Knowledge Engineering and Data Science (KEDS)
Vol 7, No 2, December 2024, pp. 139–151

pISSN 2597-4602
eISSN 2597-4637

Manifold Learning and Undersampling Approaches for Imbalanced Class Sentiment Classification

L.M. Risman Dwi Jumansyah [1,*], Agus Mohamad Soleh [2], Utami Dyah Syafitri [3]

Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University
Jalan Meranti Wing 22 Level 4 IPB Univerisity, Dramaga, Bogor, West Java 16680, Indonesia
[1] rismandwijumansyah@gmail.com; [2] agusms@apps.ipb.ac.id; [3]utamids@apps.ipb.ac.id*
** corresponding author*

ARTICLE INFO

Article history:
Received 30 September 2024
Revised 13 November 2024
Accepted 13 December 2024
Published online 24 December 2024

Keywords:
Imbalanced data
Manifold
Movie reviews
Sentiment classification
Undersampling

ABSTRACT

Movie reviews are crucial in determining a film's success by influencing audience decisions. Automating sentiment classification is essential for efficient public opinion analysis. However, it faces challenges such as high-dimensional data and imbalanced class distributions. This study addresses these issues by applying manifold learning techniques, Principal Component Analysis (PCA) and Laplacian Eigenmaps (LE) to reduce data complexity and undersampling strategies (Random Undersampling (RUS) and EasyEnsemble) to balance data and improve predictions for both sentiment classes. On reviews of The Raid 2: *Berandal*, EasyEnsemble achieved the highest average G-Mean of 0.694 using Term Frequency-Inverse Document Frequency (TF-IDF) features with a linear kernel without dimensionality reduction. RUS provided balanced but inconsistent results, while Review of Systems (ROS) combined with PCA (85% variance cumulative) improved predictions for negative reviews. Laplacian Eigenmaps were effective for negative reviews with 500 dimensions but less accurate for positive ones. This study highlights EasyEnsemble's superior performance in addressing the class imbalance, though optimization with manifold learning remains challenging.

I. Introduction

Movie reviews are a vital source of information on social media platforms, as they directly reflect viewers' opinions on various aspects of a film [1]. Movie reviews on websites are typically informal and unstructured yet effectively convey viewers' emotions [2]. The use of lengthy plot summaries and complex literary devices, such as rhetoric and sarcasm, complicates sentiment analysis, making it difficult to interpret the underlying meaning and emotions in the text accurately [3]. Sentiment analysis of movie review data plays a crucial role in evaluating films by determining the sentiment expressed in clusters of text data, such as documents, sentences, and paragraphs. Sentiments are typically classified into three categories: positive, negative, or neutral [4]. Negative reviews, in particular, can strongly influence viewers' decisions, often deterring potential audiences from watching a film [5]. There are two primary approaches to sentiment analysis. The first approach employs a sentiment lexicon, where experts compile a list of sentiment-laden words to assess the sentiment expressed in the text [6]. The second approach utilizes machine learning techniques, mainly focusing on binary classification to distinguish between positive and negative sentiments [7].

Support Vector Machine (SVM) is a widely used machine learning algorithm for sentiment classification of movie reviews. Studies [8] and [9] demonstrated the algorithm's effectiveness in classifying sentiments using movie review data from the IMDB website, showing superior performance compared to other algorithms. The data used in both studies exhibit a balanced class distribution, which only reflects the reality where the class imbalance is commonly found. Class imbalance occurs in sentiment analysis when positive or negative opinions dominate the datasets [10]. This imbalance causes classification methods to be biased toward the majority class, resulting in more frequent incorrect predictions for the minority class. Such a condition reduces overall model performance and explicitly impacts sensitivity and precision for the minority class [11]. The other issue is the high-dimensional nature of text data, where the representation of words through document-term matrices often results in a matrix of size $n \times p$, with n denoting the number of documents and p representing the number of features [12]. Here, features are typically unigram tokens, with each word

https://doi.org/10.17977/um018v7i12024p139-151

considered a separate explanatory variable, leading to a high-dimensional and sparse matrix [13]. High dimensionality increases memory usage and computational time, which can hinder the generalization capability of learning algorithms [14].

The study resolves the trade-off between computational efficiency and classification performance in high-dimensional with imbalanced classes by utilizing approaches that boost model effectiveness while simultaneously reducing processing time. Manifold learning techniques such as Principal Component Analysis (PCA) and Laplacian Eigenmaps (LE) are applied to handle the high-dimensionality problem. These methods assume that high-dimensional data can be effectively projected onto a lower-dimensional manifold, capturing and preserving the essential underlying structure of the original data clusters [15]. PCA focuses on retaining the principal components that capture the majority of the variance in the data, thereby reducing computational time [16]. Studies [17] and [18] applied PCA before implementing SVM modelling on Amazon product review data, resulting in better performance than models without PCA. Meanwhile, LE constructs a similarity graph to map data points, preserving the local similarities after dimensionality reduction and maintaining the integrity of the data [19]. A study [20] demonstrated that LE outperformed other nonlinear methods, such as kernel PCA, t-SNE, and ISOMAP, when applied prior to SVM modelling using the same dataset.

Undersampling techniques like Random Undersampling (RUS) and EasyEnsemble address class imbalance. RUS reduces the sample size of the majority class to match that of the minority class, thereby balancing the class distribution [21]. Study [22] proved effective in improving performance when applying this method to high-dimensional and imbalanced data using text data from Twitter. Compared to oversampling methods, RUS has shown competitive results regarding the Area Under the Curve (AUC) and G-Mean metrics [23]. EasyEnsemble, on the other hand, divides the majority class into multiple subsets, each of which is used to train separate estimators alongside the minority class data [24]. Study [25] combined Random Forest with EasyEnsemble for fake review classification, resulting in better performance than the alternative methods.

This research aims to improve sentiment classification performance in movie reviews by addressing high-dimensional data and class imbalance challenges. The proposed approach integrates manifold learning techniques like PCA and Laplacian Eigenmaps to reduce data dimensionality and undersampling techniques, like RUS and EasyEnsemble, to handle class imbalance. This method is expected to enhance classification accuracy, improve computational efficiency, and accelerate decision-making processes related to audience sentiment. The findings have significant implications for Indonesia's film industry, enabling a better understanding of audience sentiment to optimize marketing strategies and audience engagement. Automating sentiment analysis on platforms like TikTok, Instagram, and X could also help Indonesian filmmakers and production houses craft targeted promotional campaigns and gain insights into public reception, fostering growth and competitiveness in the global market.

II. Methods

A. Data

The data used in this study consists of reviews and ratings for the Indonesian action-thriller film The Raid 2: *Berandal*, directed by Welsh filmmaker Gareth Evans and released in 2014. The film garnered significant international attention, especially in regions like the United States, United Kingdom, Canada, and Japan, known for their intense action scenes. This broad appeal made The Raid 2 an ideal choice for analysis. The reviews were collected using R version 4.3.2 web scraping techniques, specifically through the *Rvest* version 1.0.4 package from the IMDB and Letterboxd websites.

