P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v7i4.386 Vol. 7, No. 4. September 2025
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Active Learning Query by Committee Labeling Method to Increase Accuracy and Efficiency of Sentiment Analysis Classification

Dipa Anasta Iskandar[-1], R. Mohamad Atok[-2]

Technology Management
Sepuluh Nopember Institute of Technology
6032231085@student.its.ac.id[-1], moh_atok@statistika.its.ac.id[-2]

Abstract

This study proposes the Query by Committee (QBC) labeling method to improve the accuracy of classification models—specifically XLM-RoBERTa—and to increase labeling efficiency compared to manual, supervised labeling, which generally requires more time and resources. The dataset consists of unannotated healthcare-industry application reviews scraped from Google Play. Six distinct labeling strategies were applied as input for fine-tuning XLM-RoBERTa models under identical hyperparameter settings. The six labeling approaches were evaluated namely Rating-based labeling, Lexicon-based labeling, QBC for Rating-Vader labeling, QBC for Rating-Pseudo labeling, QBC for Vader-Pseudo labeling, and QBC triplet for Rating-Pseudo-Vader labeling. Each labeled dataset was split using stratified random sampling, and class weights were set to "auto" during training to address label imbalance. All models were subsequently tested on the IndoNLU SmSA test dataset, with performance compared in terms of accuracy, precision, recall, and F1-score. Results indicate that the triplet QBC approach (combining Rating, VADER, and Pseudo labeling) outperformed all other methods, achieving an accuracy of 91.4%, a precision of 91.28%, a recall of 91.4%, and an F1-score of 91.21%. These findings demonstrate that the QBC labeling method can serve as an effective and efficient alternative to manual annotation for similar classification tasks.

Keywords: Sentiment Analysis; Labeling Method; Query by Committee; Active Learning;

Abstrak

Penelitian ini mengusulkan metode pelabelan Query by Committee (QBC) untuk meningkatkan akurasi model klasifikasi—khususnya XLM-RoBERTa—dan untuk meningkatkan efisiensi pelabelan dibandingkan dengan pelabelan manual tersupervisi, yang umumnya membutuhkan lebih banyak waktu dan sumber daya. Dataset terdiri dari ulasan aplikasi industri kesehatan yang belum dianotasi yang diambil dari Google Play. Enam strategi pelabelan yang berbeda diterapkan sebagai masukan untuk fine-tuning model XLM-RoBERTa di bawah pengaturan hyperparameter yang identik. Enam pendekatan pelabelan dievaluasi, yaitu pelabelan berbasis Rating, pelabelan berbasis Leksikon, QBC untuk pelabelan Rating-Vader, QBC untuk pelabelan Rating-Pseudo, QBC untuk pelabelan Vader-Pseudo, dan triplet QBC untuk pelabelan Rating-Pseudo-Vader. Setiap dataset yang telah dilabeli dibagi menggunakan stratified random sampling, dan bobot kelas diatur ke "auto" selama pelatihan untuk mengatasi ketidakseimbangan label. Semua model kemudian diuji pada dataset uji IndoNLU SmSA, dengan perbandingan kinerja dalam hal akurasi, presisi, recall, dan F1-score. Hasil menunjukkan bahwa pendekatan QBC triplet (menggabungkan pelabelan Rating, VADER, dan Pseudo) mengungguli semua metode lain, mencapai akurasi 91,4%, presisi 91,28%, recall 91,4%, dan F1-score 91,21%. Temuan ini menunjukkan bahwa metode pelabelan QBC dapat berfungsi sebagai alternatif yang efektif dan efisien untuk anotasi manual untuk tugas klasifikasi serupa.

Kata Kunci: Analisis Sentimen; Metode Pelabelan; Query by Committee; Pembelajaran Aktif;

INTRODUCTION

Classification models require labeled data for training and evaluation. The most prevalent approach is supervised learning, wherein each training instance is manually annotated by human experts. In this paradigm, humans serve as annotators who assign labels or categories to each data sample (Lu, Song, Arachie, & Huang, 2025; Zhao, Hong, Yang, Zhao, & Ding, 2023). However, a

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i4.386
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

significant limitation of this approach is that acquiring manually labeled data demands substantial time and resources, particularly when the dataset is large.

