Towards Trusted Textual Data Valuation: Information Entropy as the Key

1st Wenze Xiong
Business School
University of Auckland
Auckland, 1010, New Zealand
wenze.xiong@auckland.ac.nz

1st Yetong Wang
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, 215123, China
yetong.wang19@student.xjtlu.edu.cn

2nd Wanxin Li*
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, 215123, China
wanxin.li@xjtlu.edu.cn

3rd Hao Guo
School of Software
Northwestern Polytechnical University
Taicang, 215400, China
haoguo@nwpu.edu.cn

4th Jie Zhang
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, 215123, China
jie.zhang01@xjtlu.edu.cn

***Abstract*—Trusted data circulation requires reliable valuation mechanisms to ensure fair and transparent data exchange in modern information ecosystems. The absence of systematic data valuation methods creates barriers to establishing trust and fairness in data transactions across various domains, including healthcare, finance, and smart cities. This research addresses these challenges by developing an information entropy driven valuation framework for textual data that integrates information content assessment with trust evaluation mechanisms. Our approach provides comprehensive data value measurement through multi-dimensional entropy calculations and systematic trust scoring that evaluates consistency, verifiability, and integrity. Experimental validation through progressive data analysis demonstrates strong correlations between our valuation metrics and practical data utility measures, establishing the effectiveness of entropy-based approaches for trusted textual data valuation in circulation systems.**

***Keywords*-Textual Data Valuation; Information Entropy; Trust Assessment.**

I. Introduction

Trusted data circulation has become a cornerstone of modern digital ecosystems, where textual data serves as a fundamental asset driving innovation across healthcare patient records, financial transaction analyses, and smart city service optimization [1] [2]. However, the absence of systematic valuation mechanisms creates significant barriers to data exchange, as participants struggle to establish fair pricing and assess data quality prior to transactions. This challenge is particularly acute for textual data, which captures rich semantic meaning and behavioral patterns but presents unique pricing complexities due to its contextual dependency and variable utility across applications [3]. In this study, we operationalize data valuation as its measurable impact on downstream machine learning performance, thereby linking the worth of a dataset to its contribution in improving predictive accuracy and robustness.

Current data circulation systems suffer from significant trust deficits rooted in valuation uncertainties. Healthcare consortia that aim to share anonymized patient records for joint research still lack reliable criteria to evaluate and compare the contributions of different datasets [4]. Supply chain networks possessing valuable logistics information face difficulties in determining appropriate compensation for data sharing that benefits entire ecosystems [5]. Financial institutions collaborating in fraud detection face trust deficits due to transparency limitations and unclear benefit-sharing mechanisms in data exchange partnerships [6]. These scenarios illustrate how valuation inadequacies can directly impede the circulation of trusted data.

Existing approaches to data valuation present substantial limitations that hinder trusted circulation. Algorithm-based methods [7] [8] and quality-based models [9]–[11] often lack generalizability and focus insufficiently on data's trusted intrinsic characteristics. Economic approaches including auction-based methods [12], [13] frequently overlook textual data's inherent information value. While information entropy has shown promise in preliminary data pricing research [14], its application to textual data valuation for circulation systems remains largely unexplored.

This research addresses these gaps by developing an entropy-driven valuation framework for textual data circulation that integrates multidimensional entropy calculations with systematic trust assessments. Our approach enables objective valuations, fair pricing, and reduced information asymmetries in data transactions, supporting various scenarios from institutional sharing to emerging marketplaces. The key innovations include:

- We propose a combined entropy calculation method

This study is partially supported by the Jiangsu Province Science and Technology Youth Talent Promotion Program under Grant No. JSTJ-2025-144, the XJTLU Research Development Fund under Grant No. RDF-22-02-106, the Natural Science Basic Research Program of Shaanxi under Grant No. 2025JC-YBMS-688, and the Key R&D Programs of Taicang 2024 under Grant No. TC2024SF10.

integrating character-level, element-level, and structural diversity measures to provide comprehensive data content assessment beyond traditional approaches.
- We develop a trust assessment framework incorporating consistency, verifiability, and integrity dimensions that leverages entropy characteristics for robust reliability evaluation.
- We validate our methodology through experiments demonstrating strong correlations between trusted data values and machine learning performance, bridging theoretical valuation with practical utility.

