Terbit *online* pada laman: http://jurnal.utu.ac.id/JTI

Jurnal Teknologi Informasi

| ISSN (Online): 2829-8934 |

Klasifikasi Sentimen Shopee Berbasis *Weak Supervision Respons Customer Service* dan SVM

Nisrina Hanifa Setiono[1*], Bella Okta Sari Miranda[2]
[1,] Sains Data, Universitas Telkom Purwokerto, Banyumas, Indonesia
[2] Bisnis Digital, Universitas Telkom Purwokerto, Banyumas, Indonesia
Email: [1]nisrinahanifas@telkomuniversity.ac.id*, [2]bellaoktasarimiranda@telkomuniversity.ac.id

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima: 24 Agustus 2026
Revisi: 2 September 2026
Diterbitkan: 30 Oktober 2026

Kata Kunci:
Klasifikasi Sentimen
Weak supervision
Customer service
TF-IDF
Support Vector Machine

ABSTRAK

Klasifikasi sentimen ulasan e-commerce memerlukan data berlabel, sementara anotasi manual dalam jumlah besar membutuhkan waktu dan sumber daya yang tinggi. Penelitian ini mengevaluasi pemanfaatan respons *customer service* sebagai sumber *weak supervision* untuk membentuk label sentimen pada ulasan pengguna Shopee. Dataset awal berjumlah 485.320 baris; setelah seleksi, preprocessing, dan deduplikasi, penelitian ini menyisakan 264.230 ulasan unik. Penelitian ini membentuk weak label menggunakan labeling functions berbasis pola linguistik pada replyContent, lalu merepresentasikan comment_text dengan TF-IDF dan mengklasifikasikannya menggunakan Linear Support Vector Machine (SVM). Penelitian ini memvalidasi kualitas weak label melalui *human test* natural dan berstrata. Sebanyak 250.583 ulasan memperoleh label biner (coverage 94,84%). Pada *human test* natural, presisi kelas negatif dan positif masing-masing mencapai 1,000 dan 0,985, sedangkan pada skema berstrata mencapai 0,979 dan 0,983. Linear SVM mencatatkan macro-F1 sebesar 0,974 pada skema natural dan 0,973 pada skema berstrata. Performa SVM juga tetap stabil pada pengujian exact reply string holdout. Respons *customer service* berpotensi menjadi sumber *weak supervision* yang efektif untuk mengurangi ketergantungan pada anotasi manual.

Copyright © 2026 Jurnal Teknologi Informasi UTU
All rights reserved

1. Pendahuluan

Ulasan dari pengguna termasuk salah satu informasi penting untuk memahami kebutuhan, pengalaman, dan masalah yang dihadapi pengguna terhadap layanan digital tertentu. Data yang ada bukan hanya menunjukkan tingkat kepuasan pengguna, tetapi juga digunakan untuk mengidentifikasi kebutuhan, pemeliharaan, analisis kesalahan, serta pengembangan produk. Dąbrowski dkk. mengungkapkan bahwa review aplikasi telah banyak digunakan dalam berbagai aktivitas rekayasa perangkat lunak, sementara Das dkk. menunjukkan bantuan analisis dengan *Natural Language Processing* (NLP) dapat mengaitkan umpan balik pengguna dengan fitur atau elemen aplikasi yang membutuhkan perhatian [1], [2].

Dari segi *e-commerce*, sentimen analisis difungsikan untuk mengelola interaksi pengguna secara otomatis dan mengelompokkan berdasarkan kecenderungan sentimen tertentu. Hashmi dan Yayilgan membandingkan berbagai representasi teks serta metode klasifikasi pada ulasan Amazon, menggunakan TF-IDF, Word2Vec, FastText, Support Vector Machine (SVM), deep learning, dan transformer. Hasilnya menunjukkan bahwa SVM tetap bersaing dengan akurasi sekitar 0,91 pada salah satu dataset

ulasan perangkat lunak [3]. Dalam konteks Shopee Indonesia, Sikana dkk. menggunakan BERT dan mendapatkan akurasi 83,08%, precision 82,91%, recall 83,08%, serta F1-score 82,87%. Penelitian itu juga menekankan pemakaian bahasa tidak formal, ketidakjelasan makna, serta ketidaksesuaian antara konten teks dan nilai bintang [4]. Kondisi ini menunjukkan bahwa penilaian bintang tidak selalu dapat langsung digunakan sebagai gambaran sentimen teks, sehingga proses penentuan label menjadi elemen krusial.

Klasifikasi yang diawasi memerlukan data yang sudah memiliki label, sementara anotasi manual dalam jumlah besar membutuhkan waktu dan sumber daya yang cukup besar. Salah satu metode untuk mengurangi ketergantungan itu adalah *weak supervision*. Wallaart dkk. menunjukkan bahwa *weak supervision* dapat membentuk informasi pelatihan dari data yang tidak sepenuhnya memiliki *gold label* [5]. Namun, label lemah tidak bisa langsung dianggap sebagai kebenaran dasar. Zhang dkk. menekankan bahwa klasifikasi teks yang diawasi lemah menghadapi tantangan seperti label yang memiliki noisy dan ketidakseimbangan kelas, sehingga kualitas label yang diperoleh harus diperhatikan dan divalidasi [6].

