkan pada dataset *Lexicon-Based*. Namun, distribusi label yang lebih merata tidak secara otomatis membuktikan bahwa

label *Lexicon-Based* lebih benar atau lebih merepresentasikan sentimen aktual. Kualitas label dan keseimbangan distribusi merupakan dua aspek yang berbeda dan perlu dievaluasi secara terpisah.

Temuan mengenai kinerja SVM pada data teks berdimensi tinggi dapat dibahas dalam kaitannya dengan penelitian terdahulu [1][10][11] yang mengkaji penggunaan algoritma klasifikasi untuk analisis sentimen. Dalam penelitian ini, SVM menggunakan representasi TF-IDF dengan 8.739 fitur dan menghasilkan nilai F1-score tertinggi pada dataset *Lexicon-Based*. Hasil tersebut memberikan bukti empiris dalam konteks dataset yang digunakan bahwa SVM merupakan algoritma dengan kinerja terbaik berdasarkan metrik tersebut pada skema pelabelan *Lexicon-Based*. Akan tetapi, temuan ini tidak cukup untuk menyatakan bahwa SVM selalu lebih unggul pada seluruh korpus media sosial atau pada semua skema pelabelan. Kesimpulan yang lebih luas memerlukan pengujian pada dataset dan konteks yang berbeda.

Karakteristik kedua skema memperlihatkan bahwa dominasi kelas negatif pada dataset RoBERTa disertai nilai *accuracy* yang relatif tinggi pada *Neural Network*, sementara kombinasi *Lexicon-Based* dan SVM menghasilkan F1-score tertinggi di antara hasil yang dilaporkan. Perbedaan ini mengindikasikan bahwa pemilihan metrik evaluasi memengaruhi interpretasi kinerja model. Karena penelitian melibatkan tiga kelas sentimen dengan distribusi yang berbeda, F1-score dapat memberikan informasi tambahan selain *accuracy*. Namun, penggunaan F1-score sebagai acuan utama harus disertai penjelasan mengenai metode agregasi yang digunakan, misalnya *macro average* atau *weighted average*. Informasi tersebut penting agar pembaca dapat memahami apakah setiap kelas memperoleh bobot yang sama atau kontribusinya dipengaruhi oleh jumlah data pada masing-masing kelas.

Berdasarkan F1-score yang dilaporkan, kombinasi *Lexicon-Based* dan SVM menjadi kombinasi dengan kinerja tertinggi pada dataset *Lexicon-Based*, dengan nilai sebesar 0,6323. Sementara itu, kombinasi RoBERTa dan *Neural Network* memperoleh *accuracy* tertinggi sebesar 0,7333. Kedua hasil ini tidak perlu dipertentangkan sebagai satu kesimpulan tunggal karena masing-masing menonjol pada metrik yang berbeda dan menggunakan sumber label yang berbeda pula. Oleh sebab itu, temuan utama penelitian ini adalah adanya variasi kinerja algoritma berdasarkan skema *pseudo-labeling* yang digunakan, bukan pembuktian bahwa satu skema pelabelan secara mutlak lebih baik daripada skema lainnya.

Perlu ditekankan bahwa seluruh nilai evaluasi dalam penelitian ini diperoleh dengan menggunakan label *pseudo-labeling* sebagai acuan pada data uji. Dengan demikian, nilai *accuracy*, *precision*, *recall*, dan F1-score menunjukkan kesesuaian prediksi algoritma *machine learning* terhadap label otomatis yang dihasilkan RoBERTa atau *Lexicon-Based*. Nilai tersebut belum secara langsung mengukur ketepatan model dalam mengenali sentimen sebenarnya dari komentar YouTube. Keterbatasan ini penting karena masing-masing pendekatan pelabelan dapat menghasilkan kesalahan, terutama pada komentar yang mengandung negasi, ironi, sarkasme, atau ungkapan yang maknanya bergantung pada konteks.

Untuk memperkuat validitas temuan, penelitian lanjutan perlu melakukan anotasi manual pada sebagian komentar dengan pedoman pelabelan yang jelas. Label hasil anotasi manusia dapat digunakan sebagai acuan independen untuk menilai kualitas label RoBERTa dan *Lexicon-Based*, sekaligus mengukur kinerja algoritma klasifikasi terhadap sentimen yang telah ditentukan melalui anotasi tersebut. Selain itu, pengujian pada lebih dari satu video atau topik dapat dilakukan untuk menilai sejauh mana hasil perbandingan dapat diterapkan pada konteks komentar yang berbeda. Dengan batasan tersebut, hasil penelitian saat ini memberikan gambaran komparatif mengenai kinerja tiga algoritma *machine learning* pada dua skema pelabelan otomatis dalam satu konteks video YouTube, serta menyediakan dasar untuk evaluasi lebih lanjut terhadap kualitas label dan generalisasi model.

