Hybrid Generative–Discriminative Regression Models for High-Dimensional Fake News Detection

Sana Iftikhar*, Fatma Najar†, Nuha Zamzami‡, Nizar Bouguila§
*Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada
†Department of Mathematics and Computer Science, John Jay College (CUNY), New York, NY, USA
‡Department of Computer Science and Artificial Intelligence, University of Jeddah, Jeddah, Saudi Arabia
§Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, Canada
Emails: sana.iftikhar@concordia.ca, fnajar@jjay.cuny.edu, nezamzami@uj.edu.sa, nizar.bouguila@concordia.ca

***Abstract*— High-dimensional sparse count data arise in many real-world applications, including text-based fake news datasets, where representations such as bag-of-words remain well-known due to their simplicity and interpretability. However, effectively modeling such data remains challenging due to sparsity, imbalance, and complex distributional properties that are not explicitly captured by generic numerical models. Building on Multinomial Beta–Liouville (MBL) and Multinomial Scaled Dirichlet (MSD) regression models, we introduce novel exponential approximations, namely Exponential Multinomial Beta–Liouville (EMBL) and Exponential Multinomial Scaled Dirichlet (EMSD) regression models, that improve stability and scalability while preserving key distributional properties. We further propose a hybrid generative–discriminative framework in which likelihood- and score-based quantities derived from the learned model parameters are used as feature representations for linear discriminative prediction. The proposed approach is evaluated on fake news detection tasks, demonstrating the effectiveness of these representations for sparse, high-dimensional text data.**

I. Introduction

Fake news detection is a critical challenge for online information systems and public discourse [1], [2]. Despite the rapid development of deep learning approaches, many real-world pipelines still rely on high-dimensional sparse count representations such as bag-of-words and n-grams, due to their simplicity, transparency, and low data requirements. However, such representations are typically heterogeneous, overdispersed, and exhibit complex dependence structures that are poorly matched by models treating count vectors as generic numerical inputs.

Distribution-based regression offers a principled alternative by explicitly linking the parameters of a multivariate count distribution to observed covariates [3]. Flexible distributional families can capture key properties of sparse count data, including overdispersion and inter-feature dependence, but regression likelihoods based on these models are often costly to optimize in high-dimensional settings [4]. Exponential approximations have been proposed to address these challenges by improving numerical stability and enabling sparse likelihood evaluation [5], [6]. In this work, we introduce regression models based on these exponential formulations, namely Exponential Multinomial Beta–Liouville (EMBL) and Exponential Multinomial Scaled Dirichlet (EMSD), designed for scalable learning on high-dimensional sparse count data.

Beyond direct generative prediction, fitted generative regression models define informative quantities—such as likelihood- and score-based terms—that summarize how observed data align with learned distributional structure. We exploit this property by proposing a hybrid generative–discriminative framework in which distribution-based regression models serve as representation learners. Model parameters are evaluated at fixed anchor covariates to construct structured, fixed-dimensional feature representations, which are then used with a simple linear classifier for prediction. Anchors correspond to fixed covariate settings (e.g., one-hot class vectors) at which regression links are evaluated.

The main contributions of this paper are as follows:

- We introduce EMBL and EMSD regression models that enable scalable distribution-based learning for high-dimensional sparse count data.
- We propose an anchor-evaluated feature construction that transforms learned generative regression parameters into discriminative representations for linear classification.
- We empirically evaluate the proposed framework on binary and multi-class fake news detection datasets, demonstrating the effectiveness of distribution-aware representations.

II. Related Work

Early fake news detection methods for textual data largely relied on sparse count-based representations such as bag-of-words, n-grams, and TF–IDF [7]. Despite the rise of deep neural models, these representations remain widely used due to their interpretability, modest data requirements, and strong baseline performance.

Most discriminative approaches treat count vectors as generic numerical inputs, ignoring the statistical structure of multivariate count data. In practice, such data are often overdispersed and exhibit complex dependence patterns that are poorly captured by multinomial-logit or Poisson-based models. To address these limitations, Zhang et al. [3] studied regression frameworks based on flexible multivariate count distributions, including the Dirichlet-multinomial, generalized Dirichlet-multinomial, and negative multinomial. Although these models fall outside the natural exponential family, they showed that maximum likelihood and regularized estimation can be handled within a unified optimization framework,

highlighting both the expressive power and the computational challenges of distribution-based regression for high-dimensional count data, particularly in settings involving strong dependence and overdispersion.

