Asian And Pacific Economic Review
ISSN: 1000-6052

THE ROLE OF FOLK ART BASED DESCRIPTIONS AND MARKETPLACE INDICATORS IN PRICING INDIAN FOLK ART ONLINE: AN NLP-BASED TEXT ANALYTICS STUDY

Vidhya Rao
Department of Computer Applications SIES College of Management Studies, Navi Mumbai;
Research Scholar, Chhatrapati Shivaji Maharaj University, Panvel, India
https://orcid.org/0000-0001-7686-7071

Surekha Kohle
Department of Computer Science and Engineering,
Chhatrapati Shivaji Maharaj University, Panvel, India
https://orcid.org/0009-005-8588-3803

https://doie.org/10.65985/APER.2026549597

Abstract

Online marketplaces have become major platforms for the commercialization of Indian folk and traditional paintings, yet empirical evidence on how textual descriptions influence artwork pricing remains limited. This study examines how descriptive language, along with marketplace indicators such as art type, painting area, ratings, and reviews, shapes prices of online folk art listings. We introduce the Indian Painting Ecommerce Metadata (IPEM) dataset comprising 385 manually authenticated online listings of Indian paintings, including textual descriptions, prices, physical dimensions, art form categories, and market signals. Manual verification ex-cluded counterfeit and replica artworks, ensuring dataset reliability. Machine learning–based text analytics are applied using three representation techniques: Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and Word2Vec. Each representation is paired with an appropriate learning algorithm. High-dimensional and sparse BoW features are analyzed using LASSO (Least Absolute Shrinkage and Selection Operator) regression to enable feature selection and interpretability. TF-IDF representations are modeled using Random Forest, while Word2Vec embeddings are combined with XGBoost (Extreme Gradient Boosting) to exploit semantic interactions. Experimental results show that TF-IDF with Random Forest achieves the strongest predictive performance, explaining approximately 73% of the variance in log-transformed prices. The BoW + LASSO model reveals that keywords related to cultural identity, regional heritage, craftsmanship, traditional materials, and emotional aesthetics positively influence pricing, whereas decor-oriented, generic, or reproduction-related descriptors are negatively associated with value. The study provides managerial insights for sellers, marketplaces, and policymakers, emphasizing strategic text optimization to enhance visibility and pricing outcomes.

Keywords: BoW, TF-IDF, Word2Vec, LASSO, Random Forest, XGBoost, Folk art paintings

Introduction

Indian folk art encompasses traditional visual art practices engrained in local communities, expressing the social values, spiritual beliefs, and cultural identities of India's diverse regions. Transmitted across generations, these art forms include well-known traditions such as Madhubani from Bihar, Warli from Maharashtra, Gond from central India, Pattachitra from Odisha and Bengal, and Aipan from Uttarakhand, among others. Characterized by rich symbolism, narrative expressions, region specific motifs and handcrafted techniques, Indian folk art differs fundamentally from formal and contemporary art traditions.

Indian folk art has increasingly transitioned from local exhibitions and galleries to online marketplaces such as Amazon, Memeraki and Etsy, creating new opportunities as well as challenges for artisans and sellers. While digital platforms expand market reach, they also intensify competition and reduce the ability of buyers to physically assess authenticity, craftsmanship, and cultural value. As a result, pricing decisions in online folk art markets are formed largely by informational prompts embedded in folk art listings, particularly textual descriptions.

Despite the growing focus on digital business cases, empirical evidence on how textual communication influences pricing outcomes remains limited, particularly for cultural and creative products such as Indian folk art. The present study addresses this gap by examining how folk art descriptions, along with marketplace indicators such as painting dimensions, consumer ratings, reviews and observable product attributes contribute to price formation in online marketplaces. Text analytics is employed as a decision-support tool rather than as a purely technical implementation.

This study employs a text analytics based analytical framework that integrates Natural Language Processing (NLP) techniques with machine learning models to examine pricing dynamics in online marketplaces for Indian folk art. Folk art descriptions are represented using established NLP approaches, including BoW, TF-IDF and Word2Vec and are analyzed alongside observable marketplace indicators such as area of the paintings, consumer ratings and reviews. Predictive models including LASSO regression, Random Forest, and XGBoost are used to assess the relative contribution of textual and non-textual features to price variation. The emphasis of the analysis is on generating interpretable insights relevant to managerial decision making rather than on technical model comparison alone.

