Journal of Technology and Data Science (JTDS)
Vol. 4, No. 2, July–December 2026, pp. 410-424
p-ISSN: 2962-3693 | e-ISSN: 2962-7788

Sentiment Analysis of the Si Anteng Application and Classification of Trans Jateng Passenger Complaints Based on Naive Bayes

Moch. Nazril Ilham[1,*], M. Ichwan Al Hafizd[2], Zahra Zahrani[3], Laelatul Hidayah[4], Muhammad Na'afian Anugerah[5], Imam Prayogo Pujiono[6]

[1,2,3,4,5,6]*Faculty of Science and Technology, Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan, Pekalongan, Indonesia*

E-mail: moch.nazril.ilham24071@mhs.uingusdur.ac.id[*]; m.ichwan.al.hafizd24092@mhs.uingusdur.ac.id; zahra.zahrani24011@mhs.uingusdur.ac.id; laelatul.hidayah24009@mhs.uingusdur.ac.id; muhammad.naafian.anugerah24070@mhs.uingusdur.ac.id; imam.prayogo.pujiono@uingusdur.ac.id
[*]*Corresponding Author*

© 2026. Moch. Nazril Ilham, et al. Published by JTDS. *This article is open-access under the CC BY-NC-SA 4.0 license*

Article Info

Received: 1 November 2026
Revised: 24 November 2026
Published: 26 November 2026

Keywords:
sentiment analysis; Naive Bayes; TF-IDF; complaint mapping; Si Anteng

ABSTRACT

The Si Anteng application is the information gateway for Trans Jateng Bus Rapid Transit (BRT) passengers, so its Google Play Store reviews mix judgments about the app with experiences of riding the bus. This study classifies review sentiment and maps the complaint themes that emerge. A total of 443 Indonesian-language reviews of com.ngi.sianteng were collected with google-play-scraper in newest-first order. Sentiment labels were derived from star ratings (1–2 negative, 3 neutral, 4–5 positive) and neutral reviews were set aside, leaving 410 reviews (271 positive and 139 negative). The text was cleaned, weighted with unigram–bigram Term Frequency–Inverse Document Frequency (TF-IDF), and classified with Multinomial Naive Bayes using an 80:20 train–test split. On the 82 test reviews, accuracy reached 70.73%, yet recall for the negative class was only 14.29% (balanced accuracy 57.14%), and the accuracy was only slightly above a majority-class guess (65.85%); an exact McNemar test on these test data showed that the difference was not significant ($p = 0.125$). In five-fold cross-validation, negative-class recall of the baseline was only 17.20 ± 6.06%, rising to 66.85 ± 6.47% with balanced class priors (equivalent to Complement Naive Bayes for two classes) and to 90.61 ± 5.90% with the Synthetic Minority Over-sampling Technique (SMOTE). Multi-label mapping of the 139 negative reviews placed application problems first (51 reviews; 36.7%), followed by routes, stops, and travel information (32; 23.0%) and bus punctuality and service timing (27; 19.4%). The findings indicate that users judge the application and the bus service as one experience. Because the labels come from ratings, the model is better used as an initial filter for setting inspection priorities than as a human-validated measure of sentiment.

Info Artikel

Diterima: 1 November 2026
Direvisi: 24 November 2026
Diterbitkan: 26 November 2026

Kata Kunci:
analisis sentimen; Naive Bayes; TF-IDF; pemetaan keluhan; Si Anteng

ABSTRAK

Aplikasi Si Anteng menjadi pintu informasi bagi pengguna Bus Rapid Transit (BRT) Trans Jateng, sehingga ulasannya di Google Play Store memuat penilaian terhadap aplikasi sekaligus pengalaman naik bus. Penelitian ini mengklasifikasikan sentimen ulasan dan memetakan tema keluhan yang muncul. Sebanyak 443 ulasan berbahasa Indonesia untuk aplikasi com.ngi.sianteng diambil dengan google-play-scraper menurut urutan terbaru. Label sentimen diturunkan dari rating (1–2 negatif, 3 netral, 4–5 positif) dan ulasan netral tidak diikutkan, sehingga tersisa 410 ulasan (271 positif dan 139 negatif). Teks dibersihkan, dibobot dengan Term Frequency–Inverse Document Frequency (TF-IDF) unigram–bigram, lalu diklasifikasikan memakai Multinomial Naive Bayes dengan pembagian data latih dan uji 80:20. Pada 82 data uji, akurasi mencapai 70,73%, tetapi recall kelas negatif hanya 14,29% (balanced accuracy 57,14%), dan akurasi itu hanya sedikit di atas strategi menebak kelas mayoritas (65,85%); uji McNemar eksak pada data uji tersebut menunjukkan selisih itu belum signifikan (p = 0,125). Pada validasi silang lima lipatan, recall kelas negatif baseline hanya 17,20 ± 6,06%, lalu naik menjadi 66,85 ± 6,47% dengan prior kelas seimbang (setara Complement Naive Bayes pada dua kelas) dan 90,61 ± 5,90% dengan Synthetic Minority Over-sampling Technique (SMOTE). Pemetaan multi-label pada 139 ulasan negatif menempatkan kendala aplikasi sebagai tema terbanyak (51 ulasan; 36,7%), disusul rute, halte, dan informasi perjalanan (32; 23,0%) serta ketepatan bus dan waktu layanan (27; 19,4%). Temuan ini menunjukkan bahwa pengguna menilai aplikasi dan layanan bus sebagai satu pengalaman. Karena label berasal dari rating, model lebih tepat dipakai sebagai penyaring awal untuk menentukan prioritas pemeriksaan daripada sebagai ukuran sentimen yang telah divalidasi manusia.

1. INTRODUCTION

The smartphone is now the first thing passengers open before setting out. Schedules, fleet positions, fares, and payment methods are looked up through applications, so judgments about public transport are no longer formed only at the bus stop or inside the vehicle, but also on the screen. A review of 81 open-access articles from 2014–2024 shows that sentiment analysis and topic modeling are increasingly used to read the opinions of transport users [1]. A similar approach has been used to examine views on sustainable mobility in campus settings [2] and to monitor service quality at public transport nodes [3].

