Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1330

A Review of Dimension Reduction and Feature Selection: A New Perspective

Simon Ashiru[1]*, Sahalu Junaidu[2], Salisu Aliyu[2], Abubakar Yahaya[3]

[1]Department of Computer Science, Kaduna State University (KASU), Nigeria
[2]Department of Computer Science, Ahmadu Bello University Zaria, Nigeria
[3]Department of Statistics, Ahmadu Bello University Zaria, Nigeria

Corresponding Author: a.simon@kasu.edu.ng

ABSTRACT

With the rapid growth of high-dimensional data, dimensionality reduction has become essential for improving computational efficiency and model performance. This paper presents a comprehensive review of two primary approaches to dimensionality reduction: feature selection (FS) and feature extraction (FE), with particular emphasis on the Chi-Square-based feature selection technique. While Chi-Square has been widely adopted due to its simplicity and effectiveness, its dependence on document frequency introduces notable limitations. This review highlights key developments in feature selection strategies, identifies gaps in existing Chi-Square-based methods, and explores recent improvements proposed in the literature. Finally, the paper offers a new perspective on enhancing Chi-Square by incorporating contextual relevance and hybrid scoring mechanisms, setting the stage for future research.

Keywords: Feature Selection, Dimensionality Reduction, Chi-Square, Text Classification

INTRODUCTION

The rapid growth of textual data from sources such as social media, news articles, online reviews, and web content has transformed the Internet into a massive repository of information for data analysis (J. Cai et al., 2018; Chuanxin et al., 2015) consequently, the need for efficient data processing has become more critical than ever.

However, such data is often high-dimensional, containing noisy, redundant, or irrelevant features that degrade machine learning model performance (J. Cai et al., 2018; Deng et al., 2019). This challenge is particularly pronounced in tasks like text classification, where a large feature space can lead to increased complexity, overfitting, and reduced accuracy.

Dimensionality reduction techniques, particularly Feature Selection (FS) and Feature Extraction (FE), play a critical role in addressing this issue. FS focuses on identifying a subset of relevant features by eliminating noise and redundancy, whereas FE transforms data into a lower-dimensional representation. Among FS techniques, the Chi-Square test has gained popularity due to its simplicity and effectiveness in evaluating term-class dependence. Despite its advantages, Chi-Square suffers from limitations such as reliance on document frequency, which may overlook contextual or semantic relevance.

This review explores recent developments in dimensionality reduction, with a particular emphasis on Chi-Square-based feature selection. It critically examines the strengths and weaknesses of traditional approaches, reviews hybrid and enhanced methods, and highlights ongoing challenges. The paper also offers a new perspective by proposing conceptual strategies to improve the Chi-Square method, including the incorporation of contextual information and hybrid scoring

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1330

metrics. By synthesizing key findings and identifying research gaps, this review provides a foundation for future work in more effective and adaptive feature selection techniques.

OVERVIEW OF DIMENSION REDUCTION TECHNIQUES

Dimension reduction is an essential step in preprocessing high-dimensional data, especially in domains like text classification, sentiment analysis, and pattern recognition. Feature Selection (FS) and Feature Extraction (FE) remain the two primary categories. FS methods aim to identify a minimal subset of the most relevant features, whereas FE techniques like PCA and autoencoders generate new features based on mathematical transformations (Zebari et al., 2020). FS, in particular, helps reduce overfitting, speeds up training time, and improves model interpretability (Blum & Langley, 1997; Sammut & Webb, 2016).

Strengths and Limitations of Chi-Square

The Chi-Square (χ^2) test is widely used to evaluate the independence between terms and class labels. It is simple, computationally efficient, and performs well in many contexts (Bahassine et al., 2018; Sun et al., 2017). However, it is fundamentally based on document frequency, which limits its capacity to consider term frequency within classes, positional significance, or semantic context (Chuanxin et al., 2015). This causes it to potentially overlook frequently occurring and semantically meaningful terms, especially in unbalanced datasets.

Addressing Redundancy and Feature Relevance

Researchers have introduced several enhancements to χ^2 by incorporating additional statistical or semantic features:

Lu et al. (2015) introduced a method using Euclidean distance to measure feature relevance and redundancy by comparing feature vectors to category center vectors.

Wang and Zhou (2021) improved χ^2 by integrating a minimum redundancy criterion, reducing redundant terms in the final feature set.

