International Journal of Management Science and Information Technology
IJMSIT

E-ISSN: 2774-5694 P-ISSN: 2776-7388 OPEN ACCESS

Volume 6 (2), July-December 2026, 1714-1724
DOI: https://doi.org/10.35870/ijmsit.v6i2.8061

# Multi-Platform Sentiment Analysis of Diabetes Mellitus on X and TikTok Using K-Nearest Neighbor, Chi-Square Selection, and Oversampling

**Asti Devi Mutiara Khoirun Nisa [1*], Noor Latifah [2], R. Rhoedy Setiawan [3]**
[1*,2,3] Information Systems Study Program, Faculty of Engineering, Universitas Muria Kudus, Kudus Regency, Central Java, Indonesia

Email: 202253160@std.umk.ac.id [1*], noor.latifah@umk.ac.id [2], rhoedy.setiawan@umk.ac.id [3]

Article history:
Received July 17, 2026
Revised July 22, 2026
Accepted August 10, 2026

## Abstract

Diabetes mellitus is a chronic health condition that is widely discussed by the public on social media, generating a large volume of opinions that are difficult to interpret manually. This study analyzes public sentiment toward Diabetes mellitus using data collected from X (Twitter) and TikTok. Text data were preprocessed (cleaning, slang normalization, stopword removal, and stemming) and duplicate entries were removed. Sentiment labels were generated automatically through a rule-based lexicon across four categories (positive, negative, neutral, and irrelevant/discarded), and the reliability of this automatic labeling was verified through manual validation of a stratified sample of 200 data points, measured using Cohen's Kappa. Positive and negative data were then weighted using TF-IDF, reduced using Chi-Square feature selection, balanced using SMOTE, and classified using K-Nearest Neighbor (KNN). Model performance was evaluated using accuracy, precision, recall, and F1-score, and compared across four scenarios: baseline KNN, KNN with Chi-Square, KNN with SMOTE, and the combined KNN+Chi-Square+SMOTE model. The manual validation of 198 valid samples produced an agreement accuracy of 59.60% and a Cohen's Kappa of 0.459 (moderate agreement), indicating that the main source of disagreement lies at the boundary between the neutral and sentiment-bearing classes, while direct positive-negative misclassification was rare (4.5%). The combined model achieved an accuracy of 82.86%, with a macro-averaged precision of 82.95%, recall of 83.56%, and F1-score of 82.79%. Interestingly, Chi-Square feature selection alone yielded the highest accuracy among the four scenarios (85.10%), suggesting that feature selection contributed more to performance gains than class balancing in this dataset. These findings suggest that combining feature selection and oversampling techniques improves the reliability of multi-platform sentiment classification for health-related topics and can inform more effective public health communication strategies regarding diabetes.

**Keywords**:
Sentiment Analysis; Diabetes Mellitus; K-Nearest Neighbor; Chi-Square; SMOTE.

## 1. INTRODUCTION

The rapid advancement of digital technology has transformed the way society communicates and accesses information. Social media platforms now allow the public not only to receive information passively but also to actively produce and share opinions, including on health-related issues (Qadri, 2020). This shift has significantly influenced how people perceive and discuss various health conditions in digital spaces.

Among the health topics frequently discussed online, Diabetes mellitus has drawn considerable public attention due to its status as a chronic disease closely associated with everyday lifestyle (Aminuddin et al.,

2023). Social media content related to diabetes ranges from health education material and personal experiences to unverified opinions circulating among users (Wahyuni et al., 2022). While this open exchange of information increases public awareness, it also generates a large and heterogeneous volume of opinions that is difficult to interpret manually, particularly when the aim is to capture the overall tendency of public sentiment.

Platforms such as X (Twitter) and TikTok exhibit distinct communication characteristics, yet both serve as significant spaces for public expression (Qadri, 2020). X (Twitter) is widely used for direct, text-based opinion sharing, whereas TikTok emphasizes video content complemented by user comments. Combining data from both platforms enables a broader and more representative capture of public sentiment than relying on a single source.

Sentiment analysis provides a systematic approach for classifying large volumes of textual opinion into categories such as positive and negative. Among classification algorithms, K-Nearest Neighbor (KNN) has been widely applied in sentiment analysis tasks due to its simplicity and its ability to group data based on feature similarity, particularly when combined with TF-IDF weighting (Bhuana et al., 2022). However, text classification of social media data commonly faces two persistent challenges: class imbalance and the presence of high-dimensional, noisy features. The Synthetic Minority Over-sampling Technique (SMOTE) has been shown to improve classification performance by balancing minority and majority sentiment classes (Pramayasa et al., 2023), while Chi-Square feature selection helps retain only the most statistically relevant terms for classification, thereby improving model efficiency and accuracy (Bhuana et al., 2022).