B. Preprocessing

Data preprocessing is a series of essential steps in preparing raw data for classification [26]. Text processing involves several key steps to prepare textual data for analysis. First, case folding is performed to standardize all text into lowercase, ensuring consistency regardless of capitalization. Next, the cleaning text step removes unwanted elements such as punctuation, symbols, numbers, and emojis, leaving only meaningful text. Acronym expansion follows, converting abbreviations into complete forms to enhance clarity and context. Subsequently, stemming or lemmatization is applied

to transform words into their base or root forms, reducing variations and simplifying analysis. Lastly, stopword removal eliminates common words like conjunctions and pronouns, which often carry little semantic value, focusing on the text's more significant components.

The preprocessed text data undergoes feature extraction to transform the raw text into numerical representations suitable for machine learning models. Various feature extraction techniques are employed, such as TF-IDF, Word2Vec, n-grams, GloVe, and BERT. This study uses TF-IDF, Word2Vec, and their combination (Word2Vec + TF-IDF) due to their effectiveness in capturing word relationships and improving sentiment classification. Additionally, these methods are chosen for their computational efficiency compared to more complex models [27]. Term Frequency-Inverse Document Frequency (TF-IDF) is a standard information processing and data mining method. *tf* measures how often a word appears in a specific text. At the same time, *idf* assesses the importance of that word across the entire collection of documents [28]. The formulation for calculating TF-IDF as in (1).

$$w_{ij} = tf_{ij}.log\left(\frac{N}{df_i}\right) \quad (1)$$

Word2Vec is a word vector learning technique that utilizes neural networks to predict the next word. There are two main types of neural network models based on the prediction approach: the Continuous Bag of Words (CBOW) model, which predicts the next word based on a set of input words similar to the n-gram model, and the Skip-Gram (SG) model, which predicts a set of surrounding words using a single word as input [29]. This study employs the Skip-Gram model, known for its effectiveness in generating word embeddings by capturing word context and semantic relationships, where the Skip-Gram method predicts the context based on the target word [30]. The method [31] enhances feature representation, where Word2Vec and TF-IDF are combined. Specifically, the term value from TF-IDF is multiplied by its corresponding word vector, and all term vectors from the input are then summed to produce a single vector representation.

The next step in the preprocessing phase is to label the film review data sentiment based on each review's rating score. Reviews rated four or less out of 10 are categorized as unfavourable. In contrast, those with a rating of 7 or higher are classified as positive. Reviews with a rating score of 5 or 6 out of 10 are considered neutral. However, this study concentrates only on positive and negative reviews [32].

C. *Data Analysis*

Various methodological approaches are used to classify sentiment in movie review data. This study employed R software (version 4.3.2) for dimensionality reduction using the prcomp function from the stats package for PCA and the do. Lapeig function from the Rdimtools package for LE. It also addressed imbalanced data and implemented classification algorithms, explicitly using the SVM function from the e1071 package for SVM. The data analysis process is presented as a flowchart in Figure 1.

The first step is to perform dimensionality reduction on the feature-extracted data. This process is illustrated in Figure 2. This study's dimensionality reduction approach to feature extraction uses PCA and LE. PCA is a statistical method that transforms correlated variables into a new set of uncorrelated variables. This technique can produce fewer new variables while effectively explaining the variance in the data [33][34]. Conversely, LE, introduced by [35], is a spectral technique for nonlinear dimensionality reduction that preserves local data structures through graph-based Laplacian matrix decomposition. LE considers the intrinsic geometry of the data by constructing a graph based on the neighbourhood relationships among observations. In this graph, each node represents an observation, and the size of the edge connecting these nodes reflects the degree of similarity between neighbouring data points [36]. This study also compares the sentiment classification of movie reviews without applying dimensionality reduction.

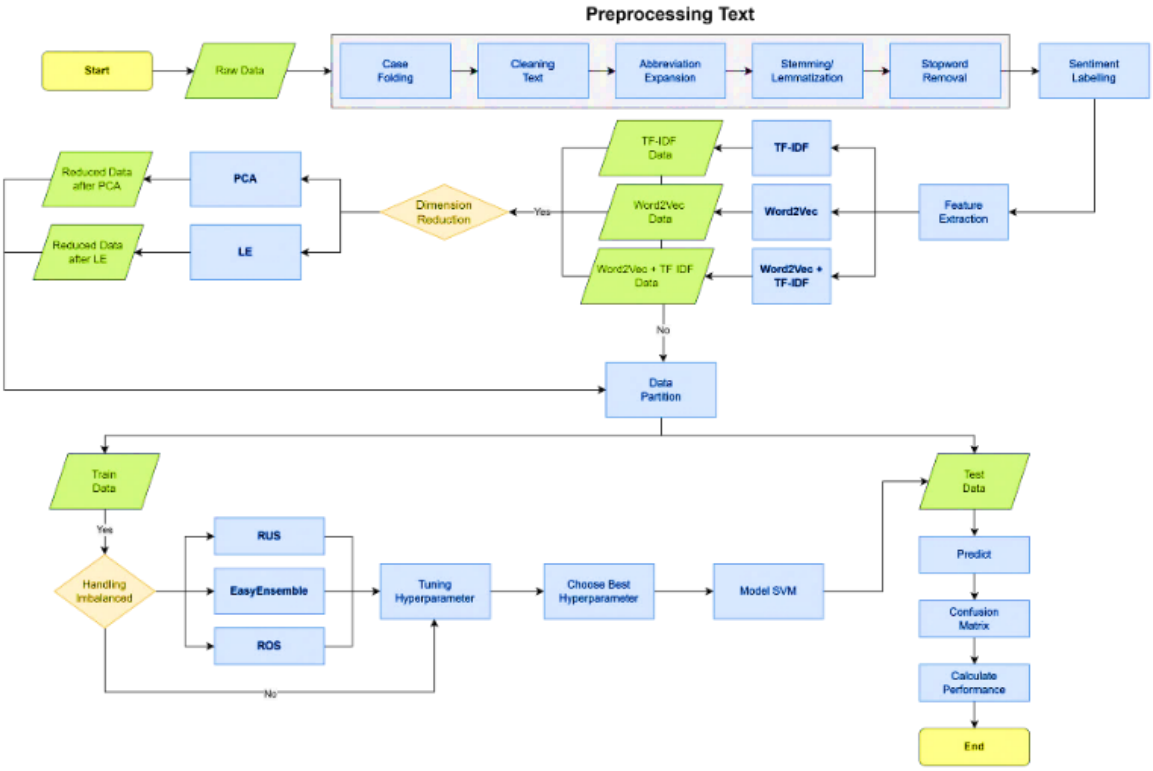

Fig. 1. Data analysis process

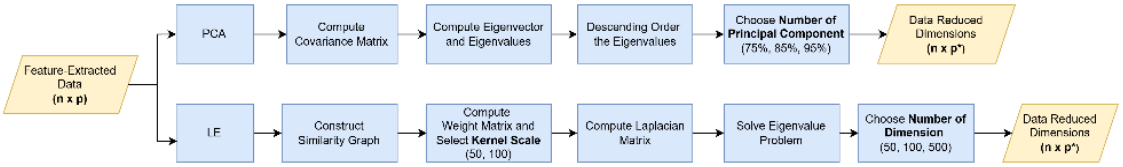

Fig. 2. Dimension reduction process

The dataset is split into 80% training, and 20% testing, with the partitioning repeated 100 times using random sampling to ensure equal representation and robustness in model evaluation. Class imbalance is addressed using undersampling techniques, namely RUS and EasyEnsemble. A scenario without handling class imbalance and oversampling with ROS is also included for comparison. RUS reduces the majority class by randomly removing samples [37]. EasyEnsemble creates multiple balanced subsets of the majority class, where each subset is used to train a model, and the results from all models are combined to make a final decision [38]. ROS increases the number of minority class samples by replicating them to improve model performance on underrepresented classes [37]. Figure 3 illustrates strategies for managing imbalanced data.

