applied sciences

MDPI

Article

NLP-Based Consumer Complaint Assessment: Synthetic Data Generation, Featurization, and Evaluation Metrics

Peiheng Gao [1,*], Chen Yang [2,3], Ning Sun [1] and Ričardas Zitikis [1,†]

[1] School of Mathematical and Statistical Sciences, Western University, London, ON N6A 3K7, Canada; nsun46@uwo.ca
[2] Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; chen.yang@mountsinai.org
[3] Institute for Health Care Delivery Science, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
* Correspondence: pgao47@uwo.ca; Tel.: +1-226-577-2163
† Deceased author.

Abstract

Machine learning (ML) has substantially advanced text classification by enabling automated understanding and categorization of complex unstructured textual data. However, accurately capturing nuanced linguistic patterns and contextual variations that are inherent in natural language, particularly within consumer complaints, remains a challenge. This study addresses these limitations by incorporating human-experience-trained algorithms that effectively recognize subtle semantic distinctions that are crucial for assessing consumer relief eligibility. Specifically, we propose synthetic data generation methods that leverage generative adversarial networks refined through expert labeling, complemented by featurization approaches for extracting textual representations. By combining expert-trained classifiers with high-quality synthetic data, this research demonstrates marked improvements in ML classifier performance, reduced dataset acquisition costs, and enhanced robustness across evaluation metrics in text classification tasks.

Keywords: consumer complaints; natural language processing; human-experience-trained algorithms; generative adversarial network

1. Introduction

ML is a discipline that is devoted to the development of algorithms and statistical models that enable computational systems to autonomously learn from data and perform tasks such as prediction, classification, and decision-making without explicit programming. In text classification, ML techniques have dramatically enhanced the capability of automated systems to process and interpret large volumes of unstructured data. Complementarily, natural language processing (NLP) is an interdisciplinary field focused on the computational understanding, generation, and analysis of human language; its advancements empower classifiers to discern nuanced linguistic patterns and support applications ranging from sentiment analysis to topic categorization. Moreover, recent developments in text generation, such as generative adversarial networks (GANs) [1], have facilitated the production of realistic synthetic textual data, further enriching training datasets and enhancing model robustness. Both ML and NLP approaches critically depend on the quality and diversity of training data to accurately capture the intricate structure of language, which frequently encompasses complexities such as slang, contextual meaning, and semantic variation.

Check for updates

Academic Editor: Christos Bouras

Received: 25 March 2026
Revised: 5 June 2026
Accepted: 8 June 2026
Published: 13 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Hence, incorporating human judgment into text classification processes is essential as it facilitates the recognition of subtle linguistic variations and complex nuances inherent in natural language, particularly when assessing whether consumer complaints warrant relief. In Gao et al. [2], human-experience-trained algorithms were developed and applied to analyze a large corpus of the Consumer Financial Protection Bureau (CFPB) records, with the findings confirming that the incorporation of such algorithms is critical to the discriminative efficacy of ML classifiers.

While human-experience-trained algorithms are indispensable for ML classifiers (see, e.g., [3]) to capture subtle linguistic nuances in consumer complaints, the exclusive reliance on real-world data, which often exhibits various imperfections, underscores the necessity of integrating synthetic data generation into text classification pipelines. The effectiveness of ML classifiers is largely contingent upon the quality and representativeness of the training data. However, real-world complaint datasets typically demonstrate significant variability in language usage, semantics, and contextual nuances.

This study proposes a two-pronged strategy to enhance the effectiveness of ML-based text classifiers. First, we integrate the expertise embedded in expert-trained ML algorithms to facilitate robust comparisons among multiple classifiers during training. Second, we leverage expert labels to generate high-quality synthetic complaint data that closely approximates real-world scenarios. Collectively, these two strategies improve classification performance while substantially reducing the costs inherent to acquiring new complaint datasets.

1.1. Related Studies

The dataset utilized in this analysis, provided by the CFPB, consisted of textual records detailing a substantial number of monetary and non-monetary relief measures granted in response to consumer complaints [4]. The CFPB offers valuable insights, particularly for understanding the underlying structure of consumer complaints. One of the major product categories is mortgage complaints. Dou and Roh [5] employed a simple linear regression model and concluded that the public disclosure of mortgage complaints enhances market discipline and improves financial protection for consumers. Begley and Purnanandam [6] examined regulatory challenges across different racial groups with respect to the quality of financial services using CFPB data. Their findings highlight an unintended adverse consequence of quantity-focused regulations on the quality of credit extended to lower-income and minority customers.

The applicability of NLP-based text classification extends across a wide range of domains, as evidenced by a growing body of empirical work. In document management, Iwaszenko et al. [7] proposed a LangGraph-based multi-agent system for automated document classification in R and D workflows, achieving a peak zero-shot accuracy of 97.3% while demonstrating that mid-sized models offer a favorable accuracy-efficiency trade-off. Alotaibi [8] proposed the HDNNM-CSA model combining Bi-LSTM, TCN, and GCN with TF-IDF embeddings for customer satisfaction classification, achieving 97.53% test accuracy. Beyond structured domains, Schwarz et al. [9] proposed a training-free zero-shot framework using Mistral-related models for disaster-related social media content, achieving a binary F1 of 0.84 for informativeness detection and a weighted F1 of 0.61 for eight-class humanitarian categorization. At the methodological level, Bazdar et al. [10] conducted a review of 36 NLP studies on traffic crash narratives, finding that topic models, encoder-based models, and decoder-based models serve complementary analytical roles, with hybrid text–tabular approaches consistently outperforming unimodal baselines. Extending these insights to institutional settings, Eyupoglu et al. [11] proposed a human-in-the-loop decision support system combining BERT-based semantic classification with a LoRA fine-tuned

LLaMA 2 generator for university student query routing, achieving 95.88% classification accuracy, a 30.8% reduction in resolution time, and a 32.6% decrease in misrouting.

Selecting an appropriate model is crucial in the task of text classification as it not only conserves computational resources but also enhances the quality of subsequent analyses. Ranganathan and Abuka [12] fine-tuned the text-to-text transfer transformer [13] (T5) model for text summarization, effectively capturing key terms from the original text and subsequently improving classification performance. Kamyab et al. [14] combined a convolutional neural network with a bidirectional long short-term memory network to develop a model that demonstrates excellent performance in classifying datasets, particularly narrative texts. Huang et al. [15] deployed the FinBERT model, a variant of BERT fine-tuned on a financial textual dataset, which yielded particularly strong performance on financial data.

Synthetic textual datasets are essential in research and practical applications, particularly when access to extensive high-quality real-world data is limited by privacy concerns, regulatory constraints, or logistical challenges. By generating synthetic data that closely approximates the structural and semantic complexities of natural language, researchers can conduct controlled experiments to rigorously evaluate and refine natural language processing models under various conditions. Shahriar [16] offered an extensive survey of recent studies utilizing GANs for the generation of visual arts, music, and literary texts. In the context of literary text generation, the long short-term memory (LSTM) architecture has been widely adopted as the generator component within the GAN framework. Broadening the scope to deep generative models, Iqbal and Qureshi [17] compared the performance of several architectures (e.g., recurrent neural networks (RNNs), LSTM, gated recurrent units, and bidirectional RNNs) and concluded that LSTM achieves superior performance in generating textual data. Imran et al. [18] extended traditional text generation architectures (e.g., RNNs, LSTM, and convolutional neural networks) by introducing the CatGAN model; their comparative analysis of imbalanced datasets and datasets containing generated reviews demonstrates that CatGAN is capable of producing higher-quality text without sacrificing diversity. Li et al. [19] introduced a feature-aware conditional GAN (FA-GAN) for controllable category-based text generation, concluding that FA-GAN not only meets the specified categorical requirements but also effectively captures the salient features of conditioned sentences, all while maintaining readability, fluency, and authenticity.

Generative adversarial networks have recently been applied in various fields. Xia et al. [20] proposed Market-GAN, a controllable generator with semantic context for the financial simulation of market features. Wang et al. [21] proposed an encoder–decoder architecture to learn the latent variable distribution from news report data, thereby providing background knowledge; the authors employed a teacher–student network to distill knowledge, refining the output of the decoder component. Yang et al. [22] presented CTGGAN, a novel model for constraint-based text generation. The model uses a language model enhanced with a bias network to produce text that meets specified criteria. CTGGAN employs a discriminator that integrates sentiment and fluency evaluations to assess the quality of the generated text and guide improvements in the generator. Dalvi et al. [23] integrated coalescent theory with a generative adversarial network to generate financial textual data and concluded that the resulting synthetic datasets are diverse and capture the complexity of authentic financial texts, thereby providing valuable resources for financial modeling and analysis. Koç et al. [24] provided a comprehensive survey of recent advances in the application of GANs to natural language processing and concluded that GANs can be effectively applied to textual data, achieving notable success in tasks such as text generation, translation, and language modeling.

Model performance is typically evaluated using a variety of metrics. When addressing imbalanced datasets, Gao et al. [2] adopted the F1 score as the primary metric. For binary classification tasks, Chicco et al. [25] recommended the Matthews correlation coefficient (MCC), noting that it provides a more comprehensive assessment of classification performance. Kolesnyk and Khairova [26] underscored the advantages of Cohen's kappa, particularly for assessing agreement levels across various NLP and text mining tasks. Khraisha et al. [27] applied Cohen's kappa to compare the performance of large language models (LLMs) with that of human reviewers in systematic reviews, concluding that LLMs are capable of rivaling human performance.

1.2. Research Gaps

Several critical research gaps have been identified in the realm of textual data analysis within consumer complaint narratives. First, although NLP techniques have advanced substantially, most methodologies are designed for general-purpose applications and fail to capture the unique linguistic characteristics of consumer complaint narratives. These narratives often feature informal language, emotional expressions, colloquialisms, and industry-specific jargon related to various products and services. Moreover, consumer complaints frequently contain non-standard abbreviations, slang, and context-dependent expressions, which present significant challenges to generic NLP models, particularly in the absence of extensive domain-specific annotated corpora.