Several prior studies have applied KNN-based sentiment analysis in different contexts. Asro'i and Februariyanti (2022) applied KNN to analyze public sentiment toward the extension of the PPKM policy on Twitter, obtaining an accuracy of 69.5%, but relied on a single social media platform and did not address class imbalance. Pramayasa et al. (2023) combined KNN with SMOTE to examine sentiment toward the Merdeka Belajar Kampus Merdeka (MBKM) program, improving accuracy from 68.81% to 76.13% after applying SMOTE, yet did not incorporate feature selection and used data from a single platform. Bhuana et al. (2022) integrated Chi-Square feature selection with KNN to analyze Indonesian public sentiment toward COVID-19 vaccination on Twitter, achieving 88.55% accuracy, but did not address data imbalance and was similarly limited to a single platform. Wijaya and Suwandhi (2024) examined sentiment in TikTok comments using KNN with two sentiment classes, reporting 91% accuracy, though the study was restricted to a single platform and a binary classification scheme without addressing feature selection or class imbalance. Candra et al. (2024) demonstrated that SMOTE improved KNN accuracy from 69.97% to 76.74% in sentiment analysis of Starlink-related YouTube comments, but again without feature selection and using only one data source. Vidya Sakta et al. (2020) combined Chi-Square feature selection with a Neighbor Weighted K-Nearest Neighbor variant to analyze tourism-related sentiment in Malang Regency, achieving 98.8% accuracy, but relied on a single platform, a single non-health-related topic, and did not address class imbalance. Beyond KNN-based approaches, Shefia et al. (2026) applied Support Vector Machine (SVM) combined with SMOTE to analyze sentiment in CapCut application reviews on Google Play Store, achieving an F1-score of 0.736 and an accuracy of 73.54%; although SMOTE improved minority-class performance, the study was limited to a single platform, did not incorporate feature selection, and did not address a health-related topic. Similarly, Ubaidillah et al. (2025) combined Chi-Square feature selection with SMOTE class balancing within a Multinomial Naïve Bayes framework to classify YouTube comment sentiment toward a political figure, achieving 67% accuracy; however, this study employed Naïve Bayes rather than KNN, relied on a single platform, and did not address a health-related topic. These findings collectively indicate that while KNN, SMOTE, and Chi-Square have each been applied individually or in partial combination --- including the joint use of Chi-Square and SMOTE with a different classifier (Ubaidillah et al., 2025) and SVM combined with SMOTE without feature selection (Shefia et al., 2026) --- no prior study has combined all three techniques (KNN, Chi-Square, and SMOTE) while simultaneously utilizing multiple social media platforms for a health-related sentiment analysis task.

Building on these gaps, this study proposes a multi-platform sentiment analysis of public opinion toward Diabetes mellitus using data collected from X (Twitter) and TikTok. The proposed approach integrates TF-IDF weighting, Chi-Square feature selection, SMOTE-based class balancing, and KNN classification within an automated pipeline. Sentiment labels were initially generated through a rule-based lexicon labeling process across four categories (positive, negative, neutral, and discarded/irrelevant content), after which non-informative and irrelevant classes were filtered out prior to classification. To assess the reliability of the automatic labeling process, a manual validation was conducted on a stratified sample of the labeled data, and agreement between the automatic system and human judgment was measured using Cohen's Kappa (Landis & Koch, 1977). Validating automatic/weak labels against a manually annotated subset, rather than the full dataset, is a common practice in recent sentiment analysis research (Setiawan & Andriyani, 2026). This validation step provides an empirical, quantifiable basis for evaluating labeling quality, an aspect largely unreported in the comparable studies discussed above.

The objective of this study is to analyze public sentiment toward Diabetes mellitus on X (Twitter) and TikTok using KNN classification supported by Chi-Square feature selection and SMOTE class balancing, and to evaluate the extent to which this combined approach improves classification performance compared to using each technique individually. The findings are expected to contribute both academically, by providing empirical evidence on the effectiveness of combining feature selection and oversampling techniques in multi-platform

sentiment analysis, and practically, by offering health authorities and related institutions a clearer picture of public perception toward diabetes that can inform more effective health communication strategies.

## 2. RESEARCH METHOD

This study employs a quantitative approach using a computational text-mining pipeline to classify public sentiment toward Diabetes mellitus. The overall research stages consist of data collection, text preprocessing, automatic sentiment labeling with manual validation, feature weighting and selection, class balancing, classification, and model evaluation.

### 2.1. Data Collection

Data were collected from two social media platforms, X (Twitter) and TikTok, using keywords related to Diabetes mellitus. Twitter data were retrieved through crawling, while TikTok data consisted of video comments obtained from a set of relevant videos. The two datasets were merged into a single corpus and labeled with a platform indicator to allow for platform-based comparison.

### 2.2. Text Preprocessing

Prior to analysis, the raw text data were preprocessed through the following steps: (1) case folding and cleaning, which removed URLs, mentions, hashtag symbols, numbers, and emojis; (2) elongated-word normalization (e.g., "sedihh" to "sedih"); (3) slang and abbreviation normalization using a custom Indonesian slang dictionary; (4) tokenization, splitting the cleaned text into individual word tokens; (5) stopword removal using the Indonesian NLTK stopword list, adjusted to retain negation words (e.g., tidak, bukan, belum), context words (e.g., masih, namun), personal pronouns, and question words that carry sentiment-relevant meaning; and (6) stemming using the Sastrawi stemmer, with a manually defined exception list to preserve medically significant terms (e.g., diabetes, insulin, komplikasi, meninggal) that would otherwise be altered by automatic stemming.