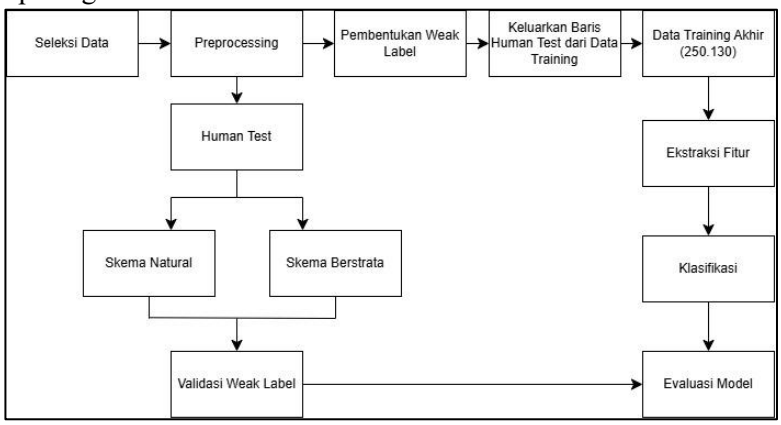
Pelabelan otomatis yang digabungkan dengan metode klasifikasi konvensional telah digunakan di beberapa penelitian di Indonesia. Amalsyah dkk. menerapkan BERT pretrained untuk melakukan pelabelan sentimen pada ulasan aplikasi Flip, kemudian menggunakan TF-IDF dan SVM untuk klasifikasi [7]. Mustakim dan Priyanta juga menggunakan SVM dalam aspect-based sentiment analysis pada ulasan KAI Access [8]. Penelitian yang disebut menunjukkan SVM masih relevan digunakan, sementara kontribusi penelitian dapat diperkuat melalui sumber informasi dan strategi pembentukan label yang berbeda.

Belum ada studi dalam daftar referensi utama yang berfokus pada penggunaan respons layanan pelanggan di ulasan Shopee sebagai sumber *weak supervision* untuk membuat label sentimen awal. Pada penelitian yang dilakukan tahun 2022 hingga 2026, ulasan dari layanan pelanggan bisa mengandung pola seperti meminta maaf, menunjukkan adanya masalah, meminta informasi tambahan, atau mengucapkan terima kasih, yang bisa jadi merupakan sinyal *weak supervision*.

Berdasarkan masalah yang ada, penelitian ini mencari tahu bagaimana respons layanan pelanggan bisa menjadi sumber *weak supervision* untuk membuat label awal dari ulasan pengguna Shopee. Ulasan akan direpresentasikan menggunakan metode TF-IDF dan kemudian diklasifikasikan menggunakan SVM. Penelitian ini bertujuan untuk menilai kualitas label weak yang didasarkan pada respons layanan pelanggan serta performa SVM dalam mengklasifikasikan sentimen. Kontribusi utama dari penelitian ini adalah melakukan eksploarasi terhadap respons layanan pelanggan sebagai sumber alternatif untuk *weak supervision* yang bisa mengurangi ketergantungan pada anotasi manual, sambil tetap menjaga proses klasifikasi yang sederhana dan efisien.

2. Metodologi Penelitian (bold, 11 pt)

Penelitian ini mengkategorikan sentimen dari ulasan pengguna Shopee dengan menggunakan respons *customer service* sebagai sumber *weak supervision*. Langkah-langkah penelitian secara umum ditampilkan pada gambar 1



Gambar 1. Langkah Penelitian

Penelitian ini diawali dengan seleksi data dan preprocessing. Setelah itu, peneliti membentuk weak label dan menyusun *human test*. *Human test* terdiri atas dua skema: natural dan berstrata, untuk memvalidasi kualitas weak label berdasarkan nilai precision dan coverage. Seluruh data yang termasuk dalam *human test* kemudian dikeluarkan dari data pelatihan demi mencegah data leakage. Diperoleh 250.130 data latih yang kemudian direpresentasikan menggunakan Term Frequency–Inverse Document Frequency (TF-IDF) dan dipakai dalam proses klasifikasi. Hasil klasifikasi dievaluasi menggunakan data *human test*, sementara hasil validasi weak label digunakan untuk menilai kualitas pelabelan yang diterapkan.

2.1. Data Penelitian

Data penelitian diambil dari dataset yang dihasilkan melalui scraping ulasan pengguna Shopee yang terdiri dari 485.320 baris data dengan contoh seperti pada gambar 2.

No.	comment_text	replyContent
1	12. 12 Tipu Tipu	Hai kak Faturrahman, mohon maaf atas ketidaknyamanannya. Terkait dengan kendala yang Kakak alami, akan kami jadikan sebagai masukan untuk menjadikan Shopee lebih baik lagi. Kami selalu mencoba memberikan yang terbaik untuk para pengguna Shopee. Apabila ada kendala kakak bisa hubungi aku melalui DM ke Media Sosial @Shopee_id ya. Di tunggu^LG
2	bebas ongkir juga ada potongan, barang yang di jual komplit dan bervariatif, barang nya bagus bagus, pengiriman aman dan cepat saya selalu belanja di shoope karena tidak mengecewakan	Hi Kak Debora Lelly makasih ya untuk reviewnya. Kuy kepoin sosial media resmi shopee biar Kakak tau update promo yang saat ini tersedia hehe. Jangan sungkan chat aku di dm IG @ShopeeCare_ID ya kalo kamu ada pertanyaan atau kendala seputar shopee. ^RB
3	Shopee, sidak kantor spx express. kota surabaya, spx tidak ada, alamatnya sama padahal sudah pindah. bagaimana mau ambil barang kalau begini? mending kembalikan kurir lain kayak jnt, JNE, indopaket dll. spx express gk	Hai kak yanti a, maaf ya udah bikin kamu ngga nyaman karena kendala yang kakak alami. Agar aku bisa bantu cek, boleh infoin username dan nomor pesanan kakak ke DM X @ShopeeID atau @ShopeeCare. Aku tunggu ^AT
4	mantapp	Hi Kak Yazid Ahmad makasih ya untuk reviewnya. Kuy kepoin sosial media resmi shopee biar Kakak tau update promo yang saat ini tersedia hehe. Jangan sungkan chat aku di dm IG @ShopeeCare_ID ya kalo kamu ada pertanyaan atau kendala seputar shopee. ^RB
5	Shopee sekarang ngirimnya lama, payah	Hi kak Rouf Fadli, maaf banget ya udah buat kamu ga nyaman. terkait kendala kamu, Aku infoin kamu bisa chat ke aku melalui DM IG @Shopeecare_ID dgn melampirkan No pesanan dan username. Sehat selalu ^RB

Gambar 2. Dataset Penelitian

Atribut utama yang dipakai adalah comment_text yang memuat ulasan pengguna dan replyContent yang berisi tanggapan dari *customer service*. Berdasarkan analisis kelengkapan data, terdapat 415.842 data yang memiliki pasangan comment_text dan replyContent yang utuh. Deduplikasi dilakukan berdasarkan teks comment_text yang telah dinormalisasi. Apabila satu teks ulasan memiliki lebih dari satu replyContent, satu respons dipilih secara deterministik sehingga setiap teks ulasan hanya direpresentasikan satu kali. Proses ini menghasilkan 264.230 teks ulasan unik dengan respons *customer service* sebagai kandidat analisis. Atribut lain dalam dataset tidak digunakan karena keberadaannya tidak konsisten di seluruh data.