One of the simplest and fastest ways to obtain sentiment labels for Google Play reviews is via Distant Supervision (Zhang & Cao, 2023), wherein the numerical review score (ranging from 1 to 5 stars) is treated as a proxy for sentiment polarity. Although this method is efficient, it often produces labels with high noise levels (Xu & Guo, 2021; Zhang & Cao, 2023), since there is frequently a mismatch between the textual content of a review and the numerical rating assigned by the user. For instance, a reviewer might write "I love this app" but still assign a one-star rating, or conversely, write a negative comment and assign a five-star rating (Hou et al., 2024). Another commonly used approach for obtaining sentiment labels is lexicon-based scoring, such as VADER (Abiola, Abayomi-Alli, Tale, Misra, & Abayomi-Alli, 2023; Barik & Misra, 2024; Budianto, Wirjodirdjo, Maflahah, & Kurnianingtyas, 2022; Isnan, Elwirehardja, & Pardamean, 2023; Ruhyana -, Salsabila Dwi Irmanti -, Agung Riyadi -, & Mardiana -, 2025) in which each word in the text is assigned a sentiment weight from a predefined lexicon. These individual word scores are then aggregated—taking into account linguistic features such as negation, intensity modifiers, punctuation, and capitalization—to calculate an overall sentiment score for the entire review. Finally, the composite score is compared against predetermined thresholds (e.g., >0.05 for positive, <–0.05 for negative, and between –0.05 and 0.05 for neutral) to generate a discrete sentiment label for each review.

Previous studies that investigate bias in Google Play reviews (Aljrees et al., 2024; Sadiq et al., 2021) confirm the existence of such inconsistencies; however, they focus primarily on predicting whether a review is biased or unbiased using unsupervised approaches (e.g., TextBlob) and Deep Learning techniques. In contrast, the present study aims to reduce noise and bias in sentiment labels while maintaining higher accuracy than Distant Supervision and greater labeling efficiency than fully supervised methods, motivated by Query by Committee Active Learning study on machine learning models (Wang, Wan, & Zhang, 2019).

To achieve this goal, we implemented six distinct labeling strategies: two convenience approaches—rating-based and lexicon-based (VADER)—and four Query by Committee (QBC) configurations that combine rating, VADER, and pseudo-labels produced by a pre-trained RoBERTa classifier. For each labeling method, the dataset was partitioned using stratified random sampling, and class weights were set to "auto" during training to address label imbalance. We then fine-tuned an XLM-RoBERTa (XLM-R) model under identical hyperparameter settings for each labeled dataset. Model performance was evaluated on the IndoNLU SmSA test set (Wilie et al., 2020) using accuracy, precision, recall, and F1-score metrics.

RESEARCH METHODS

Types of research

This study employs a quantitative research design with an experimental framework. Specifically, it evaluates the effectiveness of six distinct labeling methods—two convenience approaches (rating-based and lexicon-based) and four Query by Committee (QBC) configurations—by training and testing an XLM-RoBERTa (XLM-R) classification model. The primary objective is to compare classification performance (accuracy, precision, recall, and F1-score) across these labeling strategies, thereby assessing their impact on label noise and bias reduction.

Research Time and Location

The research was conducted in May 2025. Data preprocessing, labeling, model training, and evaluation activities were performed using private owned computational resources. All data collection (Google Play review scraping) occurred online via automated scripts; subsequent labeling, model fine-tuning, and testing were carried out in a virtual laboratory environment.

Research Target / Subject

This study employs a quantitative approach. The primary population consists of user reviews for K24Klik, a healthcare-related applications on the Google Play Store, written in either Indonesian or English. From this population, a total of 6,324 review instances were collected via automated web scraping, using a convenience sampling strategy (JoMingyu, n.d.) to include reviews that contain at least a text reviews and are associated with five-star or one-star ratings (to maximize polarity variance). After initial preprocessing, these reviews form the corpus for subsequent labeling and model training.

For the Query by Committee (QBC) configurations (Esuli & Sebastiani, 2009; Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos, Bobes-Bascarán, & Fernández-Leal, 2022), human annotators serve as "oracles." Oracles independently label only those review instances flagged as "disagreement" by the QBC committee. If

all committee members agree on a label for a review, that consensus label is accepted without oracle intervention.