II. Related Work

Data valuation spans economic, quality-oriented, and algorithmic strands. Economic approaches adapt cost and price–quantity schedule logics from asset appraisal, valuing data by production cost, expected returns, or market comparables [15], [16]. Quality-oriented methods map multidimensional quality—completeness, accuracy, rarity, and related attributes—to value [10], [11]. These signals are clear and practical, but often struggle when utility is highly task-dependent, especially for textual assets whose marginal contribution varies by application.

A related approach estimates value through marginal contributions to model performance. Algorithmic frameworks measure how individual datasets or data points impact predictive outcomes. Cong et al. [7] examine user-oriented pricing strategies for deployed machine learning models. Building on cooperative game theory, Shapley–value–based methods operationalize individual contributions. Jia et al. [17] design approximation schemes for tractable computation. While Shapley methods are widely used, they suffer from high computational costs, vulnerability to distribution shifts, and retrospective design, creating obstacles for valuing data before circulation.

Trust and governance considerations are increasingly shaping data exchange architectures in domains such as healthcare, finance, and supply chains [4]–[6]. Decentralized or consortium models can mitigate custody and compliance risks, yet valuation methods remain necessary to reduce asymmetry, certify utility, and enable transparent benefit sharing. Current governance frameworks do not provide text-specific, model-independent measures that are explainable and cost-effective for advance implementation.

Information-theoretic approaches provide useful signals for this purpose. Entropy quantifies uncertainty and informativeness, with early applications in data pricing [14]. For textual assets, entropy-like quantities (e.g., token-level uncertainty and perplexity) are natural proxies for semantic diversity and potential predictive leverage. However, their direct integration into circulation-oriented valuation remains underexplored. Prior work typically stops short of coupling entropy with trust requirements (such as provenance, consistency, and robustness) or tailoring the signal to the text's contextual and application-dependent utility.

Information entropy has also begun to influence the field of data quality assessment. By capturing uncertainty, diversity, and contextual richness, entropy provides signals that complement traditional quality dimensions such as accuracy or completeness. When integrated with trust requirements, including provenance, consistency, and robustness, entropy-driven valuation rebuilds data quality assessment from static attribute checks toward a dynamic, task-aware, and certifiable perspective.

III. Entropy-driven Data Valuation Framework

A. Information Entropy Calculation

We extend this concept for data valuation by incorporating multiple entropy dimensions for a dataset D. Our combined entropy calculation integrates character-level, element-level, and structural diversity [18], [19] measures as formulated in Equation (1), where $\alpha, \beta, \gamma, \delta$ are weighting parameters with $\alpha + \beta + \gamma = 1$.

$$\begin{aligned} H_{combined}(D) = \alpha H_{char}(D) + \beta H_{element}(D) \\ +\gamma H_{structure}(D) + \delta \cdot S(D) \end{aligned} \tag{1}$$

Character-level entropy measures individual character diversity and captures linguistic richness. Given character frequencies f_i in the dataset, the calculation follows Equation (2), which combines traditional entropy computation with a diversity adjustment term considering the ratio of unique characters to total alphabet size.

$$\begin{aligned} H_{char}(D) = -\sum_{i=1}^{n} \frac{f_i}{F_{total}} \log_2 \left(\frac{f_i}{F_{total}} \right) \\ + \min \left(\lambda_1, \frac{|\text{unique characters}|}{|\text{alphabet}|} \right) \end{aligned} \tag{2}$$

Element-level entropy evaluates content diversity at the text unit level, where varied customer reviews show higher entropy than repetitive template responses.This metric follows Equation (3), where c_j represents the count of elements j, $|D|$ is the total number of elements, and λ_1, λ_2 are diversity weighting parameters.

