

An Efficient Algorithm for Imbalanced Data Streams Management in Opinion Analysis

¹C. Ubani, ²E.O. Bennett, ³V.I.E. Anireh, ⁴D. Matthias

^{1,2,3,4}Department of Computer Science, Rivers State University, Port Harcourt, Nigeria

*E-mails: 1ubanichinyere677@gmail.com, 2bennett.okoni@ust.edu.ng, 3anireh.ike@ust.edu.ng, 4matthias.daniel@ust.edu.ng

Abstract: In many real-world streaming applications, the unequal distribution of class instances presents significant challenges to conventional machine learning algorithms, leading to biased predictions and poor detection of minority-class events. This study proposes an Efficient Algorithm for Imbalanced Data Stream Management that addresses class imbalance while maintaining high processing speed and predictive accuracy in dynamic streaming environments. Existing approaches to opinion analysis in dynamic data streams face notable limitations, particularly in managing imbalanced sentiment distributions and adapting to evolving feature spaces over time. Opinion analysis often deals with imbalanced data streams, where certain sentiments or opinions may be more prevalent than others, which can cause the model to become biased toward the more common results. Addressing imbalanced data streams is crucial to prevent biased models and ensure accurate sentiment analysis. The methodology integrates a transformer-based contextual embedding model (DistilBERT) with an adaptive autoencoder for feature compression and Synthetic Minority Over-Sampling Technique (SMOTE) to address class imbalance in streaming sentiment datasets. Concept drift in dynamic opinion data is monitored using Feature Selection Stability and Drift (FSSD) combined with the ADWIN drift detection algorithm, enabling the system to adapt to evolving sentiment patterns over time. Performance evaluation is conducted using benchmark imbalanced data stream datasets and real-time streaming data, with assessment based on classification accuracy, precision, recall, F1-score, G-mean, Area Under the ROC Curve (AUC), processing time, memory utilization, and scalability. Experimental results demonstrate that the proposed algorithm effectively enhances minority-class detection, maintains robust predictive performance under concept drift, and significantly outperforms conventional streaming classification methods in terms of efficiency and adaptability. Experimental evaluation demonstrated that the proposed system significantly improves adaptability and classification robustness compared to static feature-based streaming models, particularly in handling feature evolution and imbalanced sentiment distributions in real-time opinion analysis environments.

Keywords: Imbalanced Data Streams, Opinion Analysis, SMOTE, Concept Drift, Transformer Autoencoder, ADWIN, Feature Selection Stability and Drift.

I. INTRODUCTION

Analysing opinions in dynamic data streams with natural language processing (NLP) is an important research area. It focuses on detecting and interpreting emotions, sentiments and viewpoints expressed in written data. Sentiment analysis, often called opinion mining, is part of NLP. It aims to determine whether text conveys positive, negative, or neutral attitudes [1]. This process is essential for determining public opinion, customer feedback, and attitudes towards various topics. Researchers use machine learning, deep learning, and NLP methods to carry out sentiment analysis and interpret text [2].

A major challenge in sentiment analysis is maintaining accuracy and efficiency, since NLP tasks are often complex [3]. Researchers have explored various approaches, including the use

of advanced transformer-based language models and ensemble boosting techniques, to improve the consistency and accuracy of sentiment classification. Real-time sentiment analysis especially on platforms like Twitter has become popular because it reveals instant public reactions and opinions.

Modern digital platforms produce huge amounts of text every second from social media, news, and customer reviews. These streams capture public opinion in real time, and when analysed properly, they offer valuable insights for companies, governments and researchers. Among the associated problems, imbalanced data streams present a critical concern: opinion analysis often deals with imbalanced data streams, where certain sentiments or opinions may be more prevalent than others. Addressing imbalanced data streams is crucial to prevent biased models and ensure accurate sentiment analysis. Future research

should focus on developing specialised algorithms that can handle imbalanced data streams effectively to improve the overall performance of opinion analysis systems.

Machine learning is a fundamental technology in opinion analysis, especially in sentiment analysis tasks. Various studies have emphasised the importance of machine learning algorithms in sentiment analysis tasks. These algorithms are crucial for categorising opinions as positive, negative, or neutral, offering valuable insights into public sentiment.