The remainder of this paper is organized as follows: section 2 comprises of related work. Section 3 encompasses research gap of this study. Section 4 comprises of research objectives. Proposed work are recorded in section 5 followed by experimental setup in section 6. Results and discussion are discussed in section 7. Section 8 records ethical considerations and limitations followed by conclusion and future work in Section 9.

1. Related Work

1.1 NLP-Based Price Prediction in Digital Markets

In a study conducted by Tyagi et al. (2024), Natural Language Processing (NLP) techniques were integrated with a Random Forest regressor to predict used car prices using both structured vehicle attributes and unstructured textual descriptions. Singh et al. (2023) evaluated multiple featurization techniques with Random Forest and XGBoost models on Amazon reviews and found that TF-IDF and BoW can outperform embeddings in certain pricing-related contexts. Hegde et al. (2021) have applied statistical and deep learning models for price forecasting of agricultural commodity. They also demonstrate NLP-enabled voice interfaces in regional languages to improve information accessibility, decision making, and market participation among farmers. A recent research by Gao et al. (2025) applies natural language processing techniques to extract textual semantics from product descriptions and integrates them with six machine learning models like linear regression, neural networks, decision trees, support vector machines, random forests, and XGBoost to predict data product prices. Findings indicate that semantic embeddings such as Word2Vec perform better for continuous price prediction, whereas simpler representations like BoW and TF-IDF are more effective for price classification.

1.2 Text Representation Techniques and Machine Learning Models in NLP

Karim et al. (2025) evaluated BoW, TF-IDF, and Word2Vec with SVM classifiers for fake news detection, showing that simpler representations perform comparably to more complex approaches while incurring lower computational costs. Sallam et al. (2025) showed that Word2Vec-based semantic embeddings significantly improve emotion detection performance when compared to traditional BoW and TF-IDF representations. Jin et al. (2024) proposed a TFIDF- SP and LDA–Word2Vec hybrid model for work-order classification, achieving higher clustering accuracy and stability than conventional TF-IDF-based approaches. Tbaikhi et al. (2024) evaluated Word2Vec, FastText, and GloVe with deep learning models and reported that semantic embed- dings, particularly Word2Vec and FastText, yield superior sentiment analysis performance. Premasudha and Patil (2024) compared BoW, TF-IDF and Word2Vec representations for airline sentiment analysis and showed that Word2Vec-based models provide better balance across precision, recall, and F1-score.

Hossain et al. (2024) applied TF-IDF and Word2Vec with ensemble learning models for COVID-19 sentiment analysis and reported the best performance using TF-IDF-based stacking classifiers. Hussein and Al-Naymat (2023) demonstrated that hybrid feature representations combining TF-IDF and Word2Vec with stacking classifiers improve sentiment classification accuracy for remote work related tweets. Wang and Shi (2022) combined Word2Vec with TF-IDF to address sparsity and semantic loss in traditional text representations and demonstrated improved recommendation accuracy compared to standalone TF-IDF and BoW models. Habib et al. (2021) demonstrated that domain specific neural word embeddings trained on large scale medical corpora use Word2Vec to capture

semantic information for healthcare oriented NLP and clinical decision support applications, especially in low-resource languages such as Arabic. Yan et al. (2020) have proposed network based BoW models that incorporate structural and semantic relationships between terms, demonstrating improved text classification performance and higher efficiency compared to conventional text representation methods.

1.3 Geographical Indication (GI) tags

GI tags have been extensively studied for authenticity verification in agricultural products using deep learning, blockchain, AI, and IoT-based traceability mechanisms by Aparna et al. (2024), A study by Upputuri et al. (2024) explored the use of GI tags combined with emerging technologies to enhance product authenticity and consumer trust in e-commerce platforms. It proposed the AuthentiQ Market Space, an integrated framework leveraging blockchain for traceability, artificial intelligence for product authentication, and IoT for real-time monitoring of GI-certified goods. To improve the visibility and accessibility of GI-certified traditional art forms, an augmented reality based mobile system has been proposed for Pedana Kalamkari that employs image based recognition to deliver interactive multimedia content conveying cultural context, craftsmanship, and heritage significance by Nandini et al. (2025).