Second, the inherent heterogeneity and unstructured nature of consumer complaint narratives pose significant challenges for data preprocessing, normalization, and information extraction. Given that a diverse range of consumers submit these complaints, variations in writing style and vocabulary are common, particularly when addressing different products or services. This variability often results in inconsistencies (e.g., misspellings, divergent terminologies, and irregular syntactic structures) that complicate efforts to consolidate the data into a cohesive analytical framework. Consequently, traditional data processing pipelines, which are typically optimized for more homogeneous datasets, often struggle to effectively manage and analyze the multifaceted nature of consumer complaint narratives.

1.3. Objective

The present study examines the impact of human biases on the NLP-based classification of consumer complaints provided by the CFPB into meritorious and non-meritorious categories, with particular emphasis on performance under four different evaluation metrics. Specifically, the analysis proceeds in four steps:

1. Evaluating the fidelity of T5-generated summaries to the original complaint narratives in terms of semantic alignment and keyword-level overlap to establish a reliable textual foundation for downstream classification.
2. Proposing the FinBERT-IDF$_{\text{token}}$ featurization scheme and benchmarking it against baseline featurization approaches across multiple classifiers, including traditional models (LR and RF) and deep learning models (LSTM and FinBERT), on real complaint narratives using four evaluation metrics, namely accuracy, F1 score, MCC, and Cohen's kappa (C_κ) [28].
3. Evaluating the impact of labeling quality on classification performance by comparing expert labels derived from ground-truth CFPB labels against staff labels simulated through a Bernoulli distribution [2], assessed across accuracy, F1 score, MCC, and C_κ.
4. Developing a class-conditional GAN-based framework for synthetic complaint narrative generation and assessing the extent to which the generated data replicates the class-conditional statistical distributions of the original CFPB corpus across downstream classification tasks.

1.4. Contributions

The present study offers several practical implications for insurance companies, regulators, and related parties seeking to use large volumes of customer complaints to train classifiers that support dispute resolution. The main contributions are as follows:

1. Modifying the traditional word embedding approach to leverage its advantages while reducing computational time.
2. Building a model that yields superior performance in detecting meritorious consumer complaints.
3. Simulating a textual dataset that achieves performance that is comparable to that of the raw complaint narratives.
4. Demonstrating the utility of the proposed procedure for identifying meritorious customer complaints in the CFPB dataset through in-depth analysis.

The rest of the paper is organized as follows. Section 2 describes the subset of the CFPB consumer complaint data selected to illustrate the proposed procedure. Section 3 details the transformation of raw narratives into quantitative representations via two novel featurization approaches, with discriminative performance assessed across a diverse set of evaluation metrics, namely F1 score, Matthews correlation coefficient, Cohen's kappa, and accuracy. Section 4 presents a framework incorporating LSTM and a generative adversarial structure that is employed to synthesize textual data by integrating knowledge from human-experience-trained algorithms. Section 5 concludes the paper.

2. Data

The present study uses a dataset of a similar structure to that of Gao et al. [2]. However, by involving more claims in our analysis, we tend to work with a larger dataset. Namely, the selected complaints are restricted to those whose narratives satisfy the following three properties:

1. The claims were submitted over a five-year period spanning 1 January 2020 to 31 December 2024.
2. The claims labeled as "closed with explanation", "closed with non-monetary relief", and "closed with monetary relief" are selected. Claims resolved with monetary or non-monetary relief are labeled as "meritorious", while those resolved solely through responses are classified as "non-meritorious" [4].
3. The claims are selected with a frequency exceeding 1000 under the categorical columns (e.g., product, issue, and company).

Claims with dollar amounts outside the range \$1–\$10,000 are excluded as larger amounts are more likely to correspond to business rather than individual accounts [29]. Hence, the resulting data consist of 31,036 complaints, of which 15,052 are meritorious and the remaining 15,984 are non-meritorious. Since the number of data samples labeled as meritorious and non-meritorious is nearly equal, oversampling techniques are not considered in the following analysis. Each record contains seven features: "date received", "product", "issue", "company", "consumer complaint narrative", "dollar value", and "meritorious", where "dollar value" is a numerical variable extracted from the "consumer complaint narrative" column. Figure 1 depicts the distribution of the "dollar value" column, partitioned into twenty equal intervals of \$500 within the range $(\$0, \$10{,}000]$, along with an additional bar representing claims exceeding \$10,000. Notably, over 90% of complaints fall within the lower dispute range (i.e., $(\$0, \$1000]$), suggesting that the majority of consumer grievances involve relatively modest financial amounts. Records with dollar values exceeding \$10,000 account for a negligible proportion of the corpus, as evidenced by the markedly lower frequency of the rightmost bar relative to the retained intervals. Such high-value

claims are more likely attributable to business accounts rather than individual consumer disputes and are therefore excluded from the analysis. In Figure 1, the distribution of the dollar value is heavily right-tailed; accordingly, a logarithmic transformation is applied to this numerical column. This transformation stabilizes the variance and mitigates skewness, making the data more approximately normally distributed and improving the reliability of subsequent statistical analyses.

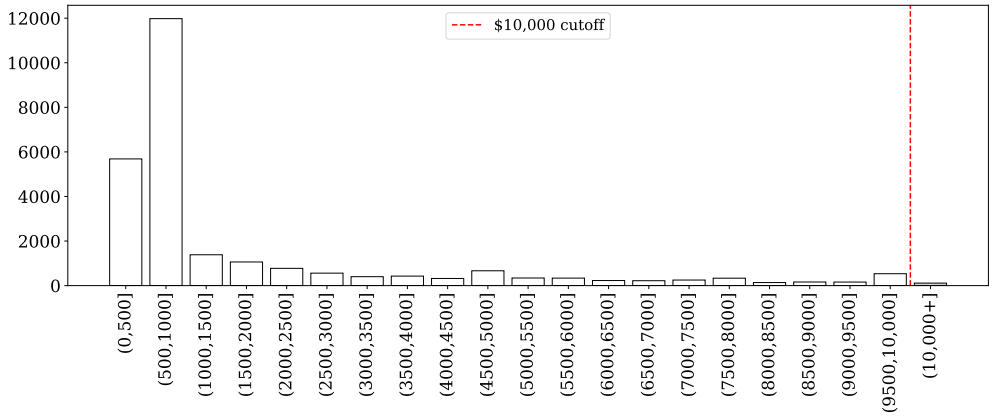


Figure 1. Frequencies of the complaint narratives in various dollar values.

The columns "product," "issue," and "company" serve as categorical features, while "meritorious" is the target variable. The column "consumer complaint narrative" provides the primary textual input for the present analysis. Hence, the following preprocessing steps are applied to clean the narratives:

1. Removing redundant whitespace, i.e., merging multiple consecutive spaces or lines into one if they do not serve a purpose.
2. Removing empty documents or texts that contain almost no useful content, i.e., extremely short texts with only one or two words.
3. Removing special characters that cannot be parsed or have no semantic meaning (i.e., "XXXX", "XXXX/XX/XX").

For better analysis of the textual dataset, corpora with shorter text lengths generally yield better performance than those with longer lengths [26,30]. For complaint data, text lengths can vary. To make each complaint comparable, we use the T5-base pretrained model provided by Google [13] to summarize the text to a specified length. Additionally, we restrict the length of each summarized text to a maximum of 128 words (150 tokens) in order to facilitate a more rigorous and detailed analysis, as recommended by Ranganathan and Abuka [12].

To assess how closely our generated summaries match the source texts in meaning, we measure the cosine similarity between their vectorized representations. Following Di Vincenzo et al. [31], we first convert the original document and its T5-base summary into bag-of-words vectors based on term frequencies. Cosine similarity then evaluates the cosine of the angle between these high-dimensional vectors, yielding a value between zero (no overlap) and one (perfect alignment). Because it depends only on the orientation of vectors, this scale-invariant metric emphasizes semantic correspondence independently of text length or absolute word counts.

Suppose we randomly select a narrative (s_o) from our corpus and use the pretrained T5-base model to summarize its corresponding textual sentence (s_s) as

$$s_o = \text{unexpected fees charged to my credit account and inaccurate statements;}$$

$$s_s = \text{unexpected fees and inaccurate credit statements.}$$

Table 1 presents the first step in computing cosine similarity: collecting all unique words in s_o and s_s and then counting the frequency of each word. Thus, the cosine similarity ($\mathcal{CS}$), defined as the cosine of the angle between the two vectors, is determined to be

$$\mathcal{CS}(s_o, s_s) = \frac{s_o \cdot s_s}{\|s_o\|\|s_s\|} = \frac{(1,1,1,1,1,1,1,1,1,1) \cdot (0,1,0,1,1,1,0,1,0,1)}{\sqrt{10}\sqrt{6}} \approx 0.775.$$

The calculated cosine similarity, approximately 0.775, indicates a high degree of semantic alignment between the summary and the original sentence. This result implies that the summary effectively captures the essential content of the original while omitting less critical details.

Table 1. Word frequencies between s_o and s_s.