2.2. Preprocessing Data

Tahap preprocessing pada teks ulasan dilakukan sebelum tahap ekstraksi fitur guna memperbaiki keseragaman data serta meminimalkan noise pada teks [9]. Proses diterapkan pada kolom comment_text dan replyContent dengan langkah-langkah yang sama, yaitu mengganti daftar kosong, menghapus tanda ^LG, menghilangkan spasi di awal dan akhir teks, mengubah semua huruf menjadi huruf kecil, serta menyatukan beberapa spasi menjadi satu spasi saja seperti pada tabel 1.

Tabel 1. Contoh Hasil Preprocessing Data

Data Mentah	Setelah Preprocessing	Atribut
Hai kak Mafia Degan, mohon maaf atas ketidaknyamanannya. Terkait kendala yg kaka alami mengenai pesanan stuck, agar bisa di cek lebih lanjut, mohon infokan detail kendalanya melalui DM ke Media Sosial @Shopee_id/@Shopeecare_id ya. Di tunggu ^LG	hai kak mafia degan, mohon maaf atas ketidaknyamanannya. terkait kendala yg kaka alami mengenai pesanan stuck, agar bisa di cek lebih lanjut, mohon infokan detail kendalanya melalui dm ke media sosial @shopee_id/@shopeecare_id ya. di tunggu	replyContent
PENGIRIMAN LAMA, NGESTUCK,SOPIR KELAMAAN NGIRIM, BUBARKAN SAJA	pengiriman lama, ngestuck,sopir kelamaan ngirim, bubarkan saja	comment_text

Pembersihan replyContent dilakukan terbatas karena atribut ini menjadi sumber pembentukan weak label. Struktur dan pola linguistik utama tetap dipertahankan, agar labeling functions masih mengenali sinyal seperti permintaan maaf, ketidaknyamanan, rating, dan ucapan terima kasih. Pada tahap klasifikasi, replyContent tidak menjadi fitur; model hanya mempelajari representasi teks dari comment_text.

2.3. Pembentukan Weak Label

Weak label terbentuk melalui pendekatan *weak supervision*: *labeling functions* yang mengenali pola linguistik pada replyContent setelah tahap normalisasi. Metode *weak supervision* melakukan pemberian label pada data yang tidak memiliki label melalui proses algoritmik, dengan memanfaatkan sumber informasi seperti kata kunci, regular expression, atau aturan heuristik [10]. Fungsi-fungsi ini menangkap tiga jenis sinyal: sinyal negatif, sinyal positif, dan istilah generik yang belum cukup menentukan polaritas sentimen. Tabel 2 merinci sinyal tersebut.

Tabel 2. Aturan weak label

Kelompok	Daftar Literal yang Dicakup	Keputusan
Negatif	maaf, ketidaknyamanan, serta variasi tidak/tak/ga/gak/nggak/ngga/gk + nyaman	Negatif, kecuali terdapat sinyal positif sehingga dikategorikan sebagai *conflict*
Positif: rating	bintang 5, bintang lima, rating 5, rating lima, 5 bintang	Positif, kecuali terdapat sinyal negatif sehingga dikategorikan sebagai *conflict*mment_text
Positif: terima kasih + konteks	makasih, makasi, mksih, mksi, terima kasih, terimakasih, thankyou, thanks, yang disertai konteks review, feedback, penilaian, rating, atau bintang	Positif; ucapan terima kasih tanpa konteks tidak cukup untuk menentukan label
Positif: emosi/loyalitas	senang, puas, setia belanja, setia berbelanja	Positif
Isu generik	kendala, masalah, error, terkendala	*Generic issue abstain* apabila tidak terdapat sinyal positif atau negatif

Istilah generik seperti kendala, masalah, error, atau terkendala tidak menentukan sentimen negatif secara langsung, karena istilah ini muncul juga pada respons informatif dan templat layanan yang netral. Sistem menandai respons semacam ini sebagai generic issue abstain, kecuali disertai sinyal positif atau negatif; kombinasi keduanya menjadi conflict. Respons tanpa sinyal memadai masuk kategori abstain. Hanya data berlabel positif atau negatif yang masuk ke tahap pelatihan model klasifikasi.

2.4. Ekstraksi Fitur dan Klasifikasi SVM

Ekstraksi fitur diterapkan pada comment_text menggunakan Term Frequency–Inverse Document Frequency (TF-IDF) yang digunakan untuk merubah teks menjadi reprsesntasi numerik yang didasarkan pada tingkatan pentingnya kata pada dokumen. Representasi TF-IDF akan menjadi masukan bagi algoritma SVM untuk klasifikasi teks [11]. Perhitungan bobot didapatkan dari hasil perkalian term freuency (TF) dengan inverse document frequency (IDF) seperti pada persamaan (1)

$$TFIDF(t,d) = TF(t,d)IDF(t) \quad (1)$$

Dimana nilai TF dihitung sebagai TF(t,d)=1+log(ft,d), dengan (f_{t,d}) berupa frekuensi kemunculan istilah (t) pada dokumen (d). Sementara itu, IDF memberikan bobot lebih tinggi pada istilah yang jarang muncul pada keseluruhan dokumen. Representasi TF-IDF pada penelitian ini menggunakan unigram dan bigram kata dengan min_df=3, max_df=0,995, frekuensi istilah sublinear, serta batas maksimum fitur 250.000.