Despite these advances, existing literature primarily focuses on classification-oriented tasks such as sentiment analysis, topic detection, and recommendation, with limited attention to pricing oriented applications. Moreover, very few studies systematically compare multiple text representations and learning algorithms within a unified experimental framework to understand how textual features influence economic outcomes such as price formation and perceived value.

3 Research Gap

While prior studies demonstrate the effectiveness of text representations such as BoW, TF-IDF, and Word2Vec in classification and sentiment analysis tasks, limited research has systematically examined their role in pricing outcomes within digital marketplaces. In particular, the influence of cultural, authenticity related and emotional textual features on folk art pricing remains underexplored. Addressing this gap, the present study proposes extended prior work by examining the role of folk art textual descriptions and marketplace indicators in pricing Indian folk art paintings using NLP based text analytics and machine learning models.

4 Research Objectives

1. **RQ1:** To evaluate the effectiveness of different text representation methods (BoW, TF–IDF and Word2Vec), with and without additional metadata such as area of painting, ratings and reviews for predicting artwork prices using R^2 and RMSE metrics.
2. **RQ2:** To identify keywords and phrases (cultural keywords, authenticity claims, and

emotional words) in folk art descriptions that significantly influence pricing.

3. **RQ3:** To provide managerial insights for sellers and artisans, online marketplaces and policy makers and cultural institutions on how text optimization can improve market visibility and pricing outcomes for folk art products.

5 Proposed Work

5.1 IPEM Dataset

This research study introduces the IPEM dataset containing textual descriptions, physical dimensions, and market indicators of Indian paintings. This dataset comprises of 385 records from authentic online sources AmazonIndia (2026), Etsy (2026), Kurumba Miniature Painting Set (2025) , Handcrafted Kurumba Painting (2026) and Memeraki (2026) which sell Indian paintings. The authenticity of each data has been manually verified. This was one of the benchmarks for the dataset to guarantee that no counterfeit images or replicas were considered for analysis. Each observation corresponds to an individual Indian painting listed on an online marketplace which includes textual title and descriptions, pricing information, art form category, and market indicators like ratings and reviews.

5.2 Variable Descriptions

The variable descriptions used for analysis is listed in Table 1.

Table 1: Variable Definitions

Category	Variable Name	Description / Operationalization
Dependent Variable	Price	Observed selling price of the artwork in Indian Rupees (INR).
Dependent Variable (Transformed)	Log Price	Natural logarithm of artwork price, used to address right-skewness and heteroscedas-ticity and applied consistently across all regression and machine learning models.
Textual Attributes	Title	Product title provided by the seller, used as unstructured textual input for NLP-based feature extraction.
Textual Attributes .	Description	Detailed folk art description text provided by the seller, used for NLP-based text analytics
Artwork Characteristics	Art Type	Categorical variable indicating the type or style of Indian folk art (e.g., Warli, Gond, Madhubani).

Artwork Characteristics	Area	Total surface area of the artwork, computed as Length × Width.
Seller Characteristics	Successful Orders	Total number of completed transactions by the seller, used as a proxy for seller experience and credibility.
Seller Characteristics	Rating	Platform-provided seller rating on a 1–5 scale, when available.
Seller Characteristics (Constructed)	Final Rating	Composite seller reputation score on a 1–5 scale. When only successful order data are available, the rating is computed using a logarithmic transformation: $1 + 4 \times \frac{\log(1 + \text{Successful Orders})}{\log(1 + 1000)}$ rounded to one decimal place. When successful order information is unavailable, the platform-provided seller rating is used.
Seller Characteristics (Constructed)	Reviews	Total count of customer reviews received by the seller, reflecting market visibility and buyer engagement.

5.3 Text Preprocessing

Textual data from the folk art title, description and art type features were combined and preprocessed prior to feature extraction. Preprocessing steps included conversion to lowercase, removal of punctuation and stop words and lemmatization. This ensured consistency and reduced noise in the textual corpus while preserving semantic meaning.