Procedure

Data collection was performed by scraping user reviews from the K24Klik application on the Google Play Store using the Google Play Library (JoMingyu, n.d.). The raw text reviews were preprocessed to ensure consistency and remove noise: URLs were filtered out; emoticons and emojis were removed; all characters were converted to lowercase; and Indonesian-language normalization was applied using a custom dictionary of slang words and a colloquial Indonesian lexicon (Aliyah Salsabila, Ardhito Winatmoko, Akbar Septiandri, & Jamal, 2018).

To enable lexicon-based scoring with VADER—which is only available for English text—a new column was added to the dataset containing English-translated versions of the preprocessed reviews. Translation was performed using googletrans python library.

Once preprocessing and translation were complete, we generated sentiment labels using two convenience methods and four Query by Committee (QBC) configurations. The first convenience method mapped Google Play star ratings (1–5) directly to sentiment classes: ratings of 1–2 were labeled as negative, a rating of 3 as neutral, and ratings of 4–5 as positive. The second convenience method applied VADER scoring to the English-translated review text, assigning labels based on standard thresholds (compound score ≥ +0.05 for positive; ≤ −0.05 for negative; and between −0.05 and +0.05 for neutral).

In addition, we introduced a "buffer" label by running a RoBERTa model—fine-tuned for sentiment classification (Wilson Wongso, 2023) on the IndoNLU SmSA dataset—over each review. Similar pseudo-labeling strategies using transformer models have been shown to enhance downstream classification performance (Kuligowska & Kowalczuk, 2021). Although this RoBERTa output serves primarily as a supplementary label, we expected it to boost overall classification accuracy, particularly within the healthcare domain addressed in our study. By combining the RoBERTa "buffer" label with the rating-based and VADER labels, our QBC configurations leverage multiple perspectives to reduce noise and bias before deferring any remaining disagreements to a human oracle.

For each of the four QBC configurations—(1) Rating + VADER, (2) Rating + Pseudo, (3) VADER + Pseudo, and (4) Triplet (Rating + VADER + Pseudo)—we compared pairs (or triplets) of labels and delegated any reviews with inconsistent labels to an oracle. Specifically, if all committee members agreed on a label for a given review, that label was accepted without further action; if they disagreed, the review was assigned to the oracle for manual sentiment annotation. The oracle (a designated human annotator) reviewed only those flagged instances and provided the final corrected label.

After labeling was complete, each of the six labeled datasets (two convenience methods and four QBC outputs) was split using stratified random sampling to preserve class proportions. 80% of each dataset was allocated to training and 20% to validation. Class weights were set to "auto" provided by Tensorflow during model training to mitigate any residual class imbalance (Fernando & Tsokos, 2022).

Model fine-tuning used an XLM-RoBERTa (XLM-R) base model with identical hyperparameters across all experiments. Specifically, we set a maximum token length of 512, a batch size of 16, three epochs of training, a learning rate of 2×10^{-5}, and a weight decay of 0.1. The model architecture was xlm-roberta-base, with three output labels (negative, neutral, positive) as defined in LABEL_DICT = {"negative": 0, "neutral": 1, "positive": 2}. All training was conducted on an NVIDIA RTX 3060 Ti GPU, and the model checkpoint corresponding to the highest F1-score on the training set was saved for subsequent evaluation.

Finally, each fine-tuned XLM-R model was evaluated on the IndoNLU SmSA test dataset (Wilie et al., 2020) which comprises balanced Indonesian-language sentiment examples. We recorded accuracy, precision, recall, and F1-score for all three classes, thereby assessing each labeling method's impact on downstream classification performance.

RESULTS AND DISCUSSION

Data were collected using the Google Play Scraper library (JoMingyu, n.d.). The target application for this study is K24Klik (package ID: com.k24klik.android), and the extraction was performed on 28 April 2025. Initially, 6,356 raw reviews—each containing review text and its corresponding star rating—were retrieved. After preprocessing, 32 reviews consisting solely of emojis or emoticons were removed, yielding a final corpus of 6,324 reviews. A new column created to contain translated preprocessed review texts to further processed as an input for VADER label.