	Account	**And**	**Charge**	**Credit**	**Fees**	**Inaccurate**	**My**	**Statements**	**To**	**Unwanted**
s_o	1	1	1	1	1	1	1	1	1	1
s_s	0	1	0	1	1	1	0	1	0	1

To complement this vector-level assessment with explicit term-level evidence, we compute the Jaccard similarity [32] between the keyword sets extracted from s_o and s_s after stopword removal. The Jaccard similarity between two sets A and B is defined as

$$\mathcal{JS}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A and B denote the keyword sets extracted from s_o and s_s, respectively. For the running example, $A = \{\text{fees, charged, credit, account, inaccurate, statements}\}$ and $B = \{\text{fees, inaccurate, credit, statements}\}$, yielding

$$\mathcal{JS}(A, B) = \frac{4}{6} \approx 0.667,$$

indicating that the majority of domain-critical keywords survive summarization, with the omitted terms (charged, account) representing peripheral rather than conceptually central content. Building on this overlap-level evidence, we additionally assess keyword importance ordering via the Spearman rank correlation [33], defined as

$$r_s = 1 - \frac{6\sum_{i=1}^{k} d_i^2}{k(k^2-1)} \tag{2}$$

where $d_i = \text{rank}_A(w_i) - \text{rank}_B(w_i)$ denotes the rank difference of the i-th shared keyword and k is the number of shared keywords. Table 2 presents the frequency ranks of the shared keywords in s_o and s_s. Since all four shared terms appear exactly once in both texts, $d_i = 0$ for all i, giving

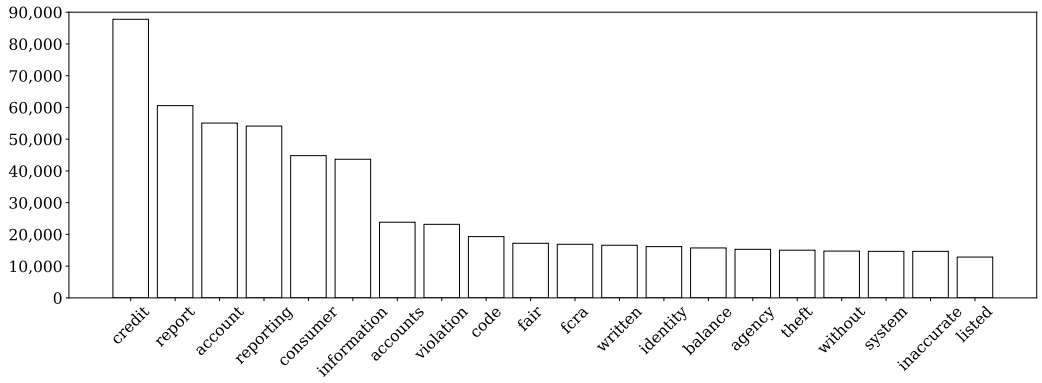
$$r_s = 1 - \frac{6 \times 0}{4(4^2-1)} = 1.000,$$

confirming that the relative importance ordering of shared keywords is perfectly preserved in this illustrative case.

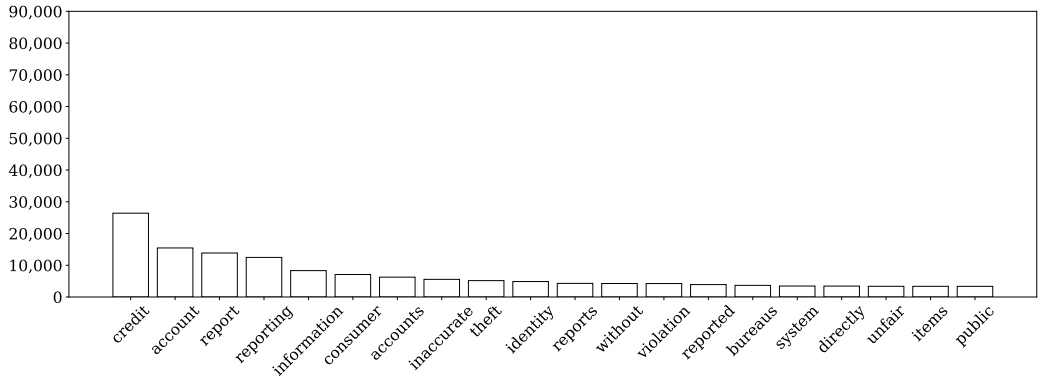
To validate these single-example findings at scale, we extend the analysis to the full CFPB corpus. Figure 2 presents the top-20 most frequent keywords extracted from the raw narratives and the T5-generated summaries, respectively, providing a visual overview of keyword preservation across the entire dataset.

Table 2. An illustrative example of keyword frequency ranks in s_o and s_s, demonstrating the computation of Spearman rank correlation.

Keyword	Freq. in s_o	Rank in s_o	Freq. in s_s	Rank in s_s
fees	1	1	1	1
credit	1	2	1	2
inaccurate	1	3	1	3
statements	1	4	1	4



(**a**) Raw narratives.



(**b**) Summarized text with maximum 128 words.

Figure 2. Comparison of the top-20 most frequent keywords before (**a**) and after (**b**) T5 summarization.

At the corpus level ($k = 20$), the Jaccard similarity between the top-20 keyword lists extracted from the raw narratives and the T5-generated summaries is 0.482, reflecting moderate keyword-level overlap, which is consistent with and expected of abstractive summarization given the inherent vocabulary compression that occurs when generating concise summaries from longer narratives. The Spearman rank correlation among the n = 13 shared keywords is $r_s = 0.868$ ($p < 0.001$), providing stronger evidence that the relative importance ordering of shared keywords is well preserved after summarization. Furthermore, the mean TF-IDF cosine similarity computed across all complaints is 0.518 ($SD = 0.238$), reflecting the vocabulary compression expected of abstractive summarization rather than a loss of semantic information.

Taken together, the single-example results (cosine similarity = 0.775, Jaccard similarity = 0.667, Spearman correlation = 1.000) and the corpus-level results (Jaccard = 0.482, Spearman $r_s = 0.868$, $p < 0.001$, cosine similarity mean = 0.518, $SD = 0.238$) collectively confirm that T5 summarization preserves semantic alignment and importance ordering to a substantial degree. As noted by the Reviewer, the Spearman correlation of $r_s = 0.868$ constitutes the stronger evidence of keyword importance preservation, indicating that the

relative discriminative importance of shared keywords is well preserved after summarization. The moderate Jaccard similarity of 0.482, while reflecting the expected reduction in exact keyword overlap under abstractive compression, is nonetheless consistent with the degree of vocabulary compression inherent in the T5 summarization process and does not indicate a loss of semantic fidelity. Collectively, these results support the use of T5 summarization as a reliable preprocessing step for the present corpus.

3. Converting Narratives into Quantitative Data

In the current section, we incorporate all features into our analysis (e.g., the textual component and the categorical components). The textual column is arguably the most critical element in our dataset as it contains valuable semantic information. We first transform the textual data into a numerical representation, which is subsequently integrated with the categorical components and used as input for our machine learning models. We then evaluate the performance of different ML classifiers using a range of evaluation metrics.

3.1. Text Featurization Approaches

Text featurization, the initial step in many classification tasks, is a fundamental process that transforms unstructured textual data into a structured representation, thereby facilitating efficient processing by ML algorithms. Extracting meaningful features uncovers the intrinsic patterns and contextual relationships within the text. We next implement the following two text featurization approaches of textual data into numerical matrices.

3.1.1. Truncated Term Frequency–Inverse Document Frequency (Truncated TF-IDF)

TF-IDF is a statistical technique that evaluates the importance of individual words within a document relative to their frequency across the entire corpus. It is widely used in information retrieval tasks (e.g., document classification, scoring, and relevance ranking), as described in, e.g., Gao et al. [2] and references therein. The dimensionality reduction step described below applies truncated singular value decomposition, equivalent to latent semantic analysis [34], and is a standard technique rather than a novel contribution.

Formally, let $\mathcal{C} = \{c_1, \ldots, c_n\}$ be the set of cleaned complaint narratives and $\mathcal{W}$ the set of all words in $\mathcal{C}$. The term frequency $\mathrm{TF}(c_i, w_j)$ [35] counts how many times word w_j appears in narrative c_i. The inverse document frequency is defined as

$$\mathrm{IDF}(w_j) = \log\left(\frac{n}{n_j}\right),$$

where $n_j = \sum_{i=1}^{n} \mathbf{1}\{\mathrm{TF}(c_i, w_j) > 0\}$ counts the number of narratives containing w_j. The TF-IDF score is then [35]

$$\text{TF-IDF}(c_i, w_j) = \mathrm{TF}(c_i, w_j) \times \mathrm{IDF}(w_j). \tag{3}$$

Using Equation (3), the cleaned narratives are represented as a matrix of n rows and $|\mathcal{W}|$ columns.

When applying TF-IDF to a large corpus (e.g., 31,036 narratives), each document is represented in a high-dimensional space where the dimensions correspond to every unique word in the entire collection. Since any individual document contains only a small subset of these words, most entries in the resulting document–term matrix are zero. This high level of sparsity is inherent to the nature of textual data, where vocabulary size grows significantly with corpus size, yet each document uses only a fraction of the total vocabulary.

To address the challenges associated with sparse matrices, we apply truncated singular value decomposition (SVD) to the TF-IDF matrix, a well-established dimensionality

reduction technique known as latent semantic analysis (LSA), or, equivalently, latent semantic indexing (LSI) [34]. Following Huang et al. [15], we reduce the dimensionality to a 768-dimensional vector. Given a TF-IDF matrix

$$\text{TF-IDF}(c_i, w_j) \in \mathbb{R}^{n \times |\mathcal{W}|},$$

the truncated SVD factorization yields:

$$\text{TF-IDF}(c_i, w_j) \approx U \Sigma V^T,$$

where $U \in \mathbb{R}^{n \times 768}$ is the matrix of left singular vectors, $\Sigma \in \mathbb{R}^{768 \times 768}$ is a diagonal matrix containing the top-768 singular values, and $V \in \mathbb{R}^{|\mathcal{W}| \times 768}$ is the matrix of right singular vectors. The resulting lower-dimensional representation, commonly referred to as truncated SVD or LSA [34], reduces the disproportionate influence of common terms and alleviates sparsity. Under this featurization approach, the dimensionality is constrained to a predetermined threshold, thereby decreasing the weight assigned to high-frequency terms. The resulting dense matrix offers a more balanced representation of text data, improving task performance.

3.1.2. FinBERT-IDF$_{\text{token}}$

TF-IDF is a widely used method for feature representation at the word level, often producing strong results for short texts with relatively small vocabularies. Its effectiveness mainly arises from its ability to measure the importance of a term in a document compared to its frequency in the entire corpus. However, as the corpus size increases, the vocabulary expands significantly, leading to higher dimensionality in the feature space and consequently increasingly sparse TF-IDF matrices. This sparsity not only complicates subsequent analyses but also presents challenges for many machine learning algorithms that find it difficult to learn effectively from high-dimensional sparse data.

Furthermore, the vocabulary size can expand dramatically when transitioning from word-level to subword-level feature representation (e.g., by breaking down each word into smaller tokens). This expansion further worsens the sparsity issue, particularly in cases where languages with complex morphological structures or subword segmentation techniques are utilized. Although dimensionality reduction techniques, such as singular value decomposition, can be used to compress the high-dimensional TF-IDF matrix, these methods are essentially linear approximations and do not address the fundamental limitation of TF-IDF as a static statistical measure.