4. KESIMPULAN

Penelitian ini membandingkan pendekatan RoBERTa dan *Lexicon-Based* dalam menghasilkan *pseudo-label* sentimen pada komentar YouTube terhadap video “Presiden Prabowo Menjawab (PART 2): Efisiensi, Pertaruhan MBG & Reformasi Kepolisian | Mata Najwa”. Dari 2.917 komentar mentah, diperoleh 2.850 data akhir setelah proses pembersihan dan *preprocessing*. Hasil *pseudo-labeling* menunjukkan adanya perbedaan distribusi label sentimen antara kedua pendekatan. RoBERTa menghasilkan distribusi yang didominasi kelas negatif sebesar 64,07%, sedangkan *Lexicon-Based* menghasilkan distribusi label yang lebih merata, yaitu positif sebesar 40,67%, netral (*neutral*) sebesar 31,30%, dan negatif sebesar 28,04%. Temuan ini menunjukkan bahwa kedua pendekatan menghasilkan pola pelabelan otomatis yang berbeda pada dataset yang dianalisis. Namun, distribusi tersebut tidak dapat secara langsung diinterpretasikan sebagai gambaran sentimen publik yang telah terverifikasi karena penelitian menggunakan komentar dari satu video dan label sentimen yang dihasilkan belum divalidasi menggunakan anotasi manusia.

Pada tahap evaluasi *machine learning*, *Neural Network* memperoleh nilai *accuracy* tertinggi pada dataset berlabel RoBERTa, yaitu 0,7333, dengan F1-score sebesar 0,6024. Sementara itu, SVM memperoleh nilai *accuracy* sebesar 0,6439 dan F1-score sebesar 0,6323 pada dataset berlabel *Lexicon-Based*. Hasil tersebut menunjukkan adanya perbedaan kinerja algoritma dalam mereplikasi label otomatis dari kedua pendekatan *pseudo-labeling*. Pada dataset RoBERTa, dominasi kelas negatif disertai perbedaan kemampuan model dalam mengenali ketiga kelas sentimen, sedangkan kombinasi *Lexicon-Based* dan SVM memperoleh F1-score tertinggi pada dataset berlabel *Lexicon-Based*. Dengan demikian, berdasarkan hasil evaluasi yang diperoleh, *Neural Network* menunjukkan *accuracy* tertinggi pada dataset RoBERTa, sedangkan SVM menunjukkan F1-score tertinggi pada dataset *Lexicon-Based*. Meskipun demikian, hasil tersebut belum membuktikan bahwa salah satu pendekatan menghasilkan label yang lebih akurat terhadap sentimen sebenarnya. Evaluasi menggunakan anotasi manusia yang independen serta pengujian pada data dari video atau topik lain diperlukan untuk menilai kualitas label dan menguji generalisasi hasil penelitian.

5. SARAN

Penelitian selanjutnya disarankan melibatkan ahli atau anotator manusia untuk memvalidasi *pseudo-labeling* secara objektif berdasarkan pedoman anotasi yang jelas. Validasi tersebut penting untuk menilai kesesuaian label yang dihasilkan oleh RoBERTa dan *Lexicon-Based* dengan sentimen sebenarnya pada komentar YouTube. Dalam penelitian ini, evaluasi model masih didasarkan pada label otomatis yang digunakan sebagai acuan, sehingga hasil evaluasi belum dapat secara langsung membuktikan ketepatan klasifikasi terhadap sentimen sebenarnya. Oleh karena itu, apabila validasi anotasi manusia belum dapat dilakukan dalam penelitian ini, keterbatasan tersebut perlu dipertimbangkan dalam menginterpretasikan hasil dan membatasi klaim penelitian pada perbandingan distribusi *pseudo-label* serta kemampuan model dalam mereplikasi label otomatis.

Selain itu, penelitian selanjutnya disarankan memperluas ukuran dan sumber *dataset*, baik dengan melibatkan komentar dari beberapa video maupun topik kebijakan publik yang berbeda, agar hasil analisis dapat diuji pada konteks yang lebih beragam. Pengembangan model juga dapat dioptimalkan melalui *fine-tuning* model *transformer* pada korpus politik berbahasa Indonesia serta penerapan strategi penanganan ketidakseimbangan kelas, seperti *resampling*, *class weighting*, atau SMOTE. Merujuk pada penelitian [12], strategi tersebut dapat dipertimbangkan sebagai alternatif untuk meningkatkan kinerja klasifikasi. Namun, efektivitas

setiap strategi perlu diuji secara empiris dengan prosedur evaluasi yang tepat dan, apabila memungkinkan, menggunakan label hasil anotasi manusia sebagai acuan independen. Dengan demikian, penelitian lanjutan diharapkan tidak hanya meningkatkan kinerja model dalam mereplikasi label otomatis, tetapi juga memperkuat validitas hasil analisis sentimen secara kontekstual.