Rating-based labeling was performed according to the following criteria: reviews with a score of 1 or 2 stars were labeled as negative; a

score of 3 stars was labeled as neutral; and scores of 4 or 5 stars were labeled as positive. As shown in Table 1, the resulting distribution is heavily skewed toward positive labels, while neutral labels constitute the smallest category.

Table 1. Rating-based Labeling Counts

Sentiment	***Count of Rating-based***
negative	2522
neutral	241
positive	3560

VADER-based labeling was applied to the English-translated reviews using standard VADER thresholds (compound score ≥ 0.05 → positive; ≤ –0.05 → negative; otherwise neutral). Table 2 summarizes the resulting label distribution. Positive labels are the most frequent, while neutral labels outnumber negative labels.

Table 2. VADER-based Labeling Counts

Row Labels	***Count of layer1_input***
negative	1680
neutral	1427
positive	3216

For the Query by Committee configurations, we compared the labels produced by the Rating, VADER, and Pseudo strategies to identify both agreement and disagreement. Table 3 summarizes the number of reviews for which all committee members agreed versus those requiring manual verification by the oracle. Notably, the QBC triplet configuration exhibited the highest disagreement rate: out of 6,324 total reviews, 2,667 (42.18%) required manual review—substantially more than any other QBC variant where QBC Rating-Pseudo exhibited the lowest disagreement rate 1,162 (18,38%) out of the total reviews.

Table 3. Query by Committee Agreement – Disagreement on Each Method

Method	*Agreement*	*Disagreement*	*Ratio to Check Manually*
QBC_rating_vader	4161	2162	34,19%
QBC_rating_pseudo	5161	1162	18,38%
QBC_pseudo_vader	4032	2291	36,23%
QBC_triplet	3656	2667	42,18%

During the manual verification phase, all reviews flagged as "disagreement" by the committee were forwarded to the oracle for final labeling. Table 4 presents the resulting sentiment-label distribution for each QBC method after oracle adjustment.

Table 4. Final Label Counts on Each QBC Method

Method	*negative*	*neutral*	*positive*
QBC_rating_vader	2601	362	3360
QBC_rating_pseudo	2674	524	3125
QBC_pseudo_vader	2593	604	3126
QBC_triplet	2674	524	3125

All six labeled datasets were then used as input for XLM-RoBERTa training under the identical hyperparameter settings described in the Procedures section. Class weights were automatically calculated using TensorFlow's class-weight balancing feature to mitigate any remaining class imbalance. Table 5 presents the computed weight for each sentiment class across all labeling methods.

Table 5. Weight on Each Class on Each Method

Labeling Method	*negative*	*neutral*	*positive*
Rating Based	0.8357	8.7455	0.5920
Vader Based	1.2545	1.4769	0.6553
QBC Rating + Pseudo	0.7882	4.0222	0.6744
QBC Vader + Pseudo	0.8128	3.4895	0.6742
QBC Rating + Vader	0.8103	5.8222	0.6272
QBC Triplet	0.7882	4.0222	0.6744

Each labeled dataset served as input to a separate XLM-RoBERTa model. Training was performed over three epochs, and the checkpoint with the highest F1-score—regardless of epoch number—was saved as the final model. Figure 1 illustrates a comparison of the training performance for the best model obtained under each labeling method. Based on the validation F1 score, VADER based model is the wost model with score of 79,77% where the best model comes from QBC Rating + Vader with validation F1 score of 93,22%.

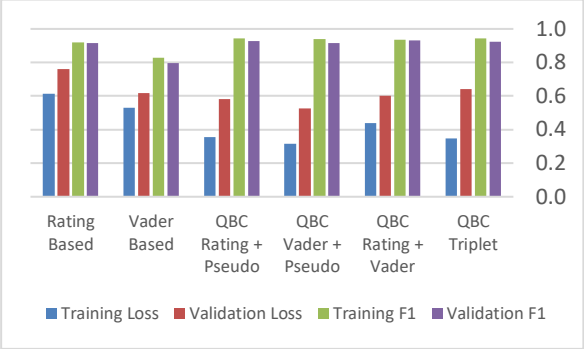

Figure 1. Training Performance for Each Labeling Method

In the testing phase, we evaluated each trained model using the IndoNLU SmSA test dataset to measure classification performance. Table 6 summarizes the accuracy, precision, recall, and F1-score for all six labeling methods.

Table 6. Test Result on Each Method.