### 2.3. Automatic Sentiment Labeling and Manual Validation

Sentiment labels were generated automatically using a rule-based lexicon labeling system developed specifically for the diabetes discussion domain. The system follows a sequential decision procedure: (1) content identified as advertising or product promotion is discarded; (2) figurative jokes or unrelated hyperbole are discarded; (3) content that does not express any opinion, experience, or reaction toward diabetes or related conditions is labeled neutral; (4) content expressing complaints, worry, or negative health outcomes is labeled negative; and (5) content expressing gratitude, improvement, or positive outcomes is labeled positive. This process produced four categories: positive, negative, neutral, and discarded.

To assess the reliability of this automatic labeling process, a stratified random sample of 200 data points was drawn proportionally from the positive, negative, neutral, and discarded classes, as well as from both platforms, and independently labeled manually by the researcher. Agreement between the manual and automatic labels was evaluated using accuracy and Cohen's Kappa coefficient, interpreted according to the standard scale proposed by Landis and Koch (1977). As in recent comparable studies that validate automatically generated labels against a manually annotated sample (Setiawan & Andriyani, 2026), this validation was conducted on a stratified subset rather than the entire dataset. Of the 200 sampled data points, 198 contained complete manual labels and were used for the agreement analysis.

Following validation, the neutral and discarded categories were excluded from the dataset, and only the positive and negative classes were retained for the subsequent classification stage.

### 2.4. Feature Weighting and Selection

The retained text data were split into training and testing sets using an 80:20 ratio with stratified sampling to preserve class proportion. Term Frequency-Inverse Document Frequency (TF-IDF) with unigram and bigram features was applied to convert text into numerical vectors. Chi-Square feature selection was then applied to retain only the features most statistically associated with the sentiment classes. The optimal number of selected features (k) was determined by comparing several candidate values using 3-fold cross-validated F1-macro score with a baseline KNN classifier.

### 2.5. Class Balancing with SMOTE

Because the number of negative and positive samples in social media sentiment data is commonly imbalanced, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to generate synthetic samples for the minority class. To prevent data leakage during hyperparameter tuning, SMOTE was embedded inside a cross-validation pipeline (using imblearn. pipeline. Pipeline) so that oversampling was performed only on each training fold, while the validation fold in each iteration remained untouched.

### 2.6. Classification with K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm was used to classify text into positive and negative sentiment. Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation, optimizing for F1-macro score, across a range of values for the number of neighbors (k = 3, 5, 7, 9, 11), weighting scheme (uniform, distance), and distance metric (Euclidean, Manhattan).

### 2.7. Model Evaluation

Model performance was evaluated on the held-out test set using accuracy, precision, recall, and F1-score (macro-averaged), along with a confusion matrix. To assess the individual and combined contribution of each technique, four scenarios were compared: (1) baseline KNN using TF-IDF features only; (2) KNN with Chi-Square feature selection; (3) KNN with SMOTE class balancing; and (4) the combined KNN + Chi-Square + SMOTE model.

## 3. RESULTS AND DISCUSSION

### 3.1. Results

#### 3.1.1. Dataset Composition

Data collection through scraping yielded a combined total of 14,337 raw entries from X (Twitter) and TikTok (10,341 and 3,996 entries, respectively). After text preprocessing, 13,620 entries remained with non-empty cleaned text. Duplicate removal based on the cleaned text (text_clean) further reduced the dataset to 12,826 unique entries, with 794 duplicate entries discarded. Automatic labeling was then applied to these 12,826 entries, of which 440 were classified as discarded/irrelevant content and excluded from further analysis. The final dataset used for exploratory analysis, comprising the positive, negative, and neutral classes, totaled 12,386 data points, as summarized in Table 1.

Table 1. Dataset Composition per Platform (Positive, Negative, and Neutral Classes)

| Platform | Total Data (Positive + Negative + Neutral) |
|---|---|
| X (Twitter) | 9,291 |
| TikTok | 3,095 |
| Total | 12,386 |

#### 3.1.2. Automatic Labeling Results

The rule-based lexicon labeler classified the dataset into four categories: positive, negative, neutral, and discarded. Table 2 presents the overall and per-platform distribution of the positive, negative, and neutral classes.

Table 2. Distribution of Sentiment Labels, Overall and per Platform

| Label | X (Twitter) n (%) |
|---|---|
| Positive | 2,484 (26.7%) |
| Negative | 3,270 (35.2%) |
| Neutral | 3,537 (38.1%) |
| Subtotal | 9,291 (100%) |
| Label | TikTok n (%) |
| Positive | 315 (10.2%) |
| Negative | 403 (13.0%) |
| Neutral | 2,377 (76.8%) |
| Subtotal | 3,095 (100%) |
| Label | Combined n (%) |
| Positive | 2,799 (22.6%) |
| Negative | 3,673 (29.7%) |
| Neutral | 5,914 (47.7%) |
| Total | 12,386 (100%) |