Hence, a deep featurization approach (e.g., FinBERT) could be necessary. FinBERT [15] is a transformer model pretrained on general text and then fine-tuned on financial documents to yield rich context-sensitive embeddings. Given an input sequence of n tokens $\mathcal{S} = \{x_1, \ldots, x_n\}$, each initial embedding of token is formed by summing three components:

$$\mathbf{x}_i \;=\; \mathbf{e}_{\text{token}}(x_i) \;+\; \mathbf{e}_{\text{pos}}(i) \;+\; \mathbf{e}_{\text{seg}}(x_i).$$

These embeddings pass through 12 transformer layers to produce contextualized token-level hidden states. Specifically, let $\mathbf{h}_i \in \mathbb{R}^{768}$ denote the hidden state of the i-th token output by the final transformer layer. The standard FinBERT representation uses the `[CLS]` token hidden state as the 768-dimensional sentence vector:

$$\mathbf{v} \;=\; H^{(12)}_{[\text{CLS}]} \;\in\; \mathbb{R}^{768}.$$

Applied to our dataset of 31,036 consumer complaints, as discussed in Section 2, this yields the feature matrix $\mathbf{V} \in \mathbb{R}^{31{,}036 \times 768}$. Although these embeddings capture deep contextual information, the `[CLS]` token representation aggregates all sequence information

into a single vector position implicitly, without any explicit distinction between high-information financial terms (e.g., liquidity risk and default) and low-information function words (e.g., the and a). Similarly, if one were to aggregate token representations via uniform mean pooling, all tokens would contribute equally regardless of their discriminative importance, causing the resulting sentence-level representation to be diluted by tokens that contribute little discriminative signal. Furthermore, while token-level IDF scores provide a corpus-statistical measure of term importance, they carry no contextual awareness and cannot capture the semantic variability of a token across different contexts. Critically, naively combining these two featurizations, whether by concatenation or weighted summation of feature vectors, does not resolve either limitation: the IDF component remains context-free, and the FinBERT component remains importance-agnostic. The respective deficiencies do not cancel each other out through simple fusion.

To address this, we propose FinBERT-IDF$_{\text{token}}$, which is designed precisely to avoid this failure mode. Rather than treating IDF as an independent feature vector to be concatenated, we use IDF scores as importance weights over the token-level hidden states produced by the FinBERT encoder prior to pooling. Crucially, the IDF weighting is applied exclusively at the output of the FinBERT encoder as a post-encoding token importance weighting strategy and does not modify the input embedding layer. The three input embedding components (token, position, and segment) and their summation feeding into the first transformer block remain entirely unchanged, preserving the pretrained weight assumptions of FinBERT. Formally, the IDF-weighted sentence-level representation is obtained as the normalized weighted mean of the final-layer token hidden states:

$$\mathbf{s} = \frac{\sum_{i=1}^{n} \text{IDF}_i \cdot \mathbf{h}_i}{\sum_{i=1}^{n} \text{IDF}_i}, \tag{4}$$

where IDF$_i$ denotes the precomputed IDF score of the i-th token derived from the full CFPB corpus, n denotes the total number of tokens in the input sequence, and $\mathbf{s} \in \mathbb{R}^{768}$ serves as the final document-level representation for downstream classification. In this formulation, IDF serves as a selection mechanism that determines how much each contextually encoded token representation contributes to the final sentence embedding. Tokens that appear rarely across the corpus, and are therefore more likely to carry domain-specific and complaint-relevant information, receive higher weights in the aggregation, while semantically uninformative but frequently occurring tokens (e.g., function words and domain-generic terms) are down-weighted. This yields a representation that is simultaneously semantically rich and importance-aware, a property that neither featurization achieves in isolation.

The IDF computation is performed offline once on the full CFPB corpus, and the per-document weighted aggregation adds only $O(n)$ operations, where n denotes the number of tokens per document. The incremental memory requirement is less than 1 MB for the CFPB corpus, representing less than 0.23% of FinBERT's 440 MB parameter footprint, confirming that the computational overhead of FinBERT-IDF$_{\text{token}}$ relative to the baseline FinBERT model is negligible. By incorporating corpus-level term discriminativeness into the encoder output representations as a post-encoding importance weighting mechanism, FinBERT-IDF$_{\text{token}}$ simultaneously reduces the influence of semantically uninformative tokens and amplifies the contribution of rare domain-specific terms, making it particularly effective for complex financial text classification.

3.2. Machine Learning Classifiers

Traditional models (e.g., logistic regression (LR) and random forest (RF)) often perform well when data are simple and structured, delivering accurate classification with minimal complexity. However, as the variability and complexity of the data increase, these basic models may struggle to capture subtle nuances and contextual dependencies. In such cases, deep learning models like LSTM and FinBERT become essential. They are designed to learn sophisticated language representations that automatically discern intricate semantic patterns. This capability enhances classification accuracy and offers greater robustness in handling diverse and complex text data.

In this study, we utilize both traditional models and deep learning architectures. Traditional models, recognized for their computational efficiency and interpretability, perform adequately on straightforward well-structured datasets; however, they often struggle to capture the nuanced semantic and syntactic relationships present in more complex textual data. In contrast, deep learning architectures like FinBERT excel at extracting intricate language patterns from unstructured inputs. Additionally, we incorporate an LSTM model, which has demonstrated strong performance in text classification tasks owing to its capacity to model sequential dependencies and retain contextual information across variable-length inputs. Furthermore, the advanced FinBERT model is implemented on a high-performance GPU, with an M4 MAX GPU used to expedite data preprocessing. Ultimately, benchmarking these varied architectures enables us to identify which approach yields superior performance when analyzing complex textual data such as consumer complaint narratives.

3.3. Evaluation Metrics

Using the truncated TF-IDF and FinBERT-IDF$_{\text{token}}$ featurization methods, textual data can be represented as a 768-dimensional numerical matrix. After selecting appropriate machine learning classifiers, the crucial final step is to choose a suitable evaluation metric (standard), which is essential for accurately assessing the performance of models.

In our prior work, we focused on accuracy as our primary evaluation standard. This standard is simple and interpretable, particularly when applied to small and balanced datasets. However, in domains like consumer complaint analysis, increasing the number of complaints and including additional subgroups often leads to significant class imbalance (for example, relatively few narratives are labeled meritorious). To address this, the present study adopts the F1 score as the main evaluation standard as it provides a more balanced assessment of classification performance under class-imbalanced settings.

Building on our understanding of accuracy and the F1 score, we introduce two additional evaluation standards to deepen our insight into the classification task. The response variable is binary, distinguishing between meritorious and non-meritorious cases. The MCC delivers the most robust assessment in binary classification by accounting for true positives, true negatives, false positives, and false negatives [36]. The MCC is computed as

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \qquad \text{MCC} \in [-1, 1] \tag{5}$$

when MCC $= 1$ indicates perfect prediction, 0 indicates random prediction, and -1 indicates complete disagreement between predictions and actual values.

As our input variables consist of a combination of categorical and textual variables, Cohen's kappa (C_κ), which measures the level of agreement between two raters while

accounting for the possibility of chance agreement, demonstrates the best performance in NLP [26,27]. The C_κ is computed as

$$C_\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad C_\kappa \in [-1, 1] \tag{6}$$

where p_o is the observed agreement (accuracy), and p_e is the expected agreement by chance (expected accuracy).

In the subsequent analysis, we employ accuracy, the F_1 score, the Matthews correlation coefficient, and Cohen's kappa as our evaluation metrics. Accordingly, we define the metric set as $\mathcal{E} = \{\text{accuracy, F1, MCC, } C_\kappa\}$. Algorithm 1 outlines the procedure by which we apply the input variables to the machine learning classifiers and compute various evaluation metrics. In the following section, we illustrate classifier performance using Algorithm 1 and employ both TF-IDF and FinBERT featurization (as suggested by Huang et al. [15]) to benchmark our proposed methods.

Algorithm 1 Performance of diverse evaluation metrics in the NLP-based assessment of textual consumer complaints

Require
1. A dataset D with true labels "meritorious" and "non-meritorious".
2. A specific classfication method.

Compute the evaluation metrics $\mathcal{E}$
1. Convert the input categorical variables to dummy variables named X_{dummy}.
2. Convert the input textual variables to a numerical matrix based on the text featurization approach, and append the numerical matrix to X_{dummy}.
3. Randomly assign $(1 - \text{r})100\%$ of the dataset D as a training set and the remaining r100% as a testing set.
4. Train an ML classifier f based on the training set.
5. Apply f on the testing set and compute the ML label.
6. Compute the evaluation metrics.

3.4. Classification Results in Real Narratives

In traditional word embedding techniques like TF-IDF, each word is treated as an independent term without considering context or word order. This method counts the frequency of words and scales them by their inverse document frequency, meaning all words are assumed to contribute equally regardless of their position or relationship with other words. In contrast, deep learning models such as LSTM process text as a sequence, capturing the contextual dependencies and subtle semantic relationships between words. This fundamental difference in how information is represented can lead to inconsistencies when combining the two approaches. For instance, feeding TF-IDF vectors into an LSTM is theoretically incompatible because the sequential information that the LSTM is designed to exploit is absent in a bag-of-words representation. As a result, the performance metrics derived from such a hybrid approach are excluded from the analysis.

We employ a comparative evaluation of different ML classifiers applied to consumer narratives, with various featurization approaches presented in Table 3. All results are reported as the mean $\pm$ standard deviation averaged over 10 independent runs with seeds $\{42, 2021, 2022, \ldots, 2029\}$ to ensure statistical reliability. To facilitate reproducibility and enable direct comparison across all models, the complete hyperparameter settings for all classifiers are summarized in Table 4. All transformer-based models, including the fine-tuned external baselines (BERT-base and RoBERTa) and the proposed FinBERT-based featurization approaches (FinBERT and FinBERT-IDF$_{\text{token}}$), share a common configuration: a learning rate of 2×10^{-5}, a batch size of 16, 5 training epochs, the AdamW optimizer with a weight decay of 0.01, a linear warmup scheduler with 10% warmup steps, and gradient clipping at 1.0. All hyperparameters

are selected based on established best practices in the transformer fine-tuning literature [37] and verified to produce stable convergence on the validation set across all 10 independent runs, with no formal hyperparameter search conducted.

