*https://doi.org/10.48047/AFJBS.6.14.2024.4983-4987*

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

# A Frame for domain based identification and insertion of missing pages using Document Classification

**V.Nirmala [1], Dr.B.Lavanya [2]**

[1] University of Madras ,Department of computer science

[2] University of Madras,Department of computer science

Volume 6, Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 15 Aug 2024

*doi: 10.48047/AFJBS.6.14.2024.4983-4987*

**Abstract**

There are numerous challenges associated with document classification; here we focus on the identification of missing pages in the given set of documents. The missing pages are identified, then, the attempt to classify them to a specific domain, using BOW of the missing pages with the BOW of the domain using classification algorithm is done. We the next phase of the proposed work is to search through all the domain specific documents available to identify the incomplete document. Appending of the identified missing pages into the incomplete document and it is made complete. Even if the inserted original document is complete, an incorrect page arrangement could prevent the file from being comprehensive. So, the next phase of internal sorting of pages is carried out, to create a comprehensive complete document. Missing pages are a common occurrence in the process of classifying documents , which can disrupt the accurate classification of domain-specific information. The missing page results in a loss of specific time and file usage. System errors and manual processing errors often leave many datasets incomplete, leading to confusion during data submission. To address this issue, this study proposes a framework that identifies missing pages within a file, classifies these missing pages within a specified domain using a Mis-BOW Naïve Bayes algorithm and then arranges the pages in sequential order using internal sorting and creates a complete document.

**Keywords:** text classification, missing page identification, Machine learning.

## 1. Introduction

In the realm of document classification, one of the significant challenges is the identification and rectification of missing pages within documents. Missing pages can severely impact the classification accuracy of domain-specific information, leading to potential data loss and misclassification. These gaps often arise due to system errors or manual processing mistakes, resulting in incomplete datasets and confusion during data submission. It is necessary to attach each label to the document and then save the

common keywords in BOW. We design the Mis-BOW Naïve Bayes algorithm to identify, classify, and organize missing pages within documents. The proposed framework involves internal sorting of the pages to establish the correct sequence, creating a fully comprehensive and complete document. The paper [1] highlights that electronic health records (EHR) improve data access and management but often suffer from missing data, leading to inaccurate analysis when using simple imputation methods and use a new method using denoising auto encoders (DAE) combined with k-nearest neighbors (KNN) and optimizes it by periodically reapplying KNN. The authors [2] examines the challenge of missing values in datasets that negatively affect machine learning classification accuracy, by using support vector machine (SVM) regression for imputing missing values, and introduce a two-level classification process to reduce false classifications. The article [3] describes algorithmic fairness and missing values in the classification. Due to missing data patterns, training a classifier on imputed data reduces group fairness and average accuracy, they use scalable and adaptive fair classification algorithms that maintain missing data patterns. The paper [4] deals the imputation of missing data in classification problems, criticizing existing k-nearest neighbor (KNN) methods for not considering class labels and using ineffective distance measures for heterogeneous data they use a new iterative KNN imputation technique using class-weighted grey distance enhanced by mutual information (MI) to improve classification. The paper [5] focuses on improving classification accuracy in data mining by effectively imputing missing values using a model based on correlation. This method works for both categorical and numerical data and enhances classification performance. The paper [6] examines handling missing values in classifiers using optimal bayesian classification. It integrates missing data into Bayesian models and develops a decision rule for Gaussian models.

## 2.Objective

- Identifying and analyzing the domain for the missing pages
- Finding the index for the missing page number in the respective document.
- Concatenate the missing pages into respective original set of documents
- Arrange the documents pages in progressive order.