Setelah menghapus semua baris yang terkait dengan pengujian *human test* dari dataset berlabel biner guna mencegah kebocoran data, tersisa 250.130 data. Data tersebut kemudian digunakan untuk pelatihan, bersama dengan 144.262 fitur TF-IDF, yang mana hasil representasi akan digunakan sebagai input dari Linear Support Vector Machine. SVM bekerja dengan dengan menemukan hyperplane yang memisahkan dua kelas dengan jarak maksimum [12] [13]. Fungsi keputusan di Linear SVM seperti dalam Persamaan (2).

$$f(x) = w\hat{}\{T\}x + b \quad (2)$$

Dengan x pada persamaan tersebut merupakan vektor fitur TF-IDF, (w) adalah vektor bobot, dan (b) adalah bias. Penelitian ini menerapkan LinearSVC dengan parameter C=1.0, class_weight='balanced', dan random_state=42. Pembobotan kelas diterapkan dengan landasan distribusi label lemah tidak seimbang antara kelas positif dan negatif. Sebagai perbandingan, penelitian

ini juga melakukan eksperimen pada Logistic Regression, Multinomial Naive Bayes, serta classifier mayoritas. Seluruh classifier memakai representasi teks dan sumber label lemah yang serupa, sehingga perbandingan dilakukan pada metode klasifikasinya, bukan pada cara pembentukan labelnya.

2.5. Evaluasi Model dan Validasi

Evaluasi mencakup dua komponen: kualitas labeling functions dan performa model klasifikasi. Weak labeling diukur dengan precision dan coverage seperti pada persamaan (3) dan (4)

$$Precision = TP \,/\, (TP \,+\, FP) \tag{3}$$

$$Coverage \;=\; N_labeled \,/\, N_total \tag{4}$$

Sedangkan performa SVM diukur menggunakan accuracy precision, recall, F1-score per kelas, dan macro-F1 yang umum digunakan untuk mengevaluasi performa model klasifikasi teks [14], balanced accuracy juga diterapkan sebagai evaluasi yang digunakan untuk distribusi kelas tidak seimbang [15], seperti pada persamaan 5

$$\begin{aligned} &Accuracy \;=\frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad Macro-F1 \;=\; (1\,/\,K) \;\times\; \Sigma\, F1_k \\ &Balanced\ Accuracy \;=\; 1/2 \;\times\; [(TP\,/\,(TP \;+\; FN)) \;+\; (TN\,/\,(TN \;+\; FP)) \\ &Recall = \frac{TP}{(TP+FN)} \qquad F1-score = 2 \times (Precision \times Recall)/(Precision + Recall) \end{aligned} \tag{5}$$

Penelitian ini menambahkan majority classifier sebagai lower-bound baseline untuk memastikan peningkatan performa tidak berasal semata dari dominasi kelas positif [16]. Untuk mencegah data leakage, teks ulasan pada data *human test* dari data pelatihan tidak digunakan exact reply template dan komentar dengan representasi token kanonik identic juga tidak digunakan. Untuk mengukur ketidakpastian macro-F1, bootstrap dijalankan sebanyak 2.000 replikasi.

Validasi kualitas weak label dilakukan lewat dua skema *human test*: skema natural melibatkan 272 sampel yang mencerminkan distribusi data nyata dan skema berstrata 256 sampel yang disusun untuk meningkatkan representasi kasus yang lebih beragam. Pada kedua skema, hasil weak labeling dibandingkan dengan label referensi manusia untuk menilai kualitas aturan pelabelan menggunakan metrik precision dan binary coverage. Seluruh ulasan yang termasuk dalam *human test* dikeluarkan dari data pelatihan guna mencegah data leakage. Selain itu, data *human test* juga dipakai sebagai data evaluasi untuk mengukur performa model klasifikasi.

3. Hasil dan Pembahasan

3.1. Spesifikasi Karakteristik Data dan Hasil *Weak Labeling*

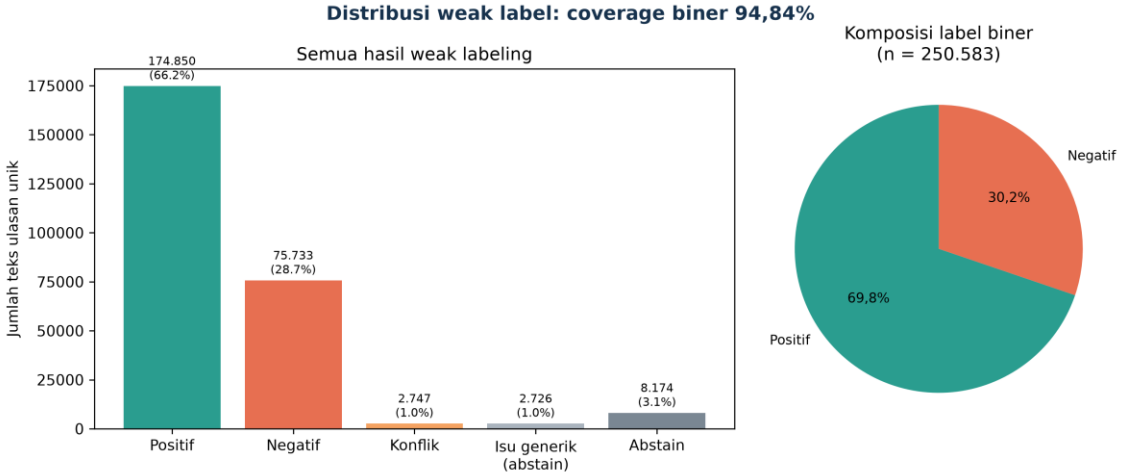
Dataset awal terdiri atas 485.320 baris ulasan Shopee. Setelah pemeriksaan kelengkapan, diperoleh 415.842 pasangan comment_text dan replyContent yang tidak kosong. Deduplikasi berdasarkan teks ulasan ternormalisasi menghasilkan 264.230 ulasan unik dengan respons *customer service*. Sebanyak 151.612 data teridentifikasi sebagai duplikasi, atau 36,46% dari pasangan lengkap. Tingkat duplikasi ini menunjukkan pentingnya deduplikasi sebelum pembentukan model: penggunaan seluruh baris mentah memberi bobot berulang pada teks yang sama dan mengubah distribusi kelas.

