Multitask Transformer Models for Demographic and Industry Profiling on Long-Form Blog Texts

Anonymous authors
Paper under double-blind review

Abstract

We address the challenge of multitask author profiling on long-form blog text by developing four transformer-based models that jointly predict gender, age group, and industry. Using a cleaned version of the Blog Authorship Corpus Anonymous (2025a), we explore document-length handling strategies that span input ranges from 192 to 500 tokens, including long-context encoding, BART-based summarization, and chunked processing with prediction fusion. Our experiments show that multitask learning consistently outperforms strong single-task baselines, with the largest gains for industry. We further find that broader input context yields more reliable predictions, while alternative representations emphasize complementary stylistic and topical cues. Taken together, these findings provide a comprehensive analysis of text-length effects in multitask author profiling and highlight the importance of contextual breadth for robust demographic inference. The dataset was preprocessed by merging industry tags into fourteen categories and applying standard text normalization.

1 Introduction

Author profiling is the task of inferring demographic traits of an individual — including gender, age, occupation, and personality — from their written text (Rangel et al., 2018). The task has practical uses in areas like forensic analysis, social-behavior research, improving personalised recommendations, and identifying misinformation, deception, or harmful accounts on social media (Ott et al., 2011; Mishra et al., 2019; Lanza-Cruz et al., 2023). Extended personal writing, such as blog posts, remains a particularly rich source of signal because it simultaneously exposes linguistic style, topical preferences, and sociolectal patterns that correlate strongly with real-world author attributes (Schler et al., 2006b; Nguyen et al., 2016).

Blog posts are long, multi-paragraph documents in which demographic cues are sparse and spread across extended spans of text; as a result, standard truncation or aggressive compression often discards a substantial portion of the usable signal. Prior work has shown that user-generated text is particularly challenging to model because of its noise and high stylistic variation (Rangel et al., 2018). Long-document studies show that preserving extended context is essential for accurate modeling, and this directly affects author profiling on long-form text (Adhikari et al., 2019). Recent analyses of transformer models confirm that both document-representation strategy and effective context length significantly influence downstream performance (Park et al., 2022). Blogs therefore constitute a demanding yet realistic testbed that exposes the limitations of short-context approaches and motivates a systematic investigation of long-document handling strategies.

Existing multitask author profiling work (Jiang et al., 2018) relies on pre-transformer architectures such as CNNs and LSTMs, whereas transformer-based studies (Thakur & Tickoo, 2023) typically predict individual demographic traits and do not include industry sector prediction. To our knowledge, no prior work has proposed a transformer-based multitask framework that jointly predicts gender, age group, and industry sector from long-form blog posts—a gap the present study directly addresses.

The main contribution of our approach is threefold: (1) We present the first transformer-based multitask framework that jointly predicts gender, age group, and a fine-grained 14-class industry sector from real blog posts, using the cleaned Blog Authorship Corpus (Anonymous, 2025b). (2) We conduct a controlled comparison of four long-document representation strategies—truncation, extended-context encoding, summarization-based compression, and chunk-based processing—and quantify their differential impact across all three profiling tasks. (3) We uncover a key novel finding: gender and age-group prediction are notably robust to aggressive truncation, summarization, and chunking, whereas industry classification suffers sharp performance degradation under the same conditions. This indicates that professional cues are far sparser and more widely distributed throughout the document than demographic signals, with important implications for long-document author profiling systems.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset, preprocessing pipeline, and proposed models. Section 4 presents experimental results, and Section 5 discusses findings and concludes the paper.

2 Related Work

The Transformer architecture replaced recurrent networks with self-attention mechanisms and became the cornerstone of modern NLP (Vaswani et al., 2017). BERT introduced bidirectional pre-training through masked language modeling, significantly boosting performance on text classification and author profiling (Devlin et al., 2019). DeBERTa further enhanced this paradigm using disentangled attention and improved pre-training objectives, achieving state-of-the-art results on numerous benchmarks (He et al., 2021). These developments established the modern foundation for encoder-based language models, redefining how large-scale text representations are learned and driving the broad advances in language understanding. Beyond surface-level sequence modeling, transformer language models have been shown to encode deep hierarchical structure, successfully learning and reasoning over context-free grammars, with internal representations aligned to algorithmic parsing procedures Allen-Zhu & Li (2025).

Multitask learning (MTL) has been widely employed in author profiling to exploit correlations between demographic traits. Early shared-task systems jointly predicted gender and age across genres using hard parameter sharing with classical machine-learning models such as SVMs and logistic regression (Rangel et al., 2016). More recent shared tasks, such as the Italian TAG-it benchmark, examine multitask prediction of age, gender, and topic on blog posts, though their topic labels represent thematic categories rather than the industry sectors modeled in our work (Cimino et al., 2020). Recent studies have also investigated the interaction between multitask learning and group-wise robustness, showing that standard multitask fine-tuning can have inconsistent effects on worst-group accuracy, motivating closer analysis of shared representations (Kulkarni et al., 2024). MTL typically employs hard parameter sharing, where lower neural network layers are shared across tasks to reduce parameters and overfitting, or soft parameter sharing, which regularizes task-specific layers for greater flexibility (Chen et al., 2024; Zouari, 2020). Challenges such as task interference, where conflicting gradients degrade performance, are mitigated through techniques like dynamic loss weighting and gradient alignment (Chen et al., 2024). Together, these works demonstrate that multitask learning is an effective approach for capturing shared linguistic signals across demographic attributes in author profiling, while highlighting the importance of architectural choices and careful optimization.