$$\begin{aligned} H_{element}(D) = -\sum_{j=1}^{m} \frac{c_j}{|D|} \log_2 \left(\frac{c_j}{|D|} \right) \\ + \lambda_2 \cdot \frac{|\text{unique elements}|}{|D|} \end{aligned} \tag{3}$$

Structural diversity entropy captures length distribution patterns and measures formatting variation across the dataset, distinguishing collections with mixed text lengths from uniform-structure data. As shown in Equation (4), n_l represents the number of elements with length l, and L is the maximum element length.

$$H_{structure}(D) = -\sum_{l=1}^{L} \frac{n_l}{|D|} \log_2 \left(\frac{n_l}{|D|} \right) \tag{4}$$

The scale factor incorporates dataset size effects and provides normalization according to Equation (5), where k is a normalization constant, λ_3, λ_4 are scale adjustment parameters, and the logarithmic term ensures appropriate scaling while preventing unbounded growth.

$$S(D) = \lambda_3 + \frac{\log_2(|D| + \lambda_4)}{k} \tag{5}$$

B. Trust Assessment Framework

Data trustworthiness is evaluated through consistency, verifiability, and integrity, measured using dataset size and entropy characteristics. The trust score $T(D)$ is calculated according to Equation (6), where $C(D), V(D), I(D)$ represent consistency, verifiability, and integrity scores of textual data, respectively, and $w_1 + w_2 + w_3 = 1$. The trust score $T(D)$ ranges from 0 to 1, ensuring meaningful integration with entropy measures for data valuation.

$$T(D) = w_1 \cdot C(D) + w_2 \cdot V(D) + w_3 \cdot I(D) \tag{6}$$

Each trust component incorporates dataset size effects through a base scoring function $B(|D|)$ defined in Equation (7). This piecewise function uses threshold parameters $\theta_1, \ldots, \theta_9$ that ensure $B(|D|) \geq 0$, providing different scaling behaviors for small, medium, and large datasets.

$$B(|D|) = \begin{cases} \theta_1 + \frac{|D|}{2\theta_2}, & |D| < \theta_2 \\ \theta_3 + \frac{(|D| - \theta_2)\theta_5}{\theta_4}, & \theta_2 \leq |D| < \theta_6 \\ \theta_7 + \min\left(\theta_8, \frac{|D| - \theta_6}{\theta_9}\right), & \text{otherwise} \end{cases} \tag{7}$$

Consistency $C(D)$ measures the stability of the entropy across subsets of data using Equation (8). Here, cv represents the coefficient of variation of entropy measurements across multiple random subsets, while η_1, η_2, η_3 are adjustment parameters for proper scaling.

$$C(D) = B(|D|) + \max(-\eta_1, \eta_2 - cv \cdot \eta_3) \tag{8}$$

Verifiability $V(D)$ assesses entropy preservation under verification processes as shown in Equation (9). Here, $H_{verified}(D)$ represents the combined entropy from a verified dataset obtained through data cleaning, including removal of duplicates, invalid entries, and noise. The parameters $\phi_1, \ldots, \phi_4$ are weighting factors that control the relative importance of base scoring and entropy preservation.

$$\begin{aligned} V(D) =& \phi_1 \cdot B(|D|) + \phi_2 \cdot \max\Big(\phi_3, \\ & 1 - \frac{|H_{combined}(D) - H_{verified}(D)|}{H_{combined}(D)} \cdot \phi_4\Big) \end{aligned} \tag{9}$$

Integrity $I(D)$ evaluates robustness to data incompleteness using Equation (10), incorporating both the base scoring function and a robustness measure $R(D)$. The robustness measure, defined in Equation (11), tests data resilience by comparing entropy before and after random data reduction, typically involving 15-20% data removal. The weights ψ_1, ψ_2, ψ_3 balance size effects and robustness characteristics, where $H_{reduced}(D)$ represents entropy from the randomly reduced dataset.

