Vol. 7, No. 2, August 2025, pp. 195~207, p-ISSN 2685-2381, e-ISSN 2715-2626
http://dx.doi.org/10.28989/avitec.v7i2.3023

Enhancing Aspect-Based Sentiment Analysis in Imbalanced Multilabel Datasets using Resampling and Classifiers for Digital Signature Applications

Efriza Cahya Narendra*, Amalia Anjani Arifiyanti, Tri Luhur Indayanti Sugata
Department of Information Systems, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia

Article Info

Article history:
Submitted May 25, 2025
Accepted June 12, 2025
Published June 23, 2025

Keywords:
Aspect-based sentiment analysis;
multilabel classification;
imbalanced data resampling;
digital signature;
e-stamp.

ABSTRACT

Amid the growing demand for digital identity solutions, applications like Privy, VIDA, and Xignature offer integrated digital signature and e-stamp services, generating extensive user feedback on platforms like Google Play Store and App Store. Extracting meaningful insights from thousands of reviews is challenging, necessitating effective sentiment analysis. Aspect-Based Sentiment Analysis (ABSA) enables detailed sentiment evaluation by linking user feedback to specific aspects and sentiments. However, ABSA faces challenges with imbalanced datasets where label distributions are uneven. This study explores the application of three resampling techniques, including MLROS, MLSMOTE, and REMEDIAL, to address this issue in multilabel classification. Using multilabel classifiers, including Binary Relevance, Label Powerset, and Classifier Chains, the study systematically evaluates their performance. Results reveal that resampling significantly enhances outcomes, with MLROS and Classifier Chains under a 70:30 split achieving the best performance, reducing Hamming Loss to 0.0401 or 95% accuracy. This marks a 34.2% improvement over baseline models without resampling or classifiers. The model generalizes well to unseen data with minimal overfitting, as indicated by validation results. These results underscore the importance of imbalanced data resampling and multilabel classification techniques in advancing ABSA, offering valuable insights for improving sentiment analysis in real-world applications.

Corresponding Author:
Efriza Cahya Narendra,
Department of Information Systems, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294
Email: *21082010059@student.upnjatim.ac.id

1. INTRODUCTION

In the rapidly advancing digital era, digital signatures and e-stamps have become crucial for enabling efficient, secure, and legally recognized document validation in both governmental and private sectors. In Indonesia, this transformation is supported by Presidential Regulation No. 95 of 2018 on Electronic-Based Government Systems (SPBE), mandating the adoption of electronic systems in government operations. A concrete of example the implementation of digital signatures can be observed in the Bogor City Government's integration of e-Letters with digital signatures, enhancing document security, efficiency, and tracking [1]. Additionally, the 2024 Civil Service Candidate (CPNS) selection process mandates the use of e-stamp for registration [2], significantly increasing demand.

Amidst the growing demand for reliable digital identity solutions, several apps, such as Privy, VIDA, and Xignature, have emerged as integrated digital signature and e-stamp service providers. These apps have been officially registered with the Electronic Certification Provider (PSrE) under the Ministry of Communications and Information Technology, which ensures that they comply with the established security standards. Available on widely used platforms, such as the Google Play Store and App Store, these applications have received varying reviews and ratings. For instance, Privy has achieved over one million downloads on the Google Play Store, accompanied by thousands of user ratings and reviews. These evaluations play a pivotal role in shaping user perceptions and can influence prospective users when deciding whether to download the application [3].

Understanding user feedback is essential for evaluating the quality of services and features provided by these applications. Unfortunately, it can be challenging to effectively understand user needs and complaints from the thousands of reviews available. Therefore, to gain deeper insights into the sentiment embedded in reviews

and ratings, sentiment analysis techniques are employed. Sentiment analysis is a domain within text mining that focuses on examining opinions, sentiments, evaluations, judgments, attitudes, and emotions expressed about specific topics, services, products, individuals, organizations, or activities. The primary goal is to classify textual data based on the polarity of sentiment it conveys [4].

In sentiment analysis, identifying the polarity within documents, sentences, or opinions is fundamental. However, a more granular approach is often necessary to derive actionable insights, especially when analyzing detailed user feedback. This calls for aspect-level sentiment analysis, also known as Aspect-Based Sentiment Analysis (ABSA). ABSA delves deeper by identifying specific aspects of a product or service that are important to users and determining the sentiment associated with each aspect. Conducting this type of analysis is crucial as it serves as a quality indicator for the services or features being evaluated [5]. By focusing on particular aspects and their corresponding sentiments, ABSA provides a nuanced understanding of user perceptions, enabling more targeted improvements and decision-making.

In this study, the output of Aspect-Based Sentiment Analysis (ABSA) takes form of multilabel classification. Multi-Label Classification (MLC) represents a broader approach to classification, different from standard single-label classification, where a single instance can simultaneously be associated with several labels [6]. In this context, each review can be associated with more than one aspect, with corresponding sentiments for each aspect. The primary challenge in MLC is the imbalance in multi-label dataset, where samples and their associated labels are unevenly distributed across the dataset. The approaches for standard single-label classification differ significantly from those used in multilabel classification. In multilabel classification, where each instance can be associated with multiple labels, the imbalance issue is compounded due to the multi-dimensional nature of the labels.

Charte et al. introduced three methods, there are Multilabel Random Oversampling (MLROS), Multilabel Synthetic Minority Over-sampling Technique (MLSMOTE), and REsampling MultilabEl datasets by Decoupling highly ImbAlanced Labels (REMEDIAL) to address data imbalance in multilabel classification. These methods have been applied in various domains, including mitigating label imbalance in predicting adverse drug reactions [7], enhancing the hierarchical multilabel classification of research papers [8], resolving mislabeling issues in Stack Overflow tagging posts [9], and refining the classification of individual hosts observed in the darknet [10]. Notably, MLROS, MLSMOTE, and REMEDIAL were also employed in the medical domain to improve deep learning predictions for chest X-ray abnormalities [11], where performance was evaluated using metrics including Hamming Loss, which aligns with this study's evaluation metric. Their findings indicated REMEDIAL achieved the best overall performance across classifiers, with VGG16 outperforming others such as DenseNet and CNN, reaching a highest accuracy of 48% and hamming loss of 0.0324. These results underscore the critical role of resampling methods in enhancing multilabel classification on imbalanced datasets.