This study introduces an adaptive opinion analysis framework that integrates transformer-based contextual learning with feature compression and imbalance-aware sampling techniques to improve sentiment classification performance in dynamic textual data streams. Unlike traditional static streaming classifiers, the proposed system incorporates an adaptive autoencoder for dimensionality reduction and SMOTE-based resampling within a sliding window stream environment to address evolving feature spaces and class imbalance. Additionally, a concept drift detection mechanism based on ADWIN enables continuous model adaptation to temporal sentiment variations. This approach ensures improved classification accuracy, stability, and adaptability in non-stationary real-time opinion mining environments.

II. RELATED WORKS

Opinion mining, also known as sentiment analysis, has become one of the most active research areas in artificial intelligence, natural language processing, and data mining due to the exponential increase in user-generated content across digital platforms. Organisations increasingly rely on opinion mining techniques to extract valuable insights from customer reviews, social media posts, blogs, and online discussion forums.

Class imbalance has become one of the most difficult problems in machine learning because many real-world datasets contain significantly fewer minority-class instances than majority-class observations. In streaming environments, this challenge is further complicated by the continuous arrival of new data and the inability to repeatedly access previously processed observations. Conventional classification algorithms generally assume balanced class distributions and therefore tend to maximize overall accuracy by favouring majority-class predictions. Although such models may report impressive accuracy values, they often fail to identify minority-class events, which are usually the most important observations in practical applications. Examples include fraudulent financial transactions, network intrusions, equipment failures, medical emergencies,

and rare environmental events. Consequently, researchers have increasingly focused on developing imbalance-aware learning algorithms capable of improving minority-class recognition while maintaining computational efficiency in continuous data streams.

Machine learning is a cornerstone technology in opinion analysis, particularly in sentiment analysis tasks. By leveraging machine learning algorithms, researchers and practitioners can extract valuable insights from textual data, accurately classify sentiments, and predict outcomes based on sentiment analysis. The continuous progress in machine learning techniques further enhances sentiment analysis capabilities, making it an indispensable tool for comprehending public opinions across diverse domains.

Moreover, sentiment analysis is not limited to English text but extends to multilingual sentiment analysis, where under-resourced languages are also being studied [4]. This expansion of sentiment analysis to diverse languages reflects the importance of understanding opinions and sentiments across different cultural contexts. Furthermore, sentiment analysis has been applied in various domains such as tourism, politics, and social media to gauge public sentiment towards specific events, policies, or products.

Machine learning algorithms have also played a pivotal role in predicting outcomes based on sentiment analysis. Additionally, [5] utilised machine learning models in customer sentiment analysis, analysing feedback on telecommunication products. The integration of machine learning with other technologies such as natural language processing and deep learning has bolstered sentiment analysis capabilities. Studies like [6] have illustrated the potential of machine learning and deep learning approaches in sentiment analysis of COVID-19 related tweets, underscoring the significance of hybrid approaches for enhanced analysis.

A key difficulty in analysing opinions in live data streams is concept drift: the way data patterns shift over time. This occurs when the underlying patterns in data shift substantially, causing previously-learned relationships to become outdated and less predictive. [7]. Existing research has highlighted the importance of detecting and adapting to concept drift in data streams to ensure the accuracy and relevance of opinion analysis models. [8] Introduced an online class imbalance learning framework that integrates resampling strategies with time-decayed metrics, allowing for real-time adjustments to the imbalance status as fresh data flows in. The approach enhances the classifier's ability

to adapt to evolving data distribution, thus improving outcomes in changing environments settings.

Furthermore, the integration of boosting algorithms with SMOTE, as discussed by [9], has been shown to yield superior performance metrics, like G-mean and AUC, within imbalanced datasets.

Further exploration into more robust and adaptive techniques to effectively handle different types of concept drifts that may occur in real-time data streams is needed [10]. Building efficient machine learning models for opinion analysis in dynamic data streams also requires addressing the dynamic nature of features and concepts. As data streams evolve, the importance of certain features changes, which can affect how well sentiment models perform [11].