The findings of such studies are fairly consistent across modes: for rail services in India, the ticketing process, delays, and staff behavior are the areas most in need of improvement [4]; Teman Bus reviews frequently mention bus tracking and punctuality [5]; and reviews of Access by KAI (Kereta Api Indonesia, the Indonesian Railway Company) are dominated by negative sentiment with the words “aplikasi” (application), “bayar” (pay), and “tiket” (ticket) [6]. For Indonesian app store reviews, Naive Bayes and Term Frequency–Inverse Document Frequency (TF-IDF) are widely used, for example on Shopee [7], iPusnas [8], and KitaLulus [9], alongside IndoBERT (an Indonesian variant of Bidirectional Encoder Representations from Transformers) [10, 11], multilabel classification [12], and deep learning [13].

The Si Anteng application, with the package ID (identifier) com.ngi.sianteng, is the object of this study. The content of its reviews shows that users place the application in a single chain with the Trans Jateng Bus Rapid Transit (BRT) service: complaints about failed account logins appear side by side with accounts of buses not stopping at stops, slipping schedules, routes perceived as incomplete, staff attitudes, and ticket payment status. A single review can therefore judge two things at once, namely the software and the transport service. If both are merged into one sentiment score, the application team and the bus operator alike will struggle to find the part they must handle.

Most studies stop at sentiment percentages and accuracy, whereas what is more useful to service managers is what is being complained about and who should follow it up: the application team, the bus operator, or field staff. Another issue concerns the labels: review sentiment is commonly derived from star ratings, yet the stars and the text do not always agree; this mismatch has been the target of

a dedicated study [14], and rating consistency was examined for the PLN (Perusahaan Listrik Negara, the State Electricity Company) Mobile application [10]. Results based on proxy labels need to be read within clear limits.

Based on this need, this study aims to (1) describe the distribution of ratings and sentiment in Si Anteng reviews; (2) build a two-class classifier based on TF-IDF and Multinomial Naive Bayes; (3) evaluate the model with per-class metrics and compare it with a majority-class guessing strategy through cross-validation, class imbalance handling, and significance testing; and (4) map the complaint themes in negative reviews in a multi-label manner. The contributions of this study are as follows.

- a side-by-side treatment of two-class sentiment classification and multi-label complaint theme mapping for a single regional transport application case, namely Si Anteng;
- the separation of complaints about the application from complaints about the physical BRT service within one review corpus;
- robustness testing of the results through stratified 5-fold cross-validation, three class imbalance handling scenarios, a McNemar test against the majority-class strategy, false negative and negation analysis, and validation of the theme mapping against researcher coding with Cohen's kappa; and
- evaluation reporting with per-class metrics, balanced accuracy, and a majority-class baseline so that accuracy is not over-interpreted, accompanied by re-runnable code.

The claims of this study are deliberately limited: the labels come from ratings rather than manual annotation, so the results are treated as a preliminary mapping.

2. LITERATURE REVIEW

2.1 Sentiment Analysis of App Reviews and Rating-Based Labeling

Sentiment analysis converts opinion text into polarity categories. For app store reviews, the most practical way to obtain labels is to use the star ratings given by users. This approach was used for iPusnas reviews labeled by rating [8], in a comparison of Naive Bayes and logistic regression on marketplace reviews [15], and for Teman Bus reviews, where rating 3 was placed in the negative class [5]. These differing rating thresholds show that the labeling rule is a researcher's decision rather than a natural fact. This study chooses to exclude rating 3 so that the classes do not mix.

The weakness of proxy labels has been demonstrated by a study that specifically searched for mismatches between review text and numeric ratings on the Google Play Store [14]. Some researchers have taken a more costly route: 1,000 of 2,000 AdaKami reviews were labeled manually by two annotators [16], 34,178 SATUSEHAT Mobile reviews were annotated manually [11], and health application reviews were annotated manually with an additional analysis of language intensity [17].

2.2 TF-IDF and Multinomial Naive Bayes

TF-IDF weights a term according to how often it appears in a document and how rarely it appears across the whole corpus. The scikit-learn implementation used here computes term weights as in (1) and (2), then normalizes each document vector with the L2 norm.

$$\text{TF-IDF}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t \tag{1}$$

$$\text{idf}_t = \ln \frac{1 + N}{1 + \text{df}_t} + 1 \tag{2}$$

In these equations, $\text{tf}(t, d)$ is the frequency of term t in document d, N is the number of documents in the corpus, and $\text{df}(t)$ is the number of documents containing term t. Multinomial Naive Bayes selects the class $\hat{c}$ that maximizes the probability in (3), with the probability of a term given a class computed using (4).

$$\hat{c} = \arg\max_{c} \; P(c) \prod_{t} P(t \mid c)\, w_{t,d} \tag{3}$$

$$P(t \mid c) = \frac{N_{tc} + \alpha}{N_c + \alpha |V|} \tag{4}$$

In these equations, $w(t, d)$ is the TF-IDF weight of term t in document d after L2 normalization, $P(c)$ is the prior probability of class c, N_{tc} is the total weight of term t in the training documents of class c, N_c is the total weight of all terms in class c, $|V|$ is the vocabulary size, and α is the Laplace smoothing parameter, set to 1.

The combination of the two is lightweight and easy to replicate, so it is commonly used as a basic benchmark for Indonesian-language app reviews, for example on Shopee [7] and KitaLulus [9]. On Access by KAI reviews, logistic regression was superior in accuracy and precision, whereas Naive Bayes was higher in recall [6]. Comparisons of classical models with deep learning have also been carried out on Indonesian-language hotel reviews [18].

2.3 Class Imbalance and Metric Selection

App store reviews generally lean toward the positive class. Under this condition a classifier easily favors the majority class, as cautioned by a study of MyXL reviews that subsequently applied the Synthetic Minority Over-sampling Technique (SMOTE) together with a Support Vector Machine (SVM) [19]; similar handling was reported for customer comments on the SiCepat courier service [20]. As a result, accuracy alone can be misleading. Differences in behavior between metrics are visible on iPusnas reviews: the accuracy of Naive Bayes was only 58%, but its recall was 81% [8]. This study responds by reporting per-class metrics, balanced accuracy, and the accuracy of the majority strategy as a baseline.