However, the resampling methods application within the domain of ABSA remains largely unexplored. Furthermore, comprehensive comparative evaluations of these methods in the ABSA context are scarce. This gap is particularly significant given the unique challenges posed by ABSA, where interactions between aspect categories and sentiment polarity can lead to pronounced label imbalances as certain aspects or sentiments may be more prevalent than others. The interdependencies between these labels also complicate the classification task, as a sentiment label for one aspect can influence, or be influenced by, the sentiment of another.

This study seeks to bridge the gap by being the first to apply MLROS, MLSMOTE, and REMEDIAL specifically to multilabel datasets derived from user feedback in the context of ABSA. For the classification of resampled datasets, this study explores the implementation of multiple classification classifiers to assess their ability to enhance classification outcomes. Through a systematic comparison of these methods, this research aims to evaluate their effectiveness in addressing data imbalance and improving multilabel classification performance. By focusing on real-world sentiment data, this study contributes both to the theoretical understanding and the practical application of imbalance-handling techniques in ABSA. Ultimately, the findings aim to provide insights into selecting appropriate resampling and classification strategies for improving sentiment analysis in highly imbalanced multilabel contexts.

2. RESEARCH METHODS

2.1 Research Model

The research model used in this study is structured into several systematic steps to ensure a comprehensive analysis and effective implementation. The process begins with data collection, data preparation, topic extraction, data labeling, data exploration, data splitting, multilabel classification model building, and model evaluation. All processes were conducted using Google Colab, a cloud-based platform that provides a flexible and scalable environment for computational tasks. The complete research workflow is illustrated in the Figure 1, outlining each stage in detail to provide a clear understanding of the methodological framework.

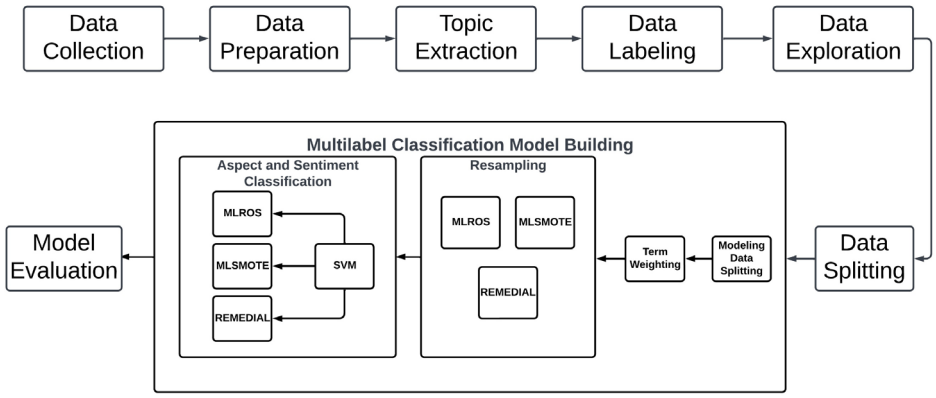

Figure 1. Research model

2.2 Data Collection

During this stage, user reviews for the Privy, Vida, and Xignature applications were gathered from the Google Play Store and App Store using web scraping techniques. For extracting reviews from the Google Play Store, the `google-play-scraper` package was utilized, while the `app-store-scraper` package facilitated data collection from the App Store. These tools ensured efficient and structured retrieval of user feedback for subsequent analysis. These tools not only facilitated efficient data retrieval but also ensured that the reviews were organized in a structured format, enabling seamless preprocessing and analysis in following stages.

2.3 Data Preparation

In this stage, data preparation is performed a series of systematic steps aimed at ensuring the data is clean, consistent, and properly formatted for detailed analysis and modeling. This phase encompasses a series of structured steps designed to ensure the data is clean, uniform, and well-suited for further processing. All data preparation tasks were implemented using the Natural Language Toolkit (NLTK) library in Python, which provides comprehensive tools for text processing and natural language analysis. Each of these processes is outlined below:

a. Data Cleaning: Ensure the quality of the dataset by identifying and removing errors or inconsistencies [12]. In this study, data cleaning involved eliminating duplicate and short reviews fewer than three words. This requirement is applied to exclude reviews that are unlikely to provide enough information for meaningful sentiment and aspect analysis.
b. Text Cleaning: Includes various techniques aimed at addressing imperfections in raw text data, such as removing unnecessary characters, symbols, or formatting issues [13]. This was achieved using regular expressions, implemented through the `re` library.
c. Case Folding: Standardizes text by converting all characters to lowercase, aims to help reduce unnecessary variations, such as distinguishing between uppercase and lowercase letters [14].
d. Tokenization: Dividing the text into smaller, meaningful units called tokens, which facilitate further analysis [14]. This study utilized the `word_tokenize` function from the NLTK library to segment user reviews into tokens.
e. Normalization: Convert non-standard text into its standard form to ensure consistency [15]. For this study, normalization is referred to the “kamusalay.csv” sourced from the GitHub repository by Nikmatun Aliyah Salsabila.
f. Stopword Removal: Eliminates common words that add little value to the analysis, ensuring the focus remains on meaningful content [14]. In this study, the stopword list is based on the `stopwords` module of NLTK, with custom additions such as "application" word and the application names.
g. Stemming: Reduces words to their root form, ensuring consistency and aiding in the recognition of similar terms [14]. This study utilized the NLTK's `PorterStemmer` module to perform stemming and `Sastrawi` library to ensure that the words generated are in accordance with the applicable Indonesian language rules.

2.4 Topic Extraction

After data preparation, topic extraction is performed to identify the dominant topics frequently mentioned in the review data. This study utilized Latent Dirichlet Allocation (LDA), an unsupervised topic modeling technique that explores the relationships among words, topics, and documents by assuming that documents are generated through a specific probabilistic model [16]. LDA enables the identification of topics or aspects commonly discussed by users in their reviews, which are then analyzed further. To determine the optimal number of topics, a coherence score is calculated. This score evaluates how well the topics generated by the LDA model align with their context in the data. In this study, the number of topics tested ranges from 2 to 11, with a

total of 10 iterations for each configuration. The LDA process was conducted using the `Gensim` library, which was installed specifically for this analytical task.