Sentiment analysis in dynamic data streams using NLP plays a vital role in understanding and interpreting opinions and emotions expressed in textual data. By leveraging deep learning models, machine learning algorithms, and NLP techniques, researchers aim to enhance the accuracy and efficiency of sentiment analysis processes. The application of sentiment analysis is diverse, spanning across different languages and domains, highlighting its significance in capturing and analysing public sentiments in real-time scenarios [12]. Sentiment analysis using NLP and machine learning techniques was used to assess student feedback on professors [13]. Their approach successfully identified key topics and sentiments from the feedback, enabling a more structured understanding of student evaluations. This study showcased the potential of sentiment analysis in educational settings, particularly in improving faculty evaluations. However, the study's narrow focus on a specific demographic—students in a particular academic institution—limits the broader applicability of its findings. Educational contexts vary widely across institutions, and the sentiments expressed by students in one setting may not necessarily apply to others, which calls for more extensive cross-institutional research. [14] introduced a novel confidence interval-based classification approach for analyzing sentiments related to online product brands, showcasing the potential of innovative methodologies in the field of sentiment analysis. By applying confidence intervals, the study aimed to enhance the reliability of sentiment classification in the context of product reviews, providing a more nuanced understanding of consumer opinions. However, the research was constrained by its focus on specific product categories, which may limit the generalizability of its approach to other types of sentiment analysis. The results, while promising, suggest that further research is needed to determine

whether this methodology can be applied effectively across different industries and contexts.

III. METHODOLOGY

The system for managing imbalanced data streams in opinion analysis is designed to improve real-time sentiment classification by combining the contextual understanding capability of transformer networks with imbalance-aware sampling techniques. The architectural diagram illustrates a robust approach to processing dynamic data streams for real-time classification. Architectural design defines the overall framework of the system, showing its main components and how they interact. Specifically, the framework delineates the flow of data from social media or other sources into the system governing how the architecture collects, stores, processes and analyzes opinions in real – time as illustrated in Figure 1.

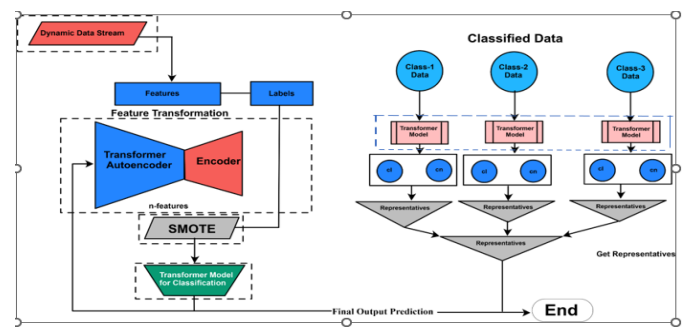


Figure 1: Architecture of the Proposed System

It begins with the continuous ingestion of raw data, which is immediately preprocessed to remove noise and extract relevant features. The data stream, denoted as D , consists of multiple feature instances $\{x_1, x_2, \dots, x_n\}$ arriving in real-time. The data points contain essential features that will be further processed for classification. Mathematically, the streaming data can be represented as:

$$D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^d, y_i \in Y; i = 1, 2, \dots, n\} \quad (\text{Eqn. 1})$$

To enhance feature representation, an Adaptive Autoencoder is employed, compressing high-dimensional data into a lower-dimensional latent space while retaining critical information. This transformation ensures that essential patterns in the data are preserved for further analysis. The transformation can be denoted as:

$$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^m \quad (\text{Eqn. 2})$$

Where m is the transformed feature space dimension, typically optimised to retain maximum variance in the data while reducing

redundancy. In the given architecture, an Adaptive Autoencoder is utilised for this transformation. The autoencoder consists of an encoder function $E(x)$ that compresses the input into a lower-dimensional latent space and a decoder function $D(z)$ that reconstructs the original input. The autoencoder is trained to minimise the reconstruction error:

$$L = \sum_i ||x_i - D(E(x_i))||^2 \quad (\text{Eqn. 3})$$

where L represents the loss function aiming to minimise the difference between the original input x_i and the reconstructed output $D(E(x_i))$.