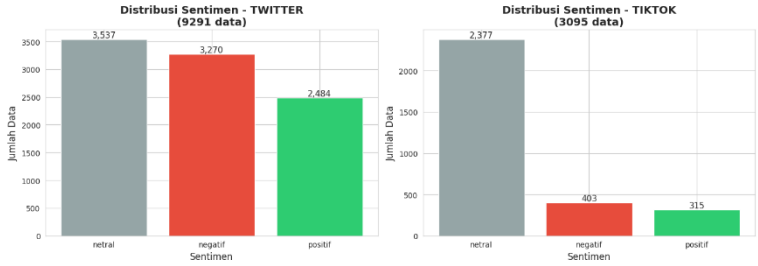


Figure 1. Sentiment Distribution per Platform (X/Twitter vs. TikTok)

As illustrated in Figure 1, the label distribution differs markedly between the two platforms: neutral content dominates TikTok comments (76.8%), whereas Twitter shows a more balanced distribution across the three classes, with a comparatively higher share of negative sentiment (35.2%).

After the neutral class was filtered out, the dataset used for classification consisted only of positive and negative labels, as shown in Table 3.

Table 3. Positive vs. Negative Data Used for Classification

| Label | Count (%) |
|---|---|
| Positive | 2,799 (43.3%) |
| Negative | 3,673 (56.7%) |
| Total | 6,472 (100%) |

### 3.1.3. Manual Validation and Inter-Rater Agreement

To evaluate the reliability of the automatic labeling system, a stratified sample of 200 data points was manually labeled and compared against the system-generated labels. Of the 200 sampled entries, 198 contained complete manual annotations and were used in the agreement analysis. The comparison yielded an overall accuracy of 59.60% and a Cohen's Kappa coefficient of 0.459, which falls into the "moderate" agreement category according to the interpretation scale of Landis and Koch (1977), as summarized in Table 4.

Table 4. Cohen's Kappa Interpretation Scale (Landis & Koch, 1977) and Result Position

| Kappa Range | Agreement Category |
|---|---|
| < 0.00 | No agreement |
| 0.01 – 0.20 | Slight |
| 0.21 – 0.40 | Fair |
| 0.41 – 0.60 | Moderate (this study: 0.459) |
| 0.61 – 0.80 | Substantial |
| 0.81 – 1.00 | Almost perfect |

The confusion matrix comparing manual and automatic labels (Table 5) shows that direct misclassification between the positive and negative classes was rare, occurring in only 9 of 198 samples (4.5%). The largest source of disagreement occurred at the boundary between the neutral class and the sentiment-bearing classes: of the 63 samples manually judged as neutral, 35 (55.6%) were classified by the automatic system as either positive or negative, most commonly involving educational or informational content that the system interpreted as sentiment-bearing.

Table 5. Confusion Matrix — Manual Label vs. Automatic Label (n = 198)

| Manual \ System | Positive / Negative / Neutral / Discarded (n) |
|---|---|
| Positive (n = 49) | 33 / 7 / 7 / 2 |
| Negative (n = 45) | 2 / 34 / 6 / 3 |
| Neutral (n = 63) | 22 / 13 / 27 / 1 |
| Discarded (n = 41) | 2 / 5 / 10 / 24 |

These results indicate that the automatic lexicon-based labeling system is reasonably reliable for distinguishing positive from negative sentiment, with a low risk of sign-flip errors that would otherwise compromise downstream classification. The main limitation lies in the boundary between neutral and sentiment-bearing content, particularly for educational or informational posts, which is acknowledged as a limitation of this study.

### 3.1.4. Chi-Square Feature Selection

Several candidate values of k (the number of selected features) were evaluated using 3-fold cross-validated F1-macro score with a baseline KNN classifier; the value that produced the highest cross-validated

F1-macro score was selected as the optimal k for the final model. The 20 features with the highest Chi-Square scores are presented in Table 6 and Figure 2.

Table 6. Top 20 Features by Chi-Square Score

| Feature | Chi-Square Score |
|---|---|
| naik | 74.28 |
| sehat | 64.65 |
| darah naik | 63.81 |
| tinggi | 59.63 |
| bantu | 57.34 |
| darah tinggi | 55.61 |
| stabil | 50.22 |
| manfaat | 40.44 |
| alhamdulillah | 39.01 |
| normal | 36.19 |
| kontrol gula | 34.30 |
| kontrol | 33.54 |
| takut | 32.79 |
| sakit | 32.12 |
| darah stabil | 30.18 |
| darah rendah | 29.00 |
| jaga | 26.78 |
| stabil gula | 25.05 |
| bantu kontrol | 22.71 |
| hasil | 19.41 |