Haipeng et al. (2017) merged synonyms of selected terms to minimize feature loss, but their method applied the synonym merge across all classes uniformly, ignoring class-specific semantics.

Bahassine et al. (2018) enhanced χ^2 by enforcing proportional feature selection across all categories, ensuring balanced class representation.

Despite these advancements, redundancy detection still lacks consistent and scalable mechanisms. Few approaches thoroughly address the correlations between terms across classes or how redundant information affects model performance.

Trends in Hybrid and Adaptive Techniques

Recent studies trend toward combining multiple criteria into a single FS framework:

GuiChuan and Shubin (2015) merged χ^2 and Mutual Information (MI) with term frequency and positional weighting, assigning different weights to terms in the title, keyword, or body.

Chuanxin et al. (2015) added term distribution (via variance) and term frequency in the main class to the χ^2 formula, improving classification accuracy but neglecting inter-class frequency analysis.

Sun et al. (2017) proposed a model that considers both intra-class dispersion and inter-class concentration, capturing how term frequency differs across categories.

L.-j. Cai et al. (2021) extended χ^2 by integrating document frequency, term frequency, variance, and coefficient of variation, creating a robust but computationally complex model.

These methods reveal a move toward context-enriched, statistically rigorous, and hybrid FS models, though they often suffer from increased computational cost and lack interpretability.

Toward Semantically-Driven Feature Selection

A recurring limitation across all reviewed FS methods is the absence of semantic understanding. Even TF-IDF and MI fail to distinguish between synonymous or contextually similar terms (e.g., "joy" and "happiness"). Techniques like term synthesis, bigram merging, or word embedding integration have not been widely incorporated into traditional χ^2 frameworks.

For example, while (Parlar et al., 2018) and (Şahin & Kılıç, 2019) improved FS with document frequency adjustments and category-specific term counts, they did not address semantic grouping or contextual inference.

Thus, a promising direction involves fusing statistical models like χ^2 with semantic models that understand term relationships, distributions, and usage patterns. This would enable the development of FS techniques that are not only statistically sound but also linguistically meaningful and context-sensitive.

EMERGING DIMENSIONALITY REDUCTION TECHNIQUES

There are recent advancement in the area of dimension reduction where hybrid and deep leaning-based techniques are used such as manifold learning, sparse coding and contrastive learning. Some of their advantages are: efficiency, noise reduction and reduced complexities.

Manifold Learning

Manifold learning is an unsupervised technique which assumes high dimensional data can be represented on a lower-dimensional manifold (Meilă & Zhang, 2024). This means that simpler structure representing the essential pattern hidden within a complex structure. For example, a curve in 3D space can be seen as a simpler curve in 1D. This technique combines Principal Component Analysis (PCA) with t-Distributed Stochastic Neighbor embedding (t-SNE) to capture both global and local structures.

Sparse Coding

This is a feature extraction technique which is used to represent data as a sparse linear combination of basic vectors. This technique assumes that most data points can be represented using a small number of non-zero features. In image processing, only the basic patterns like edges and textures are combined to create a new way of describing the image (Guo et al., 2007). An example is the Least Absolute Shrinkage and Selection Operator (LASSO), which is applied in image compression.

Contrastive Learning

This technique is a self-supervised learning where a model learns by contrasting positive pairs and negative pairs (Wu et al., 2024). It is widely applied in computer vision. This technique assumes that data points can be paired based on their similarities. By placing similar data points closer and dissimilar points further, the system improves – data organization, easier visualization, enhanced feature learning and better algorithm performance. Examples are SimCLR

(Simple Framework for Contrastive Learning of Representations), MoCo (Momentum Contrast) and BYOL (Bootstrap Your Own Latent).

Table 1: Summary of emerging dimensionality reduction techniques.

Technique	Focus	Primary goal	Applications
Manifold Learning	Discover low-dimensional structure.	Maintain geometric structure in lower dimensions.	Visualization, compression
Sparse Coding	Create sparse representations	Represent data using minimal active features	Image and signal processing
Contrastive Learning	Self-supervised representation.	Bring similar pairs closer, separate dissimilar pairs	Pre-training for vision and NLP tasks.

CHALLENGES IN DIMENSIONALITY REDUCTION

Despite significant achievements, challenges like scalability, interpretability, and domain-specific needs remain.

Scalability

As data size increases, many dimensionality reduction techniques become computationally expensive in terms of time and memory.