Given that real-world datasets often suffer from class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) is incorporated to generate synthetic data points for underrepresented classes, thereby improving classification fairness and preventing bias. Given a sample x and its nearest neighbour x_{nn} , a synthetic point x_{new} is generated as:

$$x_{new} = x + \lambda(x_{nn} - x), \quad \lambda \sim U(0,1) \quad (\text{Eqn..4})$$

where $U(0,1)$ is a uniform random variable. SMOTE ensures a more balanced dataset, improving classification performance and reducing bias toward majority classes

Once the data is preprocessed and balanced, it is passed into a transformer-based classification model, leveraging self-attention mechanisms to capture intricate dependencies among features.

The model also employs class-wise data representation, wherein transformer models extract key representative instances from each class to improve interpretability and robustness. . Given an input feature vector x , the self-attention mechanism computes:

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T / \sqrt{d_k}) V \quad (\text{Eqn. 5})$$

where Q, K, V are query, key, and value matrices obtained through learned transformations. The transformer model produces a classification output y by applying a softmax activation function over the final layer:

$$y = \text{softmax}(Wx + b) \quad (\text{Eqn 6})$$

where W and b are trainable parameters. This results in a probability distribution over the possible classes, allowing accurate prediction

The integration of these advanced techniques results in a streamlined, end-to-end pipeline capable of efficiently handling real-time data streams, ensuring high classification accuracy and adaptability to evolving data patterns.

The architectural decomposition of the Efficient Algorithm for Imbalanced Data Stream Management is further modularized into interconnected modules, each performing a specific function to ensure efficient real-time opinion mining. In the management of high-velocity and imbalanced data streams, a modular component design is critical to ensuring both computational efficiency and system adaptability. Figure 2 illustrates the internal component design of the sentiment classification module. This modular configuration allows the system to process volatile data streams in real time.

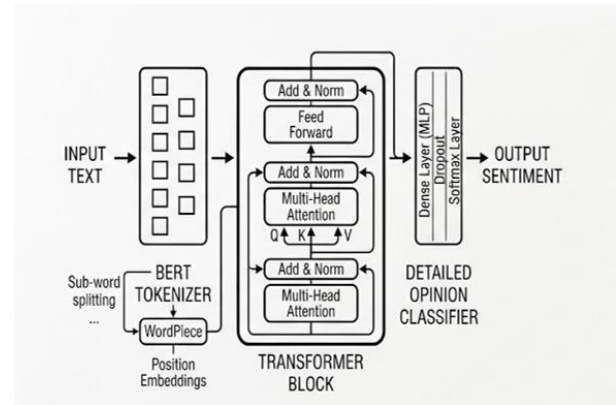


Figure 2: Component Design of the Opinion Classifier

The tokenized and embedded text sequences are routed through a deep learning pipeline consisting of a dual-stacked Transformer Block and a detailed opinion classification head.

Feature Extraction via Transformer Block

The input embeddings are first processed by the Transformer Block. The primary mechanism is the Multi-Head Attention layer, which maps Query (Q), Key (K), and Value (V) matrices to compute contextual semantic relationships. The scaled dot-product attention for a single head is mathematically expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (\text{Eqn. 7})$$

Where d_k represents the dimensional scaling factor of the keys. To capture information from different representation subspaces, the attention mechanism is extended to multiple heads.

As shown visually in the architecture block, each sub-layer is stabilized using an Add & Norm operation, which implements a residual connection followed by layer normalization:

$$\text{Output}_{\text{Norm}} = \text{LayerNorm}(x + \text{SubLayer}(X)) \quad (\text{Eqn.8})$$

Where X is the input tensor and $\text{SubLayer}(x)$ represents the transformation function of either the attention mechanism or the subsequent Feed-Forward Network (FFN). The FFN layer is applied position – wise as follows:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (\text{Eqn. 9})$$

where:

(x) represents the input feature vector,
 (W_1) and (W_2) denote trainable weight matrices,