Method	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Rating Based	0.8040	0.8353	0.8040	0.7522
Vader Based	0.8180	0.8141	0.8180	0.8072
QBC Rating + Pseudo	0.8960	0.8968	0.8960	0.8954
QBC Vader + Pseudo	0.9020	0.9035	0.9020	0.8980
QBC Rating + Vader	0.8780	0.8770	0.8780	0.8774
QBC Triplet	0.9140	0.9128	0.9140	0.9121

Among all methods, the QBC Triplet configuration achieved the highest scores across all metrics, with an accuracy of 91.40% and an F1-score of 91.21%. In contrast, the simple Rating-Based approach yielded the lowest performance, with an accuracy of 80.40% and an F1-score of 75.22%. These results confirm that combining multiple weak labelers within the QBC framework significantly improves classification performance and reduces label noise compared to convenience labeling alone.

To assess model performance in the presence of class imbalance, we plotted the Receiver Operating Characteristic (ROC) curves for each approach (Figures 2–7). Notably, the convenience-based methods (Rating-Based and VADER-Based) exhibit relatively unstable ROC curves, with AUC values hovering close to the diagonal line. In contrast, all QBC configurations demonstrate substantially improved discrimination: their ROC curves lie well above the AUC = 0.90 threshold, indicating consistently strong performance across imbalance-affected classes.

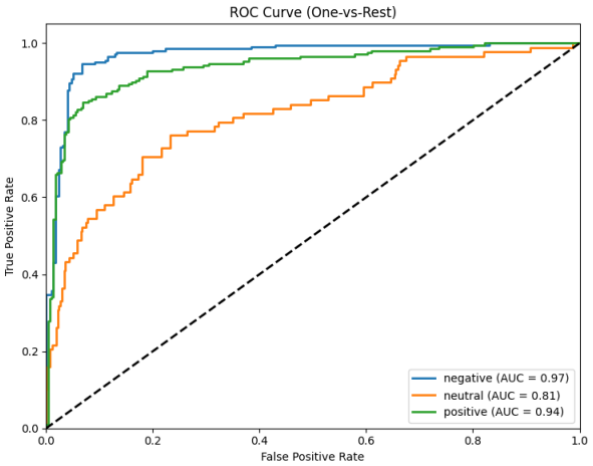

Figure 2. ROC Curve on Rating Based Model

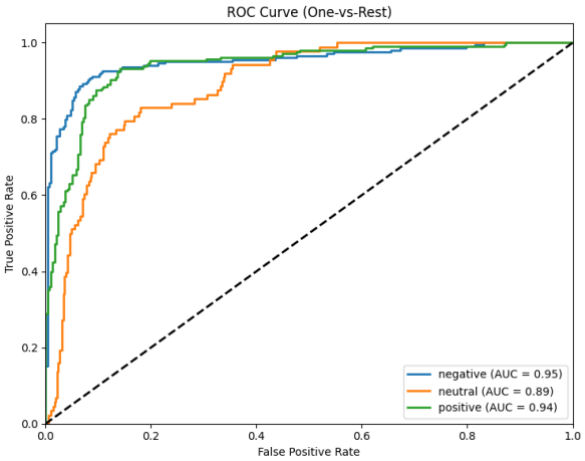

Figure 3. ROC Curve on Vader Based Model

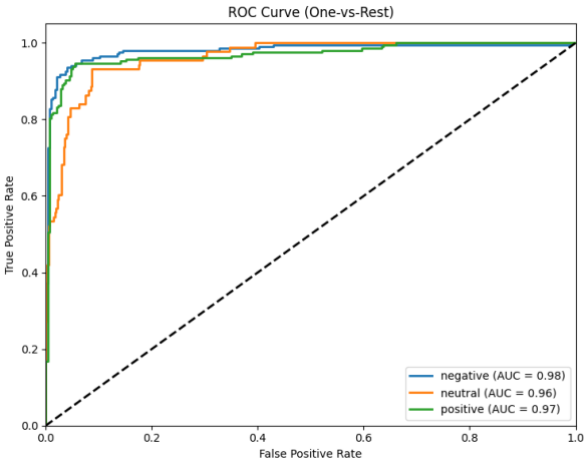

Figure 4. ROC Curve on QBC Rating-Pseudo Model

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v7i4.386
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