## 3.Related works

The paper [7] introduce a method to learn Label-Specific features for multi-label classification with Missing Labels, named LSML .Unlike traditional methods that use a single data representation for all labels, LSML creates a supplementary label matrix by learning label correlations. Then develops label-specific data representations and builds a classifier using these representations. The paper [8] addresses both missing features and labels. It uses matrix factorization to recover missing values and builds a classification model from the latent spaces of features and labels. MMFL(Multi-Label learning with missing features and labels)has an extra classifier for labeling sparse tails and uses manifold regularization to keep label correlations and instance similarity. The article [9] deals with identifying the lack of public multi-page document classification datasets, describing different classification tasks that come up in real-world situations and covering a calibration evaluation. The paper [10] discuss about missing data patterns, imputation methods, and classification algorithms interact. It finds the effectiveness of classification algorithms depends on the interplay between the imputation method and the missing data pattern, highlighting the importance of matching imputation techniques with classification algorithms based on the type of missing data. The authors [11] suggested the multimodal classification using a dataset of hand-annotated pages divided into six groups. They train two unimodel classifiers: a residual network that is optimized for images and a convolutional neural network that is optimized for text. Using a fusion method, they combine their visual and textual features. These references make it clear that the existing methodologies deals with missing values. However, observed a research gap that the system fails to account for classification due to missing pages. The presence of missing pages can lead to incomplete information, which negatively impacts the accuracy and reliability of the document classification process.

## 4.Research Methodology

The input file includes two categories of documents: one set consisting of documents imported as PDF and Word documents, and another set containing documents with missing pages. Preprocessing the content of the PDF and Word files is done, which includes converting text to lowercase, removing punctuation, and performing stemming and lemmatization. Next, assign labels to each document. Collect frequently used keywords from the PDFs and Word documents and compile them into the Input Bag of Words (Ibow). The

common words are aggregated based on the document labels and stored in the Aggregate Bag of Words (Abow). From the missing pages, gather common keywords and store it in the Missing Bag of Words (Mbow). Compare the keywords in Mbow with those in Abow. If a match is found, predict the domain of the missing page based on the aggregate keywords and organize it sequentially. If the specific keywords do not match with aggregate bag of words (Abow), the page will be set to plain mode, and the domain labels will not align with the subject. After predicting domain for the missing pages, the text data needs to be vectorized into a numerical format using Count Vectorizer. The dataset will be split into two groups one is training set and the other is test set. The training set will be used to train the Multinomial Naive Bayes model with proposed Mis-BOW Naïve bayes, and the test set will be used to evaluate its performance. The current classification system effectively categorizes documents based on missing values within the content but fails to account for missing pages, leading to classification due to incomplete information, reduced context, and inconsistent data representation. To address this, we propose an enhanced algorithm that incorporates Mis-BOW Multinomial Naive Bayes and handles missing pages. The process includes generating Bag of Words (BOW) for input documents ,missing pages, matching domain common words, and concatenating these with input documents for training. The classifier is then used to predict classes for missing pages, ensuring more accurate and reliable document classification.

## 5. Mis-BOW Naïve Bayes (Missing - Bag of words Naïve Bayes)

The Bag of Words (BOW) model is a basic technique used in text classification and natural language processing (NLP). It keeps a list of unique words in its vocabulary and it is easy to keep track of the occurrences of the words in terms of sentences or documents - Mis-BOW maintains the common keywords from the missing pages document. The Multinomial Naive Bayes classifier is appropriate for word count text classification with discrete features; however, it typically requires integer feature counts. The probability of one characteristic occurrence does not affect the other. so we use Mis-BOW multinominal Naïve Bayes which yields a good performance. The proposed Mis-BOW Naïve Bayes can be used in many domain based classification because of its robustness, ease of implementation, rapidity, and accuracy.