5.4 Text Representation Techniques and Model Specifications

To quantify textual information, three widely used text representation methods viz BoW, TF– IDF and Word2Vec were employed. These methods allow a comparative evaluation of frequency- based versus semantic text representations in explaining price variation. In a survey conducted by Patil et al. (2023) , (BoW) model represents text as word-frequency vectors, providing a simple and interpretable representation but ignoring word order and semantic relationships. TF- IDF extends BoW by weighting words according to their importance across documents, reducing the influence of common terms while remaining sparse and context independent. In contrast, Word2Vec learns dense, low dimensional word embeddings from large corpora, capturing semantic and syntactic relationships between words based on their contextual usage. Different combinations of text representation and learning algorithms were evaluated. BoW features were analyzed using LASSO regression to enable feature selection. LASSO regression is an effective shrinkage and variable selection technique

for high dimensional text regression, enabling the extraction of relevant textual features to explain continuous outcomes such as price. In a research study by Freo and Luati , BoW is extremely high dimensional, sparse and may contain many irrelevant or redundant features. Lasso regression fits BoW as it is designed to handle sparse, high dimensional feature spaces. In a study by Putri and Mukti (2025) and Shariff et al. (2025), TF-IDF representations were modeled using Random Forest to capture nonlinear relationships while remaining robust to noisy predictors. Random Forest is an ensemble learning method that handles high dimensional, sparse text representations like TF- IDF very well leading to robust classification performance without heavy tuning or overfitting. Word2Vec learns dense semantic vector representations of text that capture contextual and relational meaning among words, which when used as features with XGBoost's gradient-boosted decision trees, significantly improves predictive performance in recommendation and sentiment analysis tasks Paliwal et al. (2022). In our proposed study, Word2Vec embeddings were paired with XGBoost to exploit dense semantic representations, exploit complex feature interactions and enhance predictive performance. Control variables such as art type, artwork area, and seller reputation were included in all model specifications.

5.5 Model Evaluation

Model performance was evaluated using standard regression metrics like R^2 (coefficient of determination) and Root Mean Square Error (RMSE). Coefficients (β) are reported only for the BoW + Lasso model, as its linear and sparse structure enables direct interpretation of the direction and magnitude of keyword effects on prices. While TF-IDF + Random Forest and Word2Vec + XGBoost provide feature importance or gain values reflecting relative relevance, the primary objective of this analysis is interpretability specifically, identifying culturally and descriptively meaningful keywords influencing artwork pricing rather than prediction accuracy alone.

5.6 Workflow Diagram for preprocessing and model evaluation

Comparative analysis across text representation techniques and models was conducted to identify the most effective approach for explaining price variations. A workflow diagram of our proposed study is represented in Figure 1.

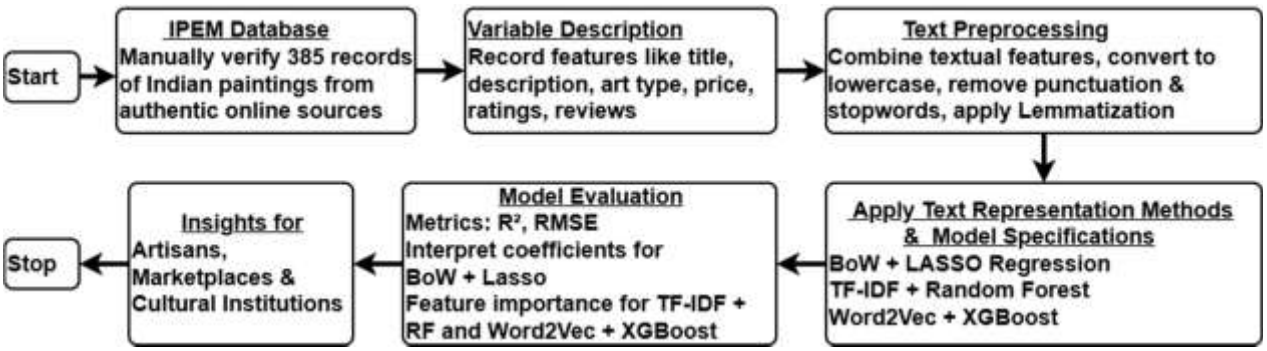

Figure 1: Workflow diagram of the proposed work.