To overcome the restrictive covariance structure of Dirichlet-multinomial models, richer families such as Multinomial Beta–Liouville (MBL) and Multinomial Scaled Dirichlet (MSD) distributions were introduced to decouple scale, dispersion, and dependence effects [8], [9], and later incorporated into regression frameworks for multivariate count data [4]. However, their exact likelihoods remain costly in sparse, high-dimensional regimes.

Exponential approximations of complex multivariate count distributions have been proposed as a principled way to simplify optimization while preserving core modeling assumptions. In particular, exponential approximations of Multinomial Beta–Liouville (MBL) and Multinomial Scaled Dirichlet (MSD) distributions alleviate the computational burden of exact likelihoods, leading to the Exponential Multinomial Beta–Liouville (EMBL) and Exponential Multinomial Scaled Dirichlet (EMSD) models [5], [6].

These formulations improve numerical stability and scalability by yielding simplified likelihoods, gradients, and sufficient statistics that are evaluated only over nonzero counts, so the dominant computation scales with $\mathrm{nnz}(X)$ rather than the full vocabulary size D. Importantly, the exponential-family form separates data-dependent and parameter-dependent terms, enabling efficient sparse likelihood evaluation while retaining the ability to model overdispersion and complex dependence structures.

From a representation standpoint, the exponential-family form enables the reuse of likelihood-derived quantities, including information-divergence measures and gradient statistics, in hybrid generative–discriminative pipelines. These quantities support kernel-based classifiers, including divergence-based kernels derived from MSD models [6] and Fisher-score representations from generative likelihoods [10].

Hybrid generative–discriminative methods use likelihood- or gradient-based quantities as features for downstream classifiers. Classical examples include Fisher kernels [10] and probabilistic generative embeddings [11], while recent text classification work extracts likelihood-based features from pretrained language models [12].

In contrast to deep generative language models [12] or kernelized generative embeddings [10], [6], our framework integrates hybrid generative–discriminative learning directly within distribution-based regression for high-dimensional sparse counts, reusing quantities already defined by the regression objective without introducing separate latent representations or feature-extraction stages.

Unlike transformer-based fake-news detectors that learn dense representations, our framework targets sparse count data and is complementary in settings where interpretability, low data requirements, and sparse computation are important.

III. Proposed Exponential Regression Models

Throughout the paper, $X \in \mathbb{N}^{N\times D}$ denotes the sparse count matrix, $x_j \in \mathbb{N}^D$ denotes the bag-of-words count vector for sample j, $z_j \in \{1,\dots,K\}$ denotes its class label, and $y_j \in \mathbb{R}^p$ denotes the covariate vector used in the regression links. In our experiments, y_j is the one-hot encoding of z_j, so $p = K$. Here, D is the vocabulary size and $\mathrm{nnz}(X)$ denotes the number of nonzero entries in X.

A. Exponential Multinomial Beta–Liouville (EMBL) Regression

The Multinomial Beta–Liouville (MBL) distribution is a flexible alternative to Dirichlet-based models for multivariate count data, capturing richer dependence and burstiness through its Beta–Liouville prior. However, the exact MBL likelihood involves Gamma function ratios that are expensive to evaluate in high-dimensional sparse settings. An exponential approximation, referred to as the Exponential Multinomial Beta–Liouville (EMBL), was introduced to improve scalability by exploiting sparsity, since the resulting likelihood depends only on non-zero counts [5].