Detailed Opinion Classifier

The dense contextual vectors generated by the final Transformer layer are passed directly into the Detailed Opinion Classifier head. This component maps the features into class probabilities across three distinct operational layers:

First, the Dense Layer (MLP) applies a linear transformation to project the high-dimensional features down to the target sentiment classes:

$$Z = X_{\text{transformer}}W_c + b_c \quad (\text{Eqn.10})$$

Where W_c and b_c denote the weight matrix and bias vector of the classification layer, respectively. To mitigate overfitting, a critical safeguard when optimizing on highly imbalanced data streams - a Dropout layer is applied to the scores by applying a masking vector:

$$Z_{\text{regularized}} = z * m \quad (\text{Eqn. 11})$$

Where m is a random masking variable with a specified dropout probability.

Finally, the features pass into the Softmax Layer, which applies an exponential normalization function the regularized score vector. This operation scales the raw outputs into a distinct probability distribution bounded between 0 and 1. The pipeline concludes at the output sentiment stage, which assigns the final sentiment label.

Algorithm 1: Opinion Analysis in Dynamic Data Streams

Require: Real-time data stream $D = \{x_1, x_2, \dots, x_n\}$

Ensure: Classified sentiment labels $Y = \{y_1, y_2, \dots, y_n\}$

- 1: Initialize streaming data ingestion module
- 2: for each data instance x_i in D do
- 3: Preprocess x_i (tokenization, stopwords removal, stemming)
- 4: Extract features F_i using Adaptive Autoencoder
- 5: $F_i = \text{Encoder}(x_i)$
- 6: if class imbalance detected then
- 7: Apply SMOTE to generate synthetic samples
- 8: end if
- 9: Detect concept drift using Feature Selection Stability and Drift (FSSD)
- 10: if drift detected then
- 11: Adjust feature representation dynamically
- 12: end if
- 13: Classify sentiment using Transformer-based model
- 14: $y_i = \text{Transformer}(F_i)$
- 15: end for
- 16: Return sentiment classifications Y

This algorithm serves as a comprehensive pipeline to handle the volatility of real-time data ingestion. It begins by streaming data instances and applying standard natural language processing (NLP) tasks, such as tokenization and stemming. The core innovation lies in its adaptive feature extraction, which utilises an Autoencoder to compress raw data into meaningful representations. To ensure the model remains robust against real-world data skews, it incorporates a SMOTE step to mitigate class imbalances. Furthermore, it employs the Feature Selection Stability and Drift (FSSD) method to monitor concept drift, allowing the system to dynamically adjust its feature representation before passing the data to a Transformer for final sentiment classification.

Algorithm 2: Online SMOTE for Data Stream Imbalance Handling

- Input: Feature set X (including new F_t), labels Y , minority class C_{min} , oversampling ratio r
- Output: Augmented features X' , augmented labels Y'
- 1: Maintain sliding-window buffer W of recent (feature, label) pairs;
 - 2: for each incoming (F_t, y_t) do

- 3: Insert (F_t, y_t) into W (drop oldest if window full);
- 4: Compute class counts in W ;
- 5: if $\text{count}(C_{\min}) < \text{desired threshold}$ then
- 6: Let $S \leftarrow \{(F_t, y_t) \in W : y_t \in C_{\min}\}$;
- 7: for $k \leftarrow 1$ to $\lceil r \cdot (\text{desired_count} - |S|) \rceil$ do
- 8: Randomly pick x_a, x_b from S ;
- 9: Generate synthetic: $x_{\text{syn}} \leftarrow x_a + \lambda(x_b - x_a)$ where $\lambda \sim U(0,1)$;
- 10: Optionally perturb and validate x_{syn} ;
- 11: Append $(x_{\text{syn}}, y_{\text{syn}} = C_{\min})$ to W ;
- 12: Set $X' \leftarrow \text{features in } W$, $Y' \leftarrow \text{labels in } W$;
- 13: return X', Y' ;