Class imbalance in Indonesian-language hotel reviews classified with Multinomial Naive Bayes and TF-IDF caused predictions to lean toward the positive class and lowered sensitivity to negative sentiment [21]. Beyond SMOTE, Complement Naive Bayes (CNB) was designed to reduce the influence of the dominant class in text classification [22] and has been compared with a linear SVM on imbalanced Indonesian-language sentiment data [23]; the simplest option is to adjust the class prior $P(c)$ in (3). The difference in accuracy against a simple baseline can be tested with the McNemar test, which uses only discordant prediction pairs and is therefore suitable for a single data split [24].

2.4 Complaint Mapping in Public Services

Sentiment answers the direction of a judgment, whereas complaint mapping answers the substance of the problem. Complaints from residents of Jambi City on Twitter have been classified with a recurrent neural network (RNN) and feature selection [25]; iPusnas review categories were linked with ratings to set improvement priorities with the RICE (Reach, Impact, Confidence, Effort) framework [26]; a study of railways in India separated ticketing, delays, and staff behavior into distinct themes [4]; and multilabel classification was applied to Indonesian mobile app reviews for sentiment and emotion simultaneously [12]. The idea is adapted here in a simple form: one negative review may contain more than one theme.

2.5 Negation Handling in Indonesian

Negation words such as “tidak” (not), “bukan” (not, no), “belum” (not yet), and “jangan” (do not) can reverse the meaning of the words they modify. In Multinomial Naive Bayes, the phrase “tidak jelek” (not bad), processed as two unrelated words, may be misclassified as negative, and syntactic analysis alone is not sufficient for negation handling [27]. The conjecture that negation is the main cause of missed negative reviews is tested in Section 4.5.

2.6 Research Gap

Among the studies reviewed, none appears to separate application complaints from physical service complaints within one review corpus of a regional transport application while also reporting a majority-class baseline so that accuracy is not over-interpreted. This study fills that gap for the Si Anteng case.

3. RESEARCH METHODOLOGY

3.1 Research Design and Workflow

This is a quantitative-descriptive study with text classification experiments. The workflow is summarized in Fig. 1 and consists of two paths that diverge at the labeling stage: a sentiment classification path for all binary reviews and a complaint mapping path for negative reviews only. The implementation uses Python on Google Colab with the google-play-scraper, pandas, scikit-learn, matplotlib, and seaborn libraries. The code referred to in this article is presented in Listing 1 and Listing 2; the data collection and text cleaning code is available in the companion notebook.

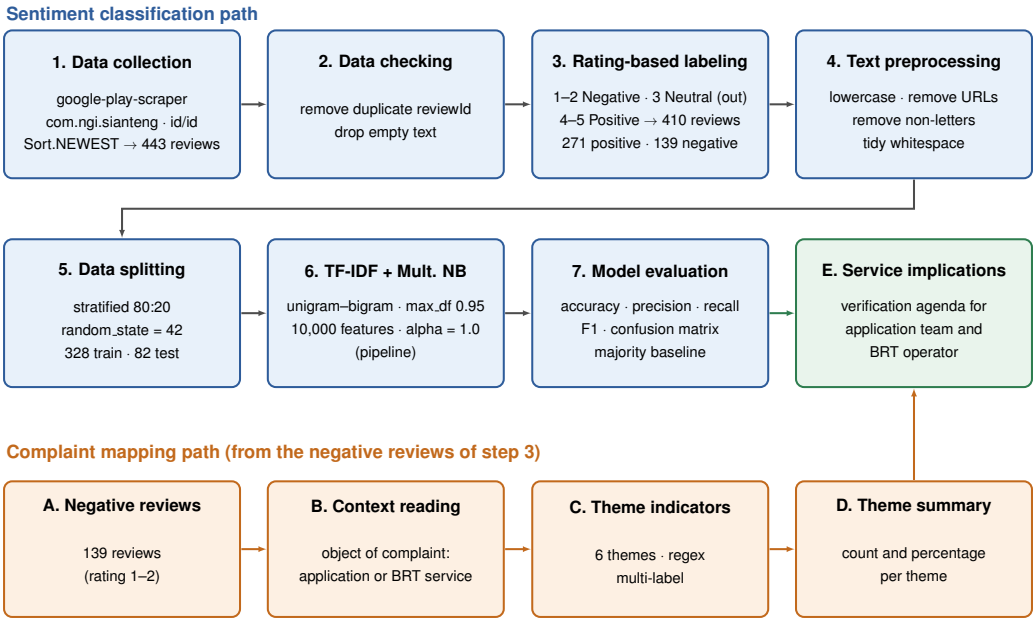

Fig. 1. Research workflow: sentiment classification path (blue) and complaint mapping path (orange).

3.2 Data Collection

The data come from public reviews of the com.ngi.sianteng application on the Google Play Store. Reviews were collected in stages using next-page tokens, with the language parameter id, country id, and newest-first order (Sort.NEWEST), with an upper limit of 3,000 reviews. The result was three pages containing 200, 200, and 43 unique reviews, giving a final corpus of 443 reviews; the most recent review in the corpus is dated 21 October 2026. Duplicates were removed based on reviewId, and rows without text were deleted.

3.3 Labeling and Preprocessing

Ratings 1 and 2 were labeled negative, rating 3 neutral, and ratings 4 and 5 positive. Neutral reviews were excluded from the two-class classification to reduce ambiguity but were still counted in the descriptive analysis of ratings. Preprocessing consisted of lowercasing, removal of Uniform Resource Locators (URLs), removal of non-letter characters, and whitespace tidying. Stemming, stop word removal, and normalization of non-standard language were not applied so that the text remains close to what users wrote.