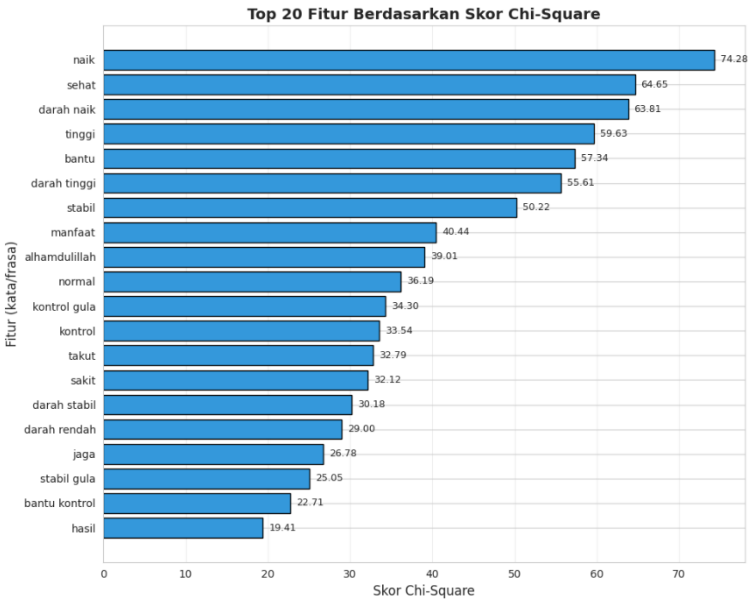


Figure 2. Top 20 Features by Chi-Square Score

The two highest-ranking features are “naik” (rise/increase) and “sehat” (healthy), representing the two dominant sentiment directions in the dataset. “Naik” is strongly associated with negative sentiment, reflecting frequent complaints about rising blood glucose levels, while “sehat,” ranked second, is associated with positive sentiment, reflecting expressions of good health condition. The third-ranked feature, “darah naik” (blood sugar rising), reinforces the negative-sentiment pattern seen in the top feature. Other positive-associated features such as “stabil” (stable), “normal”, and “alhamdulillah” (an expression of gratitude) further reflect expressions of improved health condition and gratitude. This indicates that the Chi-Square selection process successfully retained terms that are semantically meaningful and directly relevant to diabetes-related sentiment, rather than generic or noisy terms.

To complement the quantitative Chi-Square ranking, Figure 3 presents word clouds illustrating the most frequent terms for each sentiment class across both platforms.


Figure 3. Word Clouds per Platform and Sentiment Class

The word clouds show that positive-sentiment terms such as "sehat," "stabil," and "alhamdulillah" dominate both platforms, while negative-sentiment terms are dominated by "gula," "naik," and "sakit," consistent with the top-ranked Chi-Square features in Table 6. The neutral-class Word Clouds are comparatively diffuse, dominated by generic terms such as "aku" and "tidak," reflecting the broader range of topics associated with non-opinionated or informational content.

### 3.1.5. Class Balancing with SMOTE

Prior to balancing, the training data (80% of the positive-negative dataset, proportionally split by class) exhibited a class imbalance, with the negative class more frequent than the positive class, consistent with the overall distribution in Table 7. Figure 4 illustrates the training-set class distribution before and after applying SMOTE.

Table 7. Training Data Distribution Before and After SMOTE

| Class | Before SMOTE / After SMOTE |
|---|---|
| Negative | 2,938 / 2,938 |
| Positive | 2,239 / 2,938 |

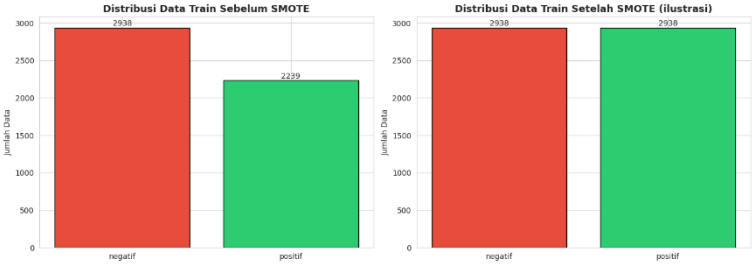


Figure 4. Training Data Distribution Before and After SMOTE

As shown in Table 7, SMOTE generated synthetic samples for the positive (minority) class so that both classes reached 2,938 samples in the resampled training data used during cross-validation, while the original test set was left untouched to preserve an unbiased evaluation.

### 3.1.6. Classification Performance

The combined KNN + Chi-Square + SMOTE model, with hyperparameters selected through GridSearchCV, was evaluated on the held-out test set, achieving the performance summarized in Table 8 and Figure 5.

Table 8. Classification Performance of the Combined KNN + Chi-Square + SMOTE Model

| Metric | Score |
|---|---|
| Accuracy | 82.86% |
| Precision (macro) | 82.95% |
| Recall (macro) | 83.56% |
| F1-score (macro) | 82.79% |