Penerapan labeling functions pada replyContent menghasilkan beberapa kelompok sinyal linguistik. Sinyal yang paling banyak ditemukan adalah ucapan terima kasih yang disertai konteks ulasan, frekuensi masing-masing sinyal ditunjukkan pada tabel 3.

Tabel 3. Frekuensi Sinyal

Sinyal	Jumlah	Presentase dari 264.230
Negatif: maaf/tidak nyaman	78.480	29,70%
Positif: Bintang/rating	95.440	36,12%
Positif: terima kasih + konteks	173.651	65,72%
Positif: emosi/loyalitas	47.175	*17,85%*
Isu generik	157.407	*59,57%*

Istilah generik seperti kendala, masalah, error, dan terkendala ditemukan pada 157.407 respons (59,57%), namun tidak langsung diperlakukan sebagai indikator negatif karena dapat muncul dalam respons yang bersifat informatif maupun template layanan. Labeling functions berbasis replyContent menghasilkan 250.583 label sentimen biner, atau 94,84% dari 264.230 ulasan unik. Sebanyak 174.850 ulasan (66,17%) memperoleh label positif dan 75.733 ulasan (28,66%) memperoleh label negatif. Sisanya, 2.747 data (1,04%) dikategorikan sebagai conflict, 2.726 data (1,03%) sebagai generic issue abstain, dan 8.174 data (3,09%) sebagai abstain. Distribusi hasil pelabelan ditunjukkan pada gambar 3.



Gambar 3. Distribusi Weak Label

Cakupan label biner sebesar 94,84% menunjukkan bahwa respons *customer service* menyediakan sinyal yang dapat dimanfaatkan pada sebagian besar data. Sinyal positif paling banyak berasal dari ucapan terima kasih yang disertai konteks ulasan, ditemukan pada 173.651 data (65,72%), sedangkan sinyal negatif berupa permintaan maaf atau ungkapan ketidaknyamanan ditemukan pada 78.480 data (29,70%). Mekanisme abstain diterapkan pada kondisi tersebut untuk mengurangi risiko pemberian label negatif hanya berdasarkan keberadaan kata terkait masalah, sejalan dengan prinsip *weak supervision* bahwa sinyal pelabelan bersifat tidak sempurna dan noise perlu ditangani secara eksplisit [5], [6].

3.2. Kualitas *Weak supervision*

Kualitas labeling functions pada tahap pengembangan dievaluasi menggunakan dua skema *human test*: sampel natural yang mengikuti distribusi data aktual, dan sampel berstrata yang sengaja meningkatkan representasi kasus diagnostik. Skema kedua dirancang untuk menguji ketahanan aturan pada contoh yang lebih beragam dan sulit. Pada *human test* natural, aturan negatif mencapai presisi 1,000 dengan 56 dari 56 prediksi negatif sesuai dengan label referensi, sedangkan aturan positif memperoleh presisi 0,985 dengan 200 dari 203 prediksi sesuai. binary coverage pada *human test* natural mencapai 0,952 atau 259 dari 272 data biner. Pada skema *human test* berstrata, presisi tetap tinggi, yaitu 0,979 untuk kelas negatif dan 0,983 untuk kelas0 positif yaitu 57 dari 58 diprediksi sesuai. Binary coverage mencapai 0,605 yang berarti 155 dari 256 data seperti pada tabel 4.

Tabel 4. Kualitas *Weak supervision*

Skema *Human test*	Presisi Negatif	Presisi Positif	Binary Coverage
Natural	1,000	0,985	0,952
Berstrata	0,979	0,983	0,605

Selisih cakupan antara kedua skema menunjukkan bahwa kinerja aturan dipengaruhi oleh sifat data yang dinilai. Pada sampel yang mengikuti distribusi natural, pola respons CS dapat memberikan label pada sebagian besar ulasan. Sebaliknya, sampel berstrata secara sengaja memperbanyak representasi kasus yang lebih menantang sehingga lebih banyak respons tidak memberikan sinyal positif atau negatif yang cukup kuat. Akan tetapi, presisi pada kedua kelas tetap di atas 0,97. Hasil ini menunjukkan bagaimana mekanisme abstain berfungsi: sistem tidak memberlakukan label pada respons yang tidak jelas untuk menjaga kualitas label yang diterapkan selama pelatihan.