3.4 Modeling and Evaluation

The binary data were split into 80% training and 20% test data with class stratification. TF-IDF and Multinomial Naive Bayes were arranged in a single pipeline so that TF-IDF weights were learned only

from the training data and no test data information leaked into the model (Listing 1). Evaluation used accuracy, precision, recall, F1-score, and the confusion matrix. Two benchmarks were added: the accuracy of the majority-class strategy, that is, the proportion of the largest class in the test data, and balanced accuracy, that is, the mean recall of the two classes. Because the labels come from ratings, all metrics are read as the ability of the model to reproduce those labels.

```
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import (accuracy_score, balanced_accuracy_score,
classification_report, confusion_matrix)
X, y = data_biner["teks_bersih"], data_biner["sentimen"]
X_train, X_test, y_train, y_test = train_test_split(
X, y, test_size=0.2, random_state=42, stratify=y)
model = Pipeline([
("tfidf", TfidfVectorizer(ngram_range=(1, 2), min_df=1,
max_df=0.95, max_features=10000)),
("naive_bayes", MultinomialNB(alpha=1.0)),
])
model.fit(X_train, y_train)
prediksi = model.predict(X_test)
print(accuracy_score(y_test, prediksi))
print(classification_report(y_test, prediksi,
labels=["Negatif", "Positif"], zero_division=0))
print(confusion_matrix(y_test, prediksi, labels=["Negatif", "Positif"]))
print(y_test.value_counts(normalize=True).max())
print(balanced_accuracy_score(y_test, prediksi))
```

Listing 1. Data Splitting, TF-IDF–Multinomial Naive Bayes Pipeline, and Evaluation.

3.5 Additional Tests: Cross-Validation, Class Imbalance, Significance, and Negation

The stability of the hold-out results was checked with stratified 5-fold cross-validation (shuffling enabled, random_state = 42) on all 410 reviews; accuracy, negative-class recall and precision, and balanced accuracy are reported as mean ± standard deviation across folds. Five scenarios were compared on the same folds: the majority-class strategy; the baseline Multinomial Naive Bayes (MNB) with $\alpha = 1.0$; MNB with class_prior $0.5 : 0.5$; Complement Naive Bayes (CNB); and the baseline with SMOTE ($k = 5$). In the SMOTE scenario, TF-IDF and SMOTE are inside the pipeline so that synthetic samples come only from the training data of each fold. All scenarios use the TF-IDF settings in Listing 1.

The difference in accuracy against the majority strategy was tested with a two-sided exact McNemar test [24] on the 82 hold-out test reviews and on the out-of-fold predictions for all 410 reviews; only discordant prediction pairs determine the p-value (binomial distribution with probability 0.5). Because the training data of different folds overlap, the out-of-fold results are read as indicative.

Negative reviews predicted as positive (false negatives) in the test data were examined one by one; quotations were shortened and cleared of links, mentions, and telephone numbers. The prevalence of negation words (“tidak”, “tak”, “bukan”, “belum”, “gak”, “ga”, “nggak”, “ngga”, “enggak”, “engga”, “jangan”, which are standard and colloquial Indonesian forms of “no”, “not”, “not yet”, and “do not”) was counted per class, and term weights were read from the difference between the log probabilities of the negative and positive classes. Two preprocessing variants were tested on the same cross-validation: tokens following a negation word were given the prefix NEG_ within a window of one or two words. The code for all of these tests is available in the companion notebook.

DOI: https://doi.org/10.59025/jtds.v4i2.64 | https://jtech.onlinelibrary.id/index.php/JTDS

3.6 Multi-label Complaint Mapping

Mapping was performed on the 139 negative reviews. The first stage was reading the reviews to find the objects of complaint. Six themes emerged from this reading: application problems; bus punctuality and service timing; routes, stops, and information; staff attitude or professionalism; tickets, fares, and payment; and grievances and responses. The second stage was matching each review against a list of indicator words for each theme using regular expressions (Listing 2). One review may fall under more than one theme, so the percentages across themes do not add up to 100%. Because a word such as "aplikasi" (application) can refer to many things, the counts are read together with the review excerpts rather than as stand-alone figures.

```
TEMA = {
"Kendala aplikasi": r"\b(?:aplikasi|apk|login|akun|otp|error|eror|bug|lag|loading|"
r"crash|update|versi|peta|tracking|pelacakan|server|internet|sinyal)\b",
"Ketepatan bus dan waktu": r"\b(?:telat|terlambat|lama|menunggu|nunggu|dilewati|"
r"melewati|dilewatin|ngetem|jadwal|molor|gak datang|tidak datang)\b",
"Rute, halte, dan informasi": r"\b(?:rute|halte|jalur|trayek|alamat|informasi|"
r"info|lengkap|tujuan)\b",
"Sikap/profesionalitas petugas": r"\b(?:petugas|kondektur\w*|sopir|supir|driver|"
r"judes|galak|ketus|kasar|ramah|sopan|profesional)\b",
"Tiket/tarif/pembayaran": r"\b(?:tiket|tarif|bayar|pembayaran|saldo|topup|"
r"top up|qris|kartu|harga|refund)\b",
"Pengaduan dan respons": r"\b(?:aduan|pengaduan|komplain|laporan|lapor|respon|"
r"respons|tanggapan|keluhan)\b",
}
negatif = data_biner[data_biner["sentimen"] == "Negatif"].copy()
for tema, pola in TEMA.items():
negatif[tema] = negatif["teks_bersih"].str.contains(pola, regex=True)
rekap = (negatif[list(TEMA)].sum().sort_values(ascending=False)
.rename("Jumlah ulasan").to_frame())
rekap["Persentase (%)"] = (rekap["Jumlah ulasan"] / len(negatif) * 100).round(1)
print(rekap)
print("Rata-rata tema per ulasan:", negatif[list(TEMA)].sum(axis=1).mean().round(2))
```

Listing 2. Indicator-Word-Based Multi-Label Complaint Theme Mapping.

The mapping was validated on a 20% random sample of negative reviews (28 of 139; random_state = 42) coded by the second author. Because the second author also helped compile the indicators and knew the earlier labeling results (non-blinded), the resulting kappa represents the level of agreement between the indicators and the researcher's judgment, not independent inter-rater reliability. The coding marked the presence or absence of each theme; Cohen's kappa [28] was computed per theme, averaged, and interpreted according to the criteria of Landis and Koch [29].

3.7 Ethics and Data Limitations

The data analyzed are public reviews; account identities were not used, results are presented in aggregate, and review excerpts are shortened. This study draws no conclusions about the truth of any allegation or the technical cause of any disruption. Because reviews can be deleted or changed by users, collecting them again at another time may yield slightly different figures.