The training data determines the optimal hyperparameters after addressing class imbalance. This process employs stratified 5-fold cross-validation, which divides the data into five folds while maintaining a balanced proportion of minority and majority classes. One fold is the validation set, and the remaining four folds are used for training [39]. Hyperparameters are selected using grid search to explore predefined parameter combinations systematically. The optimal hyperparameters are chosen based on the highest average G-Mean across the folds. The hyperparameters with the most minor cost or gamma values are selected if the average G-Mean values are identical. The specific hyperparameters adjusted in this study are listed in Table 1.

The classification method employed is the SVM. The concept of SVM involves identifying a suitable hyperplane that effectively separates the two data sets by maximizing the margin between the hyperplane and the nearest data points [40][41]. This study uses Linear and Radial Basis Functions (RBF) kernel.

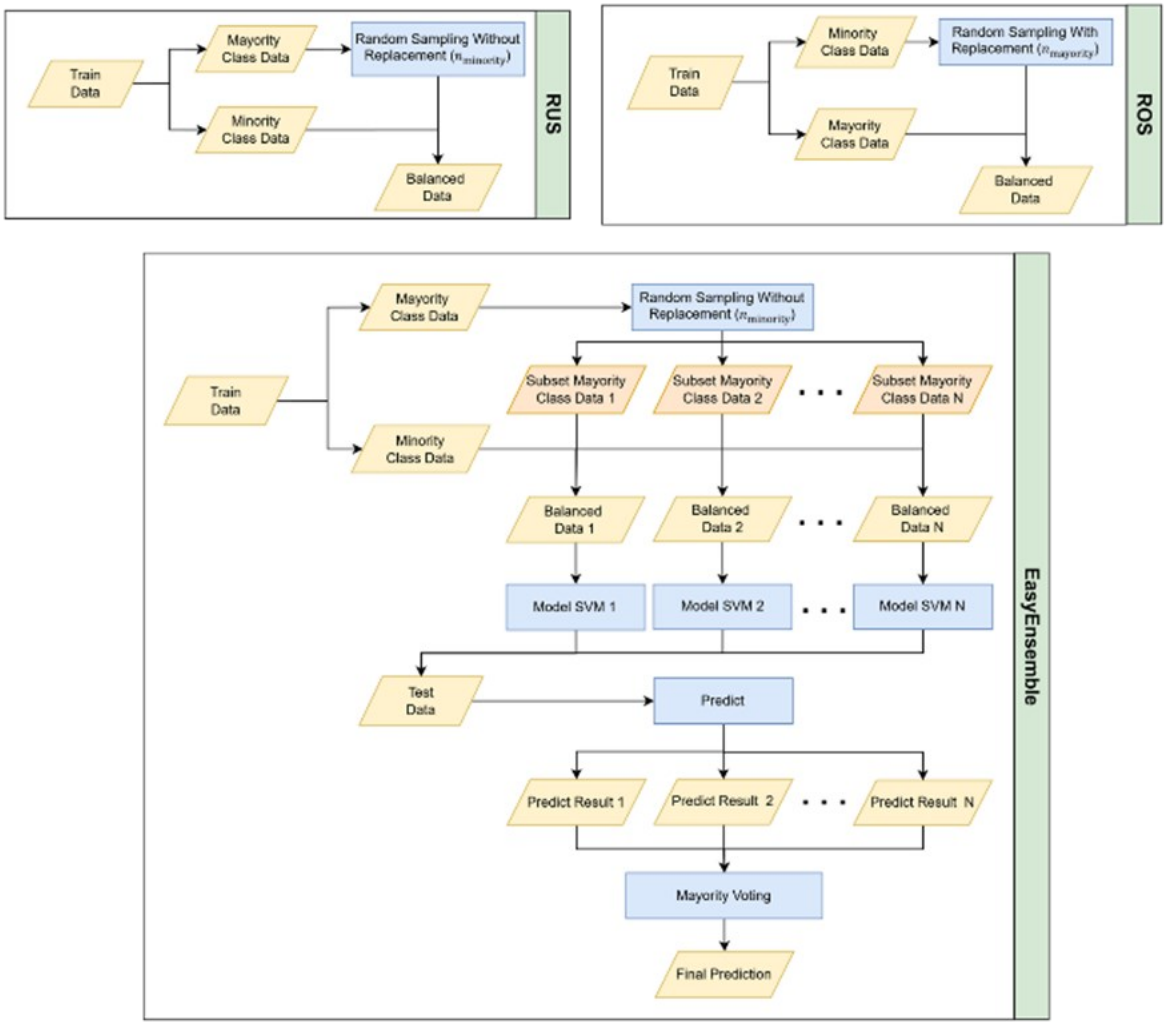

Fig. 3. Handling imbalanced class process

Table 1. Hyperparameter and values

Method	Hyperparameter	Value
PCA	Cumulative Variance Proportion	75%, 85%, and 95%
LE	Number of Neighbour (k)	5
	Number of Dimensions	50,100, and 500
	Kernel Scale	50 and 100
EasyEnsemble	Number of Estimator	101
SVM	Cost (*C*)	0.01, 0.05, 0.1, 0.5, 1, and 5
	Gamma (γ)	0.001, 0.005, 0.01, 0.05, 0.1, and 0.5

D. Evaluation

The classification model is evaluated using a Confusion Matrix, which compares the actual sentiment labels with the predicted ones, encompassing the metrics of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [24]. Model performance is assessed through Accuracy, Sensitivity, Specificity, and G-Mean. The formulas for calculating these metrics are presented as in (2) to (5). Accuracy is not a metric, as it represents the overall classification ability and can be misleading when data distribution is imbalanced. Sensitivity and specificity are essential for identifying the prediction capability for positive and negative classes, helping to assess classifier effectiveness for both majority and minority classes [42]. G-Mean is an evaluation metric that provides an overall picture of model accuracy by considering both minority and majority classes [42].

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

$$\text{G-Mean} = \sqrt{Sensitivity \; x \; Specificity} \quad (5)$$

III. Results and Discussion

The study focuses on English-language reviews to capture responses from a global audience rather than solely from Indonesia. The reviews collected include 357 from IMDB and 2749 from the Letterboxd website, totalling 3106 reviews for the film The Raid 2: *Berandal*. Neutral reviews, totalling 1273, were excluded from the analysis because the study focuses only on positive and negative reviews. Table 2 shows a highly imbalanced dataset, with 96.18% positive and only 3.81% negative reviews. This imbalance challenges classification models, which often favour the majority class.