2.5 Data Labeling

This stage is carried out after the aspects to be analyzed are determined. The keywords generated from the topic extraction process become the basis for determining whether a review belongs to a particular aspect. This labeling process is done automatically using the OR operator to associate reviews with relevant aspects. By employing this approach, the system ensures that a review can be categorized into more than one aspect if it contains at least one related keyword.

After identifying the aspects for each review, sentiment labeling is performed to classify the sentiments associated with the identified aspects. This process offers three label options: 0 for irrelevant, 1 for negative sentiment, and 2 for positive sentiment. To ensure accuracy and consistency, three independent annotators were involved. Before labeling, the annotators received a comprehensive guide detailing the procedure, the scope of each aspect, also the definition and criteria for each sentiment label. An overview of the data labeling can be seen in Table 1 as follows.

Table 1. Overview of Data Labeling

Reviews	Login and Verification	Efficiency	User Services	Responsiveness
The purchase part of the e-stamp was fine, but the time of affixing the error kept occurring. Now it's even maintenance of the system, if you can't do it, don't sell the e-stamp yet. Just answer here, do not bother using email, either it is not answered.	0	2	1	1
Your application is horrible. Very long maintenance, data verification is also complicated...	1	1	0	1
thanks for helping... even though I had to verify my face multiple times 😅	1	2	0	0

Labeling results were evaluated for reliability using Krippendorff's Alpha, a general nonparametric measure of inter-rater agreement. As defined by Hayes and Krippendorff, this metric assesses the consistency among two or more annotators evaluating the same units of analysis [17]. Krippendorff's Alpha provides a comprehensive measure of agreement by accounting for the magnitude of errors made during the labeling process [18]. Its flexibility allows it to accommodate various data types, including nominal, ordinal, interval, and ratio scales, making it a versatile tool for reliability assessment. In this study, Krippendorff's Alpha was calculated using the Python `krippendorff` library, ensured an efficient and accurate assessment of agreement levels. By employing this metric, potential biases can be minimized, ensuring that the labeled data meets high reliability standards.

Once the reliability of the labels was confirmed, the dominant sentiment for each aspect was then determined using saturation techniques. This approach identifies the most frequently occurring sentiment within a particular aspect, providing a consistent and representative classification. For reviews where no dominant sentiment could be discerned, an 'equal' label was assigned to reflect the balance in sentiment values accurately. It ensures the fairness of the sentiment classification process.

2.6 Data Exploration

After labeling, the dataset is explored to uncover underlying patterns and trends, offering valuable insights for the modeling phase. The frequency of aspects and their corresponding sentiments is visualized through bar charts generated using the `Matplotlib` library. These visualizations effectively highlight the most discussed aspects, offering a clear view of user priorities and concerns. Additionally, the charts provide an overview of whether users perceive these aspects positively or negatively, serving as a foundation for understanding the sentiment overview.

2.7 Data Splitting

At this stage, the dataset is partitioned into two different subsets to facilitate model building. The main objective is to separate the data used to train the model from the data reserved for validation. This separation is crucial to ensure that the model works effectively not only on the training data but also on previously unseen data, thereby demonstrating its generalization ability. In this study, the dataset is split using a 90:10 ratio, where 90% of the data is allocated for modeling, and the remaining 10% for validation.

2.8 Multilabel Classification Model Building

2.8.1 Modeling Data Splitting

At this stage, the modeling dataset is further divided into two subsets: training data and testing data, following the predetermined scenarios. This step is essential to ensure that the model learns from the training data while being evaluated on unseen testing data. By doing so, the model's performance can be assessed more accurately in terms of its ability to generalize to new, unseen data. In this study, two data-splitting ratios were employed: 80% for training data and 20% for testing data, as well as 70% for training data and 30% for testing data. These ratios were selected to analyze the impact of different data splits on model performance and to ensure a robust evaluation framework.

2.8.2 Term-Weighting

This stage transforms each word in the review with a weighted based on its frequency in the document and its importance in the overall dataset. The term-weighting method used in this study is Term Frequency-Inverse Document Frequency (TF-IDF), a weight statistic used to measure the importance of a word in the context of a particular document in a collection or corpus [19]. TF-IDF will be applied to both the training and testing datasets to enhance the identification of relevant words for the classification process. This technique assigns greater importance to words that frequently appear in individual documents while adjusting for their prevalence across the entire corpus. In this study, TF-IDF implementation is facilitated using the `TfidfVectorizer` module from `sklearn` library.

2.8.3 Resampling

In measuring data imbalance, it is crucial to evaluate the extent to which the dataset indicates an unbalanced distribution of labels. This can be done using specialized metrics such as Imbalance Ratio per Label (IRLbl) and Mean Imbalance Ratio (MeanIR), as proposed in relevant research [20]. IRLbl refers to the ratio between the number of data on a particular label and the number of data on the majority label, as formulated in Equation (1). Meanwhile, MeanIR is the average of all IRLbl values in a multilabel dataset, reflecting the average degree of imbalance between labels in the dataset, as described in Equation (2).

$$IRLbl(l) = \frac{\max_{l' \in L}(\sum_{i=1}^{|D|} [\![l' \in Yi]\!])}{\sum_{i=1}^{|D|} [\![l' \in Yi]\!])} \quad (1)$$

$$MeanIR = \frac{1}{|L|} \sum_{l \in L} IRLbl(l) \quad (2)$$

The results of the IRLbl and MeanIR calculations serve as a valuable basis for determining which labels to sample. These metrics provide insight into the degree of imbalance in the dataset, highlighting labels with significantly lower occurrences that may affect the performance of the classification model. According to [21], labels with IRLbl values higher than MeanIR are categorized as minor labels and become the main focus for resampling. The resampling methods employed in this study consist of several carefully selected algorithms tailored to address the imbalance in multilabel datasets. These methods, proposed by Charte et al., include:

a. MLROS

As introduced in [20], MLROS (Multilabel Random Over-sampling) identifies the minority label instance by assessing whether its IRLbl is higher than MeanIR. These instances are then added to bags of instances, from which random samples are selected, cloned, and used to generate new instances.

b. MLSMOTE

As detailed in [22], MLSMOTE (Multilabel Synthetic Minority Over-sampling Technique) generates synthetic instances by initially selecting minority instances based on its IRLbl values, identifying their nearest neighbors, then generating new feature trough interpolation. Finally, a synthetic label set is assigned to the newly created instance.

c. REMEDIAL

Described in [21], REMEDIAL (REsampling MultilabEl datasets by Decoupling highly ImbAlanced Labels) aims to minimize concurrence level in datasets by decoupling instances that exhibit high SCUMBLE values. SCUMBLE, introduced in [23], is a metric used to measure the degree of concurrence between imbalanced labels in multilabel datasets. REMEDIAL works by duplicating these instances and splitting them into two separate instances, one associated with majority labels and the other with minority labels, to reduce the concurrence level.