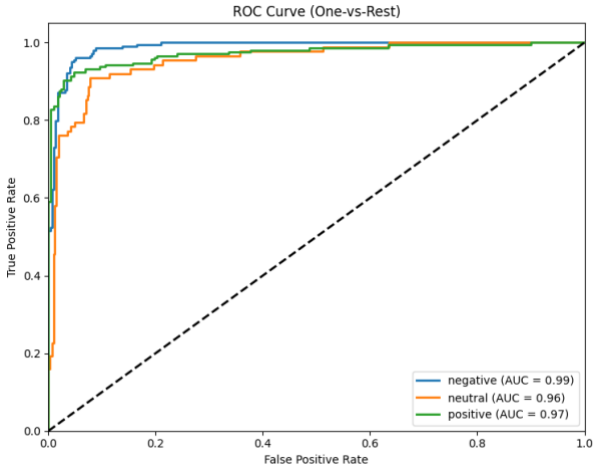

Figure 5. ROC Curve on QBC VADER-Pseudo Model

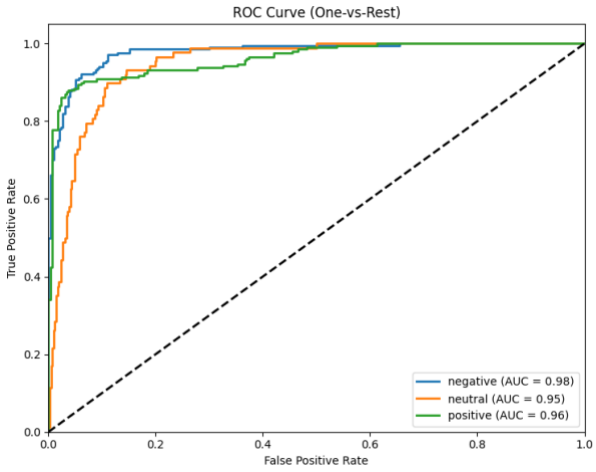

Figure 6. ROC Curve on QBC Rating-VADER Model

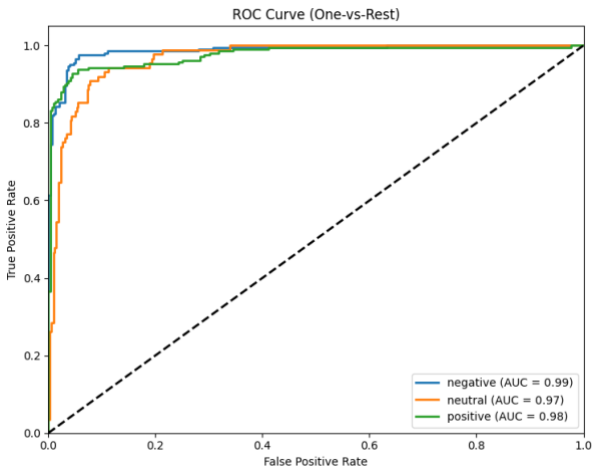

Figure 7. ROC Curve on QBC Triplet Model

Across all figures, the ROC curve for the neutral class remains below those of the negative and positive classes. We attribute this to the relatively small proportion of neutral examples in both the training and testing datasets, which likely limited the model's ability to learn and generalize for this minority class.

To investigate these discrepancies further, we conducted a focused analysis on the 2,667 reviews flagged by the QBC Triplet method—i.e., those instances where Rating, VADER, and Pseudo labels did not unanimously agree—and compared each against the oracle's manually verified label. Of these 2,667 "disagreement" cases, 669 reviews (25.1%) exhibited a mismatch between the original star rating and the oracle-assigned sentiment. In contrast, VADER failed to match the oracle label in 2,183 reviews (81.8% of the disagreement set). We attribute this high VADER misclassification rate primarily to translation errors during the preprocessing stage: subtle idiomatic expressions or colloquial phrases in the original Indonesian reviews were sometimes rendered inaccurately in English, leading VADER's lexicon-based scoring to produce incorrect polarity assignments.