4. RESULTS AND DISCUSSION

4.1 Rating and Sentiment Distribution

Rating 5 dominates with 232 reviews (52.37%), followed by rating 1 with 110 reviews (24.83%). Ratings 4, 3, and 2 account for 41 (9.26%), 31 (7.00%), and 29 (6.55%) reviews, respectively (Table I). The mean rating of 3.58 sounds reasonably good, but that figure hides a U-shaped distribution: users tend to give

either full stars or one star, and rarely anything in between.

TABLE I
Distribution of Ratings and Sentiment Labels

Rating	Number of reviews	Percentage	Label
1	110	24.83%	Negative
2	29	6.55%	Negative
3	31	7.00%	Neutral (excluded)
4	41	9.26%	Positive
5	232	52.37%	Positive
Total	**443**	**100.00%**	410 binary reviews

Under the chosen labeling rule, the binary corpus consists of 271 positive reviews (66.10%) and 139 negative reviews (33.90%), a ratio of about 1.95 : 1. The critical group is not small: 139 reviews equal 31.38% of the whole corpus, while 31 reviews with rating 3 are not included in the classification.

4.2 Classification Performance

The training data contain 328 reviews (217 positive and 111 negative) and the test data 82 reviews (54 positive and 28 negative). The model obtained an accuracy of 70.73%. The details are in Table II: of the 28 negative reviews in the test data, only 4 were predicted negative and 24 passed as positive, whereas all 54 positive reviews were predicted correctly.

TABLE II
Evaluation Metrics of Multinomial Naive Bayes on 82 Test Reviews

Class	Precision	Recall	F1-score	Support
Negative	1.00	0.14	0.25	28
Positive	0.69	1.00	0.82	54
Macro average	0.85	0.57	0.53	82
Weighted average	0.80	0.71	0.62	82

The figure of 70.73% needs to be read together with its benchmark: guessing "positive" for every review already yields 54/82 or 65.85%, so the difference is only 4 reviews (4.88 points), and a balanced accuracy of 57.14% is close to random guessing (50%). The model labeled only 4 of the 82 test reviews (4.9%) as negative although 28 (34.1%) carry a negative label; a negative-class precision of 1.00 is paid for with a recall of 0.14, meaning six out of seven criticisms go undetected.

Three causes can be considered. First, the negative class makes up only one third of the training data, so the model leans toward the majority class, as cautioned by other research [19]. Second, negative reviews are often narratives of incidents without evaluative words. Third, rating labels do not always agree with the text [10, 14], so some of the model's "errors" may be label inaccuracies. The second and third conjectures cannot yet be separated without manual annotation, whereas the first is tested in Section 4.3.

Comparison with other studies requires caution: Naive Bayes on Teman Bus reviews reached an accuracy of 82% and an SVM with a radial basis function (RBF) kernel 85% [5], whereas on iPusnas its accuracy was 58% with a recall of 81% [8]. The corpora, labeling rules, and data splitting schemes differ, so differences in figures are not differences in model quality; what is shared is the pattern that a single accuracy figure is not enough to judge the usefulness of a model, consistent with Naive Bayes on Access by KAI reviews, which lost in accuracy but was superior in recall [6].

4.3 Five-Fold Cross-Validation and Class Imbalance Handling

Table III presents the stratified 5-fold cross-validation on 410 reviews. The baseline accuracy (71.46 ± 1.24%) is almost the same as the hold-out figure (70.73%), and its negative-class recall (17.20 ± 6.06%) is low in every fold (7.4%–25.0%), so the low recall on the hold-out set is not the result of one unlucky split.

TABLE III
Results of Stratified 5-Fold Cross-Validation (Mean ± Standard Deviation, %)

Scenario	Accuracy	Recall neg.	Precision neg.	Balanced acc.
Majority-class strategy	66.10 ± 0.49	0.00 ± 0.00	0.00 ± 0.00	50.00 ± 0.00
MNB (baseline)	71.46 ± 1.24	17.20 ± 6.06	94.64 ± 6.59	58.23 ± 2.67
MNB + class_prior 50:50	82.68 ± 2.10	66.85 ± 6.47	79.72 ± 7.31	78.82 ± 2.40
Complement NB	82.68 ± 2.10	66.85 ± 6.47	79.72 ± 7.31	78.82 ± 2.40
MNB + SMOTE	82.93 ± 2.78	90.61 ± 5.90	69.08 ± 4.21	84.79 ± 2.90

Balanced priors raise negative-class recall to 66.85% (+49.65 points) and balanced accuracy to 78.82%; accuracy also rises (82.68%) because the largest error of the baseline is missed negative reviews, at the cost of negative-class precision falling from 94.64% to 79.72%. CNB is identical to balanced priors on every prediction because the two approaches are equivalent for two classes. SMOTE gives the highest negative-class recall (90.61%) with the lowest negative-class precision (69.08%). Of the three conjectures in Section 4.2, the first (the prior favors the positive class because the negative class is smaller) is best supported; the choice of scenario depends on the cost of missed critical reviews relative to positive reviews wrongly flagged.

4.4 McNemar Test against the Majority-Class Strategy

On the hold-out set, the model and the majority strategy differ on only 4 reviews, all of them negative reviews recognized by the model (a negative-class precision of 1.00 means there are none in the opposite direction). The two-sided exact McNemar test gives $p = 2 \times 0.5^4 = 0.125$, so the 4.88-point difference is not significant; at least six pairs in the same direction would be needed for $p < 0.05$. On the out-of-fold predictions for 410 reviews, the baseline differs from the majority strategy (24 correct only by the baseline, 2 only by the majority; $p < 0.001$), and balanced priors and CNB (69 versus 23; $p < 0.001$) as well as SMOTE (102 versus 55; $p = 0.0002$) differ from the baseline. The advantage of the baseline over the majority remains small (5.36 accuracy points), and the out-of-fold p-values are only indicative.