2.8.4 Aspects and Sentiment Classification

At this stage, a classification model is developed using Support Vector Machine (SVM) [24] to classify user reviews by their associated aspects and sentiment, either positive or negative. The model is trained using training data to learn underlying patterns of aspects and sentiments. SVM focuses on finding the optimal hyperplane to distinctly separate the aspect and sentiment classes, ensuring accurate predictions for testing data.

In this study, the implementation is carried out using the `LinearSVC` module from the `sklearn.svm` library. Furthermore, SVM will be paired with various multilabel classifiers according to the approaches detailed in [25], including:

a. Binary Relevance (BR)

Binary Relevance divides the multilabel learning (MLL) problem into multiple binary classification tasks, where each label is handled independently. Each label is treated independently as a standalone binary classification problem, ignoring potential dependencies among labels. The Binary Relevance implementation is facilitated using the `BinaryRelevance` module from the `skmultilearn` library.

b. Label Powerset (LP)

Label Powerset addresses multilabel classification by considering label correlations. It does so by converting the multilabel problem into single-label classification task. This approach treats every subset of labels in the training data as a unique class, called a "label set". In this study, the application of the Label Powerset approach is enabled through the `LabelPowerset` module.

c. Classifier Chain (CC)

Classifier Chain utilizes a sequence of binary classifiers to tackle multilabel classification challenges. Each classifier in the chain is responsible for predicting the label at a certain level and expanding the features with predicted results based on previous labels, thus allowing this method to effectively capture dependency among labels. The Classifier Chain approach is implemented using the `ClassifierChain` module available in the `sklearn` library.

2.9 Model Evaluation

The final step in this study is to evaluate the model with best performance using unseen data to assess its ability to classify aspects and sentiments for each review in the testing data. This evaluation utilizes the Hamming Loss metric, which is commonly used in multilabel classification. Hamming Loss is calculated as the ratio of symmetric difference between the actual and predicted labels and divided by the total labels in the dataset. The closer the Hamming Loss is to zero, the better the model performs [6]. This evaluation is conducted using the `hamming_loss` function from the `sklearn.metrics` library.

The model's susceptibility to overfitting will be assessed by comparing its Hamming Loss on training and validation data. Unlike training data, validation data is exclusively used to evaluate the model's performance without influencing the training process. A minimal difference in its percentages suggests that the model is not overfitted. The validation step is crucial to confirm that the model remains accurate in real-world scenarios.

3. RESULTS AND DISCUSSION

3.1 Data Collection Results

The reviews scraped from all targeted applications were merged into one unified dataset. This process resulted in 32,763 user review entries, which offered a solid basis for further analysis. Table 2 provides a breakdown of the data obtained at this stage.

Table 2. Data count details

Platform	Total Data
Google Play Store	12.147
App Store	20.616

3.2 Data Preparation Results

Following data collection, the dataset was prepared to meet the requirements for model building. From the initial total of 32,763 entries, the dataset was reduced to 8,250 entries after filtering for duplicates and short reviews. A detailed summary of the filtering process and its impact on the dataset size is presented in Table 3, highlighting the transformation from raw data to a refined dataset ready for subsequent stages.

Table 3. Data filtering details

Process	Total Data
Data Collection	32.763
Elimination of Data Duplication	9.613
Elimination of Short Reviews (<3 Words)	8.250

The dataset was subsequently processed through a series of text preparation steps to enhance its quality and suitability for modeling. These steps included text cleaning, case folding, tokenization, normalization, stopword removal, and stemming. The final result consists of tokens where each word is reduced to its base form. Table 4 provides a detailed comparison of the dataset before and after data preparation, showcasing the significant improvement in data.

Table 4. Comparison of data preparation results

Before Data Preparation	After Data Preparation
Very easy to use for work, but the performance needs further improvement. Thank you 😊	['easy', 'work', 'performance', 'improve', 'thank you']

3.3 Topic Extraction Results

Based on the coherence score evaluation, the optimal number of topics was identified. It highlighted the 4-topic model as the optimal choice. The results of the Latent Dirichlet Allocation (LDA) process identified four distinct topics, each associated with a list of relevant keywords. These keywords were carefully analyzed to assign descriptive and meaningful names to the aspects, ensuring alignment with the context of the user reviews. Table 5 summarizes the identified aspects and their corresponding keywords, providing a clear overview of the topics derived from the dataset.

Table 5. Identified aspects

Aspect Category	Keywords
Login and Verification	Register, Email, Sign in, Verification, Login, ID, NIK (National Identity Number), Failed, Account, Change, Photo, Password, ID card, Data, Retry, Selfie, Liveness, Camera, Register, Registration, Face
Efficiency	Easy, Assist, Sign, Signature, Good, Digital, Complicated, Doc, Work, App, Use, Application, Cool, Excellent, Recommended
User Services	Response, Fast, Service, Customer Service, Chat, Send, DM, Instagram, Help, Reply, Live, Friendly, Solution, Helpdesk, Admin, Support, Complaint
Responsiveness	Open, Update, Connection, Download, Network, App, YouTube, Error, Internet, Bug, Slow, Install, Black screen, Laggy, Loading, Server, Lag, Upgrade, Store, Maintenance

The Login and Verification aspect, identified through keywords such as "register", "email", "login", "verification", and "ID", emphasizes the functionality of account access, registration, and user authentication processes, which are pivotal to the initial user experience. For instance, reviews such as "The selfie verification keeps failing even though the photo quality is clear" or "I couldn't log in after updating my account details" highlight common challenges users face in ensuring a seamless login and verification process.

The Efficiency aspect, characterized by terms like "easy", "complicated", "signature", "digital", and "doc", reflects user evaluations of the application's ease of use and its capability to facilitate tasks efficiently. Users often express sentiments such as "The app is very easy to navigate and makes signing documents quick" or "The process is too complicated and takes longer than expected," indicating their perceptions of the app's practicality in streamlining workflows.