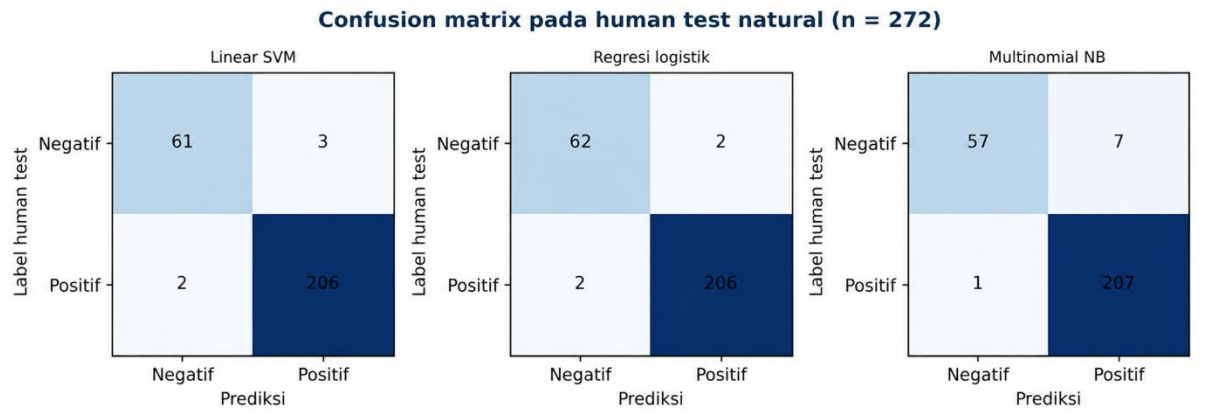
3.3 Performa Klasifikasi Sentimen

Kinerja Linear SVM dibandingkan terhadap tiga metode: Majority Classifier, Logistic Regression, dan Multinomial Naive Bayes. Majority Classifier berfungsi sebagai lower-bound baseline, untuk mengetahui apakah model memberi peningkatan dibanding sekadar memprediksi kelas paling dominan. Logistic Regression menjadi pembanding sesama model linear, sementara Multinomial Naive Bayes mewakili pendekatan probabilistik yang umum dipakai pada klasifikasi teks. Sebanyak 250.130 data yang telah diberi label digunakan sebagai data pelatihan setelah data yang digunakan dalam *human test* dikeluarkan. Keempat model diuji dengan representasi TF-IDF, sumber weak label, dan data evaluasi yang sama. Hasil evaluasi terdapat pada tabel 5.

Tabel 5. Performa Model Klasifikasi

Skema	Model	Accuracy	Balanced Accuracy	Macro F1
Natural	Majority	0,765	0,500	0,433
	Linear SVM	**0,982**	**0,972**	**0,974**
	Logistic Regression	0,985	0,980	0,980
	Multinomial NB	0,971	0,943	0,958
Berstrata	Majority	0,496	0,500	0,332
	Linear SVM	**0,973**	**0,973**	**0,973**
	Logistic Regression	0,973	0,973	0,973
	Multinomial NB	0,973	0,973	0,973

Dalam skema natural, Linear SVM memperoleh akurasi 0,982, akurasi seimbang 0,972, dan macro-F1 0,974. Di tingkat kelas, precision, recall, dan F1-score dicatat masing-masing sebesar 0,968, 0,953, dan 0,961 untuk sentimen negatif, sementara untuk sentimen positif adalah 0,986, 0,990, dan 0,988. Tingginya nilai balanced accuracy dan F1-score untuk kelas negatif menunjukkan bahwa kinerja model tidak hanya dipengaruhi oleh dominasi label positif — hal ini terlihat dari perbandingan dengan classifier mayoritas, yang hanya mencapai accuracy 0,765 dengan balanced accuracy 0,500 dan macro-F1 0,433. Confusion matrix Linear SVM, Logistic Regression, dan Multinomial Naive Bayes pada skema natural ditunjukkan pada gambar 4, memperlihatkan pola kesalahan prediksi masing-masing model secara lebih rinci.



Gambar 4. Confusion Matrix *Human test* Natural

Linear SVM pada skema natural mengklasifikasikan dengan benar 61 data negatif dan 206 data positif, dengan lima kesalahan: tiga negatif diprediksi positif, dan dua positif diprediksi negatif (Gambar 7). Logistic Regression mencatat empat kesalahan, sementara Multinomial Naive Bayes delapan. Urutan ini konsisten dengan macro-F1 masing-masing: 0,980 untuk Logistic Regression, 0,974 untuk Linear SVM, dan 0,958 untuk Multinomial Naive Bayes. Logistic Regression memperoleh macro-F1 tertinggi pada skema natural, meski selisihnya dengan Linear SVM kecil. Pada skema berstrata, ketiga model justru memperoleh macro-F1 yang sama, 0,973. Hasil ini SVM tetap kompetitif dibandingkan classifier linear lain, serta representasi TF-IDF dari weak label berbasis respons *customer service* terbukti efektif dipakai oleh beberapa metode klasifikasi sekaligus. Hasil SVM juga sesuai dengan penelitian sebelumnya yang menunjukkan bahwa SVM tetap unggul dalam klasifikasi sentimen berdasarkan ulasan [3], [8]. Interval kepercayaan macro-F1 Linear SVM melalui 2.000 replikasi

bootstrap berada pada 0,949–0,995 pada skema natural, mengindikasikan performa yang konsisten pada sampel pengembangan.

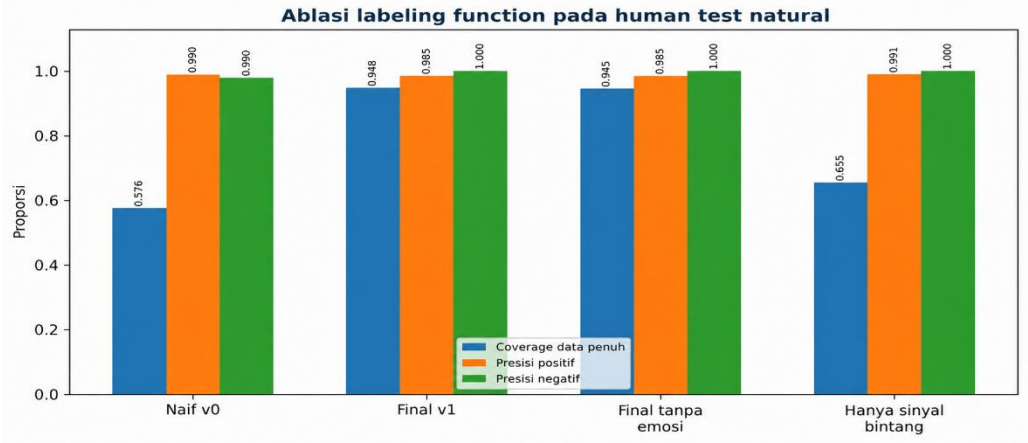
3.4 Robustness dan Pengaruh Desain Labeling Function

Salah satu risiko memanfaatkan respons *customer service* sebagai sumber *weak supervision* adalah kemungkinan model secara tidak langsung mendapatkan manfaat dari pola respons yang serupa. Untuk memverifikasi hal tersebut, dilakukan pengujian *exact reply string holdout*, yaitu dengan mengeluarkan pola respons yang terdapat pada data *human test* dari data latih. Linear SVM mempertahankan nilai macro-F1 sebesar 0,974 baik sebelum maupun sesudah template holdout, sehingga tidak ada penurunan kinerja yang terdeteksi. Logistic Regression mengalami penurunan ringan dari 0,980 menjadi 0,975, sedangkan Multinomial NB tetap di 0,958. Hasil ini menunjukkan bahwa kinerja classifier tidak hanya tergantung pada adanya templat respons yang identik dalam data pelatihan dan evaluasi, seperti pada tabel 6.