In many data streams, certain categories (like "rare" sentiments) appear much less frequently than others, which can cause the model to become biased toward the more common results. This algorithm solves that problem using an Online SMOTE technique. It maintains a "sliding window" of recent data and checks if the minority class is underrepresented. If it finds an imbalance, it creates synthetic samples by looking at existing minority examples and creating new, similar data points to fill the gap. This ensures the training buffer remains balanced, allowing the sentiment classifier to learn how to identify rare sentiments just as accurately as common ones

IV. RESULTS AND DISCUSSION

The proposed system for managing imbalanced data streams in opinion analysis was evaluated using a continuously streamed customer review dataset. Experiments were conducted in Python (Anaconda/Jupyter) on a workstation with 4.5 GHZT CPU and 8GB RAM. The transformer model was trained using the AdamW optimiser with a learning rate of 2×10^{-5} over ten epochs, achieving a validation accuracy of 94.50%, precision of 0.93, recall of 0.94, F1-score of 0.94, and ROC-AUC score of 0.79 after drift adaptation. Figures 3 and 4 illustrate the class imbalance before and after applying SMOTE where minority opinions were equalized to match majority class distribution. The confusion matrix (Figure 5) confirms improved detection of minority sentiments, reducing false negatives. These results highlight the efficiency of the algorithm in managing imbalanced data streams, ensuring fairness in opinion classification while maintaining high overall accuracy. Compared to static feature-based streaming models, the proposed approach demonstrates stronger adaptability to feature evolution and robustness against

imbalance, validating its suitability for real-time opinion analysis environments.