For EMBL regression, y_j corresponds to a class-indicator vector used to condition the regression links. The model is parameterized by positive word-level parameters α_{jd} and document-level Beta–Liouville parameters (a_j, b_j), linked to covariates through exponential link functions:

$$\begin{aligned} \alpha_{jd} &= \exp(y_j^\top A_d), \\ a_j &= \exp(y_j^\top A), \\ b_j &= \exp(y_j^\top B). \end{aligned} \tag{1}$$

We define the sample-specific concentration term

$$s_j^{(\alpha)} = \sum_{d=1}^{D} \alpha_{jd}.$$

Under the exponential approximation, the EMBL probability mass function for a count vector x_j is

$$\begin{aligned} p_{\mathrm{EMBL}}(x_j \mid \alpha_j, a_j, b_j) \propto \frac{\Gamma(s_j^{(\alpha)})}{\Gamma(s_j^{(\alpha)} + n_j)} \prod_{d:x_{jd}>0} \frac{\alpha_{jd}}{x_{jd}} \\ \times \frac{\Gamma(a'_j)\,\Gamma(b'_j)}{\Gamma(a'_j + b'_j)}. \end{aligned} \tag{2}$$

where (a'_j, b'_j) denote aggregated Beta–Liouville parameters.

Let $m_j = \sum_{d=1}^{D} x_{jd}$. EMBL introduces a padding bucket for unobserved mass, yielding an augmented $(D+1)$-dimensional count vector. In experiments, $n_j = m_j + \delta$, so $x_{j,D+1} = \delta$:

$$x_{j,D+1} = n_j - m_j, \qquad n_j = \sum_{d=1}^{D} x_{jd} + x_{j,D+1}. \tag{3}$$

Under the exponential approximation, padding affects the likelihood only through the aggregated Beta–Liouville parameters

$$a'_j = a_j + m_j, \qquad b'_j = b_j + x_{j,D+1}. \tag{4}$$

The resulting sparse EMBL log-likelihood optimized during training is

$$\mathcal{L}_{\mathrm{EMBL}}(\Theta) = \sum_{j=1}^{N} \Big(\log\Gamma(m_j+1) + \log\Gamma\big(s_j^{(\alpha)}\big) - \log\Gamma\big(s_j^{(\alpha)} + n_j\big) + \log\Gamma(a'_j) + \log\Gamma(b'_j) - \log\Gamma(a'_j + b'_j) + \log a_j + \sum_{d\in\mathcal{I}_j} \big(\log\alpha_{jd} - \log x_{jd}\big)\Big). \tag{5}$$

$$\mathcal{I}_j = \{d \in \{1,\dots,D\} : x_{jd} > 0\}.$$

The regression parameters $\Theta = \{A_d\}_{d=1}^{D}, A, B$ are estimated by maximizing the regularized conditional log-likelihood

$$\max_{\Theta} \sum_{j=1}^{N} \log p_{\mathrm{EMBL}}(x_j \mid \alpha_j, a_j, b_j) - \lambda_{\mathrm{reg}}\|\Theta\|_2^2. \tag{6}$$

Owing to the exponential formulation, the objective admits efficient gradient-based optimization and scales well to high-dimensional sparse count data.

B. Exponential Multinomial Scaled Dirichlet (EMSD) Regression

The Multinomial Scaled Dirichlet (MSD) distribution was introduced as a flexible alternative to Dirichlet-based models for multivariate count data by decoupling scale and dispersion effects across dimensions. While MSD provides expressive modeling of overdispersed counts, its likelihood involves Gamma function ratios that can be computationally expensive to evaluate in high-dimensional sparse settings. To address this limitation, an exponential approximation of MSD, referred to as the Exponential Multinomial Scaled Dirichlet (EMSD), has been proposed to improve scalability while preserving the core modeling behavior [5], [6].

For EMSD regression, let $x_j = (x_{j1},\dots,x_{jD}) \in \mathbb{N}^D$ denote the count vector for sample j, with total count $m_j = \sum_{d=1}^{D} x_{jd}$ and associated covariates $y_j \in \mathbb{R}^p$. EMSD is parameterized by positive word-level parameters λ_{jd} and ν_{jd}. We define an EMSD regression model by linking these parameters to covariates through exponential link functions:

$$\lambda_{jd} = \exp(y_j^\top C_d), \qquad \nu_{jd} = \exp(y_j^\top D_d), \tag{7}$$