DAFTAR PUSTAKA

- [1] Santoso, B. A., Nugroho, B. I., & Asyfiya, D. U. (2025). Perbandingan algoritma Naïve Bayes, Support Vector Machine, dan Random Forest untuk analisis sentimen komentar politik YouTube. *TIN: Terapan Informatika Nusantara*, 6(4), 391–400. <https://ejurnal.seminar-id.com/index.php/tin/article/view/8326>
- [2] Sihombing, K. E., & Pakereng, M. A. I. (2025). Analisis sentimen komentar YouTube terhadap wawancara Presiden Prabowo menggunakan machine learning dan Orange Data Mining. *STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer*, 4(4), 327–336. <https://doi.org/10.55123/storage.v4i4.6478>
- [3] Najiyah, I., & Rizal, M. (2025). Analisis sentimen media sosial X terhadap kebijakan Presiden Republik Indonesia Prabowo Subianto. *Jurnal FASILKOM*, 15(3), 443–455. <https://doi.org/10.37859/jf.v15i3.10385>
- [4] Rizkia, A. S., Wufron, & Roji, F. F. (2025). Analisis sentimen Coretax: Perbandingan pelabelan data manual, transformers-based, dan lexicon-based pada performa IndoBERT. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 5(3), 1037–1048. <https://doi.org/10.57152/malcom.v5i3.2151>
- [5] Putri, N. A. R., & Ardiansyah. (2023). Analisis sentimen terhadap kemajuan kecerdasan buatan di Indonesia menggunakan BERT dan RoBERTa. *Jurnal Sains dan Informatika*, 9(2), 137–146.
- [6] Rusmanto, A., Ardiansyah, M. A., Hizbulloh, D. C., & Rosyani, P. (2026). Sentiment analysis tweet Covid-19 menggunakan RoBERTa. *JRIIN: Jurnal Riset Informatika dan Inovasi*, 3(8), 2270–2275. <https://jurnalmahasiswa.com/index.php/jriin/article/view/3295>
- [7] Aripriyanto, S., Tukino, Sufyan, A., & Nandaputra, R. (2022). Sentimen analisis Twitter Ibu Kota Negara Nusantara menggunakan Long Short-Term Memory dan lexicon-based. *EXPERT: Jurnal Manajemen Sistem Informasi dan Teknologi*, 12(2), 119–125. <https://doi.org/10.36448/expert.v12i2.2821>
- [8] Rustanto, I., & Rakhmawati, N. A. (2020). Media sentiment analysis of East Java Province: Lexicon-based vs machine learning SVM. *IPTEK Journal of Proceedings Series*, 6, 203–208.
- [9] Subagio, A. W., Sari, A. P., & Sihananto, A. N. (2024). Klasifikasi lexicon-based sentiment analysis tragedi Kanjuruhan pada Twitter menggunakan algoritma Convolutional Neural Network. *Jurnal Ilmiah Sistem Informasi dan Ilmu Komputer*, 4(1), 166–177. <https://doi.org/10.55606/juisik.v4i1.759>

- [10] Hanif, K. H., Muntiari, N. R., Harto, D., & Wiranata, D. S. (2026). Perbandingan analisis sentimen komentar mahasiswa Prodi Teknik Komputer menggunakan algoritma Decision Tree, Support Vector Machine (SVM), dan Random Forest. *INSECT (Informatics and Security): Jurnal Teknik Informatika*, 12(1), 24–33. <https://doi.org/10.33506/insect.v12i01.5144>
- [11] Nurmadewi, D., Jailani, Z. F., Rafi, H., Anggoro, D. A., & Setiowati, D. (2026). Perbandingan Support Vector Machine dan Naïve Bayes untuk klasifikasi sentimen ulasan e-commerce. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 6(2), 923–933. <https://journal.irpi.or.id/index.php/malcom/article/view/2648>
- [12] Singgalen, Y. A. (2023). Analisis sentimen wisatawan terhadap kualitas layanan hotel dan resort di Lombok menggunakan SERVQUAL dan CRISP-DM. *Building of Informatics, Technology and Science (BITS)*, 4(4), 1870–1882. <https://doi.org/10.47065/bits.v4i4.3199>
- [13] Santosa, B., Magda, K., Budiman, E., & Suryono, R. R. (2026). Analisis komparatif sentimen publik terhadap liputan media terkait aksi Menteri Keuangan menggunakan algoritma SVM dan RoBERTa. *SENTRI: Jurnal Riset Ilmiah*, 5(2), 2318–2326. <https://doi.org/10.55681/sentri.v5i2.5858>