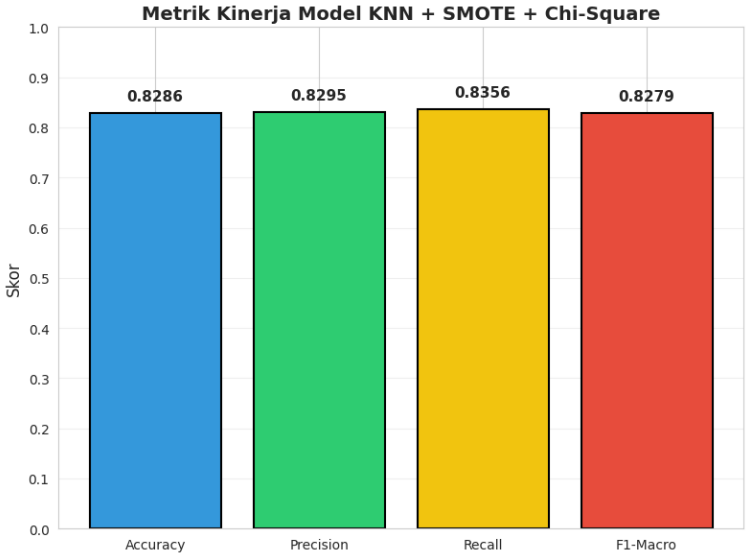


Figure 5. Performance Metrics of the Combined KNN + Chi-Square + SMOTE Model

As shown in Figure 5, precision, recall, and F1-score are closely aligned, all above 82%, with recall (83.56%) slightly exceeding precision (82.95%), indicating that the combined model is marginally more effective at correctly identifying true sentiment instances than at avoiding false positive predictions.

The confusion matrix of the combined model on the test set is presented in Figure 6. As shown in Figure 6, approximately 22% of negative reviews were misclassified as positive (recall of 78% for the negative class), compared to only about 11% of positive reviews misclassified as negative (recall of 89% for the positive class), indicating that the model is somewhat more prone to over-predicting positive sentiment than under-predicting it.

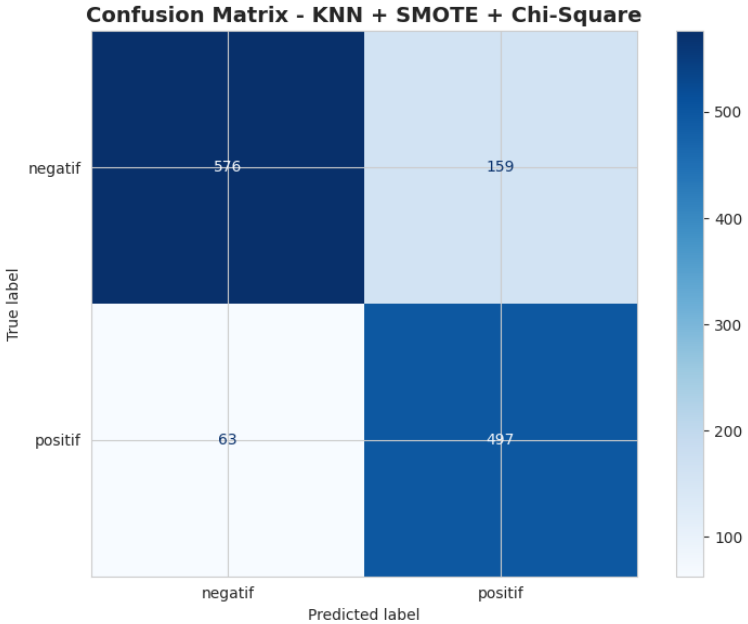


Figure 6. Confusion Matrix of the Combined Model (Test Set)

### 3.1.7. System Implementation

To operationalize the sentiment analysis pipeline described in Section 2, a web-based interface was developed, allowing Twitter/X and TikTok data files (in .csv format) to be uploaded and processed through the full pipeline — preprocessing, rule-based labeling, TF-IDF and Chi-Square feature weighting/selection, SMOTE class balancing, and KNN classification with grid search — without manual scripting, as shown in Figure 7.

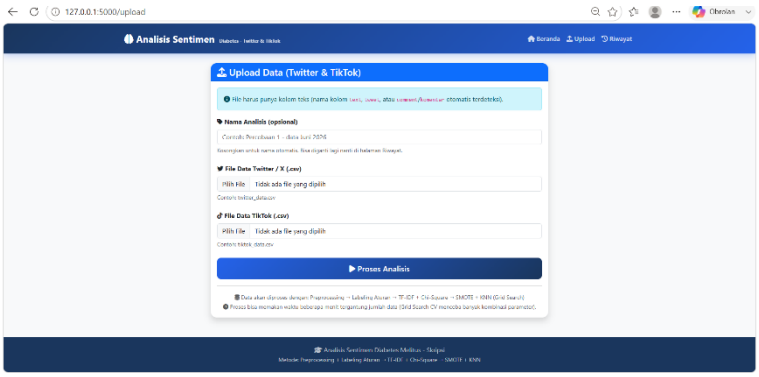


Figure 7. Data Upload Interface for Twitter/X and TikTok Sentiment Analysis

As shown in Figure 7, the interface provides separate upload fields for Twitter/X and TikTok CSV files, an optional analysis name for record-keeping, and a single "Proses Analisis" button that triggers the entire pipeline end-to-end — from preprocessing through model evaluation — without requiring the user to run individual code cells manually.

### 3.1.8. Comparison of Four Scenarios

To evaluate the individual and combined contribution of Chi-Square feature selection and SMOTE class balancing, four scenarios were compared: baseline KNN (TF-IDF features only), KNN with Chi-Square, KNN with SMOTE, and the combined KNN + Chi-Square + SMOTE model. The results are summarized in Table 9 and Figure 8.