Interpretability

Transforming original features into a new space often leads to a loss of semantic meaning, making it harder for domain experts to interpret the results.

Domain-Specific Needs

General-purpose techniques may not suit specific data types (e.g., graphs, sequences). Custom methods may be needed to effectively capture and represent such data.

Feature Relevance and Redundancy

According to (John et al., 1994), features can be categorized as strongly relevant, weakly relevant, and irrelevant. (Yu & Liu, 2004) further classified them into four groups: noisy/irrelevant, redundant/weakly relevant, weakly relevant non-redundant, and strongly relevant. Strongly relevant features are critical for optimal performance, while irrelevant features introduce noise and inefficiency (Sammut & Webb, 2016).

Redundant features, though individually relevant, repeat information found in other features and increase training complexity. Identifying and removing such features is crucial for efficiency. The best feature sets have maximum relevance and minimum redundancy (Ding & Peng, 2005). Methods like mutual information, chi-square, information gain, and correlation coefficients are often used for this purpose.

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1330

OBSERVATIONS FROM REVIEWED PAPERS

Term Frequency Limitations

Many studies use term frequency to determine document categories (Bahassine et al., 2018; L.-j. Cai et al., 2021; Chuanxin et al., 2015). However, high frequency doesn't always mean high relevance. Unique terms that appear in only one class may be more informative.

Sentence Relevance

Just as some features are more relevant than others, some sentences carry more informative content. Not all sentences in a document contribute equally to classification. Selecting the most relevant ones (e.g., first and last) can reduce complexity and processing time.

Feature Expansion

To balance dimensionality reduction with performance, highly ranked features can be expanded with directly related ones. This enriches the feature set and enhances model learning.

Redundancy Simplification

Many existing methods rely on complex mathematics. A simpler, heuristic method for identifying redundant features would be more accessible and practical for broader audiences.

PROPOSED DIMENSIONALITY REDUCTION FRAMEWORK

The proposed framework includes two key techniques: Sentence Selection and Term Synthesis. Together, they reduce dimensionality and enhance relevant feature identification.

Sentence Selection (SS)

SS involves selecting key sentences (typically the first and last) from documents, which often provide contextual or summarizing information. This method is aim at achieving – Reduces dimensionality, Enhances relevant features and Improves interpretability.

Mathematical Representation:

Let $D = \{d_1, d_2, ..., d_n\}$ be the dataset of documents.
Let $S_i = \{s_{i1}, s_{i2}, ..., s_{imi}\}$ be the sentences in document d_i.
Let $F(d_i) = s_{i1}$ (first sentence), $L(d_i) = s_{imi}$ (last sentence).

Selected sentences for document d_i:

$S(d_i) = \{F(d_i), L(d_i)\}$

Selected sentences for dataset D:

$S(D) = \cup \ S(d_i)$ for i = 1 to n

Term Synthesis (TS)

TS identifies frequently co-occurring word pairs (bigrams) and merges them into compound terms, e.g., "artificial" and "intelligence" become "artificial-intelligence". This will lead to reduction in noise, compresses feature space and eventually captures meaningful term associations

Mathematical Representation:

Let $D = \{d_1, ..., d_n\}$ be the dataset.

Let $V = \{t_1, ..., t_m\}$ be the vocabulary.

Let freq(b) = $\sum$ count_d(t_i, t_j) where (t_i, t_j) co-occur in d
If freq(b) $\geq \theta$, replace (t_i, t_j) with $t_{ij} = t_i$-'-t_j
Transformed document d' = replaced(d, b, t_{ij})

CONCLUSION

Data complexity grows with dimensionality. Reducing dimensionality while retaining relevant features enhances model efficiency and accuracy. Traditional techniques benefit from enhancement through term frequency, uniqueness, and sentence selection. Processing only informative sentences and synthesizing relevant term pairs lead to better representation. Additionally, simple heuristic methods for identifying redundant features can improve accessibility and system performance. The proposed framework, combining Sentence Selection and Term Synthesis, provides a practical approach to feature optimization in text classification tasks.

REFERENCES

Bahassine, Said; Madani, Abdellah; Al-Sarem, Mohammed; & Kissi, Mohamed. (2018). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences, 32*(2), 225-231. doi: https://doi.org/10.1016/j.jksuci.2018.05.010

Blum, A.L.; & Langley, P. (1997). Selection of Relevant Features and Examples in Machine Learning. *Artif. Intell, 97*, pp. 245–271.