6 Experimental setup

All experiments were implemented in Python using established natural language processing and machine learning libraries. Textual features were extracted using BoW, TF-IDF, and Word2Vec representations, while predictive modeling was performed using LASSO regression, Random Forest, and XGBoost. The experimental pipeline followed standard preprocessing, training, and evaluation procedures to ensure robustness and reproducibility of results.

7 Results and Discussion

This analysis directly addresses the first research objective by comparing alternative text representations under a consistent evaluation framework. The pricing of online folk art is shaped not only by tangible attributes such as painting area, ratings, and reviews, but also by how artworks are described in textual folk art descriptions. This study examines three widely used text representation methods—BoW, TF-IDF and Word2Vec to assess their effectiveness in explaining price variation in Indian folk art listings, both with and without the inclusion of additional metadata. Table 2 compares the predictive performance of alternative text representations and machine learning models. Among the evaluated models, TF-IDF combined with Random Forest shows the best predictive performance, explaining about 73% of the variation in artwork prices (logprice) and achieving the lowest prediction error (RMSE ≈ 1.011). This suggests that words used in artwork titles and descriptions, along with art-type information, contain strong signals for predicting prices.

Overall, textual features are the most important predictors of price. Adding structured information such as area of a painting, ratings and reviews leads to moderate improvements, mainly for linear models. In particular, the performance of the BoW + Lasso model increases when these numeric features are included ($R^2 \approx 0.69$). However, the TF-IDF + Random Forest model shows a small decline in performance after adding metadata, likely because text already captures most of the useful information in this relatively small dataset. Therefore, TF-IDF + Random Forest is best suited for price prediction, while BoW + Lasso is more appropriate for explaining how different features influence the prices of Indian folk art.

Table 2: Comparison of text representation methods and machine learning models in predicting online folk art prices

Dependent variable: LogPrice					
Text Representation Methods	**ML Models**	**Title, Description, ArtType**		**Title, Description, ArtType, Area, Rating, Reviews**	
		R^2	**RMSE**	R^2	**RMSE**
BoW	Lasso	0.6037	1.2270	**0.6864**	**1.0910**
TF–IDF	Random Forest	**0.7310**	**1.0110**	0.6525	1.1489
Word2Vec	XGBoost	0.4517	1.4430	0.4866	1.3966

To address the second research objective, coefficients (β) are reported exclusively for the BoW + Lasso model, where linear assumptions allow direct interpretation of direction and magnitude. For TF-IDF with Random Forest and Word2Vec with XGBoost, feature importance and gain values are reported to reflect relative relevance without directional inference. However, the primary objective of this analysis was not solely price prediction, but the identification and interpretation of culturally and descriptively meaningful keywords influencing artwork pricing. For this purpose, the BoW + Lasso model was selected due to its linear structure and sparse coefficient representation, which allows direct interpretation of feature directionality and magnitude. Table 3 summarizes thematically grouped keywords and their estimated price associations (BoW + Lasso).

As shown in Table 3, positive keywords emphasize cultural depth, craftsmanship, traditional materials, and emotional aesthetics, while negative keywords signal commercial framing, mass production, or generic description, reducing perceived artistic and cultural value. The presence of negative coefficients for folk art terms such as Gond, Aipan, Khovar, and Santhal, which represent strong cultural depth, does not indicate lower cultural or artistic value. Rather, these coefficients reflect market-level pricing patterns observed on online platforms, where such styles are widely produced and frequently sold as decorative or functional items, resulting in lower average listing prices. Consequently, the model captures commercial pricing behavior rather than intrinsic artistic merit or heritage significance. These folk art terms are negative because they are commercially underpriced despite being culturally rich, and the model learns market behavior not artistic value.

Table 3: Interpretation of Keyword Coefficients from BoW + Lasso Model

Keyword Category	Representative Keywords (Coefficient β)	Interpretation
Cultural & Religious Identity	mythology (0.4171), mahal (0.4742), thangka (0.4694), pabuji (0.3137), mata (0.2490), mewar(0.2554), mysore (0.4039), cheriyal (0.2278),pattachitra (0.1946), mughal (0.1715), tanjore(0.1620), india (0.1393), lippan (0.1340)	References to mythology, regional schools, religious figures, and culturally rooted traditions increase prices by adding symbolic depth, heritage value, and cultural authenticity.
Authenticity & Craftsmanship	vintage (0.6127), original (0.2628), detailing (0.2858), fine (0.1722), work (0.1506), timeless(0.1854), create (0.2795), feature (0.2104), body(0.4122), surface (0.3607).	These terms highlight originality, skilled labor, and careful craftsmanship, signaling uniqueness and artistic effort that lead to higher valuation,
Traditional Techniques & Materials	watercolor (0.6237), gouache (0.1554), gold (0.1806), mica (0.1572), textile (0.3236), thread(0.1670), wood (0.1708), canvas (0.1507)	Mentions of traditional media and materials suggest established artistic techniquesand material quality, positively influencing perceived worth.