We define the sample-specific concentration term

$$s_j^{(\lambda)} = \sum_{d=1}^{D} \lambda_{jd}.$$

Under the exponential approximation, the EMSD probability mass function for a count vector x_j is

$$p_{\mathrm{EMSD}}(x_j \mid \lambda_j, \nu_j) = \frac{m_j!}{\prod_{d:\,x_{jd}>0} x_{jd}} \frac{\Gamma(s_j^{(\lambda)})}{\Gamma(s_j^{(\lambda)} + m_j)} \times \prod_{d:\,x_{jd}>0} \frac{\lambda_{jd}}{\nu_{jd}^{x_{jd}}} \tag{8}$$

Algorithm 1 Generative Regression Training (MSD / EMSD / MBL / EMBL)

Input: Sparse counts $\{x_j\}_{j=1}^N$, covariates $\{y_j\}_{j=1}^N$, model $\mathcal{M}$, regularization λ_{reg}, learning-rate schedule $\{\rho_t\}$, max epochs T, tolerance ε
Output: Fitted parameters $\hat{\Theta}_{\mathcal{M}}$, final log-likelihood $\mathcal{L}_{\mathcal{M}}$
Initialize $\Theta_{\mathcal{M}}$
Compute $\mathcal{L}_{\mathcal{M}}^{(0)}$
for $t = 1$ to T **do**
  Evaluate links to obtain local parameters (model-dependent)
  Compute sparse gradients $\nabla_\Theta \mathcal{L}_{\mathcal{M}}$
  Update $\Theta_{\mathcal{M}} \leftarrow \Theta_{\mathcal{M}} + \rho_t\big(\nabla_\Theta \mathcal{L}_{\mathcal{M}} - 2\lambda_{\mathrm{reg}}\,\Theta_{\mathcal{M}}\big)$
  Compute $\mathcal{L}_{\mathcal{M}}^{(t)}$
  if $|\mathcal{L}_{\mathcal{M}}^{(t)} - \mathcal{L}_{\mathcal{M}}^{(t-1)}| < \varepsilon$ **then**
    break
  end if
end for
$\hat{\Theta}_{\mathcal{M}} \leftarrow \Theta_{\mathcal{M}}$; $\mathcal{L}_{\mathcal{M}} \leftarrow \mathcal{L}_{\mathcal{M}}^{(t)}$

Under the above parameterization, the EMSD likelihood admits a sparse formulation that depends only on non-zero counts.

$$\mathcal{L}_{\mathrm{EMSD}}(\Theta) = \sum_{j=1}^{N} \Big(\log\Gamma(m_j+1) - \sum_{d:x_{jd}>0} \log x_{jd} + \log\Gamma\big(s_j^{(\lambda)}\big) - \log\Gamma\big(s_j^{(\lambda)} + m_j\big) + \sum_{d:x_{jd}>0} \log\lambda_{jd} - \sum_{d:x_{jd}>0} x_{jd}\log\nu_{jd}\Big). \tag{9}$$

The regression parameters $\Theta = \{C_d, D_d\}_{d=1}^{D}$ are estimated by maximizing the regularized conditional log-likelihood

$$\max_{\Theta} \sum_{j=1}^{N} \log p_{\mathrm{EMSD}}(x_j \mid \lambda_j, \nu_j) - \lambda_{\mathrm{reg}}\|\Theta\|_2^2, \tag{10}$$

where $p_{\mathrm{EMSD}}(\cdot)$ denotes the exponential approximation to the MSD likelihood. Owing to the exponential formulation, the objective admits efficient gradient-based optimization and scales well to high-dimensional sparse count data.

IV. Hybrid Generative–Discriminative Framework

We employ the proposed generative regression models as structured representation learners within a two-stage hybrid framework that decouples generative parameter estimation from discriminative prediction. The overall training and inference procedure is summarized in Algorithms 1 and 2.

A. Anchor-Based Feature Construction

Rather than using the fitted generative models directly for prediction, we construct discriminative representations by evaluating the learned regression links at a fixed set of anchor covariates corresponding to target classes or conditions. Let $\{Y^{(1)},\dots,Y^{(K)}\}$ denote anchor covariates, where $Y^{(k)} \in \mathbb{R}^p$ corresponds to the k-th class. In our experiments, $p = K$ and each $Y^{(k)}$ is a one-hot class vector.