Table 3. Performance comparison (mean ± standard deviation, %, averaged over 10 independent runs with seeds $\{42, 2021, 2022, \ldots, 2029\}$) of proposed featurization approaches paired with ML classifiers and fine-tuned external baselines (`bert-base-uncased` and `roberta-base`) on T5-summarized consumer complaint narratives from the CFPB dataset, evaluated across accuracy, F1 score, MCC, and Cohen's κ.

Evaluation Metrics	Featurization Approaches	ML Classifiers				External Baselines Fine-Tuned	
		LR	RF	LSTM	FinBERT	BERT-Base	RoBERTa
accuracy	TF-IDF	59.78 ± 0.45	58.20 ± 0.52	-	-	60.47 ± 0.51	58.69 ± 0.62
	Truncated TF-IDF	59.54 ± 0.48	56.38 ± 0.61	-	-		
	FinBERT	65.80 ± 0.38	72.31 ± 0.44	78.05 ± 0.44	77.62 ± 0.31		
	FinBERT-IDF$_{\text{token}}$	65.00 ± 0.42	71.81 ± 0.49	78.42 ± 0.51	76.87 ± 0.51		
F1 score	TF-IDF	59.77 ± 0.44	58.21 ± 0.50	-	-	60.21 ± 0.67	58.59 ± 0.64
	Truncated TF-IDF	59.53 ± 0.46	56.36 ± 0.58	-	-		
	FinBERT	65.81 ± 0.40	72.25 ± 0.42	77.98 ± 0.44	77.51 ± 0.33		
	FinBERT-IDF$_{\text{token}}$	63.51 ± 0.51	69.88 ± 0.55	78.27 ± 0.59	76.78 ± 0.52		
MCC	TF-IDF	19.63 ± 0.55	16.43 ± 0.68	-	-	20.97 ± 1.13	20.47 ± 0.60
	Truncated TF-IDF	19.08 ± 0.58	12.69 ± 0.72	-	-		
	FinBERT	31.55 ± 0.62	44.56 ± 0.75	56.09 ± 0.92	55.32 ± 0.74		
	FinBERT-IDF$_{\text{token}}$	29.89 ± 0.68	43.56 ± 0.81	56.95 ± 1.07	54.72 ± 0.42		
C_κ	TF-IDF	19.60 ± 0.54	16.42 ± 0.66	-	-	20.72 ± 1.20	19.64 ± 0.72
	Truncated TF-IDF	19.07 ± 0.56	12.68 ± 0.70	-	-		
	FinBERT	31.55 ± 0.60	44.46 ± 0.73	55.87 ± 0.92	55.01 ± 0.60		
	FinBERT-IDF$_{\text{token}}$	29.89 ± 0.66	43.45 ± 0.78	56.82 ± 1.07	54.61 ± 0.46		

Table 4. Hyperparameter settings for all models.

Hyperparameter	BERT-Base	RoBERTa	FinBERT	FinBERT-IDF$_{\text{token}}$
Model	`bert-base-uncased`	`roberta-base`	`yiyanghkust/finbert-tone`	`yiyanghkust/finbert-tone`
Max sequence length	512	512	512	512
Batch size	16	16	16	16
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Weight decay	0.01	0.01	0.01	0.01
Epochs	5	5	5	5
Optimizer	AdamW	AdamW	AdamW	AdamW
Scheduler	Linear warmup	Linear warmup	Linear warmup	Linear warmup
Warmup steps	10% of total	10% of total	10% of total	10% of total
Gradient clipping	1.0	1.0	1.0	1.0
Dropout (text)	0.1	0.1	0.5	0.5
Dropout (categorical)	-	-	0.3	0.3
Dropout (combined)	-	-	0.3	0.3
IDF weighting	-	-	-	✓
IDF computation	-	-	-	Offline (preprocessing)
Embedding dim (cat)	-	-	16	16
CLS hidden size	768	768	768	768
Categorical dense	-	-	32	32
Combined dense	-	-	64	64
Loss function	Cross-entropy	Cross-entropy	Cross-entropy	Cross-entropy
Train/val/test split	72%/8%/20%	72%/8%/20%	72%/8%/20%	72%/8%/20%
Split strategy	Stratified	Stratified	Stratified	Stratified
Random seeds	10 different random seeds (e.g., $\{42, 2021, 2022, \ldots, 2029\}$).			

Note: ✓ indicates that IDF weighting is applied as a post-hoc token importance weighting strategy over the final-layer hidden states of the FinBERT encoder (Equation (4)), with IDF scores precomputed offline from the full CFPB corpus prior to training.

In Table 3, entries corresponding to theoretically incompatible featurization classifier combinations are marked with a dash "–", indicating that the combination was not evaluated as it does not accurately reflect the underlying capability of the models. Traditional methods such as TF-IDF and truncated TF-IDF, which treat each word as an independent

term, are only compatible with simpler classifiers like LR and RF. The results, averaged over 10 independent runs, indicate that these methods exhibit moderate performance, achieving accuracy figures in the upper 50% range and corresponding F1 scores while producing relatively low values for the MCC and C_κ metrics. Furthermore, the truncated TF-IDF featurization approach yields comparable performance to the traditional TF-IDF method, which is derived from TF-IDF by implementing singular value decomposition.

The two fine-tuned external baselines, BERT-base ($60.47 \pm 0.51\%$) and RoBERTa ($58.69 \pm 0.62\%$), achieve accuracy levels comparable to those of the TF-IDF-based approaches. This pattern is consistent with the interpretation that generic pretrained models, without domain-specific adaptation to financial complaint text, may not provide discriminative features beyond those captured by surface-level lexical statistics for this task. All FinBERT-based featurization approaches substantially and consistently outperform both external baselines across all evaluation metrics and all 10 independent runs ($p < 0.001$, McNemar test), a result that is robust across all random seeds and is consistent with the practical value of domain-specific pretraining for consumer complaint classification.

All four evaluation metrics exhibit a consistent pattern: as model accuracy increases, comparable trends are observed across the other three evaluation metrics (F1, MCC, and C_κ). When employing both FinBERT and FinBERT-IDF$_{\text{token}}$ for featurization, significant variability is observed across classifiers. For simpler models such as LR and RF, the FinBERT featurization approach achieves higher performance, with mean accuracies of $65.80 \pm 0.38\%$ and $72.31 \pm 0.44\%$, respectively, compared to $65.00 \pm 0.42\%$ and $71.81 \pm 0.49\%$ for FinBERT-IDF$_{\text{token}}$. For deep learning models such as LSTM and FinBERT classifiers, FinBERT-IDF$_{\text{token}}$ attains comparable performance, yielding mean accuracies of $78.42 \pm 0.51\%$ and $76.87 \pm 0.51\%$, respectively. The pairwise McNemar test [38] results across the 10 independent runs confirm that FinBERT and FinBERT-IDF$_{\text{token}}$ attain statistically comparable performance with one another, with neither variant demonstrating a statistically consistent advantage across all seeds ($p > 0.05$ in the majority of runs). These findings suggest that FinBERT-IDF$_{\text{token}}$ performs on par with the baseline FinBERT model when deep learning classifiers are employed while exhibiting lower performance when paired with simpler classifiers such as LR and RF, consistent with the IDF weighting mechanism being better suited to deep learning architectures that can more effectively exploit importance-aware token representations.

It is worth noting that FinBERT-IDF$_{\text{token}}$ operates as a post hoc token importance weighting strategy applied to the token-level hidden states of the FinBERT encoder, leaving the underlying model architecture and training procedure unchanged. The IDF weights are computed offline during corpus preprocessing, and the per-document IDF-weighted aggregation incurs only $\mathcal{O}(n)$ additional operations, where n denotes the number of tokens per document. The incremental memory requirement for the CFPB corpus is less than 1 MB, representing under 0.23% of FinBERT's 440 MB parameter footprint. Given this negligible computational overhead, FinBERT-IDF$_{\text{token}}$ offers a theoretically motivated and computationally efficient alternative to the baseline FinBERT model, achieving statistically comparable performance at negligible additional cost, and may be particularly advantageous in settings where domain-specific term weighting is of primary importance.

The results collectively demonstrate that contextualized embeddings tailored to the financial domain, namely FinBERT and FinBERT-IDF$_{\text{token}}$, are substantially more effective in capturing the nuances of consumer complaint narratives than traditional TF-IDF representations and generic pretrained baselines. Accordingly, the FinBERT featurization approach is adopted as the primary featurization method as it exhibits stronger and more consistent performance across the full range of classifiers evaluated, including both simpler and deep learning models.

4. Text Generation in Human-Experience-Trained Algorithms

In Section 3.2, we have demonstrated that integrating deep learning models into textual analysis yields significantly enhanced classification results. Although our initial experiments are conducted on an existing dataset, the use of real-world data in practice is often constrained by high acquisition costs, data privacy concerns, and inconsistencies in data quality. To address these limitations, we have opted to generate a text dataset. This approach circumvents the practical challenges associated with real-world data and affords greater control over experimental conditions, thereby facilitating systematic evaluations and reproducibility of results.

4.1. Text Generation

Text generation is a subfield of NLP that involves creating coherent and contextually relevant sequences of text based on learned patterns from training data. The process typically begins with converting raw textual data into a numerical representation, which enables machine learning models to process and understand the data. Not all numerical representation methods are suitable for generation. For instance, techniques like TF-IDF are primarily designed for feature extraction. They compute term importance based on frequency and inverse document frequency without accounting for word order or contextual dependencies. TF-IDF operates under a bag-of-words assumption that disregards the sequential structure of language, rendering it incapable of capturing the syntactic and temporal relationships necessary for generating coherent text.

Two principal approaches to converting raw textual data into a numerical representation for text generation have traditionally been pursued: statistical n-gram model and methods based on text embeddings. An N-gram model operates by computing the conditional probability of a word given a fixed window of preceding words, thereby offering a straightforward and interpretable means to generate text based solely on local context frequencies. Bigram models have been shown to play an important role in feature selection from short text [39].