The User Services aspect, highlighted by keywords such as "response", "fast", "service", "help", and "customer service", focuses on user interactions with customer support and the quality of assistance provided. Feedback like "The customer service team was very helpful and resolved my issue quickly" or "It takes too long to get a response from the support team" exemplifies the range of experiences users have with service quality and responsiveness.

Meanwhile, the Responsiveness aspect, indicated by terms such as "open", "update", "connection", "error", and "download", captures technical challenges related to the application's performance and stability. Users frequently report issues such as "The app keeps lagging and doesn't load properly on my device" or "After the update, the connection is much smoother," reflecting their experiences with system reliability and responsiveness.

3.4 Data Labeling Results

The aspect labeling process was followed by sentiment labeling conducted independently by each annotator. The reliability of these annotations was assessed using Krippendorff's Alpha. The Krippendorff's Alpha values were calculated for each aspect, with the results presented in Table 6.

Table 6. Krippendorff's alpha value per aspect

Aspect Category	Krippendorff's Alpha
Login and Verification	0.9793
Efficiency	0.9728
User Services	0.9746
Responsiveness	0.9762

Since all alpha values exceeded 0.97, this indicates a high level of agreement among the three annotators in assigning sentiment labels. It can thus be concluded that the labeling guidelines provided were both clear and

consistently applied. Therefore, the labeled data is considered trustworthy for advancing to the next stage of the research. Upon validating the reliability of the sentiment annotations, the next process was to assign the dominant sentiment label as the final label for each review.

3.5 Data Exploration Results

A visualization of sentiment distribution for each aspect was conducted to provide insights into potential data imbalances. Besides helping to highlight which aspects are most frequently discussed and whether users view them positively or negatively, this step also identifies whether certain sentiment labels dominate or are underrepresented within specific aspects. By revealing these imbalances, the visualization offers a clearer understanding of the dataset's composition, which is crucial for ensuring balanced and fair training during model building. The results of this distribution analysis are presented in Figure 2, illustrating the proportions of each sentiment label across the defined aspects and their significance.

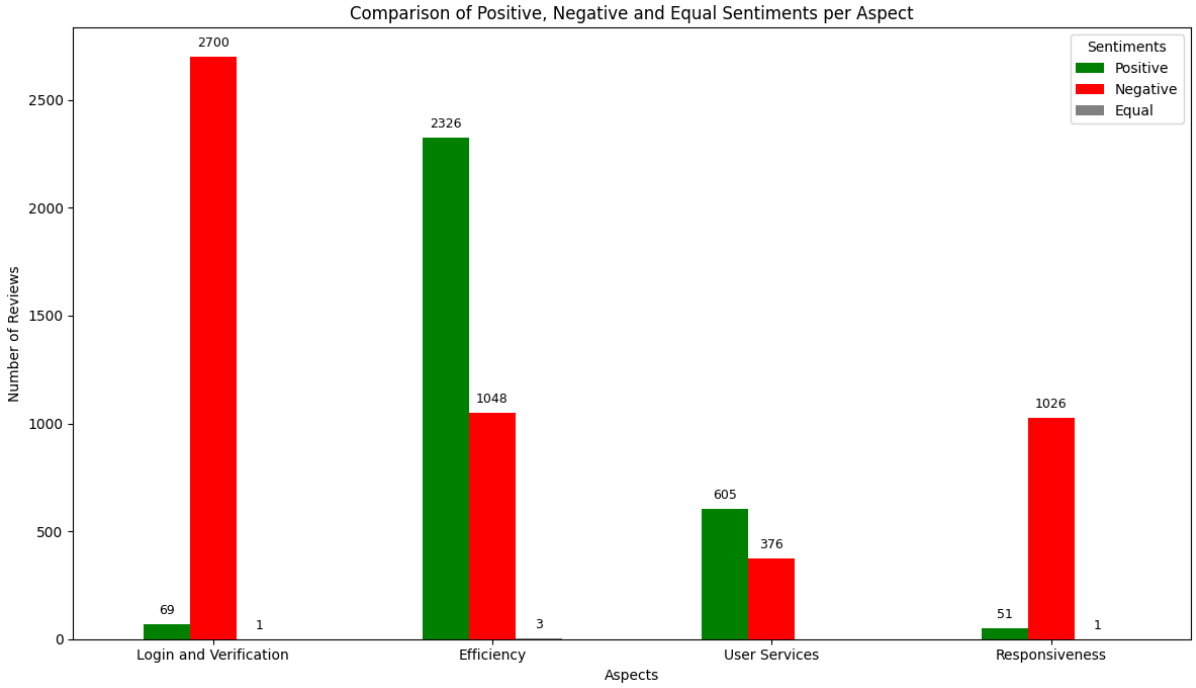

Figure 2. Sentiment distribution chart per aspect

The visualization reveals that, across all aspects, the negative sentiment label is notably dominant, indicating a significant imbalance in the dataset. Additionally, the dataset includes an "equal" label, which reflects instances where the labeled review lacks a dominant sentiment. This situation occurs when annotators have differing opinions about the sentiment expressed in a review. Given the relatively small number of reviews assigned the "equal" label, it was determined that these instances would be excluded from the analysis.

3.6 Data Splitting Results

In this study, the holdout method is applied to divide the dataset into 90% model data and 10% validation data. The validation data is stored in a separate CSV file and is later used as test data during the model evaluation phase. Both datasets are then split into feature columns and target columns. The feature columns include “review” and “stemmed_review,” while the target columns consist of the labeled data, namely “login_and_verification,” “efficiency,” “user_service,” and “responsiveness.” To preserve the label proportions and minimize significant imbalances in the validation data, an iterative splitting technique is utilized. The detailed distribution of data allocated for the model and validation sets is presented in Table 7.

Table 7. Detailed data distribution

Data Type	Total Data
Modeling Data	6.268
Validation Data	691

3.7 Multilabel Classification Model Building Results

The prepared modeling data was then split into two scenarios for the model building phase, there are 80% training data and 20% testing data, as well as 70% training data and 30% testing data. In both scenarios, the feature columns were processed using term weighting with the TF-IDF technique. This process transforms the

feature columns into numerical representations as needed for model building, emphasizing the relevance of terms based on their frequency and importance in the dataset.