Table 7 presents a small sample of reviews from the disagreement set, illustrating how Rating, VADER, and oracle labels differ in practice. For instance, in the first example below, the Indonesian-to-English translation ("Facilitating Anti-Complicated Fast") led VADER to assign a negative label—even though both the original rating (positive, 4–5 stars) and the oracle annotation concurred on a positive sentiment. Similarly, several other examples show how VADER's inability to handle translation artifacts skews its polarity score, or how a reviewer's star rating does not always align with their written intent.

Table 7. Reviews with Disagreement

Translated Review	***Rating label***	***Vader label***	***Actual***
Facilitating Anti - Complicated Fast	positive	negative	positive
The application service is bad	positive	negative	negative
really good only expensive postage	positive	negative	negative
the application cannot be opened error continues	negative	positive	negative
Not quite complete is also quite expensive	negative	neutral	negative

These findings underscore two key points: (1) star ratings alone can be noisy when users express sentiments that contradict their numeric score, and (2) lexicon-based methods like VADER—especially when applied to machine-translated text—may misinterpret contextual nuances, resulting in a disproportionately large number of neutral or incorrect labels. By using the QBC Triplet

framework to defer only these ambiguous cases to a human oracle, we effectively combine the speed of heuristic labeling with the accuracy of manual annotation.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study has demonstrated that employing a Query by Committee (QBC) framework—combining star-rating, VADER lexicon, and RoBERTa-based pseudo-label sources—significantly improves sentiment-classification performance for Indonesian healthcare-app reviews compared to convenience labeling alone. Among all methods evaluated, the QBC triplet configuration achieved the highest accuracy (91.40%) and F1-score (91.21%) on the IndoNLU SmSA test set, outperforming both rating-based (80.40% accuracy, 75.22% F1) and VADER-based (81.80% accuracy, 80.72% F1) approaches. ROC-curve analysis further confirmed that QBC models consistently exceed an AUC of 0.90 for positive and negative classes, whereas convenience methods often remain close to the diagonal line. Disagreement-case analysis revealed that 42.18% of reviews required manual oracle annotation under the triplet configuration rather than all with supervised approach; of these, 669 instances (25.1%) stemmed from rating-review mismatches and 2,183 instances (81.8%) were misclassified by VADER—primarily due to translation artifacts. By deferring only these ambiguous cases to a human oracle, the QBC method effectively balances labeling efficiency with high-quality annotations, thereby reducing overall label noise and bias.

Suggestion

To further improve the proposed QBC-based labeling framework, future researchers should first focus on enhancing translation quality by employing domain-specific machine-translation models trained on Indonesian-English parallel corpora or using VADER based on Indonesian Lexicon; this would help reduce misclassifications introduced by VADER's reliance on translated text. Additionally, it is important to address the underrepresentation of neutral-sentiment reviews—since the neutral class consistently showed lower ROC performance—by collecting more neutral examples or applying data-augmentation techniques such as back-translation or controlled paraphrasing, thereby improving class balance and model generalization. Expanding the committee to include additional weak labelers (for instance, an Indonesian-adapted sentiment lexicon or alternative multilingual transformer models) could also help decrease disagreement rates; experimenting with weighted voting schemes or confidence thresholds might further refine consensus before deferring to human annotation. Researchers should also examine the cost-benefit trade-offs associated with manual annotation, quantifying oracle time and expense against gains in classification accuracy to determine optimal oracle-involvement rates. Finally, integrating the QBC methodology into a real-time active-learning system would enable dynamic selection of the most informative reviews for annotation, potentially reducing overall annotation workload and accelerating model convergence in practical applications.

REFERENCES

Abiola, O., Abayomi-Alli, A., Tale, O. A., Misra, S., & Abayomi-Alli, O. (2023). Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. *Journal of Electrical Systems and Information Technology*, *10*(1). https://doi.org/10.1186/s43067-023-00070-9

Aliyah Salsabila, N., Ardhito Winatmoko, Y., Akbar Septiandri, A., & Jamal, A. (2018). Colloquial Indonesian Lexicon. *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 226–229. https://doi.org/10.1109/IALP.2018.8629151

Aljrees, T., Umer, M., Saidani, O., Almuqren, L., Ishaq, A., Alsubai, S., … Ashraf, I. (2024). Contradiction in text review and apps rating: prediction using textual features and transfer learning. *PeerJ Computer Science*, *10*, e1722. https://doi.org/10.7717/PEERJ-CS.1722