A bigram (e.g., n = 2) model is a statistical language model that predicts the next word in a sequence based on its immediate predecessor. It achieves this by analyzing the frequency of word pairs in a given corpus and computing the corresponding conditional probabilities. For instance, we select one of the summarized sentences from the CFPB:

unexpected fees and inaccurate credit statements

when processed using a bigram approach, it can be partitioned into the following pairs: (“unexpected”,“fees"), (“fees”, “and"), and so on. Table 5 presents the frequencies and probabilities derived from the example sentence. Each row in the table represents a unique bigram (a pair of consecutive words) along with its frequency of occurrence and the corresponding conditional probability $P(w_i \mid w_{i-1})$. For instance, the bigram (“unexpected”, “fees") occurs once in the corpus, and the probability of “fees” following “unexpected” is calculated as 1.0. This table serves as the foundation for a bigram language model, enabling the prediction of the next word in a sequence based on the previous word. By analyzing such probabilities, the model can generate coherent text or estimate the likelihood of word sequences across a range of NLP tasks.

The second approach is text embeddings, which transform discrete textual elements into continuous vector representations that capture semantic and syntactic nuances. Three primary sub-areas are widely recognized under this umbrella: character embeddings, word embeddings, and subword tokenization. Character embeddings represent individual characters as vectors, providing fine-grained control and robustness to morphological variations and out-of-vocabulary terms. Word embeddings, on the other hand, assign a unique dense vector to each word, effectively encapsulating overall semantic and syntactic properties.

Subword tokenization is a technique that decomposes words into smaller semantically meaningful units. This method enables language models to handle rare or unseen words efficiently and reduces input sequence lengths, enhancing computational performance and preserving essential linguistic nuances. This approach underpins many state-of-the-art pretrained models, such as FinBERT and T5, which employ subword tokenization as a core architecture component. These models can be fine-tuned for specific tasks or domains, enabling rapid adaptation and improved performance in applications ranging from financial sentiment analysis to text generation.

Table 5. Bigram frequencies and probabilities for one sentence from the CFPB.

Bigram	Frequency	$P(w_i \mid w_{i-1})$
(“unexpected”, “fees”)	1	1.0
(“fees”, “and”)	1	1.0
(“and”, “inaccurate”)	1	1.0
(“inaccurate”, “credit”)	1	1.0
(“credit”, “statements”)	1	1.0

After converting raw textual data into a numerical representation, the subsequent step involves selecting an LSTM architecture for text generation. LSTM networks constitute a recurrent neural network specifically designed to process sequential data. They employ memory cells that preserve crucial information over time and incorporate gating mechanisms to regulate which details are retained or discarded.

By providing numerical representations as inputs to both an LSTM-based text generation model and a statistical bigram model, we generate brief complaints (i.e., those containing 10 words) in which every sentence commences with the word “credit”. Table 6 demonstrates that the subword tokenization FinBERT model produces more consistent outputs than the statistical bigram model. Subword segmentation of FinBERT effectively handles out-of-vocabulary issues and captures complex compound expressions in financial language. In contrast, the bigram model’s reliance on fixed word pair frequencies can lead to occasional inconsistencies and a weaker contextual understanding.

Table 6. Narrative generation via statistical bigram and subword tokenization models with meritorious indicators.

Model	Narrative Generation	Meritorious
Statistical bigram	credit report fraudulent accounts obtain credit card account remains unresolved	Yes
Statistical bigram	credit file dont belong would removed credit card account ending	No
Subword tokenization FinBERT	credit legitimacy complies infringed primary amounts thought carolina represents inquire	Yes
Subword tokenization FinBERT	credit bureaus tax audit acquiring aims correspondence neglecting location changed	No

4.2. Generative Adversarial-Based Human-Experience-Trained Algorithms

Exposure bias, the discrepancy between training on ground-truth tokens and inference on model-generated outputs, is a long-recognized challenge in sequence generation [40,41]. Established approaches to mitigating this issue include scheduled sampling [41], which progressively replaces ground-truth tokens with model outputs during training; the MIXER framework [40], which incorporates sequence-level reinforcement learning signals to optimize task-relevant objectives; and the adversarial synchronization approach of Lamb et al. [42],

which aligns teacher-forced and free-running network states through adversarial training. In contrast, the GAN-based framework adopted in the present study addresses exposure bias implicitly through the adversarial training signal provided by the discriminator, which evaluates entire generated sequences rather than individual token-level predictions, thereby offering a complementary perspective on the train–inference discrepancy without requiring explicit curriculum scheduling or reinforcement learning objectives.

The GAN is the gaming (with respect to the expectation of the log utility) between the generator and discriminator, namely

$$\min_{\boldsymbol{\Theta},\boldsymbol{\psi}} \max_{\boldsymbol{\phi}\in\mathbb{R}^{769}} V(\mathrm{LSTM}_{\boldsymbol{\Theta},\boldsymbol{\psi}}, D_{\boldsymbol{\phi}})$$

where $D_{\boldsymbol{\phi}} : \mathbb{R}^{768} \mapsto (0,1)$ is defined as the sigmoid function on the inner product of a 769-dimensional vector $\boldsymbol{\phi}$ and another augmented 768+1-dimensional vector $\begin{bmatrix}\boldsymbol{h}^\intercal & 1\end{bmatrix}^\intercal$, where the appended scalar 1 serves as a bias term:

$$D_{\boldsymbol{\phi}}(\boldsymbol{h}) := \sigma\Big(\begin{bmatrix}\boldsymbol{h}^\intercal & 1\end{bmatrix}\boldsymbol{\phi}\Big) \triangleq \frac{1}{1+\exp\Big(-\begin{bmatrix}\boldsymbol{h}^\intercal & 1\end{bmatrix}\boldsymbol{\phi}\Big)}$$

and

$$V(\mathrm{LSTM}_{\boldsymbol{\Theta},\boldsymbol{\psi}}, D_{\boldsymbol{\phi}}) = \mathbb{E}_{\boldsymbol{x}}\Big[\log D_{\boldsymbol{\phi}}\Big(\boldsymbol{h}^{(x)}\Big)\Big] + \mathbb{E}_{\boldsymbol{z}}\Big[\log\Big(1 - D_{\boldsymbol{\phi}}\Big(\boldsymbol{h}^{(z)}\Big)\Big)\Big],$$

where $\boldsymbol{h}^{(x)}$ denotes the hidden state produced by the LSTM when processing the observed real narrative $\boldsymbol{x}$, and $\boldsymbol{h}^{(z)}$ denotes the hidden state produced when processing the generated narrative $\boldsymbol{z}$, as defined below. The hidden layer output $\boldsymbol{h}^{(x)}$ of the observed narrative $\boldsymbol{x} = \{x_t\}_{t=1}^{T_x}$ is determined recursively through

$$\boldsymbol{h}_t^{(x)} = \mathrm{LSTM}_{\boldsymbol{\Theta},\boldsymbol{\psi}}(x_t, \boldsymbol{h}_{t-1}^{(x)})$$

for $t = 1, 2, \ldots, T_x$ and $\boldsymbol{h}_0^{(x)} = \mathbf{0}$, while the hidden layer output $\boldsymbol{h}^{(z)}$ of the generated narrative $\boldsymbol{z} = \{z_t\}_{t=1}^{\tau}$ is determined through

$$\boldsymbol{h}_t^{(z)} = \mathrm{LSTM}_{\boldsymbol{\Theta},\boldsymbol{\psi}}(z_t, \boldsymbol{h}_{t-1}^{(z)})$$

with

$$p(z_t|\boldsymbol{h}_{t-1}^{(z)}) = \mathrm{Softmax}\Big(\Big[\Big(\boldsymbol{h}_{t-1}^{(z)}\Big)^\intercal \quad 1\Big]\boldsymbol{\psi}\Big)$$

for $t = 1, \ldots, \tau$, where τ is some stopping time.

The proposed framework consists of two independent class-conditional GAN pairs, designated for the meritorious and non-meritorious classes, respectively. Within each pair, the staff model (ST_M or ST_N) functions as the generator G, while the expert model (EX_M or EX_N) serves as the discriminator D. These two components are trained adversarially using the binary cross-entropy loss. The term "expert" refers to the use of ground-truth human-provided labels from the original CFPB dataset as the source of real training samples for the discriminator, whereas "staff" denotes the generator whose outputs approximate the class-conditional distribution of the real data. The adversarial training objectives governing the meritorious class pair are defined as

$$\mathcal{L}_{\mathrm{EX}_M} = -\mathbb{E}_{x\sim p^M_{\mathrm{real}}}[\log D_M(x)] - \mathbb{E}_{z\sim p_z}[\log(1 - D_M(G_M(z)))] \tag{7}$$

$$\mathcal{L}_{\mathrm{ST}_M} = -\mathbb{E}_{z\sim p_z}[\log D_M(G_M(z))] \tag{8}$$

and for the non-meritorious class pair:

$$\mathcal{L}_{\mathrm{EX}_N} = -\mathbb{E}_{x\sim p^N_{\mathrm{real}}}[\log D_N(x)] - \mathbb{E}_{z\sim p_z}[\log(1 - D_N(G_N(z)))] \quad (9)$$

$$\mathcal{L}_{\mathrm{ST}_N} = -\mathbb{E}_{z\sim p_z}[\log D_N(G_N(z))] \quad (10)$$

The training procedure alternates between updating each discriminator to minimize $\mathcal{L}_{\mathrm{EX}_M}$ and $\mathcal{L}_{\mathrm{EX}_N}$ while holding the generators fixed, and updating each generator to minimize $\mathcal{L}_{\mathrm{ST}_M}$ and $\mathcal{L}_{\mathrm{ST}_N}$ while holding the discriminators fixed. Each GAN pair synthesizes 5000 samples per class, resulting in a total of 10,000 synthetic complaint narratives for downstream classification experiments. Human experience is operationalized through statistical probability distributions derived from the original labeled dataset, with the prior distribution of the noise vector, $z \sim p_z$, informed by the empirical token frequency distribution and TF-IDF weighting scheme estimated from the real corpus.

Subsequently, an LSTM structure is integrated into the generative adversarial framework as the generator G, which is further augmented through a tailored structural modification to embed a human-experience-trained algorithm, as illustrated in Figure 3.