Table 2. Number and percentage of reviews

Label	Amount	Percentage (%)
Positive	1763	96.181
Negative	70	3.819
Total	1833	100

A. Preprocessing Data

The writing styles in the collected review data exhibit significant differences, including the use of capital letters, prefixes, slang words, abbreviations, conjunctions, punctuation, and emojis. Preprocessing transforms the text into its base form and removes punctuation marks, emojis, and stopwords. This process is crucial for minimizing noise and enhancing model performance. Table 3 provides examples of the results from this preprocessing.

Table 3. Results preprocessing data

Original text	Results
Some awesome fight scenes that are sadly bogged down with an overly convoluted and frankly uninteresting gangster plot. Inferior to the first film...	The awesome fight scene, sad bog convolute frank, uninteresting gangster plot infer
And just as enjoyable with relentless action and bloody, gory violence.	Enjoy relentless action, blood, gore, and violent.
I hated this movie. It's a fucking endless slog, and the action exists w/ no stakes. The plot is the dumbest fucking thing in the world, and I can't believe I forced myself through 2. to 5 hours of this shit. fuck you.	Hate fuck endless slog action exist stake plot dumb fuck force shit fuck

Figure 4 presents both review classes' positive and negative unigram word cloud visualization, with positive reviews on the left and negative reviews on the right. The size of the text in the word cloud reflects the frequency of word occurrences throughout the entire review text. The larger the word size, the more frequently the word appears. This visualization aids in understanding the most used words by users within each class. Some words frequently appear in both positive and negative reviews for this film, which are similar. Words such as "*Action*," "Raid," "Good," "Fight," "Scene," "Story," and "*Character*" have high frequencies of occurrence. However, these words are the same. Their intended meanings in the context of the reviews are not necessarily the same.

For example, "*scene*" and "*fight*" appear positively and negatively. In a negative review, one might say, "*I fell asleep during a fight scene in a raid film*." In contrast, a positive review could state, "*The fight scenes are excellent—well-choreographed and brutal—but 'I wish this were longer and more like The Departed' is not a thought I had after watching the first film*." From these examples, it can be concluded that there are viewers who find the fight scenes boring, while others highly enjoy and praise them. This highlights the subjectivity of film reviews, as individual preferences and experiences greatly influence how different audiences perceive the same content.

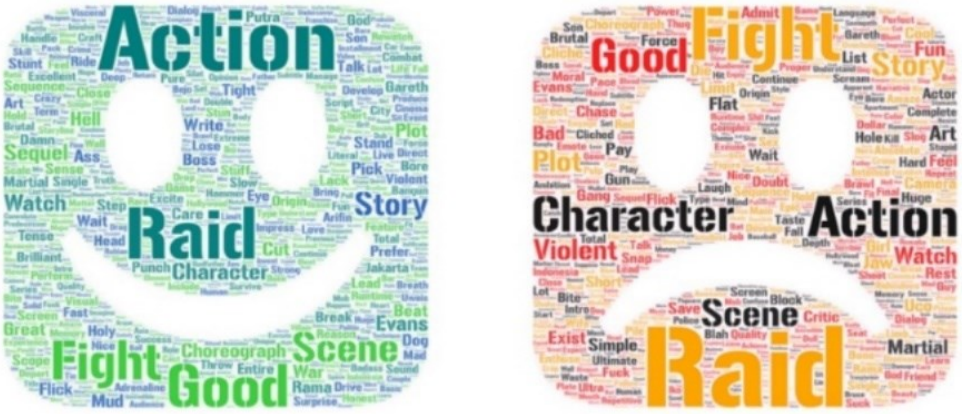

Fig. 4. Wordcloud of the raid 2: *Berandal* Reviews

The tokenized text from the preprocessing stage is transformed into structured data in numerical form. Three feature selection methods, namely TF-IDF, Word2Vec, and Word2Vec combined with TF-IDF, are applied to convert the textual data into numerical data, all using unigram tokens. This film review dataset contains 4836 unique words. Table 4 presents an example of the TF-IDF feature extraction results from three sample reviews. A high TF-IDF value indicates that a word is unique within a document and rarely appears in other documents.

Table 4. TF-IDF results

Text	Label	Words						
		Action	***Sound***	***Enjoy***	***…***	***Drama***	***Extreme***	***Supreme***
Enjoy relentless action, Blood, gore, violent	Positive	0.172	0	0.709	…	0	0	0
Enjoy extreme tense Organise sound camerawork, love	Positive	0	0.783	0.608	…	0	0.783	0
Good fight scene drama sound Understand the plot, Connect characters rudy Supreme	Negative	0	0.498	0	…	0.397	0	0.774

In contrast, a low TF-IDF value suggests that the word occurs frequently throughout the corpus. If the TF-IDF value is 0, the word is absent from the document because its term frequency (TF) is 0. For example, the word "sound" has a high value in positive reviews, indicating that the sound element is highly valued in the context of positively rated films. At the same time, its contribution is lower in negative reviews, suggesting that the sound aspect is not a significant factor in negative evaluations.

Word2Vec generates a vector for each word, with each vector containing 1000 values. The cleaned text data from the reviews is trained using the Skip-Gram model. After modelling, predictions are made for each word within a given review. Table 5 illustrates the results of the Word2Vec representation for one review, showcasing the generated vectors for the words contained in that review. This approach allows for the semantic meaning of words to be captured based on their context in the training data. The closer the word vectors are to each other in the vector space, as measured by cosine similarity, the more likely they share similar contexts in the reviews, such as themes, sentiments, or usage patterns. This proximity indicates that words frequently appearing together in similar contexts during training are more semantically related.

Table 5. Word2Vec results

Words	V1	V2	V3	…	V1000
Enjoy	0.125	1.526	-0.663	…	1.639
Relentless	-0.613	0.797	-0.189	…	0.824
Action	0.497	0.957	-0.179	…	1.194
Blood	-0.898	-0.261	0.929	…	-0.459
Gore	-0.362	0.098	0.638	…	-0.394
Violent	-0.251	0.292	0.469	…	0.012
Enjoy relentless action, Blood, gore, violent	-0.250	0.568	0.167	…	0.469

The study examines the integration of Word2Vec with TF-IDF to transform text into a more accurate and detailed numerical representation. Word2Vec captures the semantic meaning of words based on their surrounding context. At the same time, TF-IDF assigns higher weights to words that are rare across the corpus but relevant within a specific document. During the integration stage, each row of the Word2Vec embedding matrix for each review is multiplied by the corresponding word's TF-IDF value. The product of all words in each row is then summed to produce the final review representation. Figure 5 illustrates this calculation process using the TF-IDF values from the first review in Table 4 and the Word2Vec embeddings in Figure 5.