Cai, Jie; Luo, Jiawei; Wang, Shulin; & Yang, Sheng. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300*, pp.70-79. doi: https://doi.org/10.1016/j.neucom.2017.11.077

Cai, Liang-jing; Lv, Shu; & Shi, Kai-bo. (2021). Application of an Improved CHI Feature Selection Algorithm. *Discrete Dynamics in Nature and Society, vol. 2021*, pp. 1-8. doi: 10.1155/2021/9963382

Chuanxin, Jin; Tinghuai, Ma; Rongtao, Hou; Meili, Tang; Yuan, Tian; Abdullah, Al-Dhelaan; & Mznah, Al-Rodhaan. (2015). Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization. *IETE Journal of Research, 61*, pp.1-12.

Deng, Xuelian; Li, Yuqing; Weng, Jian; & Zhang, Jilian. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications, 78*(3), pp.3797-3816. doi: 10.1007/s11042-018-6083-5

Ding, Chris; & Peng, Hanchuan. (2005). Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics and Computational Biology, 3*(2), pp. 185–205.

GuiChuan, Feng; & Shubin, Cai. (2015, 2015/11). *An Improved Feature Extraction Algorithm Based on CHI and MI.* Paper presented at the Proceedings of the 2015 4th International Conference on Computer, Mechatronics, Control and Electronic Engineering.

Guo, Cheng-en; Zhu, Song-Chun; & Wu, Ying Nian. (2007). Primal sketch: Integrating structure and texture. *Computer Vision and Image Understanding, 106*(1), pp. 5-19. doi: doi.org/10.1016/j.cviu.2005.09.004

Haipeng, Yao; Chong, Liu; Peiying, Zhang; & Luyao, Wang. (2017). A Feature Selection Method Based on Synonym Merging in Text Classification System. *EURASIP Journal on Wireless Communications and Networking, 2017*. doi: 10.1186/s13638-017-0950-z

John, H; Kohavi, R; & Pfleger, K. (1994). *Irrelevant feature and the subset selection problem.* Paper presented at the In Proceedings of the Eleventh International Conference on Machine Learning.

Meilă, Marina; & Zhang, Hanyu. (2024). Manifold Learning: What, How, and Why. *Annual Review of Statistics and Its Application 11*(2024), pp. 393-417. doi: 10.1146/annurev-statistics-040522-115238

Parlar, Tuba; Özel, Selma Ayşe; & Song, Fei. (2018). QER: a new feature selection method for sentiment analysis. *Human-centric Computing and Information Sciences, 8*(1), 10. doi: 10.1186/s13673-018-0135-8

Şahin, Durmuş Özkan; & Kılıç, Erdal. (2019). Two New Feature Selection Metrics for Text Classification. *Journal for Control, Measurement, Electronics, Computing and Communications, 60*(2), pp. 162-171. doi: 10.1080/00051144.2019.1602293

Sammut, C.; & Webb, G.I. (2016). Feature Selection *Machine Learning and Data Mining* (pp. pp. 1-9). New York 2016: Springer Science+Business Media

Sun, Jian; Zhang, Xiang; Liao, Dan; & Chang, Victor. (2017). *Efficient method for Feature Selection in Text Classification.* Paper presented at the International Conference on Engineering and Technology (ICET).

Wang, Yuxian; & Zhou, Changyin. (2021). *Feature Selection Method Based on Chi-Square Test and Minimum Redundancy.* Paper presented at the Emerging Trends in Intelligent and Interactive Systems and Applications. Proceedings of the 5th International Conference on Intelligent, Interactive Systems and Applications (IISA2020).

Wu, Jiantao; Mo, Shentong; Feng, Zhenhua; Atito, Sara; Kitler, Josef; & Awais, Muhammad. (2024). Rethinking Positive Pairs in Contrastive Learning. *arXiv*, pp. 1-18. doi: 10.48550/arXiv.2410.18200

Yu, Lei; & Liu, Huan. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research - JMLR, 5*(2004), pp. 1205–1224.

Zebari, Rizgar R.; Abdulazeez, Adnan Mohsin; Zeebaree, Diyar Qader; Zebari, Dilovan Asaad; & Saeed, Jwan Najeeb. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends, 1*(2), pp. 56-70. doi: 10.38094/jastt1224