4.5 Frequent Words in Negative Reviews

In the 139 negative reviews, function words dominate the frequencies: “di” (at/in; 66), “tidak” (not; 44), and “ada” (there is; 36). The prominent content words are “aplikasi” (application; 34), “bus” (28), “halte” (bus stop; 27), “penumpang” (passenger; 19), “naik” (board/ride; 19), and “bis” (colloquial spelling of bus; 18); when “bus” and “bis” are combined (46), vehicle vocabulary appears more often than “aplikasi”. Criticism on the application page is therefore not solely about the software, although the word “aplikasi” may refer to login, maps, or bus information, so reading the context is still necessary.

4.5.1 False Negative Examples and the Role of Negation Words

Of the 28 negative reviews in the test data, 24 were predicted positive. Negation words (listed in Section 3.5) occur in 66 of the 139 negative reviews (47.5%) but in only 38 of the 271 positive reviews (14.0%), and the model captures this signal: the log ratios of “tidak” (not), “tidak bisa” (cannot), and “tidak ada” (there is no) are +0.43, +0.69, and +0.28, respectively (leaning negative). That signal is

DOI: https://doi.org/10.59025/jtds.v4i2.64 | https://jtech.onlinelibrary.id/index.php/JTDS

outweighed by the class prior (−0.67 in log terms; the positive class is 1.95 times more frequent), so that 12 of the 24 false negatives (50%) actually contain negation. Table IV presents four quotations (anonymized and shortened, with English translations) together with their suspected causes based on the model's term weights.

TABLE IV
False Negative Quotations from the Test Data (Shortened and Anonymized)

No.	Review quotation (rating 1–2)	Suspected cause
1	"Kernet dan supir tidak ramah" *(The conductor and driver are not friendly)*	"ramah" (friendly) leans positive (−0.68); the bigram "tidak ramah" is only −0.13, so negation does not flip the weight
2	"Tidak bisa di download" *(Cannot be downloaded)*	Three words; "tidak bisa" (+0.69) is offset by "bisa" (can) (−0.50) and the positive prior (−0.67)
3	"aplikasi sudah bagus, tapi pelayanan nya kurang memuaskan . . . " *(The app is already good, but the service is less than satisfactory . . .)*	The strong positive word "bagus" (good) (−2.26) is mixed with criticism
4	"PELAYANAN YANG SANGAT BURUK" *(VERY BAD SERVICE)*	"sangat" (very; −2.37) outweighs "buruk" (bad; +0.50; only 4 occurrences in the training data)

4.5.2 Effect of Negation Marking in Preprocessing

Negation marking with a one-word window raises negative-class recall by 3.59 points (17.20% to 20.79 ± 5.10%), accuracy by 0.98 points (72.44 ± 0.98%), and balanced accuracy by 1.61 points (59.84 ± 2.24%); a two-word window is similar (recall 20.03 ± 8.28%). The increase is smaller than the standard deviation of the baseline recall (6.06 points), so it cannot yet be distinguished from variation between folds. Negation is therefore not the main cause of missed negative reviews; the class prior is more decisive (+49.65 points).

4.6 Complaint Theme Mapping

The results of the multi-label mapping of the 139 negative reviews are presented in Table V. The total number of theme labels is 155, or an average of 1.12 themes per review. A total of 34 reviews (24.5%) contain two or more themes, for example when the application fails to display the bus position and the user ends up waiting a long time, whereas 31 reviews (22.3%) are not yet covered by any indicator (Section 4.6.1).

Application problems are the most frequent theme (51 reviews; 36.7%), covering difficulty logging into accounts, a slow or unresponsive application, and trouble connecting to the internet; this theme is still a large group that needs to be broken down by interaction stage. Routes, stops, and information (32; 23.0%) relate to routes perceived as incomplete or destinations that cannot be found. Bus punctuality and service timing (27; 19.4%) include buses that are late or pass by waiting passengers; the boundary between application and operations is blurred here, since a wrong bus position on screen may originate from Global Positioning System (GPS) updates, route changes, or congestion, and similar patterns were reported for Teman Bus reviews [5] and rail services in India [4]. Without GPS logs or official schedules, the causes of delays are not concluded. Staff attitude (24; 17.3%) indicates that some bad experiences

TABLE V
Frequency of Complaint Themes in 139 Negative Reviews (Multi-label)

Rank	Complaint theme	Number of reviews (% of 139)
1	Application problems	51 (36.7%)
2	Routes, stops, and information	32 (23.0%)
3	Bus punctuality and service timing	27 (19.4%)
4	Staff attitude/professionalism	24 (17.3%)
5	Tickets, fares, and payment	16 (11.5%)
6	Grievances and responses	5 (3.6%)

occur off screen [4]; tickets, fares, and payment (16; 11.5%) concern money and trust; grievances and responses (5; 3.6%) are the rarest.

4.6.1 Distribution of the Number of Themes per Review

The average is 1.12 themes per review (155 labels over 139 reviews): 53.2% of reviews contain one theme (74), 16.5% two themes (23), 6.5% three themes (9), 1.4% four themes (2), and 22.3% (31 reviews) are not yet covered by any indicator. The indicators therefore need to be expanded, and the counts in Table V are read as lower bounds.

4.6.2 Validation of the Mapping against Researcher Coding

A total of 28 reviews were coded by an author (results in Table VI). Agreement ranges from 78.6% to 96.4% (mean 88.1%) with a mean kappa of 0.50 (moderate; highest for application problems at 0.85, lowest for routes, stops, and information at 0.28). The indicators under-capture staff attitude (3 versus 8 reviews according to researcher coding) because they do not cover affixed forms and attached words such as “petugasnya” (the officer), “supirnya” (the driver), and “kepetugas” (to the officer), nor reviews that describe driver behavior without a keyword. Conversely, the indicator for routes, stops, and information is too loose (6 versus 4 reviews) because the words “halte” (bus stop) and “lengkap” (complete) also capture reviews that merely mention a stop as a location or whose meaning is unclear. The grievances and responses theme was not flagged by the indicators in this sample (two reviews according to researcher coding; kappa 0.00), and because the theme appears in only 5 of 139 reviews, its kappa is unstable.