Tabel 6. Hasil Robustness Model terhadap Exact Reply String Holdout

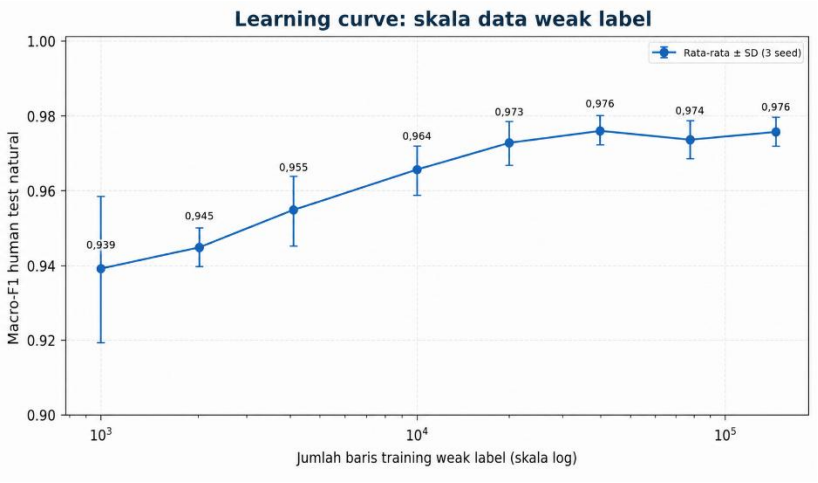
Model	Sebelum Holdout Macro F1	Sesudah Holdout Macro F1	Macro F1
Linear SVM	0,974	0,974	0,000
Logistic Regression	0,980	0,975	-0,005
Multinomial Naïve Bayes	0,958	0,958	0,000

Setelah kestabilan model terhadap kemiripan data diperiksa, analisis berikutnya difokuskan pada pengaruh desain labeling function terhadap coverage dan jumlah conflict weak label. Hasil ablasi pada Gambar 10 menunjukkan bahwa rancangan aturan awal (Naive v0) hanya menghasilkan coverage sebesar 57,59% dan menimbulkan 106.324 kasus conflict. Setelah aturan disempurnakan melalui pemisahan sinyal negatif, rating positif, ucapan terima kasih yang disertai konteks, emosi atau loyalitas, serta istilah isu generik, coverage meningkat menjadi 94,84% dan jumlah conflict menurun menjadi 2.747 data. Sebaliknya, ketika hanya sinyal rating atau bintang yang digunakan, coverage turun menjadi 65,47%.



Gambar 5. Ablasi Labeling *Human test* Natural

Hasil penelitian mengindikasikan bahwa peningkatan cakupan tidak dicapai dengan menghapus semua aturan, tetapi melalui kombinasi beberapa sinyal yang lebih relevan dan memberikan abstain pada pola yang kurang informatif. Penambahan data weakly labeledd apat digunakan oleh model dengan efektif, seperti yang terlihat pada learning curve gambar 6.



Gambar 6. Learning Curve Weak Label

Dengan keseluruhan 1.000 data latihan, nilai rata-rata macro-F1 mencapai 0,939. Nilai tersebut naik menjadi 0,973 saat jumlah data mencapai 20.000 dan sekitar 0,976 saat mencapai 40.000 data. Setelah ukuran itu, peningkatan menjadi cukup kecil, dengan macro-F1 sekitar 0,974–0,976 pada ukuran data 80.000 hingga 150.332. Pola ini menunjukkan bahwa keuntungan maksimal dari penambahan data terjadi pada ukuran data yang rendah hingga menengah, sementara setelah mencapai puluhan ribu data, model mulai mencapai kondisi plateau.

Hasil penelitian menunjukkan bahwa respons layanan pelanggan berpotensi menjadi sumber *weak supervision* karena dapat menghasilkan label biner pada banyak ulasan dengan presisi yang tinggi pada kedua skema *human test*, sedangkan TF-IDF dan SVM dapat mempelajari pola sentimen dari comment_text tanpa memanfaatkan replyContent sebagai fitur klasifikasi. Kontribusi utama dari penelitian ini bukan pada penciptaan classifier baru, tetapi pada penggunaan informasi interaksi antara pengguna dan layanan pelanggan sebagai sumber untuk membentuk label. Pendekatan ini berbeda dari studi yang memanfaatkan model pralatih sebagai sumber label [7] dan memperluas penerapan *weak supervision* dengan menggunakan sinyal yang secara alami ada dalam interaksi layanan [5].

4. Kesimpulan

Penelitian ini menunjukkan bahwa respons *customer service* pada ulasan Shopee dapat dimanfaatkan sebagai sumber *weak supervision* untuk membentuk label sentimen secara otomatis. Labeling functions yang dikembangkan mampu menghasilkan 250.583 label biner atau mencakup 94,84% dari 264.230 ulasan unik. Validasi pada *human test* natural menghasilkan presisi sebesar 1,000 untuk kelas negatif dan 0,985 untuk kelas positif dengan binary coverage 0,952, sedangkan pada skema berstrata diperoleh presisi 0,979 untuk kelas negatif dan 0,983 untuk kelas positif dengan coverage 0,605. Mekanisme abstain terbukti mempertahankan ketepatan label pada kasus yang lebih sulit dengan tidak memaksakan pemberian label ketika sinyal sentimen tidak cukup kuat.