### 5.1 Algorithm for Mis-BOW Naïve Bayes

**Input**:

$D = \{d_1, d_2, \ldots\ldots d_n\}$ Where $d_1, d_2, \ldots\ldots d_n$ set of all documents
$L = \{l_1, l_2, \ldots.. l_k\}$ where $l_1, l_2, \ldots.. l_k$ set of all labels
$M = \{M_1, M_2, \ldots. M_m\}$ where $M_1, \ldots. M_m$ set of all missing pages
$I = \{I_1, I_2, \ldots. I_n\}$ where $I_1, I_2, \ldots. I_n$ set of all input documents
Let $Ibow_1, Ibow_2, \ldots.. Ibow_n$ be BOW of all Input document
Let $Abow_1, Abow_2, \ldots.. Abow_k$ be BOW of all document label wise
Let $Mbow_1, Mbow_2, \ldots. Mbow_m$ be BOW of all Input missing pages
Let $Dcw_1, Dcw_2, \ldots Dcw_m$ be Domain common words of M
Let Ibow_sorted be sorted BOW in all input documents
Let clf contains the trained Multinomial Naive Bayes classifier
Let $X_{train}$ be the feature matrix for the Input documents
Let $Y_{train}$ be the labels
Let $X_{test}$ be the feature matrix for the I and M
Let $Y_{pred}$ be the prediction for every M

**Output:** $I \oplus M$

Step 1:Input I,y,M, $y \subseteq L$

Step 2: $\forall I_1, I_2, \ldots. I_n$ ,Tabulate $Ibow_1, Ibow_2, \ldots.. Ibow_n$

$\forall L$ ,Tabulate $Abow_1, Abow_2, \ldots.. Abow_k$

Step 3: $\forall M_1, \ldots. M_m$ ,Generate $Mbow_1, Mbow_2, \ldots. Mbow_m$

Step 4: $Dcw_1, \ldots Dcw_m = match[(Mbow_1, Mbow_2, \ldots. Mbow_m), (Abow_1, Abow_2, \ldots.. Abow_k)]$

Step 5: Concatenate$(Dcw_1, Dcw_2, \ldots Dcw_m) \oplus (I, y)$

Step6: clf =MultinomialNB($Dcw_{1\ldots\ldots m}$)

Step7: $Y_{pred}$=clf.predict($X_{test}$)
Step8: I⊕M
Step9: Internal sort(I)

In the first step, the algorithm begins by initializing the input documents I, the missing pages M, Labels (L),where Input(y) is a subset of the entire set of labels (L). In the second step for all input documents(I) from $I_{1,}$…….$I_n$ collect the common words and tabulate in $Ibow_1$, $Ibow_2$, …, $Ibow_{n.}$,and for all Labels(L) from $l_1$, $l_2$, …, $l_k$ collect the common words and tabulate in $Abow_1$, $Abow_2$,...$Abow_{k.}$.In the third step for all Missing Labels(M) from $M_1$, $M_2$, …,$M_m$ collect the common words and generate in $Mbow_1$, $Mbow_2$,...,$Mbow_{m.}$.In the fourth step, to predict the domain for the missing pages select the keywords from $Mbow_1$, $Mbow_2$,...,$Mbow_m$ and matches with $Abow_1$, $Abow_2$,...$Abow_k$ store the matched page in corresponding domain.In the fifth step, Now Combine the domain common words page with the input documents and their corresponding labels. In the sixth step, initialize and train the Multinominal Naïve Bayes classifier in $Dcw_{1……m}$.In the seventh step,Use the trained classifier to predict the labels for the missing pages. In the eight step ,combine the input documents(I) and the Missing page (M) .In the Nineth step, By applying internal sorting page numbers will be sorted sequentially .

## 6.Result

The documents gathered from the Google Scholar with various categories of PDF and Word files. A total of 500 documents have been collected in both PDF and Word (docx) formats: 350 documents in PDF format, sourced from various categories such as journals from Google Scholar, subject notes, question papers, resumes, and Journals from science department and the 150 documents contain word-format files taken as Subject notes followed in PDF and collection of resumes. Notably, the "Subject Notes" file category is available in both PDF and Word formats. The accuracy of "Subject Notes" in PDF format is 100%, while in Word format it is 86%. The proposed Mis-BOW Naïve Bayes algorithm achieves better performance for PDF than for the word files. Table1, shows the details of Document type, File type, Number of labels, Number of documents and proposed Mis-BOW Naïve Bayes accuracy.