$$I(D) = \psi_1 \cdot B(|D|) + \psi_2 \cdot R(D) \tag{10}$$

$$R(D) = \max\left(\psi_3, 1 - \frac{|H_{combined}(D) - H_{reduced}(D)|}{H_{combined}(D)}\right) \tag{11}$$

C. Trusted Data Valuation

The final trusted data value integrates entropy and trust assessments defined as:

$$Value(D) = H_{combined}(D) \times T(D) \tag{12}$$

This formulation ensures that data value reflects both information content (entropy) and reliability (trust), providing a comprehensive valuation metric for various data-driven applications. The non-negative property of both components guarantees meaningful and interpretable valuation results.

IV. Experiments

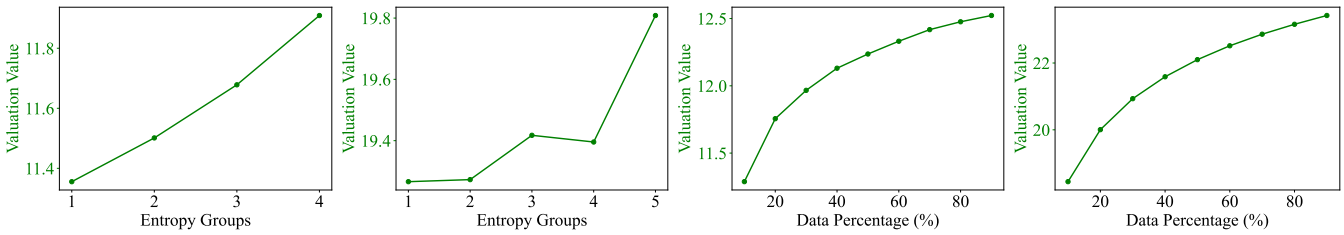
A. Experimental Setup

1) Datasets: To test the method's applicability, we selected two textual datasets: the **SMS Spam Collection** [20] and the **Gender by Name** [21] dataset. The former represents binary text classification, while the latter involves gender prediction from proper names with different linguistic features. We used one text attribute as the predictor and the other as the target variable. Detailed information is presented in Table I.

Table I: Detailed information of selected textual datasets.

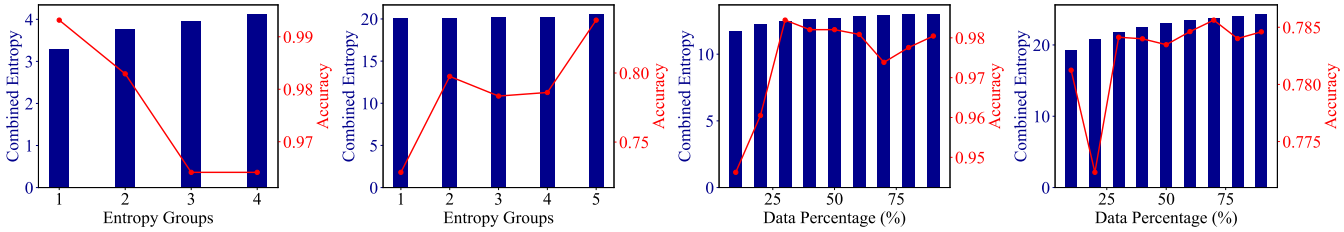
Dataset	Data Types	Records	Attributes
SMS Spam Collection	Text	5572	1
Gender by Name	Text+Num	147269	3

2) Data Division Strategies: We adopted two strategies for subset construction. **Entropy-Based Grouping**: Each record's Shannon entropy is computed, and the dataset is sorted and divided into equal-sized groups, yielding subsets with comparable informational complexity. **Cumulative Percentage Sampling**: After randomly shuffling the dataset, subsets are formed by sequentially taking the first 10%, 20% up to 90%, mimicking a scenario of incremental data growth.