TABLE VI
Cohen’s Kappa of Theme Mapping on a Sample of 28 Negative Reviews (Researcher Coding vs Indicators)

Theme	Themed reviews (researcher / indicator)	Agreement (%)	κ	Interpretation
Application problems	11/13	92.9	0.85	Almost perfect
Bus punctuality and service timing	6/6	85.7	0.58	Moderate
Routes, stops, and information	4/6	78.6	0.28	Fair
Staff attitude/professionalism	8/3	82.1	0.46	Moderate
Tickets, fares, and payment	3/4	96.4	0.84	Almost perfect
Grievances and responses	2/0	92.9	0.00	Slight
Mean	–	88.1	0.50	Moderate

DOI: https://doi.org/10.59025/jtds.v4i2.64 | https://jtech.onlinelibrary.id/index.php/JTDS

4.7 Implications for Service Management

The complaint pattern points to three inspection agendas (Table VII): the application team traces the account login flow, stability, and consistency of stop and route data; the BRT operator compares punctuality complaints with schedules, GPS logs, and records of buses passing stops; and complaints about staff and grievances go into field service evaluation and the grievance channel. This order is a verification agenda based on theme frequency, not a risk ranking, and the division of responsibilities is a proposal that needs to be confirmed with the service managers. The baseline (negative-class recall 14.29%) is not yet fit to serve as an automatic filter; balanced priors or SMOTE are more worth considering, with the consequence that more positive reviews will also be flagged.

TABLE VII
Findings, Responsible Parties, and Verification Follow-up (Proposed)

Finding (number of reviews)	**Responsible party**	**Verification follow-up**
Application problems (51; 36.7%)	Application development team	Trace login failures, load time, and version stability.
Routes, stops, and information (32; 23.0%)	Route planners and application content team	Check completeness of route and stop data.
Bus punctuality and service timing (27; 19.4%)	BRT operator and fleet data manager	Compare with schedules and GPS logs.
Staff attitude/professionalism (24; 17.3%)	BRT operator (staff development)	Evaluate field service and training.
Tickets, fares, and payment (16; 11.5%)	Ticketing manager and application team	Reconcile payment status and transactions.
Grievances and responses (5; 3.6%)	Grievance channel manager	Show follow-up status to users.

4.8 Limitations

This study has five limitations. The labels come from ratings and have not been fully validated by human annotators; the theme mapping was validated through cross-checking on a 20% random sample coded directly by an author using a structured and explicit codebook to maintain consistency and objectivity of interpretation. The corpus of 443 reviews and hold-out test set of 82 reviews make per-class metrics sensitive; cross-validation reduces, but does not remove, that dependence. Imbalance handling was limited to class_prior, CNB, and SMOTE (class-weighted models, IndoBERT, and threshold tuning have not been tested). The theme indicators were compiled by the researchers, 22.3% of negative reviews are not yet covered by any indicator, and agreement was checked on only 28 reviews. Finally, the reviews come from users who chose to write and therefore do not represent all Trans Jateng BRT passengers, and the out-of-fold p-values are indicative because the training data of different folds overlap.

5. CONCLUSION

This study analyzed 443 Google Play Store reviews of the Si Anteng application; after reviews with rating 3 were excluded, 410 reviews (271 positive, 139 negative) were used for two-class classification. The unigram–bigram TF-IDF and Multinomial Naive Bayes pipeline produced an accuracy of 70.73% on 82 test reviews, only 4.88 points above the majority-class guessing strategy, and the difference was not significant in the exact McNemar test ($p = 0.125$). A negative-class recall of 14.29% and a balanced accuracy of 57.14% show that the model is not yet able to filter user criticism. In five-fold cross-validation, the negative-class recall of the baseline was 17.20 ± 6.06%; balanced class priors raised it to 66.85 ± 6.47% (identical to Complement Naive Bayes for two classes) and SMOTE to 90.61 ± 5.90%, at

DOI: https://doi.org/10.59025/jtds.v4i2.64 | https://jtech.onlinelibrary.id/index.php/JTDS

the cost of negative-class precision falling from 94.64% to 79.72% and 69.08%; negation marking added only about 3.6 recall points and could not be distinguished from variation between folds. Multi-label mapping of the 139 negative reviews placed application problems first (36.7%), followed by routes, stops, and travel information (23.0%) and bus punctuality and service timing (19.4%); staff attitude (17.3%) indicates that some complaints lie beyond the reach of software fixes. The evaluation of Si Anteng therefore needs to connect application quality with Trans Jateng BRT operations.

Future research is advised to produce manual annotations by at least two annotators together with their agreement measure, test class-weighted models and decision threshold tuning on a larger corpus, compare IndoBERT-based models, normalize non-standard language, and link complaint themes with schedules, GPS logs, and grievance records so that suspected causes can be tested.

AUTHOR CONTRIBUTIONS

Moch. Nazril Ilham: Conceptualization, Methodology, Data Collection, Data Curation, Data Processing, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing. M. Ichwan Al Hafizd: Data Collection, Formal Analysis. Zahra Zahrani: Investigation. Laelatul Hidayah: Data Collection. Muhammad Na'afian Anugerah: Investigation, Writing – Review & Editing. Imam Prayogo Pujiono: Supervision, Guidance, Validation, Writing – Review & Editing. All authors have read and approved the final version of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest in this research.

DECLARATION ON THE USE OF ASSISTIVE TOOLS

During the research and manuscript preparation, the authors used Google Colab to write code and run tests, and used Grammarly for language editing. The authors have reviewed and verified all outputs from these tools and take full responsibility for the accuracy and content of the manuscript.

REFERENCES

[1] E. C. M. Torres and L. G. de Picado-Santos, "Sentiment analysis and topic modeling in transportation: A literature review," *Applied Sciences*, vol. 15, no. 12, p. 6576, Jun. 2025.

[2] E. C. M. Torres and L. G. de Picado-Santos, "Using sentiment analysis to study the potential for improving sustainable mobility in university campuses," *Sustainability*, vol. 17, no. 14, p. 6645, Jul. 2025.

[3] M. M. Rahimi, E. Naghizade, M. Stevenson, and S. Winter, "SentiHawkes: a sentiment-aware Hawkes point process to model service quality of public transport using Twitter data," *Public Transport*, vol. 15, no. 2, pp. 343–376, Jun. 2023.