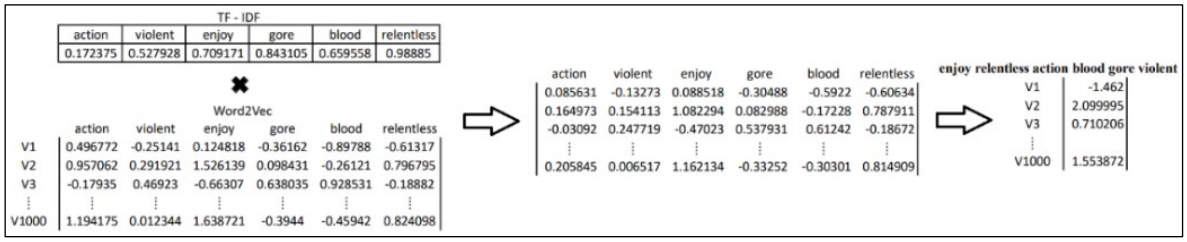

Fig. 5. Word2vec + TF-IDF results

B. *Model Performance Evaluation*

The model evaluation uses accuracy, sensitivity, specificity, and G-Mean metrics to assess performance, mainly focusing on the balance between majority and minority class accuracy. The study compares models with and without class imbalance handling and dimensionality reduction techniques. Feature extraction methods include TF-IDF, Word2Vec, and a combination of both. Techniques like PCA, LE, ROS, RUS, and EasyEnsemble are analyzed for their effectiveness in improving imbalanced data classification. Each method combination is tested 100 times, ensuring comprehensive evaluation across the dataset. The accuracy result is presented in Figure 6.

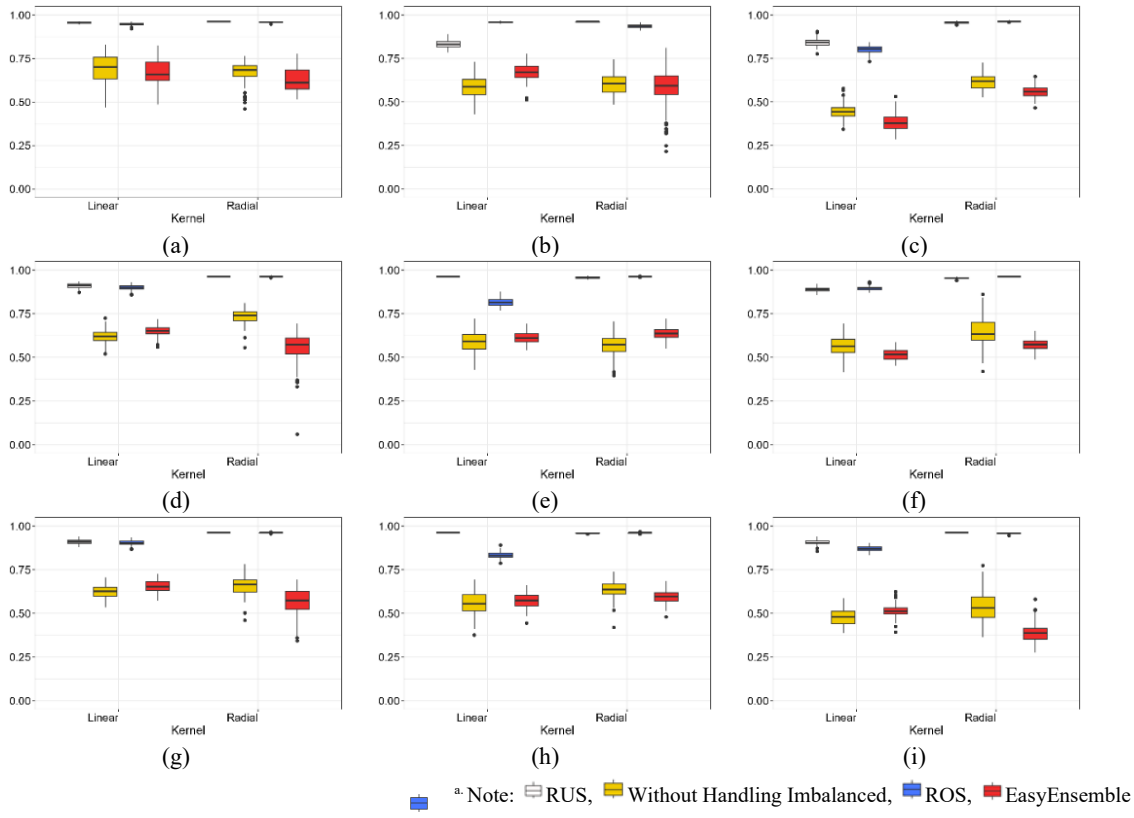

Fig. 6. Accuracy results for (a) TF-IDF, (b) TF-IDF PCA, (c) TF-IDF LE, (d) Word2Vec, (e) Word2Vec PCA, (f) Word2VecLE, (g) Word2Vec+TF-IDF, (h) Word2Vec+TF-IDF PCA, and (i) Word2Vec+TF-IDF LE

The accuracy results in Figure 6 show that the model achieves the highest accuracy without handling imbalanced data and when using ROS. EasyEnsemble outperforms RUS with a linear kernel.

RUS performs better with a radial kernel—dimensionality reduction using LE, which results in lower accuracy than models without dimensionality reduction or those utilizing PCA. PCA with Word2Vec feature extraction achieves accuracy comparable to models without dimensionality reduction, making it competitive in some scenarios. However, accuracy alone can be biased, especially with imbalanced data, so sensitivity and specificity should also be considered. Figure 7 presents the sensitivity results, reflecting the prediction accuracy for positive reviews.