The labeling result's IRLbl and MeanIR metrics are calculated as a preliminary step of the resampling process. This evaluation aimed to determine which labels exhibited imbalances and required specific interventions. As the positive and negative labels for each aspect are combined into a single column, these calculations were conducted separately for each aspect. The detailed results are provided in Table 8.

Table 8. Imbalance ratio metric calculation results

Metric		Result
MeanIR		15.76914
IRLbl	Login and Verification	39.13043
	Efficiency	2.21946
	User Services	1.60904
	Responsiveness	20.11764

The MeanIR across all labels or aspects is 15.76914, suggesting that on the average, the dataset contains more than 15 times as many samples in the majority label as in the minority label. Based on the calculation results, it is evident that the Login and Verification and Responsiveness aspects have IRLbl values exceeding the MeanIR, indicating a significant imbalance in these aspects. On the other hand, the Efficiency and User Services aspects display values that are relatively closer to being balanced.

After identifying the imbalanced aspects, the next step involves resampling and implementing resampling algorithms. During this stage, major and minor labels are classified, and resampling is performed on the minor labels. Following this stage, the process proceeds with model building based on the predetermined scenarios, ensuring each scenario is thoroughly evaluated for performance.

3.8 Multilabel Classification Model Results

3.8.1 Scenario 1: Baseline Model Without Implementing Resampling Methods and Classifiers

In this baseline scenario, the SVM algorithm was directly applied to the training and testing datasets for both split ratios, 80:20 and 70:30. This configuration is designed to serve as a baseline reference, providing a fundamental benchmark for assessing the impact of additional scenarios such as resampling methods or multilabel classifiers. The detailed evaluation results for this baseline configuration are presented in Table 9.

Table 9. Model performance in Scenario 1

Split Scenarios	Hamming Loss
80:20	0.0625
70:30	**0.0609**

The results of the baseline evaluation for the split scenario of 80:20 dan 70:30 reveal key insights into the model’s performance. In both scenarios, the Hamming Loss remains low, at 0.0625 for the 80:20 split and 0.0609 for the 70:30 split, demonstrating good model performance in predicting labels with minimal errors, with the 70:30 split achieving slightly better result.

3.8.2 Scenario 2: Resampled Model Without Implementing Classifiers

The performance of the model built through integrating various combinations of data split scenarios and resampling methods are evaluated here. By systematically testing different configurations, this study explores how adjustments in split ratios and resampling methods can enhance the model's ability to handle imbalanced datasets and predict multilabel outputs effectively. The detailed outcomes can be found in Table 10.

Table 10. Model performance in Scenario 2

Split Scenarios	Resampling Methods	Hamming Loss
80:20	MLROS	0.0548
	MLSMOTE	0.0521
	REMEDIAL	**0.0510**
70:30	MLROS	0.0518
	MLSMOTE	0.0508
	REMEDIAL	**0.0501**

The evaluation results for various split scenarios and resampling methods demonstrate significant improvements in model performance compared to the baseline scenario in scenario 1. The table shows that models using resampling methods consistently achieve lower Hamming Loss. For the 80:20 split scenario, REMEDIAL achieves the best results with a Hamming Loss of 0.0510, outperforming both MLROS and MLSMOTE. In comparison with the baseline model, the Hamming Loss decreased, representing an 18.4%

improvement after applying resampling. Similarly, in the 70:30 split scenario, REMEDIAL also provides the best performance, with a Hamming Loss of 0.0501, slightly lower than MLSMOTE and MLROS. In contrast compare to baseline model, the Hamming Loss resulted in a 17.7% improvement.

The comparison highlights that all resampling methods significantly enhance the model's ability to handle imbalanced data, as reflected in reduced Hamming Loss. Among the resampling methods, REMEDIAL consistently outperforms MLROS and MLSMOTE in both split scenarios, making it the most effective approach for this scenario combinations. These findings confirm the importance of resampling in improving model performance, especially in scenarios with imbalanced datasets.

3.8.3 Scenario 3: Model Implementing Resampling Methods and Classifiers

This scenario presents an evaluation of the models built using a combination of split scenarios, resampling methods, and multilabel classifiers. The objective of this evaluation is to explore the impact of these combined strategies on model performance, particularly in addressing the challenges posed by imbalanced datasets and multilabel classification tasks. The results are summarized in Table 11.

Table 11. Model performance in Scenario 3

Split Scenarios	80:20			70:30		
Resampling Methods	MLROS	MLSMOTE	REMEDIAL	MLROS	MLSMOTE	REMEDIAL
BR	0.0548	0.0521	0.0510	0.0518	0.0508	0.0501
LP	0.2937	0.2940	0.2940	0.2992	0.2992	0.2994
CC	**0.0420**	0.0438	0.0446	**0.0401**	0.0440	0.0422

The results across different split scenarios, resampling methods, and multilabel classifiers demonstrate varying performances, with the combination of the 70:30 split, MLROS resampling, and the Classifier Chains (CC) classifier emerging as the best-performing model. In this configuration, the model achieved the lowest Hamming Loss of 0.0401. This result highlight the model's ability to predict correct labels with minimal errors. For the 80:20 split, the CC classifier also performed exceptionally well under the MLROS resampling method, achieving a Hamming Loss of 0.0420. Although slightly higher than the 70:30 split, this configuration still outperformed all other combinations within the 80:20 scenario.

In contrast, the Label Powerset (LP) classifier consistently delivered the weakest performance across all resampling methods and split scenarios. For instance, under MLROS with the 70:30 split, LP recorded a Hamming Loss of 0.2992. This significant gap highlights LP's limitations in handling the dataset's complexity compared to CC and Binary Relevance (BR).

Among the three resampling techniques, MLROS demonstrated clear superiority over MLSMOTE and REMEDIAL across all classifier and split configurations. This finding is particularly significant in the ABSA domain, where intricate dependencies between aspect categories and sentiment polarity pose unique challenges for balancing multilabel datasets. MLROS's random oversampling strategy appears to better preserve these complex label correlations. By effectively addressing the label imbalance while maintaining natural label dependencies, MLROS enhances classification performance in this challenging context.

Compared to the baseline in scenario 1, the best-performing model that using the Classifier Chains (CC) classifier combined with MLROS resampling shows improvements in both split scenarios. For the 80:20 data split, the model reduces Hamming Loss by approximately 32.8%, indicating significantly fewer label prediction errors. Similarly, in the 70:30 split, the improvements are even more pronounced with the Hamming Loss decreases by 34.2%. These gains demonstrate that the proposed model effectively handles the multilabel classification challenge, significantly outperforming the baseline without resampling and advanced classifiers.