Table 9. Performance Comparison Across Four Scenarios

| Scenario | Accuracy / F1-macro |
|---|---|
| KNN (Baseline) | 78.22% / 77.50% |
| KNN + Chi-Square | 85.10% / 84.82% |
| KNN + SMOTE | 64.79% / 63.37% |
| KNN + Chi-Square + SMOTE (combined) | 82.86% / 82.79% |

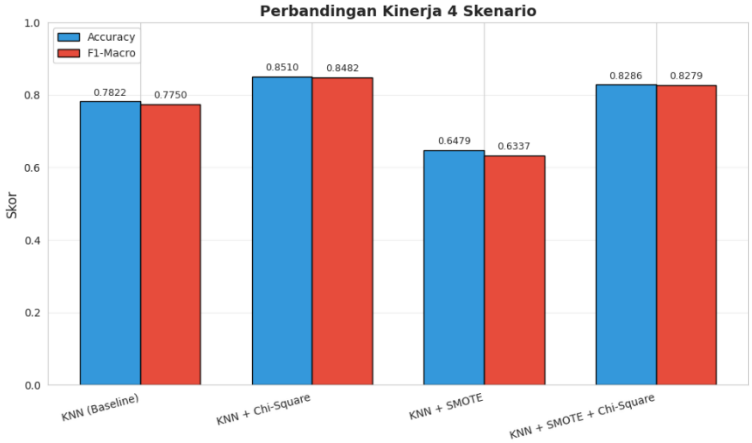


Figure 8. Performance Comparison Across Four Scenarios

Interestingly, applying Chi-Square feature selection alone (85.10% accuracy) outperformed both the baseline KNN model (78.22% accuracy) and the fully combined model (82.86% accuracy), while SMOTE alone (64.79% accuracy) performed worse than the baseline. This pattern, and its implications for how feature selection and oversampling interact with a distance-based classifier such as KNN

### 3.2. Discussion

The results show that Chi-Square feature selection substantially improved classification performance over the baseline KNN model (78.22% accuracy), confirming that reducing the high-dimensional, noisy TF-IDF feature space benefits a distance-based classifier such as KNN. In contrast, applying SMOTE alone, without prior feature selection, resulted in lower accuracy than the baseline (64.79% accuracy), indicating that generating synthetic minority-class samples in the original, unreduced feature space can introduce noise and increase class overlap, which is particularly detrimental to a nearest-neighbor algorithm that relies directly on distances between raw feature vectors. When SMOTE was combined with Chi-Square feature selection, the fully combined model (82.86% accuracy) substantially outperformed the SMOTE-only scenario and still exceeded the baseline, though it did not exceed the Chi-Square-only scenario. This indicates that, in this dataset, feature selection is the primary driver of improved performance, and that SMOTE is only beneficial once noise has already been reduced through feature selection — applying oversampling directly to a high-dimensional, unfiltered feature space can instead be counterproductive for KNN.

Compared to prior studies using KNN-based sentiment analysis, the combined model in this study achieved an accuracy of 82.86%, which is higher than Asro'i and Februariyanti (2022) at 69.5% and Pramayasa et al. (2023) at 76.13%, comparable to Bhuana et al. (2022) at 88.55%, and lower than Wijaya and Suwandhi (2024) at 91% and Vidya Sakta et al. (2020) at 98.8%.

Unlike these prior studies, which either relied on a single social media platform, did not address class imbalance, or did not report a measurable inter-rater agreement for their labeling process, this study combines multi-platform data collection, Chi-Square feature selection, SMOTE class balancing, and a quantitatively validated labeling procedure (Cohen's Kappa = 0.459) within a single pipeline. This is also consistent with the observation that studies restricted to a single, more homogeneous platform (e.g., Wijaya & Suwandhi, 2024, using only TikTok comments) or a narrower topical scope tend to report higher accuracy, whereas multi-platform, more heterogeneous data — as used in this study — introduces additional linguistic and contextual variation that makes classification comparatively more challenging, but arguably more representative of real-world public sentiment.

The manual validation results further indicate that the automatic lexicon-based labeling approach is most reliable in distinguishing clear-cut positive and negative sentiment, but tends to overclassify educational or informational content as sentiment-bearing rather than neutral. This finding is consistent with known limitations of lexicon-based sentiment labeling reported in prior sentiment analysis literature, and is acknowledged as a limitation of the present study rather than a flaw unique to this dataset. Addressing this limitation may require incorporating context-aware disambiguation rules or a hybrid lexicon-machine learning labeling approach in future implementations.

## 4. CONCLUSION

This study analyzed public sentiment toward Diabetes mellitus using data collected from X (Twitter) and TikTok, combining TF-IDF weighting, Chi-Square feature selection, SMOTE class balancing, and K-Nearest Neighbor classification. Sentiment labels were generated through a rule-based lexicon system and validated manually on a stratified sample of 200 data points, yielding an accuracy of 59.60% and a Cohen's Kappa of 0.459 (moderate agreement), with direct positive-negative misclassification occurring in only 4.5% of cases. The combined KNN + Chi-Square + SMOTE model achieved an accuracy of 82.86% and an F1-macro score of 82.79%, outperforming both the baseline KNN model (78.22% accuracy) and the SMOTE-only scenario (64.79% accuracy), although Chi-Square feature selection alone achieved the single highest accuracy (85.10%) among the four scenarios tested. This indicates that feature selection contributed more substantially to performance improvement than class balancing in this dataset — applying SMOTE without prior feature selection was in fact counterproductive, whereas combining SMOTE with Chi-Square feature selection still offered a more balanced classification across both sentiment classes at only a small cost to overall accuracy. This finding itself constitutes a key contribution of this study: rather than assuming that combining more techniques necessarily yields the best performance, this research empirically demonstrates the conditions under which each technique helps or hinders KNN-based classification — an aspect not addressed in prior studies that applied these techniques individually or in partial combination.