Keyword Category	Representative Keywords (Coefficient β)	Interpretation
Commercial /Decorative framing	suitable (−0.1039), home (−0.1159), living (−0.3577), cafe (−0.2748), top (−0.3434), theatre (−0.4229)	These words frame artworks as decor utility items rather than collect art, reducing perceived exclusivity artistic seriousness.
Reproduction & Material Cues	mdf (−0.4737), framed (−0.1241), framing (−0.5664), magnet (−0.2676), material (−0.1868), durability (−0.3507), without (−0.3513)	Such terms signal manufactured c ponents or functional construct indicating reproduction rather handcrafted originality.
Generic / Overused Descriptors	painted (−0.1018), vibrant (−0.1623), perfect (−0.2641), typical (0.2390), story (−0.1189), region(−0.1534), ancient (−0.2692), creating (−0.5537)	Broad or vague descriptors lack tura specificity and fail to different artisti value, leading to lower price influence

Table 4 shows that TF-IDF + Random Forest identifies high-importance keywords (e.g., room, living, tribal, warli) that strongly influence price predictions, though these values are non-directional, while Word2Vec + XGBoost captures semantically meaningful keywords (e.g. uplifting, craft) reflecting emotional and cultural value. Low-relevance semantic keywords (e.g., size, print, frame, gift) have minimal impact, indicating that generic descriptors contribute little to price differentiation.

Table 4: Interpretation of Keyword Influence Across Models

Model \& Keyword Effect Type	Representative Keywords (Value)	Interpretation
TF-IDF + Random Forest - High importance (non-directional)	room (Imp = 0.175), living (Imp = 0.065), perfect (Imp = 0.049), tribal (Imp = 0.032), warli (Imp = 0.028), frame (Imp = 0.022), wall (Imp = 0.018), gold (Imp = 0.017), vibrant (Imp = 0.015), work (Imp = 0.015)	Identifies keywords that strongly influence price prediction; values indicate relative importance only and do not specify whether a term increases or decreases price.
Word2Vec + XGBoost - High relevance (semantic)	thought , warming, uplifting, dedication, ritual, calmness, wonder, craft	Captures emotional and experiential semantics associated with perceived artistic value rather than literal material or cultural markers.
Word2Vec + XGBoost - Low relevance (semantic)	size, print, inch, frame, wall, gift, home	Generic physical or commercial descriptors contribute minimally to price discrimination in semantic space.

Finally, the third research objective addresses how optimizing folk art descriptions can guide sellers and artisans, online marketplaces and policy makers and cultural Institutions to improve market visibility and achieve better pricing outcomes.

- **For Sellers and Artisans:** The keyword-level results demonstrate that text optimization plays a critical role in improving both market visibility and pricing outcomes for folk art paintings. Descriptions that emphasize cultural identity (e.g., mythology, regional schools, religious traditions), authenticity and craftsmanship (e.g., vintage, original, detailing), and traditional materials and techniques (e.g., watercolor,

textile, gold) are consistently associated with higher prices, suggesting that buyers value heritage rich and skill intensive narratives. In contrast, framing artworks using decorative, commercial, or reproduction oriented terms (e.g., home décor, framed, MDF, magnet) significantly reduces perceived value by positioning the artwork as a utilitarian product rather than a collectible cultural artifact. These findings provide actionable guidance for artists, sellers, and online platforms to strategically optimize listing text by foregrounding cultural depth, craftsmanship, and material authenticity while avoiding generic or décor-centric language, thereby enhancing both visibility and economic returns in digital marketplaces.