3.9 Model Validation Results

During the model validation stage, the best-performing model was assessed using a designated validation dataset, which was deliberately set aside during the initial data preparation phase to ensure unbiased evaluation. The validation dataset consists of 692 rows and serves as unseen data to measure the generalization ability of the model and verify that the model is not overfitting. In this study, the evaluation of the validation process focuses on the Hamming Loss metric. Table 12 presents a detailed comparison of model accuracy when applied to validation and test data. The prediction success rate, derived from the calculation of $1 - Hamming\ Loss$, is also highlighted.

Table 12. Comparison of model accuracy

Data Type	Hamming Loss	Success Rate
Validation Data	0.0656	93%
Test Data	0.0401	95%

Based on the comparison presented in tabel, the 2% difference between the validation data and the test data shows that despite the slight performance degradation on the unseen data, the model maintains strong predictive ability and does not show significant signs of overfitting. These findings underscore the model's readiness to handle new data with a high degree of reliability, ensuring its practical applicability in real-world scenarios.

3.10 Discussion

Our findings indicate a significant advancement in addressing multilabel classification challenges, particularly in the domain of Aspect-Based Sentiment Analysis (ABSA). These findings align partially with [11], which also employed MLROS, MLSMOTE, and REMEDIAL to adress mutilabel classification challenges, specifically in chest X-ray abnormality detection. While their study reported a Hamming Loss of 0.0324 with REMEDIAL combined with VGG16, this study achieved a lowest Hamming Loss of 0.0401 using MLROS with Classifier Chains (CC). The observed contrast can be attributed to differences in dataset structures, algorithms, and classifiers. Unlike [11], which leveraged neural networks, this study utilized SVM with multilabel classifiers, such as Classifier Chains, Binary Relevance, and Label Powerset. This study was chosen for comparison as the comprehensive comparative evaluations of these methods in the ABSA context are scarce.

Despite this difference in results, MLROS demonstrated significant effectiveness in reducing misclassification errors while maintaining the natural dependencies among labels. This is evidenced by the lowest Hamming Loss result across all scenarios, achieving a notable improvement over the baseline model. Specifically, the baseline model, which serves as a fundamental reference point, represents the performance of a model trained on the original dataset without applying resampling techniques or advanced multilabel classifiers. In this configuration, the model recorded its lowest Hamming Loss at 0.0609. In contrast, the control experiment produced the lowest Hamming Loss of 0.0401 by integrating MLROS as the resampling method combined with the Classifier Chain. This reduction of 0.0208 in Hamming Loss reflects a substantial 34.2% improvement in reducing misclassification errors.

These results underscore the study’s innovative integration of resampling techniques with multilabel classifiers, marking a significant contribution to advancing ABSA methodologies. By effectively addressing label imbalance, the proposed approach enhances the accuracy of multilabel predictions, which is essential for capturing the nuanced interplay between aspect categories and sentiment polarities. This advancement not only improves classification performance but also provides a robust framework for tackling complex multilabel dependencies in ABSA.

In ABSA, where interactions between aspect categories and sentiment polarities create intricate dependencies, the ability to generalize effectively is vital. The validation results also demonstrate the model's capacity to address these challenges by accurately capturing and processing multilabel relationships. Its balanced performance across datasets reinforces its adaptability to the domain's unique complexities.

The practical implications of this study are substantial. The model offers a dependable tool for enhancing customer experience evaluations, analyzing product feedback, and generating actionable insights from sentiment-driven data. By effectively generalizing across datasets and preserving label dependencies, the model positions itself as a robust solution for real-world applications, particularly in domains that require nuanced understanding of multilabel sentiment interactions. These findings underscore the model's potential to drive improved decision-making and strategic planning in customer-centric industries.

While the results demonstrate significant contributions, certain limitations must be acknowledged. The dataset employed in this study is domain-specific, focusing on feedback related to digital signature and eMeterai applications. This specificity may limit the generalizability of the findings to other domains with different label distributions or dependencies. Additionally, while MLROS effectively reduced misclassification errors, its reliance on synthetic data generation raises the potential for noise or bias, especially in scenarios involving sparse labels. These considerations present opportunities for further refinement and validation in future research.

4. CONCLUSION

In conclusion, this study demonstrates the effectiveness of incorporating resampling techniques, particularly MLROS, alongside advanced multilabel classifiers in addressing the unique challenges of ABSA using real-world user feedback. The findings highlight that resampling methods such as MLROS, MLSMOTE, and REMEDIAL significantly enhance the model's ability to manage imbalanced datasets, with MLROS consistently outperforming the other methods across all classifier and split configurations. Among all scenarios, the combination of MLROS and the Classifier Chains (CC) classifier with a 70:30 data split emerged as the best-performing approach, achieving the lowest Hamming Loss of 0.0401, equivalent to a prediction accuracy of 95%. This configuration demonstrated substantial improvements over the baseline, with a reduction in Hamming Loss by 34.2%. These findings underscore the model’s ability to manage pronounced label imbalances while preserving the intricate dependencies between aspect categories and sentiment polarities inherent in ABSA datasets. The validation results further confirm the model’s strong predictive ability and generalization capacity, ensuring its practical applicability for real-world applications such as customer feedback analysis, product

evaluation, and sentiment-driven decision-making. This study provides a foundation for selecting effective resampling and classification strategies to enhance the precision and reliability of ABSA systems in highly imbalanced contexts.