The proposed methodology is divided into two distinct segments. The upper segment is dedicated to generating text from narratives classified as non-meritorious, while the lower segment focuses on narratives classified as meritorious. Evaluation metrics for each subgroup are computed as described in the preceding section. Relative to the original GAN architecture, we introduce additional noise into the staff-trained LSTM module. In real-world scenarios, data are rarely pristine and often contain inherent noise. Adding noise to the generated fake samples enables the LSTM module to better accommodate such stochastic variations, thereby enhancing the performance of the model on real-world data. Furthermore, many real-world systems invariably include uncontrollable stochastic factors. By injecting noise into the fake data, we can partially emulate these influences, aligning the GAN framework more closely with real-world system dynamics. The primary objective of Figure 3 is to develop a generative model that produces accurately labeled textual data for both categories while maintaining performance metrics comparable to those of the original dataset.

As illustrated in Figure 3, the proposed framework employs two independent class-specific GAN pairs: one trained exclusively on meritorious complaints ($\mathrm{ST}_M/\mathrm{EX}_M$) and the other on non-meritorious complaints ($\mathrm{ST}_N/\mathrm{EX}_N$). This architecture falls under the paradigm of class-conditional generation, wherein each generator is tasked with faithfully reproducing the linguistic distribution of a single class, with the class proportions specified by the experimental design rather than learned from the data. Consequently, the question of unconstrained output proportions does not apply to this framework as class membership is determined prior to generation rather than inferred by the model. Specifically, we generate a dataset of 5000 textual entries per class, each comprising approximately 50 tokens, with labels assigned as either meritorious or non-meritorious, yielding a total of 10,000 synthetic complaint narratives.

The choice to generate 5000 samples per class is consistent with the natural class distribution of the real data, in which meritorious and non-meritorious complaints account for approximately 48.5% and 51.5% of the corpus, respectively. Since the real data is already nearly balanced, the equal-sized generation introduces a negligible deviation from the true class proportions, and no oversampling technique was deemed necessary. The equal-sized generation additionally serves as an experimental control, ensuring that any observed differences in downstream classification performance can be attributed to the quality of the generated text rather than to class imbalance.

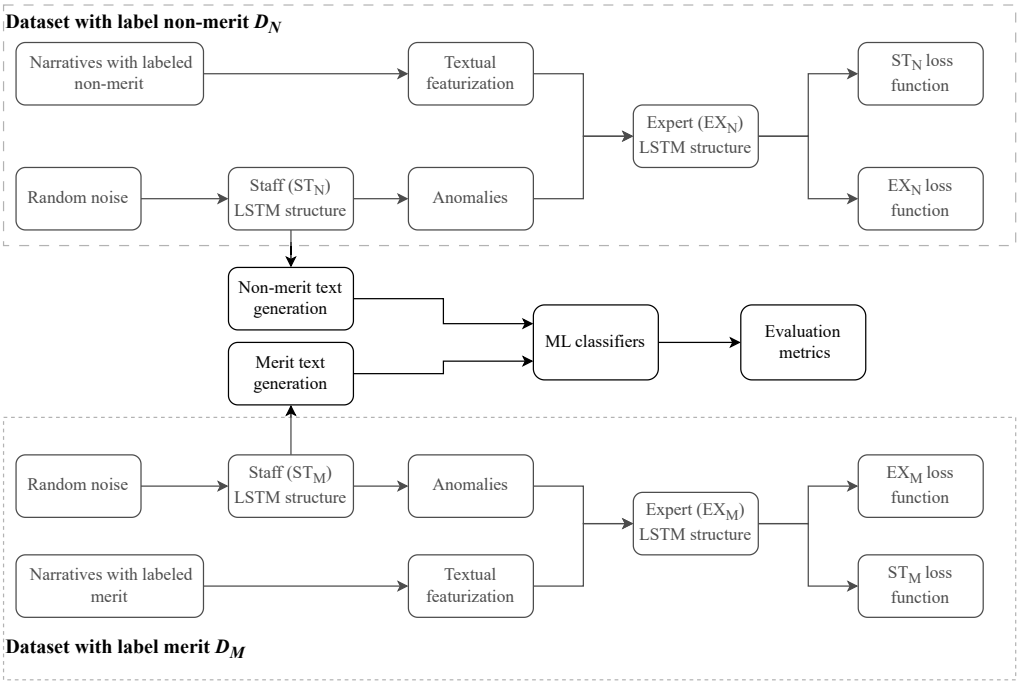


Figure 3. Generative adversarial network with human-experience-trained framework.

We then apply the evaluation metrics defined in Section 3.3 to assess the capacity of the proposed model to predict the meritorious labels for the generated textual dataset. Subsequently, we benchmark the performance by comparing the generated dataset against the actual dataset to demonstrate that our generative model successfully replicates the patterns observed in CFPB.

Building upon the framework presented in Table 3, which demonstrates that the FinBERT featurization approach achieves superior performance when integrated with deep learning models, Table 7 provides a corresponding analysis using the simulated data illustrated in Figure 3. All results are reported as mean ± standard deviation, averaged over 10 independent runs, consistent with the evaluation protocol applied to the real data in Table 3.

Table 7. Performance summary (mean ± standard deviation, %, averaged over 10 independent runs) of ML classifiers on simulated data with Figure 3 framework via FinBERT featurization approach, evaluated across accuracy, F1 score, MCC, and Cohen's κ.

Evaluation Metrics	LR	RF	LSTM	FinBERT
accuracy	52.13 ± 0.48	54.32 ± 0.55	72.31 ± 0.62	77.35 ± 0.38
F1 score	51.47 ± 0.51	53.88 ± 0.53	72.10 ± 0.59	76.88 ± 0.42
MCC	18.13 ± 0.68	17.95 ± 0.74	49.83 ± 0.98	50.09 ± 0.65
C_κ	18.33 ± 0.65	17.59 ± 0.71	49.74 ± 0.95	49.97 ± 0.61

A comparison between Table 3 and Table 7 reveals a heterogeneous pattern in classification performance across classifier types. The FinBERT classifier exhibits minimal performance degradation on the synthetic data (real: $77.62 \pm 0.31\%$, synthetic: $77.35 \pm 0.38\%$, difference: $\approx$0.27 percentage points), suggesting that domain-specific deep learning models can effectively exploit the class-conditional token distributions preserved in the GAN-generated narratives. The LSTM classifier shows a moderate performance gap (real: $78.05 \pm 0.44\%$, synthetic: $72.31 \pm 0.62\%$, difference: $\approx$5.74 percentage points), while simpler classifiers such as LR and RF exhibit substantially larger gaps (LR: $65.80 \pm 0.38\%$ vs. $52.13 \pm 0.48\%$; RF: $72.31 \pm 0.44\%$ vs. $54.32 \pm 0.55\%$). The F1 score, MCC, and C_κ metrics exhibit analogous patterns across all

classifier types. The performance degradation observed for simpler classifiers is consistent with the known sensitivity of surface-level statistical models to the reduced lexical diversity and limited surface fluency of GAN-generated text [24]. Notably, the rank ordering of classifiers is fully preserved across both datasets, with FinBERT consistently achieving the highest performance on both real and synthetic data.

These findings suggest that the proposed GAN-based pipeline may serve as a useful data augmentation tool under regulatory access constraints, particularly when paired with domain-specific deep learning classifiers such as FinBERT. However, the substantially larger performance gaps observed for simpler classifiers indicate that the utility of the synthetic data is classifier-dependent, and that the current generative framework does not fully replicate the statistical properties of the real corpus for all model types. The computation of direct distributional metrics such as n-gram divergence and MAUVE [43] scores on the GAN-generated narratives is identified as a direction for future work to provide stronger evidence of synthetic data fidelity.

5. Conclusions

This research demonstrates that expert labeling consistently outperforms staff labeling, as evidenced by marked improvements across multiple evaluation metrics, including accuracy, F1 score, MCC, and C_K. The findings suggest that implementing expert labels constitutes a more effective strategy for achieving superior accuracy and reliability, underscoring the importance of leveraging domain expertise to enhance the overall quality of the labeling process.

Based on expert labels, the textual dataset is subsequently summarized for a downstream text classification task. Crucially, classification experiments conducted on both the original and summarized texts using ML models yield comparable performance, confirming that the summarization process is advantageous due to its reduced feature space and associated computational overhead. Furthermore, three distinct word embedding methods are benchmarked, with the final results demonstrating the superiority of the FinBERT model across all the evaluation metrics. In parallel, a GAN-based generative framework is developed to synthesize datasets that faithfully exhibit patterns that are analogous to those of real-world complaint narratives. This approach mitigates the practical barriers to high-quality data acquisition, including high costs, privacy concerns, and inconsistent data quality, while providing a controlled and reproducible experimental testbed. Collectively, these findings advocate for a hybrid methodology that leverages expert labels, deep learning architectures, and synthetic data generation to advance the field of text classification.

5.1. Limitations

The present study has several limitations that should be acknowledged. The dataset is drawn exclusively from the CFPB complaint database, meaning that only consumers who actively engage with the formal complaint process are represented. This self-selection mechanism may introduce systematic bias in the observed language patterns, and the findings may not extend straightforwardly to the broader population of consumer financial grievances. Furthermore, the expert labeling procedure, while demonstrably superior in classification performance, is resource-intensive by nature and may prove difficult to scale to larger datasets or a wider range of product categories without substantial additional investment in domain expertise.

The generalizability of the proposed pipeline beyond the CFPB corpus also remains an open question. FinBERT and the IDF-weighted aggregation scheme were evaluated solely on English-language mortgage and financial product complaints; their applicability to other

regulatory corpora, such as insurance claims or SEC filings, or to non-English financial texts has not been assessed. Some degree of domain-specific retraining or recalibration would likely be necessary before the methodology could be transferred to such settings.

On the generative side, the GAN-based framework produces synthetic narratives that broadly reflect the distributional characteristics of the original complaint data, yet limitations in grammatical fidelity and semantic coherence persist. These shortcomings are partly attributable to the discrete nature of textual data, which complicates gradient-based optimization during GAN training. Addressing these issues motivates the exploration of alternative generative approaches, such as variational autoencoders and diffusion models, as outlined in the next section.