Representasi TF-IDF yang dilatih menggunakan weak label juga dapat dimanfaatkan oleh model klasifikasi. Linear SVM memperoleh macro-F1 sebesar 0,974 pada skema natural dan 0,973 pada skema berstrata, serta tetap menunjukkan kinerja yang stabil setelah pengujian exact reply string holdout. Analisis ablasi mengindikasikan kombinasi beberapa sinyal linguistik menghasilkan cakupan label yang lebih baik dibandingkan penggunaan aturan yang lebih sederhana, sedangkan learning curve memperlihatkan penambahan jumlah data weakly labeled memberikan peningkatan performa hingga mencapai kondisi yang relatif stabil pada ukuran data yang lebih besar. Dengan cara ini, respons *customer service* dapat mengurangi kebutuhan akan anotasi manual untuk klasifikasi sentimen. Penelitian selanjutnya bisa memperluas pola fungsi penandaan dan menguji metode ini pada platform atau domain ulasan yang berbeda.

Daftar Pustaka

[1] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, “Analysing app reviews for software engineering: a systematic literature review,” *Empir. Softw. Eng. 2021 272*, vol. 27, no. 2, pp. 43-, Jan. 2022, doi:

10.1007/S10664-021-10065-7.

[2] S. Das, N. Deb, N. Chaki, and A. Cortesi, "Driving the Technology Value Stream by Analyzing App Reviews," *IEEE Trans. Softw. Eng.*, vol. 49, no. 07, pp. 3753–3770, Jul. 2023, doi: 10.1109/TSE.2023.3270708.

[3] E. Hashmi, · Sule, Y. Yayilgan, E. Hashmi, and S. Y. Yayilgan, "A robust hybrid approach with product context-aware learning and explainable AI for sentiment analysis in Amazon user reviews," *Electron. Commer. Res. 2024 256*, vol. 25, no. 6, pp. 5139–5171, Aug. 2024, doi: 10.1007/S10660-024-09896-5.

[4] N. Sikana, S. Winardi, G. -, G. F. Situmorang, and R. Lubis, "Analisis Sentimen untuk Ulasan Produk E-Commerce Shopee Menggunakan BERT," *J. Sifo Mikroskil*, vol. 26, no. 2, pp. 2025–223, Oct. 2025, doi: 10.55601/JSM.V26I2.1796.

[5] O. Wallaart, F. Frasincar, and F. van der Knaap, "Weakly-supervised sentence-based aspect category and sentiment classification," *Knowledge-Based Syst.*, vol. 317, p. 113403, May 2025, doi: 10.1016/J.KNOSYS.2025.113403.

[6] W. Zhang, Y. Zhou, S. Liu, Y. Zhang, and X. Shang, "Weakly supervised text classification framework for noisy-labeled imbalanced samples," *Neurocomputing*, vol. 610, p. 128617, Dec. 2024, doi: 10.1016/J.NEUCOM.2024.128617.

[7] M. R. Amalsyah, D. Kurniawan, A. Rifai, and P. Sari, "Analisis Sentimen Ulasan Pengguna Aplikasi Fintechmenggunakan Framework CRISP-DM dalam PenentuanPrioritas Pengembangan Produk," *Sist. J. Sist. Inf.*, vol. 14, no. 2, pp. 813–825, 2025, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[8] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 16, no. 2, pp. 113–124, Apr. 2022, doi: 10.22146/IJCCS.68903.

[9] A. T. Wibowo, A. Fanani, M. Ridwan, and B. K. Aji, "NLP-driven hate speech detection on TikTok: a case study from UIN Sunan Ampel Surabaya," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 24, no. 4, pp. 1157–1167, Aug. 2026, doi: 10.12928/TELKOMNIKA.V24I4.27419.

[10] L. M. Chen, B. X. Xiu, and Z. Y. Ding, "Multiple *weak supervision* for short text classification," *Appl. Intell. 2021 528*, vol. 52, no. 8, pp. 9101–9116, Jan. 2022, doi: 10.1007/S10489-021-02958-3.

[11] A. Hermawan, R. Yusuf, B. Daniawan, and Junaedi, "Enhanced sentiment analysis and emotion detection in movie reviews using support vector machine algorithm," *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 23, no. 1, pp. 138–146, Feb. 2025, doi: 10.12928/TELKOMNIKA.V23I1.26377.

[12] K. Belajar Pemrograman Author, S. Astiti, B. Satria, Y. Gusla Nengsih, and S. Fadilah, "Penerapan Model Support Vector Machine dalam Prediksi Keberhasilan Belajar Pemrograman: Application of Support Vector Machine Model in Predicting Programming Learning Success," *Edu Cendikia J. Ilm. Kependidikan*, vol. 6, no. 01, pp. 351–365, Apr. 2026, doi: 10.47709/EDUCENDIKIA.V6I01.8061.

[13] R. Fachuzi *dkk.*, "PENERAPAN SUPPORT VECTOR MACHINE (SVM) UNTUK DETEKSI HOAKS PADA BERITA ONLINE," vol. 10, no. 4, pp. 6782–6787, 2026.

[14] D. Harist Saifulloh and B. Perdian, "Analisis Sentimen Ulasan Produk E-Commerce Berbasis Teks Dengan Pendekatan Probabilistik Naive Bayes," *J. Sist. Inf. Komput. (SIKOM)*, vol. 2, no. 1, pp. 61–74, 2025, Accessed: Aug. 24, 2026. [Online]. Available: https://journal.beaninstitute.id/index.php/sikom/article/view/151

[15] R. Anadra, H. Wijayanto, and K. Sadik, "Sentiment Analysis of Tokopedia Customer Reviews Using BiLSTM and IndoBERT with Comparative Analysis of Preprocessing and Labeling Methods," *Int. J. Adv. Data Inf. Syst.*, vol. 6, no. 3, pp. 773–788, Dec. 2025, doi: 10.59395/IJADIS.V6I3.1458.

[16] I. Al Khoiry, "Aspect-Level User Sentiment Patterns in Shopee Mobile App Reviews Using TF-IDF and Logistic Regression," vol. 8, no. 1, pp. 31–42, 2026.