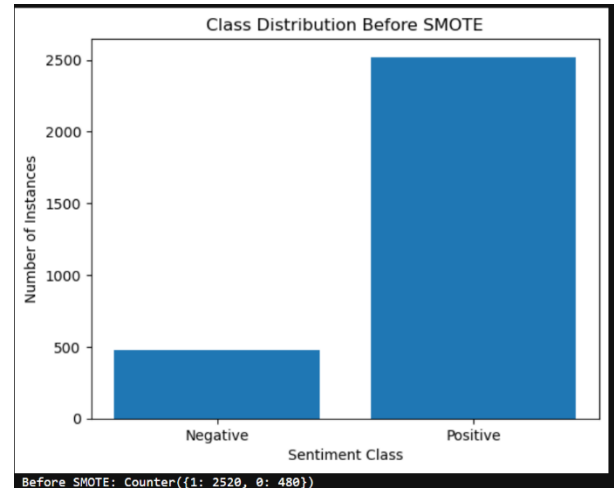


Figure 3: Data Distribution before SMOTE

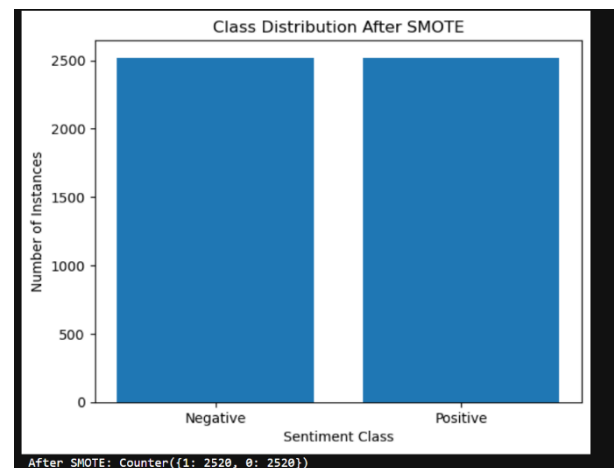


Figure 4: Data Distribution After SMOTE

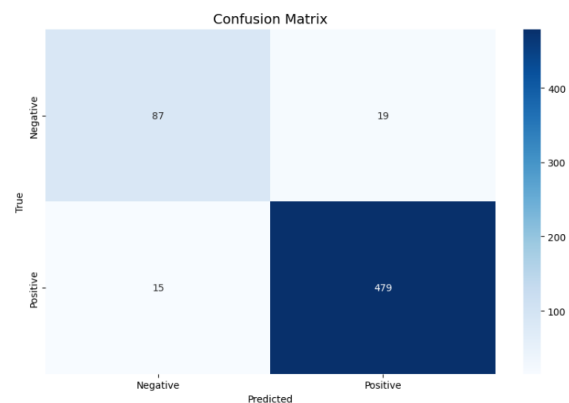


Figure 5: Confusion matrix of the model

V. CONCLUSION AND FUTURE SCOPE

This study designed and implemented an efficient algorithm for managing imbalanced data streams in opinion analysis. A key contribution is the integration of the ADWIN (Adaptive Windowing) algorithm into the class imbalance using SMOTE, which dynamically generates synthetic samples for minority sentiment classes during model updates. This minimised prediction bias toward majority classes and improved recognition of underrepresented opinions. Continuous monitoring of prediction error established a dynamic baseline for detecting distributional changes, while Feature Selection Stability and Drift (FSSD) enabled adaptive weighting of features as new opinion expressions emerged.

Although the proposed framework achieved encouraging performance, further improvements can be explored in future studies. The developed system can be applied in social media monitoring, e-commerce review analysis, and financial market sentiment analysis, where timely detection of evolving opinions provides valuable insights for decision-making in dynamic environments.

REFERENCES

- [1] Zekaoui, N. (2023). Analysis of the evolution of advanced transformer-based language models: experiments on opinion mining. *IAES International Journal of Artificial Intelligence*, 12(4), 1995-2010.
- [2] Tuama, M. (2023). Beyond polarity: the potential applications and impacts of sentiment analysis and emotion detection. *AKJS*, 1(2), 44-51.
- [3] Dang, C., García, M., & Prieta, F. (2020). Sentiment analysis based on deep learning: a comparative study. *Electronics*, 9(3), Article 483.
- [4] Mabokela, K., Çelik, T., & Raborife, M. (2023). Multilingual sentiment analysis for under-resourced languages: a systematic review of the landscape. *IEEE Access*, 11, 15996-16020.
- [5] Zhang, D., Li, S., Zhu, Q., & Zhou, G. (2020). Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning. *IEEE Access*, 8, 22945-22954.
- [6] Redjeki, S. & Widyarto, S. (2022). Comparison of seven machine learning algorithms in the classification of public opinion. *Tech-E*, 5(2), 143-149.
- [7] Chung, S. & Aring, D. (2018). Integrated real-time big data stream sentiment analysis service. *Journal of Data Analysis and Information Processing*, 06(02), 46-66.
- [8] Wang, S., Minku, L., & Yao, X. (2015). Resampling-based ensemble methods for online class imbalance learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(5), 1356-1368.
- [9] Pangastuti, S., Fithriasari, K., Iriawan, N., & Suryaningtyas, W. (2021). Data mining approach for educational decision support. *Eksakta Journal of Sciences and Data Analysis*, 33-44.
- [10] Bechini, A., Bondielli, A., Ducange, P., Marcelloni, F., & Renda, A. (2021). Addressing event-driven concept drift in twitter stream: a stance detection application. *IEEE Access*, 9, 77758-77770.
- [11] Korycki, L. & Krawczyk, B. (2021). Concept drift detection from multi-class imbalanced data streams. *arXiv*.
- [12] Bahar, K. (2023). Exploring somali sentiment analysis: a resource-light approach for small-scale text classification. *International Conference on Applied Engineering and Natural Sciences*, 1(1), 620-628.
- [13] Vargas-Calderón, V., Flórez, J., Ardila, L., Parra-A., N., Camargo, J., & Vargas, N. (2020). Learning from students' perception on professors through opinion mining. In *communications in Computer and Information Science* (330-344).
- [14] Guesmi, T. (2023). Efficient social media sentiment analysis using confidence interval-based classification of online product brands. *International Journal of Advanced and Applied Sciences*, 10(10), 94-102.



Citation of this Article:

C. Ubani, E.O. Bennett, V.I.E. Anireh, & D. Matthias. (2026). An Efficient Algorithm for Imbalanced Data Streams Management in Opinion Analysis. *Journal of Artificial Intelligence and Emerging Technologies (JAJET)*. 3(8), 1-8. Article DOI: <https://doi.org/10.47001/JAIET/2026.308001>

***** End of the Article *****