5.2. Future Works

Four directions of future work are identified. First, the current results demonstrate strong performance in text classification and generation tasks using the FinBERT pretrained model. FinBERT is based on the transformer architecture introduced by Google's BERT and has been further pretrained on a large corpus of financial texts. This specialized training enables FinBERT to better capture and analyze financial language. In the next phase, we aim to develop a BERT-based pretrained model by incorporating additional complaint and insurance data, with the expectation of achieving further performance gains in deep-model analysis.

Second, regarding text generation, the GAN-based framework has proven effective for synthesizing textual data; however, it may not always produce the most faithful representations of the original complaint data. Therefore, exploring and implementing additional text simulation methods is essential to generate synthetic data that more accurately reflects the nuances and complexities of the original dataset. Methods such as variational autoencoders and diffusion models offer promising alternatives for future work. Variational autoencoders learn a continuous latent representation of complaint narratives and can generate diverse text by sampling from this latent space. In contrast, diffusion models iteratively refine random noise into coherent text, achieving high diversity and fidelity to the original data distribution. Additionally, retrieval-augmented generation combines retrieval techniques with generative models to condition the output on real complaint examples. This ensures that the synthesized text preserves essential semantic and stylistic details that are crucial for robust classification.

Third, the grammatical correctness of generated text is likely influenced by both the data quality and the gradient propagation properties of the model when processing discrete text data. Hence, it necessitates the integration of contemporary natural language generation (NLG) techniques that focus on producing coherent text from structured data through linguistic rules and contextual information. These advanced NLG models, typically trained on large-scale datasets, have demonstrated the capacity to substantially enhance text quality, yielding narratives that are both grammatically well formed and semantically coherent.

Fourth, the evaluation of synthetic data fidelity in the present study relies primarily on downstream classification performance and indirect lexical similarity measures (e.g., Jaccard similarity, Spearman rank correlation, and TF-IDF cosine similarity). Future work will incorporate direct distributional divergence metrics, including n-gram divergence and MAUVE [43], applied directly to the GAN-generated narratives, to provide more rigorous and interpretable evidence of the degree to which the synthetic data replicates the statistical properties of the original CFPB corpus. Additionally, a systematic human evaluation of generated text quality, including assessments of grammatical correctness, semantic coherence, and domain-specific terminology, will be conducted to complement the quantitative distributional measures.

https://doi.org/10.3390/app16125992

Author Contributions: P.G.: conceptualization, data curation, formal analysis, investigation, software, visualization, writing—original draft. C.Y.: conceptualization, methodology, investigation, formal analysis. N.S.: project administration, investigation, writing—review and editing, validation. R.Z.: conceptualization, resources, supervision, validation, methodology, writing—original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by the NSERC, Canada Discovery Grant RGPIN-2022-04426.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: CFPB data is publicly available. The CFPB website is available online: https://www.consumerfinance.gov/data-research/consumer-complaints/ (accessed on 14 June 2025).

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014; Volume 27, pp. 2672–2680.
2. Gao, P.; Yang, C.; Sun, N.; Zitikis, R. Predicting classification errors using NLP-based machine learning algorithms and expert opinions. *Mach. Learn. Appl.* **2025**, *19*, 100630. [CrossRef]
3. Gribkova, N.; Zitikis, R. Functional correlations in the pursuit of performance assessment of classifiers. *Int. J. Pattern Recognit. Artif. Intell.* **2020**, *34*, 2051013. [CrossRef]
4. Hayes, R.M.; Jiang, F.; Pan, Y. Voice of the customers: Local trust culture and consumer complaints to the CFPB. *J. Account. Res.* **2021**, *59*, 1077–1121. [CrossRef]
5. Dou, Y.; Roh, Y. Public disclosure and consumer financial protection. *J. Financ. Quant. Anal.* **2024**, *59*, 2164–2198. [CrossRef]
6. Begley, T.A.; Purnanandam, A. Color and credit: Race, regulation, and the quality of financial services. *J. Financ. Econ.* **2021**, *141*, 48–65. [CrossRef]
7. Iwaszenko, S.; Czaja, S.; Kozłowski, A. NLP system for automation of document workflow in a research and development organization–a case study. *Appl. Sci.* **2026**, *16*, 4562. [CrossRef]
8. Alotaibi, S. Hybrid deep neural network with natural language processing techniques to analyze customer satisfaction with delivery platform manager responses. *Appl. Sci.* **2026**, *16*, 4359. [CrossRef]
9. Schwarz, F.; Schwarz, K.D.; Arias Aranda, D.; Bollens, K.; Shivananjappa, N.; Creutzburg, R.; Dimitrova, V. Zero-shot social media crisis classification: A training-free multimodal approach. *Appl. Sci.* **2026**, *16*, 2192. [CrossRef]
10. Bazdar, M.; Kim, H.; Dimitrijevic, B.; Lee, J. The state of practice in application of natural language processing in transportation safety analysis. *Appl. Sci.* **2026**, *16*, 4223. [CrossRef]
11. Eyupoglu, A.; Jazayeri, K.; Çelebi, E. Knowledge-driven human-in-the-loop decision support for student services using active learning and large language models. *Appl. Sci.* **2026**, *16*, 1802. [CrossRef]
12. Ranganathan, J.; Abuka, G. Text summarization using transformer model. In Proceedings of the 9th International Conference on Social Networks Analysis, Management and Security (SNAMS), Milan, Italy, 29 November–1 December 2022; pp. 1–5.
13. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
14. Kamyab, M.; Liu, G.; Adjeisah, M. Attention-based CNN and Bi-LSTM model based on TF-IDF and GloVe word embedding for sentiment analysis. *Appl. Sci.* **2021**, *11*, 11255. [CrossRef]
15. Huang, A.H.; Wang, H.; Yang, Y. FinBERT: A large language model for extracting information from financial text. *Contemp. Account. Res.* **2023**, *40*, 806–841. [CrossRef]
16. Shahriar, S. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. *Displays* **2022**, *73*, 102237. [CrossRef]
17. Iqbal, T.; Qureshi, S. The survey: Text generation models in deep learning. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 2515–2528. [CrossRef]
18. Imran, A.S.; Yang, R.; Kastrati, Z.; Daudpota, S.M.; Shaikh, S. The impact of synthetic text generation for sentiment analysis using GAN based models. *Egypt. Inform. J.* **2022**, *23*, 547–557. [CrossRef]

19. Li, X.; Mao, K.; Lin, F.; Feng, Z. Feature-aware conditional GAN for category text generation. *Neurocomputing* **2023**, *547*, 126352. [CrossRef]
20. Xia, H.; Sun, S.; Wang, X.; An, B. Market-GAN: Adding control to financial market data generation with semantic context. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 15996–16004.
21. Wang, Z.; Ren, Y.; Zhang, X.; Wang, Y. Generating long financial report using conditional variational autoencoders with knowledge distillation. *IEEE Trans. Artif. Intell.* **2024**, *5*, 1669–1680. [CrossRef]
22. Yang, Z.; Huang, Y.; Chen, Y.; Wu, X.; Feng, J.; Deng, C. CTGGAN: Controllable text generation with generative adversarial network. *Appl. Sci.* **2024**, *14*, 3106. [CrossRef]
23. Dalvi, A.; Pingulkar, S.; Tiwary, A.; Divekar, D.; Siddavatam, I.N.; More, N. Integration of coalescent theory and generative adversarial network (GAN) for synthesizing high-fidelity textual financial data. *Procedia Comput. Sci.* **2025**, *252*, 593–602. [CrossRef]
24. Koç, C.; Özyurt, F.; Iantovics, L.B. Survey on latest advances in natural language processing applications of generative adversarial networks. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2025**, *15*, e70004. [CrossRef]
25. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
26. Kolesnyk, A.S.; Khairova, N.F. Justification for the use of Cohen's kappa statistic in experimental studies of NLP and text mining. *Cybern. Syst. Anal.* **2022**, *58*, 280–288. [CrossRef]
27. Khraisha, Q.; Put, S.; Kappenberg, J.; Warraitch, A.; Hadfield, K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res. Synth. Methods* **2024**, *15*, 616–626. [CrossRef] [PubMed]
28. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
29. Gao, P.; Sun, N.; Wang, X.; Yang, C.; Zitikis, R. NLP-based detection of systematic anomalies among the narratives of consumer complaints. *J. Oper. Risk* **2024**, *19*, 79–104. [CrossRef]
30. Siddharthan, A. A survey of research on text simplification. *ITL Int. J. Appl. Linguist.* **2014**, *165*, 259–298. [CrossRef]
31. Di Vincenzo, D.; Greselin, F.; Piacenza, F.; Zitikis, R. A text analysis of operational risk loss descriptions. *J. Oper. Risk* **2023**, *18*, 63–90. [CrossRef]
32. Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **1912**, *11*, 37–50. [CrossRef]
33. Spearman, C. The proof and measurement of association between two things. In *Studies in Individual Differences: The Search for Intelligence*; Jenkins, J.J., Paterson, D.G., Eds.; Appleton-Century-Crofts: New York, NY, USA, 1961; pp. 45–58.
34. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
35. Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]
36. Chicco, D.; Warrens, M.J.; Jurman, G. The Matthews correlation coefficient (MCC) is more informative than Cohen's kappa and Brier score in binary classification assessment. *IEEE Access* **2021**, *9*, 78368–78381. [CrossRef]
37. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
38. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef]
39. Garg, M. UBIS: Unigram bigram importance score for feature selection from short text. *Expert Syst. Appl.* **2022**, *195*, 116563. [CrossRef]
40. Ranzato, M.; Chopra, S.; Auli, M.; Zaremba, W. Sequence level training with recurrent neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, PR, USA, 2–4 May 2016.
41. Bengio, S.; Vinyals, O.; Jaitly, N.; Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1171–1179.
42. Lamb, A.; Goyal, A.; Zhang, Y.; Zhang, S.; Courville, A.; Bengio, Y. Professor forcing: A new algorithm for training recurrent networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 4601–4609.
43. Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Welleck, S.; Choi, Y.; Harchaoui, Z. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 4816–4828.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.