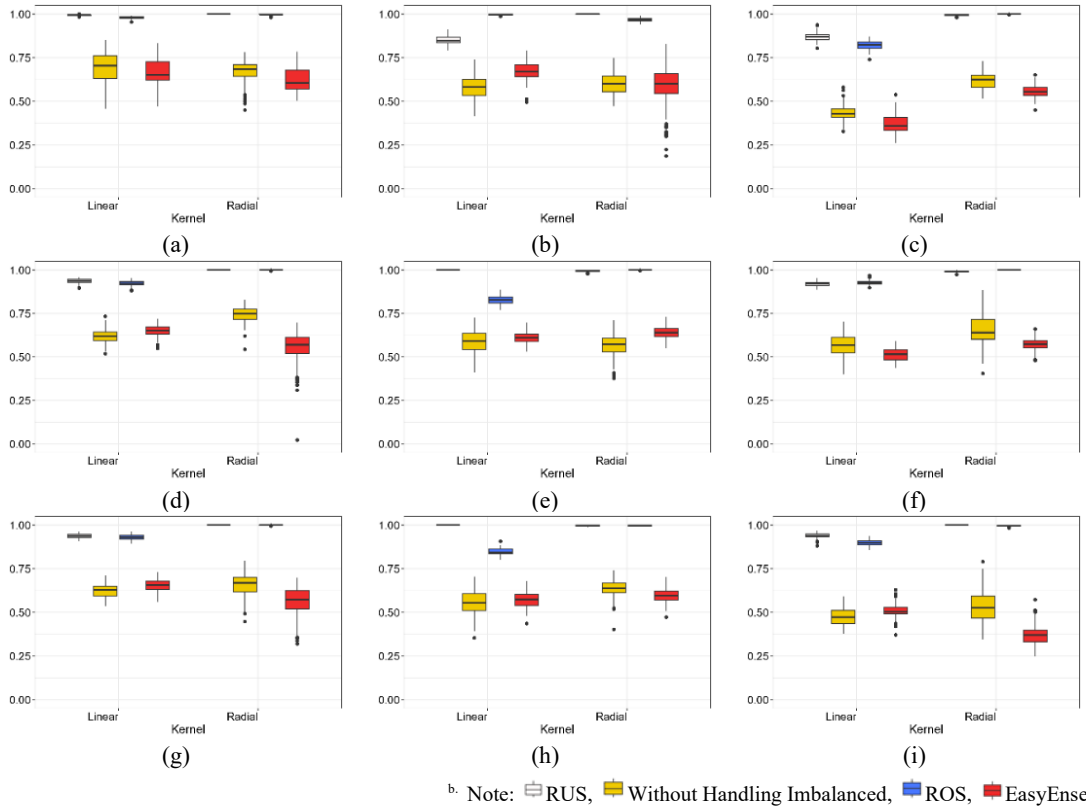

Fig. 7. Sensitivity results for (a) TF-IDF, (b) TF-IDF PCA, (c) TF-IDF LE, (d) Word2Vec, (e) Word2Vec PCA, (f) Word2VecLE, (g) Word2Vec+TF-IDF, (h) Word2Vec+TF-IDF PCA, and (i) Word2Vec+TF-IDF LE

The application of oversampling and the model without handling imbalanced data, as shown in the sensitivity results in Figure 7, indicates a tendency for both methods to predict positive reviews in both linear and radial kernels. Meanwhile, for the two undersampling methods, it is observed that RUS outperforms EasyEnsemble in predicting the positive class in some method combinations. Applying dimensionality reduction techniques like PCA and LE in models without handling imbalanced data and ROS shows similar results to models without dimensionality reduction. Conversely, using dimensionality reduction in undersampling methods has yet to improve sensitivity performance.

Figure 8 presents the specificity results, showing the prediction accuracy for negative reviews and evaluating model performance. Methods without imbalanced data handling and ROS accurately predict positive reviews but struggle with the negative class. ROS improves negative review prediction when combined with PCA and Word2Vec or Word2Vec + TF-IDF. In contrast, undersampling methods perform consistently well across all combinations, regardless of dimensionality reduction or feature extraction methods. EasyEnsemble outperforms RUS in predicting negative reviews, with the best results in several combinations that correctly predict all negative reviews. Additionally, applying LE dimensionality reduction in EasyEnsemble with TF-IDF and Word2Vec + TF-IDF feature extraction shows high specificity, enhancing negative review classification accuracy.

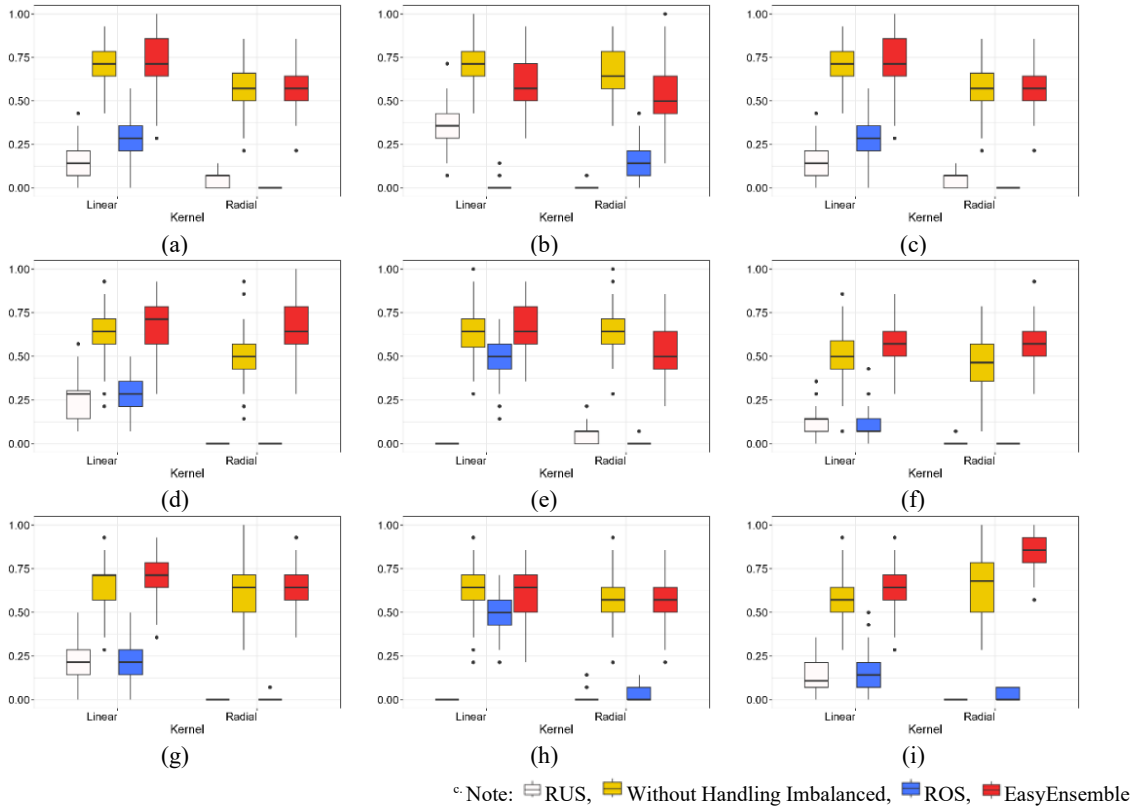

Fig. 8. Specificity results for (a) TF-IDF, (b) TF-IDF PCA, (c) TF-IDF LE, (d) Word2Vec, (e) Word2Vec PCA, (f) Word2VecLE, (g) Word2Vec+TF-IDF, (h) Word2Vec+TF-IDF PCA, and (i) Word2Vec+TF-IDF LE

The performance G-Mean, as shown in Figure 9, demonstrates that the EasyEnsemble method outperforms other approaches in several conditions. It performs best with TF-IDF feature extraction, producing a G-Mean value of 0.694, ranging from 0.476 to 0.796. While RUS performs better than ROS, its results are more balanced than those of EasyEnsemble. Although RUS reaches a maximum G-Mean of 0.8 with TF-IDF and a linear kernel, its broader boxplot range indicates inconsistent performance due to undersampling performed only once, which may result in the loss of important information. In contrast, EasyEnsemble builds models using 101 data subsets, providing more stable predictions by reducing bias and improving minority class representation.

The dimensionality reduction results in Table 6 and G-Mean performance in Figure 9 show that PCA with 85% cumulative variance (15 principal components) performs best when combined with ROS for handling data imbalance. The optimal results were achieved with Word2Vec and Word2Vec + TF-IDF feature extraction, outperforming RUS and EasyEnsemble. ROS improves classification by adding more representative data to the minority class, allowing PCA to capture more variation. However, PCA's performance with TF-IDF is less optimal due to high sparsity, as noted in a study [44]. On the other hand, LE is more effective when combined with undersampling and SVM with a nonlinear approach, preserving local relationships in nonlinear data.