Given fitted generative parameters $\hat{\Theta}_{\mathcal{M}}$, anchor-specific parameters are obtained by evaluating the model-specific

regression links,

$$\Theta^{(k)} = g_{\mathcal{M}}(Y^{(k)}; \hat{\Theta}_{\mathcal{M}}), \qquad k = 1, \ldots, K,$$

where $g_{\mathcal{M}}(\cdot)$ denotes the link functions associated with model $\mathcal{M}$. Each anchor induces a feature block, and concatenating these blocks yields the final representation Φ_j for sample j. The construction of each block depends on the underlying generative model. In our experiments, covariates correspond to class-indicator vectors, so anchors coincide with one-hot class covariates.

1) Score-Based Features (MBL / EMBL): For MBL and EMBL regression models, the likelihood does not induce a normalized vocabulary-level prior. Feature blocks are therefore constructed using digamma-based score terms derived from the log-likelihood, computed only over observed dimensions to preserve sparsity:

$$\phi_{jd}^{(k)} = \psi(\alpha_d^{(k)} + x_{jd}) - \psi(\alpha_d^{(k)}), \qquad d : x_{jd} > 0.$$

To capture global coupling effects arising from normalization and the Beta–Liouville component, a small number of anchor-specific scalar features are appended to each block. Each anchor block has dimension $D + 3$, resulting in

$$\Phi_j \in \mathbb{R}^{K(D+3)}.$$

The three appended scalars correspond to (i) a normalization/coupling score term and (ii–iii) Beta–Liouville (a,b) score terms under the anchor. We implement the D-dimensional part sparsely (nonzeros only at observed indices), but the conceptual dimensionality is D.

2) Distributional Log-Prior Features (MSD / EMSD): For MSD and EMSD regression models, we construct an anchor-specific positive mass vector from the fitted parameters. For anchor k, we define

$$\eta_w^{(k)} = \frac{\lambda_w^{(k)}}{\nu_w^{(k)}}, \qquad w = 1, \ldots, D, \tag{11}$$

which is strictly positive since $\lambda_w^{(k)}, \nu_w^{(k)} > 0$ under the exponential links. Normalizing in log-space yields an anchor-specific log-prior,

$$\log p^{(k)}(w) = \log \eta_w^{(k)} - \log \sum_{v=1}^{D} \eta_v^{(k)}.$$

Given empirical word frequencies $\hat{p}_{jw} = x_{jw}/m_j$, the anchor feature block is constructed as

$$\phi_{jw}^{(k)} = \hat{p}_{jw} \, \log p^{(k)}(w).$$

Concatenating all anchor blocks produces

$$\Phi_j \in \mathbb{R}^{KD}.$$

B. Prediction and Normalization

To ensure comparable scaling across anchors, a blockwise ℓ_2 normalization is applied to Φ. Final predictions are obtained using a linear classifier or one-vs-rest regressor applied to the normalized representation. A linear predictive head is intentionally employed to isolate the representational contribution of the anchor-induced features while preserving computational efficiency: after sparse feature construction, prediction requires only a linear model rather than end-to-end deep fine-tuning.

Algorithm 2 Discriminative Learning with Anchor-Induced Features

Input: Sparse counts $\{x_j\}_{j=1}^N$, covariates $\{y_j\}_{j=1}^N$, labels $\{z_j\}_{j=1}^N$, anchors $\{Y^{(k)}\}_{k=1}^K$, fitted generative parameters $\hat{\Theta}_{\mathcal{M}}$
Output: Trained linear predictor $\hat{\beta}$ and predicted labels $\{\hat{z}_j\}$
for $k = 1$ to K **do**
  Evaluate regression links at anchor $Y^{(k)}$ using $\hat{\Theta}_{\mathcal{M}}$
  Store anchor-specific parameters
end for
for $j = 1$ to N **do**
  for $k = 1$ to K **do**
    Construct anchor feature block $\phi_j^{(k)}$ (model-dependent)
  end for
  Form feature vector $\Phi_j = [\phi_j^{(1)} \ \cdots \ \phi_j^{(K)}]$
end for
Apply blockwise ℓ_2 normalization to $\{\Phi_j\}$
Train linear predictor $\hat{\beta}$ on $\{(\Phi_j, z_j)\}_{j=1}^N$
for $j = 1$ to N **do**
  Predict $\hat{z}_j = \arg\max_c \left(\beta_{0,c} + \boldsymbol{\beta}_c^\top \Phi_j\right)$
end for