These findings highlight the importance of understanding how feature selection and oversampling interact with a given classification algorithm, rather than assuming that applying more preprocessing techniques automatically leads to better performance. The main limitation of this study lies in the boundary between neutral and sentiment-bearing content, where educational or informational posts are sometimes misclassified by the automatic labeling system. Future research is recommended to refine the lexicon rules for distinguishing informational from opinion-bearing content, to explore alternative classification algorithms such as Support Vector Machine (SVM), deep learning architectures, and Transformer-based models (e.g., BERT), and to extend the data collection period and platform coverage to improve the generalizability of the findings.

## REFERENCES

Aminuddin, A., Sima, Y., Izza, N. C., Lalla, N. S. N., & Arda, D. (2023). Edukasi Kesehatan Tentang Penyakit Diabetes Melitus bagi Masyarakat. Abdimas Polsaka, 7–12. https://doi.org/10.35816/abdimaspolsaka.v2i1.25

Asro'i, A., & Februariyanti, H. (2022). Analisis Sentimen Pengguna Twitter Terhadap Perpanjangan PPKM Menggunakan Metode K-Nearest Neighbor. Jurnal Khatulistiwa Informatika, 10(1), 17–24. https://doi.org/10.31294/jki.v10i1.12624

Bhuana, K., Indriati, & Muflikhah, L. (2022). Analisis Sentimen Masyarakat Indonesia tentang Vaksin Covid-19 di Twitter dengan menggunakan Metode K-Nearest Neighbors dan Seleksi Fitur Chi Square. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 6(3), 1395–1401.

Candra, C., Chandra, K. W., & Irsyad, H. (2024). Efektifitas SMOTE dalam Mengatasi Imbalanced Class Algoritma K-Nearest Neighbors pada Analisis Sentimen terhadap Starlink. Jurnal Ilmu Komputer dan Informatika, 4(1), 31–42. https://doi.org/10.54082/jiki.132

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Pramayasa, K., Maysanjaya, I. M. D., & Indradewi, I. G. A. A. D. (2023). Analisis Sentimen Program MBKM Pada Media Sosial Twitter Menggunakan KNN dan SMOTE. SINTECH (Science and Information Technology) Journal, 6(2), 89–98. https://doi.org/10.31598/sintechjournal.v6i2.1372

Qadri, M. (2020). Pengaruh Media Sosial Dalam Membangun Opini Publik. Qaumiyyah: Jurnal Hukum Tata Negara, 1(1), 49–63. https://doi.org/10.24239/qaumiyyah.v1i1.4

Setiawan, I., & Andriyani, W. (2026). Perbandingan Kinerja dan Efisiensi Model NLP pada Analisis Sentimen Ulasan Aplikasi Layanan Publik Digital. Jurnal Sistem Komputer dan Informatika (JSON), 7(4), 1505–1517. https://doi.org/10.30865/json.v7i4.9774

Shefia, F. A., Setiaji, P., & Triyanto, W. A. (2026). Analisis Sentimen Ulasan Aplikasi CapCut pada Google Play Store menggunakan Support Vector Machine dengan Teknik SMOTE. Sistemasi: Jurnal Sistem Informasi, 15, 658–668. https://doi.org/10.32520/stmsi.v15i2.5948

Ubaidillah, M., Fatah, D. A., & Negara, Y. D. P. (2025). Penerapan SMOTE dan Chi-Square Feature Selection untuk Meningkatkan Akurasi Model Multinomial Naïve Bayes dalam Analisis Sentimen Video "Presiden Prabowo Menjawab" di YouTube. Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI), 8(5). https://doi.org/10.32672/jnkti.v8i5.9807

Vidya Sakta, P., Indriati, & Marji. (2020). Analisis Sentimen Pariwisata di Kabupaten Malang dengan Menggunakan Metode BM25F, Neighbor Weighted K-Nearest Neighbor dan Seleksi Fitur Chi-Square. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 4(10), 3659–3666.

Wahyuni, S., Arisani, G., Riani, R., & Hanipah, H. (2022). Peran Media Sosial Sebagai Upaya Promosi Kesehatan. Jurnal Forum Kesehatan: Media Publikasi Kesehatan Ilmiah, 11(2), 86–96. https://doi.org/10.52263/jfk.v11i2.233

Wijaya, R., & Suwandhi, A. (2024). Sentimen Komentar Universitas Pelita Harapan Pada TikTok Menggunakan Metode K-Nearest Neighbor. JDMIS: Journal of Data Mining and Information Systems, 2(1), 26–36. https://doi.org/10.54259/jdmis.v2i1.2418