(a) SMS - Entropy (b) Gender - Entropy (c) SMS - Percentage (d) Gender - Percentage

Figure 1: Data valuation using entropy-based methods across datasets with different grouping methods.



(a) SMS - Entropy (b) Gender - Entropy (c) SMS - Percentage (d) Gender - Percentage

Figure 2: Relationship between combined entropy and classification accuracy across datasets with different grouping methods.

3) Evaluation Protocol: We calculate a **Valuation Value** for each data subset and validate its effectiveness by training ridge regression models on each subset. We then analyze the correlation between our Valuation Value and the resulting model accuracy. A strong correlation indicates that our metric effectively predicts data utility.

B. Experimental Results and Analysis

1) Data Valuation Results: Our information entropy-based valuation method produces consistent results across both datasets and grouping strategies, as shown in Figure 1.

For entropy-based grouping, the SMS dataset shows valuation values from 11.36 to 11.91 across four entropy groups (Figure 1a), demonstrating sensitivity to textual complexity variations. The Gender dataset exhibits a wider range of 19.27 to 19.81 across five groups (Figure 1b), reflecting greater diversity in name structures. These results support our theoretical framework that higher entropy content receives higher valuation due to increased information density. Percentage-based grouping reveals dataset size effects on valuation. The SMS dataset increases steadily from 11.29 to 12.52 as data percentage grows from 10% to 90 % (Figure 1c), while the Gender dataset shows a higher increase from 18.45 to 23.42 (Figure 1d). This scaling behavior confirms our method correctly incorporates that larger, more comprehensive datasets possess greater value for machine learning applications.

2) Validation Through Machine Learning Performance: The correlation analysis between valuation results and classification accuracy validates our method, as shown in Figure 2 and Table II. The results reveal distinct patterns: a strong negative correlation (-0.919) for SMS entropy-based grouping correctly identifies the inverse relationship between entropy complexity and classification ease, while positive correlations for gender entropy grouping (0.787) and percentage-based analyses (0.803 for SMS, 0.595 for Gender) confirm that our valuation appropriately captures data utility relationships across different splitting strategies.

For entropy-based grouping, the SMS dataset shows the strongest correlation, with accuracy decreasing from 99.3% to 96.4% as entropy increases (Figure 2a). This suggests high-entropy messages, while informationally valuable, pose greater classification challenges due to linguistic complexity. Conversely, the Gender dataset exhibits a strong positive correlation (0.787), with accuracy increasing from 72.8% to 83.8% (Figure 2b), indicating higher entropy names provide more discriminative features. The percentage-based analysis confirms the expected positive relationship between dataset size and performance, with correlation coefficients of 0.803 and 0.595. SMS accuracy improves consistently from 94.6% to 98.2% as data volume increases (Figure 2c), while the Gender dataset shows fluctuations around 78% with a peak at 60% data inclusion (Figure 2d). These systematic correlations validate that our entropy-driven valuation method

Table II: Valuation Value-Accuracy Correlation Analysis

Experiment	**Correlation Coefficient**
(a) SMS Split by Entropy Group	-0.919
(b) Gender Split by Entropy Group	0.787
(c) SMS Split by Percentage	0.803
(d) Gender Split by Percentage	0.595

successfully captures the relationship between data characteristics and practical utility, establishing our framework's effectiveness.

V. Conclusion

This paper presents a data valuation framework combining information entropy with trust assessment through consistency, verifiability, and integrity measures. Our multi-dimensional entropy approach demonstrates strong correlations with data utility in progressive analysis experiments, establishing quantitative standards for transparent data pricing and providing a reliable foundation for data trading platforms. To address current limitations including static pricing models and limited data type coverage, future research directions include integrating reinforcement learning for dynamic pricing, developing advanced entropy measures for complex data types, implementing federated learning for privacy-preserving valuation, and exploring blockchain implementations for enhanced trust mechanisms, enabling transformation from static to adaptive valuation systems while preserving fundamental entropy-based principles.