Barik, K., & Misra, S. (2024). Analysis of customer reviews with an improved VADER lexicon classifier. *Journal of Big Data*, *11*(1), 10. https://doi.org/10.1186/s40537-023-00861-x

Budianto, A. G., Wirjodirdjo, B., Maflahah, I., & Kurnianingtyas, D. (2022). Sentiment Analysis Model for KlikIndomaret Android App During Pandemic Using Vader and Transformers NLTK Library. *2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 0423–0427. IEEE. https://doi.org/10.1109/IEEM55944.2022.9989577

Esuli, A., & Sebastiani, F. (2009). Active Learning Strategies for Multi-Label Text Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5478 LNCS*, 102–113. https://doi.org/10.1007/978-3-642-00958-7_12

Fernando, K. R. M., & Tsokos, C. P. (2022). Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(7), 2940–2951. https://doi.org/10.1109/TNNLS.2020.3047335

Hou, D., Zhang, Z., Zhao, M., Zhang, W., Zhao, Y., & Yu, J. (2024). Sentence-level Distant Supervision Relation Extraction based on Dynamic Soft Labels. *Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024*, 3194–3199. https://doi.org/10.1109/CSCWD61410.2024.10580472

Isnan, M., Elwirehardja, G. N., & Pardamean, B. (2023). Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model. *Procedia Computer Science*, *227*, 168–175. Elsevier B.V. https://doi.org/10.1016/j.procs.2023.10.514

JoMingyu. (n.d.). Google Play Scraper. Retrieved October 14, 2024, from https://github.com/JoMingyu/google-play-scraper

Kuligowska, K., & Kowalczuk, B. (2021). Pseudo-labeling with transformers for improving Question Answering systems. *Procedia Computer Science*, *192*, 1162–1169. https://doi.org/10.1016/J.PROCS.2021.08.119

Lu, Y., Song, W., Arachie, C., & Huang, B. (2025). Weakly supervised label learning flows. *Neural Networks*, *182*, 106892. https://doi.org/10.1016/J.NEUNET.2024.106892

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2022). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review 2022 56:4*, *56*(4), 3005–3054. https://doi.org/10.1007/S10462-022-10246-W

Ruhyana -, N., Salsabila Dwi Irmanti -, K., Agung Riyadi -, A., & Mardiana -, T. (2025). SENTIMENT ANALYSIS OF USER REVIEWS BRI MOBILE APPLICATION WITH GRADIENT BOOST METHOD. *Jurnal Riset Informatika*, *7*(2), 1–7. https://doi.org/10.34288/JRI.V7I2.342

Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Discrepancy detection between actual user reviews and numeric ratings of Google App store using deep learning. *Expert Systems with Applications*, *181*, 115111. https://doi.org/10.1016/J.ESWA.2021.115111

Wang, X., Wan, L., & Zhang, J. (2019). An Active Learning Framework Based on Query-By-Committee for Sentiment Analysis. *Proceedings of 2019 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2019*, 327–331. https://doi.org/10.1109/ICAICA.2019.8873452

Wilie, B., Vincentio, K., Indra Winata, G., Cahyawijaya, S., Li, X., Lim, Z. Y., … Bandung, I. T. (2020). *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*. Retrieved from https://arxiv.org/pdf/2009.05387

Wilson Wongso. (2023). *Indonesian RoBERTa Base Sentiment Classifier*. Hugging Face. Retrieved from https://huggingface.co/w11wo/indonesian-roberta-base-sentiment-classifier

Xu, M., & Guo, L. Z. (2021). Learning from group supervision: the impact of supervision deficiency on multi-label learning. *Science China Information Sciences*, *64*(3), 1–13. https://doi.org/10.1007/S11432-020-3132-4/METRICS

Zhang, J., & Cao, M. (2023). Distant supervision for relation extraction with hierarchical attention-based networks. *Expert Systems with Applications*, *220*, 119727. https://doi.org/10.1016/J.ESWA.2023.119727

Zhao, S., Hong, X., Yang, J., Zhao, Y., & Ding, G. (2023). Toward Label-Efficient Emotion and Sentiment Analysis. *Proceedings of the IEEE*, *111*(10), 1159–1197. https://doi.org/10.1109/JPROC.2023.3309299