[4] D. N. Mishra and R. K. Panda, "Decoding customer experiences in rail transport service: application of hybrid sentiment analysis," *Public Transport*, vol. 15, no. 1, pp. 31–60, Mar. 2023.

[5] F. A. Soraya and A. D. Indriyanti, "Perbandingan algoritma Naive Bayes dan Support Vector Machine dalam analisis sentimen aplikasi Teman Bus," *Journal of Emerging Information Systems and Business Intelligence (JEISBI)*, vol. 5, no. 2, pp. 118–125, Jun. 2024.

[6] R. Damanhuri and V. A. Husein, "Analisis sentimen pada ulasan aplikasi Access by KAI berbahasa Indonesia menggunakan word-embedding dan classical machine learning," *Jurnal Masyarakat Informatika*, vol. 15, no. 2, pp. 97–106, Nov. 2024.

[7] N. Agustina, D. H. Citra, W. Purnama, C. Nisa, and A. R. Kurnia, "Implementasi algoritma Naive Bayes untuk analisis sentimen ulasan Shopee pada Google Play Store," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 1, pp. 47–54, Apr. 2022.

[8] H. N. Zuhdi and B. Prasetiyo, "Analisis sentimen pada ulasan aplikasi iPusnas di Google Play Store menggunakan Naive Bayes Classifier," *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE)*, vol. 5, no. 1, pp. 12–19, Mar. 2025.

[9] D. S. Nurrochmah, N. Rahaningsih, R. D. Dana, and C. L. Rohmat, "Penerapan algoritma Naive Bayes dalam analisis sentimen ulasan aplikasi KitaLulus di Google Play Store," *Jurnal Informatika Terpadu*, vol. 11, no. 1, pp. 1–11, Mar. 2025.

[10] Y. Asri, D. Kuswardani, W. N. Suliyanti, Y. O. Manullang, and A. R. Ansyari, "Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 38, no. 1, pp. 677–688, Apr. 2025.

[11] D. Ananda, I. Budi, A. B. Santoso, and A. A. Qureshi, "Sentiment analysis of public health app reviews using IndoBERT and XLM-RoBERTa: A study on SATUSEHAT mobile app," *JIKO (Jurnal Informatika dan Komputer)*, vol. 8, no. 3, pp. 161–172, Nov. 2025.

[12] R. Riccosan and K. E. Saputra, "Multilabel classification sentiment analysis on Indonesian mobile app reviews," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, no. 5, pp. 4226–4234, Oct. 2025.

[13] P. H. C. Samanmali and R. A. H. M. Rupasingha, "Sentiment analysis on Google Play Store app users' reviews based on deep learning approach," *Multimedia Tools and Applications*, vol. 83, no. 36, pp. 84 425–84 453, 2024.

[14] S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Discrepancy detection between actual user reviews and numeric ratings of Google App Store using deep learning," *Expert Systems with Applications*, vol. 181, p. 115111, Nov. 2021.

[15] S. A. H. Bahtiar, C. K. Dewa, and A. Luthfi, "Comparison of Naive Bayes and logistic regression in sentiment analysis on marketplace reviews using rating-based labeling," *Journal of Information Systems and Informatics*, vol. 5, no. 3, pp. 915–927, Aug. 2023.

[16] W. L. Azzahra, J. Indra, R. Rahmat, and S. Faisal, "Sentiment analysis of user reviews of the AdaKami online loan app from the app store using SVM and Naive Bayes," *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 838–850, Jun. 2025.

[17] L. Varghese, R. R. Pai, G. Savitha, S. Girisha, and N. Chetty, "Manual annotation based sentiment analysis of user feedback in health and wellness app reviews," *Scientific Reports*, vol. 15, no. 1, 2025.

[18] R. Kusumaningrum, I. Z. Nisa, R. P. Nawangsari, and A. Wibowo, "Sentiment analysis of Indonesian hotel reviews: from classical machine learning to deep learning," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 3, p. 292, Nov. 2021.

[19] Badriyah, T. Chamidy, and Suhartono, "Application of SMOTE in sentiment analysis of MyXL user reviews on Google Play Store," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 10, no. 1, pp. 74–86, Jan. 2025.

DOI: https://doi.org/10.59025/jtds.v4i2.64 | https://jtech.onlinelibrary.id/index.php/JTDS

[20] E. M. Wicaksana and N. Rijati, "Analyzing sentiment of SiCepat Express user reviews," *Journal of Applied Informatics and Computing*, vol. 9, no. 1, pp. 235–240, Jan. 2025.

[21] M. K. Ridwan, Y. Irawan, and R. R. Setiawan, "Mapping digital sentiment landscapes of hotel reviews: A machine learning-based cross-platform analysis," *Edumatic: Jurnal Pendidikan Informatika*, vol. 10, no. 1, pp. 110–119, Apr. 2025.

[22] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of Naive Bayes text classifiers," in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, Washington, DC, USA, 2003, pp. 616–623.

[23] R. F. Shollis and U. P. Sanjaya, "Comparison of LinearSVC and Complement Naive Bayes on Coretax sentiment analysis using IndoBERT pseudo-labeling and ADASYN," *JIKO (Jurnal Informatika dan Komputer)*, vol. 9, no. 2, pp. 100–107, Jul. 2026.

[24] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.

[25] E. Rasywir, Y. Pratama, I. Irawan, and M. Istoningtyas, "Public complaint tweet data feature analysis for sentiment classification," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 6, pp. 4366–4377, 2024.

[26] F. W. Prabowo, "Faktor penyebab rating rendah perpustakaan digital iPusnas: Analisis konten ulasan di Google Play Store," *Anuva: Jurnal Kajian Budaya, Perpustakaan, dan Informasi*, vol. 9, no. 2, pp. 367–377, Jun. 2025.

[27] N. Yusliani, M. D. Marieska, E. Lestari, M. R. P. Sulfa, and W. Arimurti, "The development of Indonesian sentiment analysis with negation handling," in *Proc. Sriwijaya Int. Conf. Inf. Technol. Appl. (SICONIAN 2019)*. Paris, France: Atlantis Press, 2020, pp. 67–72.

[28] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[29] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977.