At inference time, test samples are transformed using the same anchor-based feature construction and normalization prior to prediction. Predictions are computed as

$$\hat{z}_j = \arg\max_c \ f_c(\Phi_j),$$

where $f_c(\cdot)$ denotes the linear decision function associated with class c. For one-vs-rest regression heads, scores $f_c(\Phi_j)$ are the per-class regression outputs.

V. Experiments

We evaluate the proposed hybrid generative–discriminative framework on two text-based fake news detection benchmarks with markedly different characteristics: BS-Detector [13], [14] and the ISOT Fake News dataset [15]. Both datasets are represented using high-dimensional sparse bag-of-words vectors, where each document contains only a small subset of the full vocabulary. This makes them suitable for assessing the scalability and representational value of distribution-based regression models.

Across both datasets, four generative regression models—MBL, EMBL, MSD, and EMSD— are evaluated under three prediction regimes: (i) linear predictive heads trained on anchor-induced feature maps, (ii) standard linear baselines trained directly on raw bag-of-words features, and (iii) direct Bayesian prediction using maximum a posteriori (MAP) inference under the fitted generative models. Performance is reported using accuracy and macro-averaged precision, recall, and F1 score to account for class imbalance.

A. BS-Detector

The BS-Detector dataset contains news articles from 244 sources labeled into eight categories, including *fake news*, *satire*, *extreme bias*, *conspiracy theory*, and *junk science*.

TABLE I

EVALUATION OF HYBRID GENERATIVE–DISCRIMINATIVE MODELS ON THE BS-DETECTOR DATASET. FOR EACH GENERATIVE REGRESSION MODEL, THE STRONGEST LINEAR CLASSIFIER AND ONE-VS-REST REGRESSION HEAD ARE REPORTED. THE OVERALL BEST-PERFORMING MODEL BY MACRO-AVERAGED F1 SCORE IS HIGHLIGHTED IN BOLD.

Generative Model	Head Type	Linear Predictive Head	Acc.	F1 (Macro)	Prec. (Macro)	Recall (Macro)
MBL	Classifier	SGD (hinge)	91.73	66.26	90.50	57.87
	Regression	OVR SGD ElasticNet	89.85	65.98	69.38	65.83
EMBL	Classifier	SGD (hinge)	**91.27**	**66.39**	**81.58**	**59.84**
	Regression	OVR SGD ElasticNet	89.52	66.27	68.00	67.23
MSD	Classifier	Linear SVC	90.66	57.75	77.67	49.02
	Regression	OVR Ridge	89.46	58.37	71.66	51.29
EMSD	Classifier	SGD (hinge)	90.98	62.99	72.90	57.94
	Regression	OVR Ridge	90.20	59.26	75.10	51.15

TABLE II

EVALUATION OF HYBRID GENERATIVE–DISCRIMINATIVE MODELS ON THE ISOT FAKE NEWS DATASET.

Generative Model	Linear Predictive Head	Acc.	F1 (Macro)	Prec. (Macro)	Recall (Macro)
MBL	SGD (hinge)	97.17	97.16	97.21	97.13
EMBL	SGD (log-loss)	97.71	97.70	97.70	97.71
MSD	SGD (log-loss)	97.14	97.13	97.14	97.12
EMSD	Linear SVC	**99.59**	**99.59**	**99.58**	**99.59**

TABLE III

DIRECT BAYESIAN (MAP) PREDICTION ON BS-DETECTOR. PRECISION, RECALL, AND F1 ARE MACRO-AVERAGED.

Model	Acc.	F1	Prec.	Rec.
MBL	73.63	35.02	30.49	50.80
EMBL	80.86	37.43	34.43	49.19
MSD	88.03	15.19	23.50	14.53
EMSD	88.03	15.19	23.50	14.53

The dataset is highly imbalanced and exhibits substantial vocabulary overlap across classes, making it a challenging multi-class setting. All discriminative heads therefore use class-balanced objectives.