Dimensionality reduction in this study did not improve prediction performance for both classes compared to models without it. LE only showed high performance for the negative class. At the same time, PCA was more effective in predicting the positive class, offering better computational efficiency. The limited data for negative reviews likely hindered the model's ability to capture relevant patterns despite addressing class imbalance. A study [20] shows that applying LE on a larger dataset improves the model's learning from negative reviews, improving the separation of positive and negative reviews.

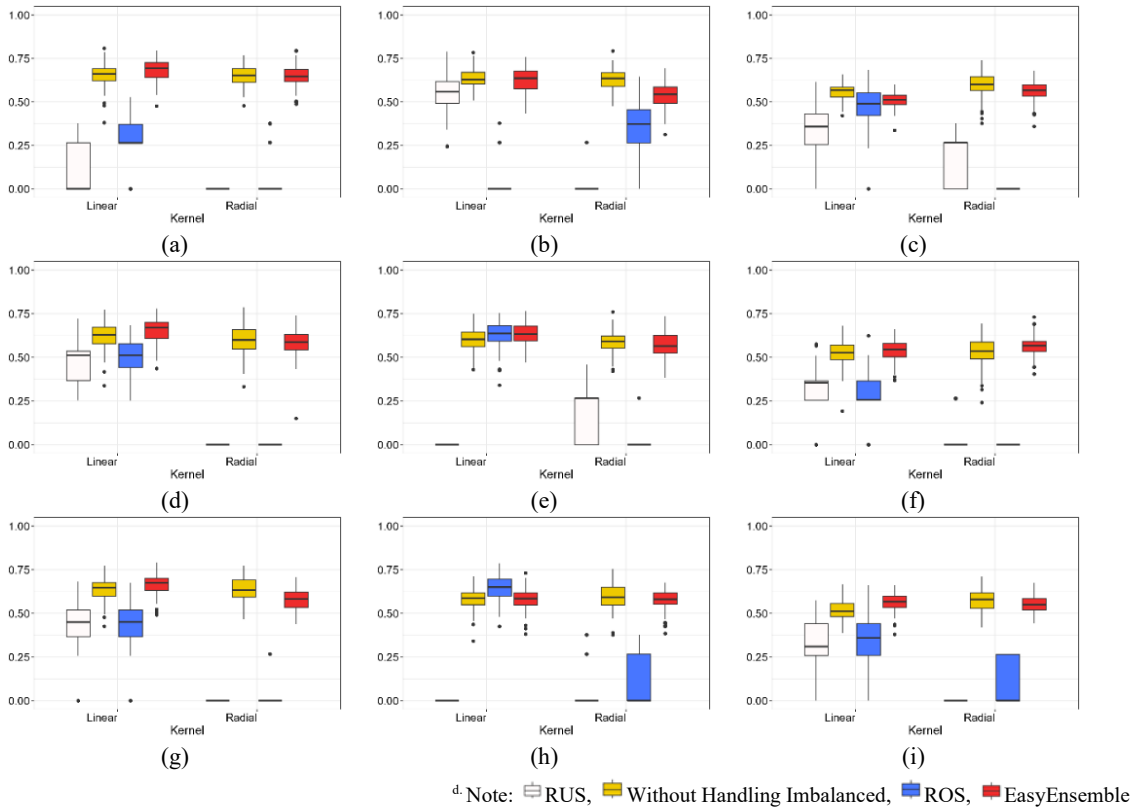

Fig. 9. G-mean results for (a) TF-IDF, (b) TF-IDF PCA, (c) TF-IDF LE, (d) Word2Vec, (e) Word2Vec PCA, (f) Word2VecLE, (g) Word2Vec+TF-IDF, (h) Word2Vec+TF-IDF PCA, and (i) Word2Vec+TF-IDF LE

Table 6. Number of features after reduction based on best G-Mean

Feature Extraction	Number of Features	Classification Method	Reduction Dimension	Reduced Features
TF-IDF	4836	RUS + SVM Radial	PCA	374
TF-IDF	4836	RUS + SVM Radial	LE	50
Word2Vec	1000	ROS + SVM Linear	PCA	15
Word2Vec	1000	EasyEnsemble + SVM Radial	LE	100
Word2Vec + TF-IDF	1000	ROS + SVM Linear	PCA	15
Word2Vec + TF-IDF	1000	RUS + SVM Radial	LE	500

IV. Conclusions

The study demonstrates that combining Manifold Learning techniques and undersampling strategies for movie review data from The Raid 2: *Berandal* has not performed best in sentiment classification. The best performance was achieved using EasyEnsemble without dimensionality reduction on an SVM with a linear kernel, showing superior specificity and average G-Mean results across all three feature extraction methods compared to other techniques. While RUS showed a relatively balanced performance with EasyEnsemble, the results were inconsistent. Using PCA combined with the ROS method improved classification performance, particularly in predicting negative reviews when paired with Word2Vec and Word2Vec + TF-IDF. Dimensionality reduction techniques such as LE did not significantly improve classification performance. However, when combined with EasyEnsemble, they did improve the prediction of negative reviews. A key limitation of this study was the low frequency of negative reviews, which affected model training and the time required for SVM hyperparameter selection. Future research is recommended to optimize SVM

hyperparameter selection and apply more advanced feature extraction techniques, such as BERT or GPT, to improve model accuracy.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering and Informatics - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] Z. Fan, Y. Guo, Z. Zhang, and M. Han, "Sentiment analysis of movie reviews based on dictionary and weak tagging information," J. Comput. Appl., vol. 38, no. 11, pp. 3048–3088, 2018.
[2] T. P. Sahu and S. Ahuja, "Sentiment analysis of movie reviews: A study on feature selection & classification algorithms," in 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), Jan. 2016, pp. 1–6.
[3] M. Govindarajan, "Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm," Int. J. Adv. Comput. Res., vol. 3, no. 4, p. 139, 2013.
[4] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts," Expert Syst. Appl., vol. 40, no. 10, pp. 4065–4074, Aug. 2013.
[5] Q. Wang, G. Zhu, S. Zhang, K. C. Li, X. Chen, and H. Xu, "Extending emotional lexicon for improving the classification accuracy of Chinese film reviews," Conn. Sci., pp. 1–20, 2020.
[6] T. A. Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. B. M. Su'ud, "Sentiment analysis using support vector machine and random forest," J. Informatics Web Eng., vol. 3, no. 1, pp. 67–75, Feb. 2024.
[7] S. Lin, R. Zhang, Z. Yu, and N. Zhang, "Sentiment analysis of movie reviews based on improved word2vec and ensemble learning," in Journal of Physics: Conference Series, Dec. 2020, vol. 1693, no. 1.
[8] A. Pandey, R. Yadav, A. Pathak, N. Shivani, B. Garg, and A. Pandey, "Sentiment Analysis of IMDB Movie Reviews," in 2024 First International Conference on Software, Systems and Information Technology (SSITCON), Oct. 2024, pp. 1–6.
[9] S. Matsumoto, H. Takamura, and M. Okumura, "Sentiment classification using word sub-sequences and dependency sub-trees," in 9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005. Proceedings 9, 2005, pp. 301–311.
[10] M. Lango, "Tackling the problem of class imbalance in multi-class sentiment classification: an experimental study," Found. Comput. Decis. Sci., vol. 44, no. 2, pp. 151–178, Jun. 2019.
[11] O. B. Deho, S. Joksimovic, J. Li, C. Zhan, J. Liu, and L. Liu, "Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why," IEEE Trans. Learn. Technol., vol. 16, no. 4, pp. 560–572, Aug. 2023.
[12] A. S. Ghareb, A. A. Bakar, and A. R. Hamdan, "Hybrid feature selection based on enhanced genetic algorithm for text categorization," Expert Syst. Appl., vol. 49, pp. 31–47, May 2016.
[13] R. Kumbhar, S. Mhamane, H. Patil, S. Patil, and S. Kale, "Text document clustering using k-means algorithm with dimension reduction techniques," in 2020 5th International Conference on Communication and Electronics Systems (ICCES), Jun. 2020, pp. 1222–1228.
[14] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Feature selection for high-dimensional data," Prog. Artif. Intell., vol. 5, no. 2, pp. 65–75, May 2016.
[15] M. Rodríguez-Ibáñez, F. J. Gimeno-Blanes, P. M. Cuenca-Jiménez, C. Soguero-Ruiz, and J. L. Rojo-Álvarez, "Sentiment analysis of political tweets from the 2019 Spanish elections," IEEE Access, vol. 9, pp. 101847–101862, 2021.
[16] S. Baehera, U. D. Syafitri, and A. M. Soleh, "Evaluasi perbandingan kinerja algoritma Cheng and church biclustering terhadap algoritma clustering klasik k-means untuk mengidentifikasi pola distribusi barang ekspor Indonesia," J. Stat. dan Apl., vol. 7, no. 2, pp. 149–161, Dec. 2023.
[17] G. Vinodhini and R. M. Chandrasekaran, "Sentiment mining using SVM-based hybrid classification model," in In Computational Intelligence, Cyber Security and Computational Models: Proceedings of ICC3, 2014, pp. 155–162.
[18] P. Verma, T. Bhardwaj, A. Bhatia, and M. Mursleen, "Sentiment Analysis 'Using SVM, KNN and SVM with PCA,'" Springer, 2023, pp. 35–53.