Table 1: Performance of Mis-Bow Naïve Bayes

| S.No | Document type | File type | #.of Labels | #.of documents | Accuracy% |
|---|---|---|---|---|---|
| 1 | journals from google scholar | PDF | 5 | 50 | 100 |
| 2 | Subject notes | PDF | 12 | 120 | 100 |
| 3 | Question papers | PDF | 3 | 35 | 100 |
| 4 | Resume | PDF | 6 | 25 | 100 |
| 5 | Journals from Science department | PDF | 13 | 120 | 100 |
| 6 | Subject notes | docx | 12 | 120 | 86 |
| 7 | Collection of resumes | docx | 11 | 30 | 85 |

## 7. Conclusion

The objective is to accurately classify missing page documents, identify the correct index for these pages within the original document sequence, and arrange them sequentially to complete the document. The Mis-BOW Naïve Bayes algorithm aims to enhance the accuracy and efficiency of missing page classification while ensuring dataset integrity and usability. The algorithm achieves 100% accuracy with PDFs and around 86% accuracy with Word files. The accuracy gap between PDFs and Word files can be attributed to differences in text extraction quality and document formatting. PDFs offer cleaner and more consistent text extraction, which improves the algorithm's performance. In contrast, Word files often have varying formatting, embedded objects, and uneven text structures, which complicate text extraction and introduce variability, resulting in lower classification performance.

## 8. Reference

[1] Psychogyios, Konstantinos, et al. "Missing value imputation methods for electronic health records." *IEEE Access* 11 (2023): 21562-21574

[2] Palanivinayagam, Ashokkumar, and Robertas Damaševičius. "Effective handling of missing values in datasets for classification using machine learning methods." *Information* 14.2 (2023): 92.

[3] Feng, Raymond, Flavio Calmon, and Hao Wang. "Adapting fairness interventions to missing values." *Advances in Neural Information Processing Systems* 36 (2024).

[4] Choudhury, Arkopal, and Michael R. Kosorok. "Missing data imputation for classification problems." arXiv preprint arXiv:2002.10709 (2020).

[5] Panda, B. S., and Rajesh Kumar Adhikari. "A method for classification of missing values using data mining techniques." *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*. IEEE, 2020.

[6] Dadaneh, Siamak Zamani, Edward R. Dougherty, and Xiaoning Qian. "Optimal Bayesian classification with missing values." *IEEE Transactions on Signal Processing* 66.16 (2018): 4182-4192.

[7] Huang, Jun, et al. "Improving multi-label classification with missing labels by learning label-specific features." *Information Sciences* 492 (2019): 124-146.

[8] Hao, Xiuyan, et al. "Multi-label learning with missing features and labels and its application to text categorization." *Intelligent Systems with Applications* 14 (2022): 200086.

[9] Van Landeghem, Jordy, et al. "Beyond Document Page Classification: Design, Datasets, and Challenges." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024.

[10] Garciarena, Unai, and Roberto Santana. "An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers." *Expert Systems with Applications* 89 (2017): 52-65.

[11] Luz de Araujo, Pedro H., et al. "Sequence-aware multimodal page classification of Brazilian legal documents." *International Journal on Document Analysis and Recognition (IJDAR)* 26.1 (2023): 33-49.

## Authors' background

| Your Name | Title* | Research Field | Personal website |
|---|---|---|---|
| V.Nirmala | A Frame for domain based identification and insertion of missing pages using Document Classification | Data mining | vnirmala7488@gmail.com |
| Dr.B.Lavanya | A Frame for domain based identification and insertion of missing pages using Document Classification | Data mining | lavanmu@gmail.com |