Each article is represented as a sparse bag-of-words count vector $x_j \in \mathbb{N}^D$, where each dimension corresponds to the frequency of a vocabulary term, yielding over 24k dimensions with approximately 92% sparsity. We use a stratified $75/25$ train–test split. Generative models are trained using class-indicator vectors as covariates, and linear classifiers or one-vs-rest regressors are trained on the resulting anchor-induced feature maps.

Tables III and I compare direct Bayesian MAP prediction and hybrid feature-map prediction on the same BS-Detector split. While several MAP models achieve moderate to high accuracy, macro-averaged F1 and recall reveal a strong bias toward majority classes. In particular, MSD and EMSD exhibit extremely low recall, indicating that direct MAP inference collapses to dominant-class predictions when class distributions overlap heavily in the bag-of-words space.

These results highlight a key limitation of using generative regression models as standalone classifiers in highly imbalanced, multi-class settings. Likelihood values alone are insufficient to induce discriminative decision boundaries in high-dimensional sparse domains, especially when minority classes share substantial vocabulary with majority ones.

In contrast, hybrid generative–discriminative models trained on anchor-induced feature maps achieve substantial gains across all macro-averaged metrics, as shown in Table I. For all four generative families, macro-F1, the primary metric under class imbalance, improves substantially relative to direct MAP inference, with particularly large gains in recall. This indicates that the feature maps enable linear heads to recover minority-class structure that is otherwise suppressed under generative inference.

Among the models, MBL- and EMBL-based representations yield the strongest overall performance on BS-Detector. Their score-based feature maps explicitly encode word-level concentration and coupling effects, which appear especially beneficial in this imbalanced multi-class setting. The exponential approximation in EMBL preserves these properties while improving numerical stability, resulting in the highest macro-F1 score overall.

For reference, logistic regression trained directly on raw bag-of-words features achieves 83.36% accuracy and a macro-F1 score of 50.60 on BS-Detector, further highlighting the benefit of distribution-aware feature construction in this challenging setting.

The improvements on BS-Detector are substantial across macro-averaged metrics: compared with direct MAP inference, the hybrid feature maps yield large gains in macro-F1 and recall, and they also outperform the raw bag-of-words logistic regression baseline in macro-F1. This indicates that the proposed feature maps are especially beneficial in the imbalanced multi-class setting.

TABLE IV

DIRECT BAYESIAN (MAP) PREDICTION ON THE ISOT FAKE NEWS DATASET. PRECISION, RECALL, AND F1 ARE MACRO-AVERAGED.

Model	Acc.	F1	Prec.	Rec.
MBL	96.08	96.07	96.06	96.09
EMBL	**96.12**	**96.11**	**96.10**	**96.12**
MSD	93.69	93.66	93.79	93.59
EMSD	93.61	93.58	93.72	93.51

B. ISOT Fake News

The ISOT Fake News dataset is a binary classification benchmark consisting of approximately balanced real and fake news articles from reliable and misinformation sources. Compared to BS-Detector, the dataset exhibits cleaner class separation and is closer to linearly separable under standard bag-of-words representations. We use an $80/20$ stratified split and $K = 2$ anchors. As in BS-Detector, articles are represented using sparse bag-of-words count vectors extracted from article titles and body text.

Tables IV and II compare direct Bayesian MAP prediction and hybrid feature-map prediction on the same ISOT split. In contrast to BS-Detector, generative models achieve strong MAP performance in this setting, reflecting the simpler binary class structure.

Hybrid results on ISOT are shown in Table II. All feature-map based models achieve near-saturated performance, with EMSD-based representations attaining almost perfect accuracy and macro-F1. These results closely match those obtained by strong linear baselines trained directly on raw bag-of-words features, indicating that the dataset is largely linearly separable. Overall, the ISOT results indicate that while the proposed framework does not outperform strong discriminative baselines on simple, linearly separable tasks, it remains competitive without degrading performance. In contrast, its benefits are most pronounced in complex, high-dimensional, and imbalanced settings such as BS-Detector, where distribution-aware representations substantially improve minority-class performance.