REFERENCE

[1] D. K. Bogor, "Implementasi E-Surat dan Tandatangan Digital di Pemerintah Kota Bogor [Implementation of E-Letters and Digital Signatures in Bogor City Government]," [Online]. Accessed: Oct. 06, 2024. [Online]. Available: https://kominfo.kotabogor.go.id/index.php/post/single/603

[2] BBC NEWS, "BBC NEWS Indonesia," Online. Accessed: Oct. 06, 2024. [Online]. Available: https://www.bbc.com/indonesia/articles/c6234qw4wzgo

[3] G. Radiena and A. Nugroho, "Analisis Sentimen Berbasis Aspek Pada Ulasan Aplikasi Kai Access Menggunakan Metode Support Vector Machine [Aspect-Based Sentiment Analysis on Kai Access App Reviews Using Support Vector Machine]," Jukanti, vol. 6, no. 1, pp. 1 – 10, 2023. https://doi.org/10.37792/jukanti.v6i1.836 (In Indonesian)

[4] F. Zamachsari, Gabriel Vangeran Saragih, Susafa'ati, and Windu Gata, "Analysis of Sentiment of Moving a National Capital with Feature Selection Naive Bayes Algorithm and Support Vector Machine," *Jurnal RESTI*, vol. 4, no. 3, pp. 504–512, Jun. 2020. https://doi.org/10.29207/resti.v4i3.1942

[5] R. Wahyudi and G. Kusumawardana, "Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine [Sentiment Analysis on Grab App in Google Play Store Using SVM]," *Jurnal Informatika*, vol. 8, no. 2, pp. 200–207, Sep. 2021. https://doi.org/10.31294/ji.v8i2.9681 (In Indonesian)

[6] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognit.*, vol. 122, 2022. https://doi.org/10.1016/j.patcog.2021.107965

[7] P. Das, J. W. Sangma, V. Pal, and Yogita, "Predicting Adverse Drug Reactions from Drug Functions by Binary Relevance Multi-label Classification and MLSMOTE," in *Machine Learning, Image Processing, Network Security and Data Sciences*, Springer, 2022, pp. 165–173. https://doi.org/10.1007/978-3-030-86258-9_17

[8] A. Masmoudi, H. Bellaaj, K. Drira, and M. Jmaiel, "A co-training-based approach for the hierarchical multi-label classification of research papers," *Expert Syst*, vol. 38, no. 4, Jun. 2021. https://doi.org/10.1111/exsy.12613

[9] A. Umparat and S. Phoomvuthisarn, "Improving Pre-Trained Models for Multi-Label Classification in Stack Overflow: A Comparison of Imbalanced Data Handling Methods," in *Proc. 20th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jun. 2023, pp. 464–469. https://ieeexplore.ieee.org/document/10202012

[10] E. d'Andréa, J. François, O. Festor, and M. Zakroum, "Multi-label Classification of Hosts Observed through a Darknet," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, IEEE, May 2023, pp. 1–6. https://doi.org/10.1109/NOMS56928.2023.10154356

[11] H. Tsaniya, C. Fatichah, and N. Suciati, "Comparison of sampling methods for handling imbalance data in deep learning-based predictions of chest X-ray abnormality tags," in *Proc. 7th Int. Conf. Med. Health Inform (ICMHI)*, New York, NY, USA: ACM, May 2023, pp. 6–10. https://doi.org/10.1145/3608298.3608300

[12] S. K. Singh and Dr. R. K. Dwivedi, "Data Mining: Dirty Data and Data Cleaning," *SSRN Electronic Journal*, 2020. https://doi.org/10.32614/RJ_2021_046

[13] A. Upadhye, "A Comprehensive Survey of Text Data Cleaning Techniques: Challenges, Methods, and Best Practices," *Available online www.jsaer.com Journal of Scientific and Engineering Research 205 Journal of Scientific and Engineering Research*, vol. 2020, no. 8, pp. 205–210

[14] R. Lourdusamy and S. Abraham, "A Survey on Text Pre-processing Techniques and Tools," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 3, pp. 148–157, 2018. https://doi.org/10.26438/ijcse/v6si3.148157

[15] A. R. Lubis and M. K. M. Nasution, "Twitter Data Analysis and Text Normalization in Collecting Standard Word," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 4, no. 2, pp. 855–863, Jun. 2023. https://doi.org/10.37385/jaets.v4i2.1991

[16] W. Wahyudin, "Aplikasi Topic Modeling Pada Pemberitaan Portal Berita Online Selama Masa Psbb Pertama [Topic Modeling Application on Online News During First PSBB Period]," *Seminar Nasional Official Statistics*, vol. 2020, no. 1, pp. 309–318, Jan. 2021. https://doi.org/10.34123/semnasoffstat.v2020i1.579 (In Indonesian)

[17] J. Hughes, "krippendorffsalpha: An R Package for Measuring Agreement using Krippendorff's Alpha Coefficient." https://doi.org/10.1111/exsy.12613

[18] J. Alshehri, M. Stanojevic, E. Dragut, and Z. Obradovic, "On Label Quality in Class Imbalance Setting -A Case Study," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Dec. 2022, pp. 1666–1671. https://ieeexplore.ieee.org/document/10012345

[19] Y. Kustiyahningsih and Y. Permana, "Penggunaan Latent Dirichlet Allocation (LDA) dan Support-Vector Machine (SVM) untuk Menganalisis Sentimen Berdasarkan Aspek Dalam Ulasan Aplikasi EdLink [Using LDA and SVM to Analyze Aspect-Based Sentiment in EdLink App Reviews]," *Teknika*, vol. 13, no. 1, pp. 127–136, Mar. 2024. https://doi.org/10.34148/teknika.v13i1.746 (In Indonesian)

[20] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, Sep. 2015. https://doi.org/10.1016/j.neucom.2014.08.091

[21] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Tackling Multilabel Imbalance through Label Decoupling and Data Resampling Hybridization," *Knowl.-Based Syst*., vol. 89, pp. 385–397, Feb. 2018. https://doi.org/10.1016/j.knosys.2015.07.019

[22] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowl Based Syst*, vol. 89, pp. 385–397, Nov. 2015. https://doi.org/10.1016/j.knosys.2015.07.019

[23] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "Concurrence among Imbalanced Labels and Its Influence on Multilabel Resampling Algorithms," in *Proc. 10th Int. Conf. Hybrid Artif. Intell. Syst*. (HAIS), Springer, 2014, pp. 110–121. https://doi.org/10.1007/978-3-319-07617-1_10

[24] A. R. Abelard and Y. Sibaroni, "Multi-aspect sentiment analysis on netflix application using latent dirichlet allocation and support vector machine methods," *J. Infotel*, vol. 13, no. 3, pp. 128–133, Aug. 2021. https://doi.org/10.20895/infotel.v13i3.670

[25] A. Hafeez *et al.*, "Addressing Imbalance Problem for Multi Label Classification of Scholarly Articles," *IEEE Access*, vol. 11, pp. 74500–74516, 2023. https://doi.org/10.1109/ACCESS.2023.3293852