References

[1] W. Chen, F. Guo, and F.-Y. Wang, "A survey of traffic data visualization," *IEEE Transactions on intelligent transportation systems*, vol. 16, no. 6, pp. 2970–2984, 2015.

[2] J. Manyika, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute*, vol. 1, 2011.

[3] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, "Text recognition in the wild: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–35, 2021.

[4] H. Subramanian, "A decentralized marketplace for patient-generated health data: design science approach," *Journal of medical internet research*, vol. 25, p. e42743, 2023.

[5] A. R. Harish, X. Liu, M. Li, R. Y. Zhong, and G. Q. Huang, "The new supply chain information sharing renaissance through crypto valuation mechanism of digital assets," *Transportation Research Part E: Logistics and Transportation Review*, vol. 195, p. 103962, 2025.

[6] E. O. Udeh, P. Amajuoyi, K. B. Adeusi, and A. O. Scott, "The role of big data in detecting and preventing financial fraud in digital transactions," *World Journal of Advanced Research and Reviews*, vol. 22, no. 2, pp. 1746–1760, 2024.

[7] Z. Cong, X. Luo, J. Pei, F. Zhu, and Y. Zhang, "Data pricing in machine learning pipelines," *Knowledge and Information Systems*, vol. 64, no. 6, pp. 1417–1455, 2022.

[8] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," *Journal of the ACM (JACM)*, vol. 62, no. 5, pp. 1–44, 2015.

[9] M. Li, H. Feng, F. Chen, and J. Kou, "Optimal versioning strategy for information products with behavior-based utility function of heterogeneous customers," *Computers & Operations Research*, vol. 40, no. 10, pp. 2374–2386, 2013.

[10] J. Yang, C. Zhao, and C. Xing, "Big data market optimization pricing model based on data quality," *Complexity*, vol. 2019, no. 1, p. 5964068, 2019.

[11] W. Xiong, Y. Wang, W. Li, Y. Zhang, J. Zhang, and H. Guo, "An advanced pricing mechanism for nonfungible tokens (nfts) based on rarity and market dynamics," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 6, pp. 7671–7684, 2024.

[12] X. Cao, Y. Chen, and K. R. Liu, "Data trading with multiple owners, collectors, and users: An iterative auction mechanism," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 268–281, 2017.

[13] J. Duan, L. Tian, J. Mao, and J. Li, "Optimal social welfare: A many-to-many data transaction mechanism based on double auctions," *Digital Communications and Networks*, vol. 9, no. 5, pp. 1230–1241, 2023.

[14] X. Li, J. Yao, X. Liu, and H. Guan, "A first look at information entropy-based data pricing," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 2053–2060.

[15] Y. Li and J. Li, "Research on the method of evaluating the value of internet enterprise data assets," *Econ. Res. Guide*, vol. 14, pp. 104–107, 2017.

[16] S. Mehta, M. Dawande, G. Janakiraman, and V. Mookerjee, "How to sell a data set? pricing policies for data monetization," *Information Systems Research*, vol. 32, no. 4, pp. 1281–1297, 2021.

[17] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, "Towards efficient data valuation based on the shapley value," in *The 22nd International Conference on Artificial Intelligence and Statistics*, vol. 89. PMLR, 2019, pp. 1167–1176.

[18] Y. Shi and L. Lei, "Lexical richness and text length: An entropy-based perspective," *Journal of Quantitative Linguistics*, vol. 29, no. 1, pp. 62–79, 2022.

[19] C. Shaib, J. Barrow, J. Sun, A. F. Siu, B. C. Wallace, and A. Nenkova, "Standardizing the measurement of text diversity: A tool and a comparative analysis of scores," *arXiv preprint arXiv:2403.00553*, 2024.

[20] T. Almeida and J. Hidalgo, "SMS Spam Collection," UCI Machine Learning Repository, 2011, DOI: https://doi.org/10.24432/C5CC84.

[21] "Gender by Name," UCI Machine Learning Repository, 2020, DOI: https://doi.org/10.24432/C55G7X.