- **For Online Marketplaces**: Online marketplaces can improve the fair valuation of folk art by offering guided description templates that systematically separate cultural information from logistical details such as shipping, framing, and taxes. Templates should prompt sellers to highlight verifiable cultural attributes, including the art form, regional origin, traditional materials, symbolic meaning, artist background, and authentication or certification details, before listing size, framing, or decor related information. Providing structured fields for artist verification, digital authenticity certificates, and customization options can further strengthen buyer trust and perceived artistic value. In parallel, text- analytics systems can be deployed to identify listings where culturally rich artworks are potentially under priced due to dominant emphasis on decorative or functional descriptors. Similarly, analytics can flag overpriced listings that lack cultural, material, or artist level substantiation. Together, these interventions can help platforms balance transparency, cultural integrity, and equitable pricing in digital folk art markets.
- **For Policy Makers and Cultural Institutions**: Data-driven pricing insights can help protect artisans from undervaluation and support sustainable commercialization of cultural heritage. It can also support artists to earn supplementary income. data-driven pricing insights can help identify under- valuation of folk art and support fair, sustainable commercialization of cultural heritage. Promoting and expanding Geographical Indication (GI) certification for folk art forms can strengthen authentic- ity, improve market recognition, and enable artisans to earn stable and supplementary income while safeguarding cultural identity.

8 Ethical Considerations and Limitations

All data used in this study were collected from publicly available sources, with no personal or identifiable customer information involved. The study acknowledges that price may not fully capture the intrinsic cultural or artistic value of folk art. Additionally, platform specific differences and missing seller feedback variables may limit direct generalization of results.

9 Conclusion and Future Work

This paper shows that pricing of Indian folk and traditional paintings in online marketplaces is shaped not only by physical and market attributes but, more importantly, by the textual framing of folk art listings. We introduce the IPEM dataset comprising 385 manually authenticated online listings with textual descriptions, prices, art form categories, physical and market indicators such as ratings and reviews. Manual verification ensured the exclusion of counterfeit or replica artworks. A comparative analysis of BoW, TF-IDF, and Word2Vec representations indicate that TF-IDF combined with Random Forest delivers the highest predictive performance, explaining approximately 73% of the variance in log-transformed prices. For interpretability, however, the BoW + Lasso model reveals that keywords related to cultural identity, regional heritage, craftsmanship, traditional materials, and emotional aesthetics positively influence prices, while generic, decor oriented, and reproduction related terms are negatively associated with value. Negative coefficients for culturally significant folk art terms reflect commercial pricing dynamics rather than diminished artistic or heritage value. The findings offer practical insights for artists and sellers, marketplace platforms and policy makers and cultural institutions on text optimization and pricing strategies, and focus on the policy relevance of data-driven valuation and expanded GI certification for protecting artisans. Future work may integrate image based features and extend the framework to other cultural and creative products.

References

AmazonIndia, 2026. Indian paintings. https://www.amazon.in/s?k=indian+paintings. Retrieved January 2026.

Aparna, B., Shoran, P., et al., 2024. Geographical indications for fruits quality and their potential for rural development using machine learning, in: 2024 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM), IEEE. pp. 1–4.

Etsy, 2026. Indian folk art paintings. https://www.etsy.com/in-en/search?q=Indianfolkartpaintings& ref=search_bar. Retrieved February 2026.

Freo, M., Luati, A., 2024. Lasso-based variable selection methods in text regression: the case of short texts. AStA Advances in Statistical Analysis 108, 69–99.

Gao, R., Xiao, F., Li, J., Cui, S., 2025. Textual semantics and machine learning methods for data product pricing. arXiv preprint arXiv:2511.22185 .

Habib, M., Faris, M., Alomari, A., Faris, H., 2021. Altibbivec: a word embedding model for medical and health applications in the arabic language. IEEE Access 9, 133875–133888.

Handcrafted Kurumba Painting, 2026. Handcrafted kurumba painting (12.8 inches). https://www.mystore. in/en/product/handcrafted-kurumba-painting-12-8-inches-8. Retrieved March 2026.

Hegde, G., Hulipalled, V.R., Simha, J., 2021. Price prediction of agriculture commodities using

machine learning and nlp, in: 2021 second international conference on smart technologies in computing, electrical and electronics (ICSTCEE), IEEE. pp. 1–6.