VI. CONCLUSION

This work introduced Exponential Multinomial Beta–Liouville (EMBL) and Exponential Multinomial Scaled Dirichlet (EMSD) regression models for high-dimensional sparse count data, along with a hybrid generative–discriminative framework based on anchor-induced feature maps.

Experiments on multi-class and binary fake news benchmarks show that likelihood- and score-based representations derived from distribution-based regression models improve discriminative performance when paired with simple linear heads. In particular, while direct generative inference can collapse in highly imbalanced and overlapping settings, the same generative structure remains informative when expressed through anchor-evaluated feature maps.

These results suggest that flexible generative models are most valuable in high-dimensional sparse domains as representation learners rather than standalone classifiers. Exponential approximations enable this framework by preserving interpretability and distributional structure while improving scalability and numerical stability.

By decoupling generative modeling from discriminative prediction, the proposed approach provides an efficient alternative to end-to-end deep models for sparse text data. Future work may explore temporal covariates, multimodal inputs, and adaptive anchor selection.

REFERENCES

[1] B. Hu, Z. Mao, and Y. Zhang, "An overview of fake news detection: From a new perspective," *Fundamental Research*, vol. 5, no. 1, 2024.

[2] A. O. Ojo, F. Najar, N. Zamzami, H. T. Himdi, and N. Bouguila, "Smoothdetector: A smoothed dirichlet multimodal approach for combating fake news on social media," *IEEE Access*, vol. 13, pp. 39 289–39 305, 2025.

[3] Y. Zhang, H. Zhou, J. Zhou, and W. Sun, "Regression models for multivariate count data," *Journal of Computational and Graphical Statistics*, vol. 26, no. 1, pp. 1–13, 2017.

[4] P. Koochemeshkian, N. Zamzami, and N. Bouguila, "Flexible distribution-based regression models for count data: Application to medical diagnosis," *Cybernetics and Systems*, vol. 51, no. 4, pp. 442–466, 2020.

[5] N. Zamzami and N. Bouguila, "High-dimensional count data clustering based on an exponential approximation to the multinomial beta-liouville distribution," *Information Sciences*, vol. 524, pp. 116–135, 2020.

[6] ——, "Hybrid generative–discriminative approaches based on multinomial scaled dirichlet mixture models," *Applied Intelligence*, vol. 49, no. 6, pp. 2153–2169, 2019.

[7] N. F. Baarir and A. Djeffal, "Fake news detection using machine learning," in *Proceedings of the 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, 2021.

[8] N. Bouguila, "Count data modeling and classification using finite mixtures of distributions," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 186–198, 2010.

[9] N. Zamzami and N. Bouguila, "Text modeling using multinomial scaled dirichlet distributions," in *Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2018, pp. 69–80.

[10] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*, vol. 11, 1998. [Online]. Available: https://papers.nips.cc/paper_files/paper/1998/hash/db1915052d15f7815c8b88e879465a1e-Abstract.html

[11] A. Agarwal and H. Daumé, "Generative kernels for exponential families," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 15, 2011, pp. 85–92. [Online]. Available: https://proceedings.mlr.press/v15/agarwal11b.html

[12] S. Pawar, N. Ramrakhiyani, A. Sinha, M. Apte, and G. Palshikar, "Why generate when you can discriminate? a novel technique for text classification using language models," in *Findings of the European Chapter of the Association for Computational Linguistics*, 2024, pp. 1099–1114. [Online]. Available: https://aclanthology.org/2024.findings-eacl.74/

[13] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, ser. Lecture Notes in Computer Science, vol. 10618. Springer, 2017, pp. 127–138.

[14] ——, "Detecting opinion spams and fake news using text classification," *Journal of Security and Privacy*, vol. 1, no. 1, 2018.

[15] M. Risdal, "Getting real about fake news," 2016. [Online]. Available: https://www.kaggle.com/dsv/911

[0]Additional derivations, implementation details, and feature-map construction details are available at: `https://github.com/Sanah27/Hybrid-Generative-Discriminative-Regression-Models-for-High-Dimensional-Fake-News-Detection/blob/main/Supplementary_Material.pdf`.