[19] Y. Sun and F. Zhang, "Optimization of classification results on gene expression datasets using dimensionality reduction," in CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms, Jun. 2022, pp. 1–11.
[20] K. Kim and J. Lee, "Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction," Pattern Recognit., vol. 47, no. 2, pp. 758–768, Feb. 2014.
[21] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and N. A. Khan, "Enhancing sentiment analysis via random majority under-sampling with reduced time complexity for classifying tweet reviews," Electronics, vol. 11, no. 21, p. 3624, Nov. 2022.
[22] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Using random under sampling to alleviate class imbalance on tweet sentiment data," in 2015 IEEE International Conference on Information Reuse and Integration, Aug. 2015, pp. 197–202.
[23] T. Komamizu, Y. Ogawa, and K. Toyama, "An ensemble framework of multi-ratio undersampling-based imbalanced classification," J. Data Intell., vol. 2, no. 1, pp. 30–46, Mar. 2021.
[24] J. Zhou, F. Chen, A. Khattak, and S. Dong, "Interpretable ensemble-imbalance learning strategy on dealing with imbalanced vehicle-bicycle crash data: A case study of Ningbo, China," Int. J. Crashworthiness, pp. 1–14, Mar. 2024.
[25] X. Ren, Z. Yuan, and J. Huang, "Research on fake reviews detection based on feature construction and easyensemble-rf," in 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Nov. 2021, pp. 478–482.
[26] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," Glob. Transitions Proc., vol. 3, no. 1, pp. 91–99, Jun. 2022.
[27] M. George, "Improving sentiment analysis of financial news headlines using hybrid Word2Vec-TFIDF feature extraction technique," Procedia Comput. Sci., vol. 244, pp. 1–8, 2024.
[28] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "LSTM, VADER and TF-IDF based hybrid sentiment analysis model," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 7, pp. 265–275, 2021.
[29] M. Jain, P. Goel, P. Singla, and R. Tehlan, "Comparison of Various Word Embeddings for Hate-Speech Detection," 2021, pp. 251–265.
[30] H. Wang, "Word2vec and SVM fusion for advanced sentiment analysis on Amazon reviews," Highlights Sci. Eng. Technol., vol. 85, pp. 743–749, Mar. 2024.
[31] M. Razzaghnoori, H. Sajedi, and I. K. Jazani, "Question classification in Persian using word vectors and frequencies," Cogn. Syst. Res., vol. 47, pp. 16–27, Jan. 2018.
[32] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif. Intell. Rev., vol. 55, no. 7, pp. 5731–5780, Oct. 2022.
[33] S. Nanga et al., "Review of Dimension Reduction Methods," J. Data Anal. Inf. Process., vol. 09, no. 03, pp. 189–231, 2021.
[34] N. Pospelov, A. Tetereva, O. Martynova, and K. Anokhin, "The Laplacian eigenmaps dimensionality reduction of fMRI data for discovering stimulus-induced changes in the resting-state brain activity," Neuroimage: Reports, vol. 1, no. 3, p. 100035, Sep. 2021.
[35] G. Srivastava and M. Jangid, "Multi-view Sparse Laplacian Eigenmaps for nonlinear Spectral Feature Selection," in 2023 International Conference on System Science and Engineering (ICSSE), Jul. 2023, pp. 548–553.
[36] V. R. P. Borges, "Visualizing multidimensional data based on Laplacian Eigenmaps projection," in 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Oct. 2014, pp. 1654–16593.
[37] R. Bertolini, S. J. Finch, and R. H. Nehm, "Quantifying variability in predictions of student performance: Examining the impact of bootstrap resampling in data pipelines," Comput. Educ. Artif. Intell., vol. 3, p. 100067, 2022.
[38] T. Wang, C. Lu, W. Ju, and C. Liu, "Imbalanced heartbeat classification using EasyEnsemble technique and global heartbeat information," Biomed. Signal Process. Control, vol. 71, p. 103105, Jan. 2022.
[39] J. G. Moreno-Torres, J. A. Saez, and F. Herrera, "Study on the Impact of Partition-Induced Dataset Shift on k-Fold Cross-Validation," IEEE Trans. Neural Networks Learn. Syst., vol. 23, no. 8, pp. 1304–1312, Aug. 2012.
[40] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," Decis. Anal. J., vol. 3, p. 100071, Jun. 2022.
[41] J. Bektaş, "EKSL: An effective novel dynamic ensemble model for unbalanced datasets based on LR and SVM hyperplane-distances," Inf. Sci. (Ny)., vol. 597, pp. 182–192, Jun. 2022.
[42] V. P. K. Turlapati and M. R. Prusty, "Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19," Intell. Med., vol. 3–4, p. 100023, Dec. 2020.
[43] B. Mirza, D. Haroon, B. Khan, A. Padhani, and T. Q. Syed, "Deep Generative Models to Counter Class Imbalance: A Model-Metric Mapping With Proportion Calibration Methodology," IEEE Access, vol. 9, pp. 55879–55897, 2021.
[44] R. Drikvandi and O. Lawal, "Sparse principal component analysis for natural language processing," Ann. Data Sci., vol. 10, no. 1, pp. 25–41, Feb. 2023.