Hossain, M.A., Rahman, A., Rabbi, M.F., 2024. Covid emotionnet: A machine learning approach to unravel- ing pandemic sentiments, in: 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), IEEE. pp. 505–510.

Hussein, N., Al-Naymat, G., 2023. Sentiment analysis of remote worker tweets during covid-19, in: 2023 24th International Arab Conference on Information Technology (ACIT), IEEE. pp. 1– 6.

Jin, D., Wang, D., Hu, Y., Zhou, S., Huang, G., Hou, B., Gao, J., 2024. Research on intelligent classification of work orders based on tfidf_sp algorithm and lda-word2vec model, in: 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE), IEEE. pp. 342–346.

Karim, A.A.J., Asad, K.H.M., Azam, A., 2025. Strengthening false information propagation detection: Leveraging svm and sophisticated text vectorization techniques in comparison to bert, in: 2025 Interna- tional Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), IEEE. pp. 1–6.

Kurumba Miniature Painting Set, 2025. Kurumba miniature painting set. https://tntribalmart.com/ products/kurumba-miniature-painting-set. Retrieved January 2026.

Memeraki, 2026. Indian folk art paintings. https://www.memeraki.com/search?products[query] =indianfolkartpaintings. Retrieved February 2026.

Nandini, C.D., Rupa, C., Chandana, B., 2025. Ar for gi protection: Exploring kalamkari art for digital heritage and cultural education, in: 2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE. pp. 258–263.

Paliwal, S., Mishra, A.K., Mishra, R.K., Nawaz, N., Senthilkumar, M., 2022. Xgbrs framework integrated with word2vec sentiment analysis for augmented drug recommendation .

Patil, R., Boit, S., Gudivada, V., Nandigam, J., 2023. A survey of text representation and embedding techniques in nlp. IEEe Access 11, 36120–36146.

Premasudha, B., Patil, V., 2024. Enhanced sentiment analysis of airline twitter review using hybrid machine learning and deep learning models, in: 2024 First International Conference on Innovations in Communi- cations, Electrical and Computer Engineering (ICICEC), IEEE. pp. 1–8.

Putri, N.A., Mukti, B.P., 2025. Leveraging tf-idf and random forest to uncover genre patterns in google books metadata. International Journal for Applied Information Management 5, 168–178.

Sallam, A.A., Saif, W.Q., Rassem, T.H., Mohammed, B.A., Alanazi, W., Alshammari, M.K., 2025. Detecting emotional sentiments in textual data using various machine learning and deep learning techniques, in: 2025 IEEE International Conference on Computation, Big-Data and Engineering (ICCBE), IEEE. pp. 411–416.

Shariff, T., Siddique, S.A., Zaid, S., NR, D., et al., 2025. Harnessing random forest classifiers and tf-idf vectorization: A text-based approach to detect stress and analyze mental health using artificial intelligence, in: 2025 3rd International Conference on Smart Systems for applications in Electrical Sciences (ICSSES), IEEE. pp. 1–6.

Singh, R., Kumar, A., Ray, M., 2023. Performances of machine learning models and featurization techniques on amazon fine food reviews. Optimization Techniques in Engineering: Advances and Applications , 187–199.

Tbaikhi, S., Jakha, H., ElHoussaini, S., ElHoussaini, M.A., ElKafi, J., 2024. New approach based on word embedding and deep learning algorithms to optimize the sentiment analysis performance in social business intelligence, in: 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), IEEE. pp. 1–7.

Tyagi, S., Sirohi, S., Singh, Y., Vishwakarma, A., et al., 2024. Hybrid model for predicting used car prices: Integrating natural language processing with random forest regressor, in: 2024 Second International Conference on Advances in Information Technology (ICAIT), IEEE. pp. 1–6.

Upputuri, B., Noorullah, R., Kolluru, R., 2024. Authentiq market space: Leveraging smart city technologies for authentic e-commerce system of geographical indication products, in: 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), IEEE. pp. 1413–1418.

Wang, R., Shi, Y., 2022. Research on application of article recommendation algorithm based on word2vec and tfidf, in: 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), IEEE. pp. 454–457.

Yan, D., Li, K., Gu, S., Yang, L., 2020. Network-based bag-of-words model for text classification. IEEE Access 8, 82641–82652.