The Blog Authorship Corpus (Schler et al., 2006a) has long served as a standard benchmark for author profiling, providing blog posts annotated with gender, age, and industry. Subsequent study has primarily targeted gender and age prediction, progressing from early stylistic and lexical features to deep learning architectures using Blog Authorship Corpus (Thakur & Tickoo, 2023). Although a few works have investigated multitask learning using hierarchical document representations on the same dataset (Jiang et al., 2018), they primarily focus on other prediction tasks, and industry prediction has received minimal attention.

Unlike prior studies that largely ignored industry prediction, our work jointly models gender, age group, and 14-class industry using a single multitask transformer with extended 500-token context. Our work is the first to systematically demonstrate that long contiguous sequences are critical for industry inference while gender and age remain robust under aggressive truncation or summarization, resulting in improved performance across the tasks.

Table 1: Distribution of Gender, Age Group, and Industry in the dataset.

Attribute	Category	Count	%
Gender	Male	51,002	53.02
	Female	45,197	46.98
Age Group	13–17	17,471	18.16
	18–29	55,377	57.57
	30–48	23,351	24.27
Industry	Education	10,322	10.73
	Technology	9,560	9.94
	Arts	9,728	10.11
	Business & Cons.	6,371	6.62
	Other (10 categories)	50,225	51.60

3 Methodology

This section describes the dataset cleaning and preprocessing pipeline, the single-task baselines, and our four multitask transformer models with different document representation strategies.

3.1 Dataset Preprocessing

We use the Blog Authorship Corpus, a collection of individual blog posts written before 2004 (Tatman, 2017), and rely on the cleaned and standardized version introduced in (Anonymous, 2025b). In total, our cleaned dataset includes 96,199 posts and 43 million words authored by 5,477 bloggers—averaging 17.6 posts and approximately 7,900 words per author. The dataset is refined for author profiling and NLP tasks by applying the preprocessing steps summarized in Fig. 1, which include text cleaning and normalization, merging semantically related industry labels, and splitting long posts into 200–1,500-word chunks. Stopwords and emojis are retained to preserve stylistic and affective cues and to maintain the contextual integrity required by transformer-based models. Summaries are generated using the BART-Large-CNN model for efficient downstream processing (Lewis et al., 2020). Table 1 summarizes the final distributions of gender, age, and industry categories retained in the dataset. Code and configurations are available at (Anonymous, 2025a).

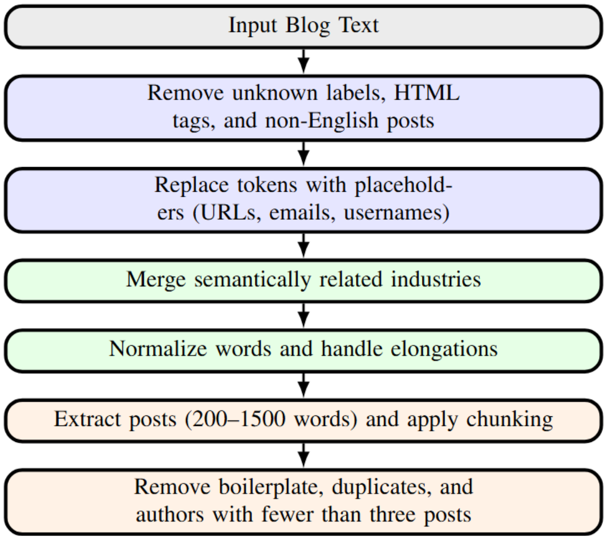

Input Blog Text
Remove unknown labels, HTML tags, and non-English posts
Replace tokens with placeholders (URLs, emails, usernames)
Merge semantically related industries
Normalize words and handle elongations
Extract posts (200–1500 words) and apply chunking
Remove boilerplate, duplicates, and authors with fewer than three posts

Figure 1: Dataset preprocessing pipeline. Blue = cleaning, green = normalization, orange = filter- ing.

3.2 Model Architectures and Training

We first introduce the single-task baselines used for comparison. Logistic Regression, SVM, LSTM, and BERT are trained independently to predict gender, three age groups, and fourteen industries using the cleaned Blog Authorship Corpus. To investigate the benefits of joint learning and extended context, we introduce four multitask learning (MTL) transformer variants that share a single encoder followed by three task-specific classification heads.

The variants are designed to systematically compare input processing strategies within an otherwise identical MTL framework:

RoBERTa-192: RoBERTa-base (Liu et al., 2019) processing 192-tokens of each post.

DeBERTa-500: DeBERTa-v3-base taking full sequences up to 500 tokens.

DeBERTa-Summary: DeBERTa-v3-base applied to single BART-large-CNN summaries truncated to 256 tokens.

DeBERTa chunking-model that segments long posts to capture full-document information.

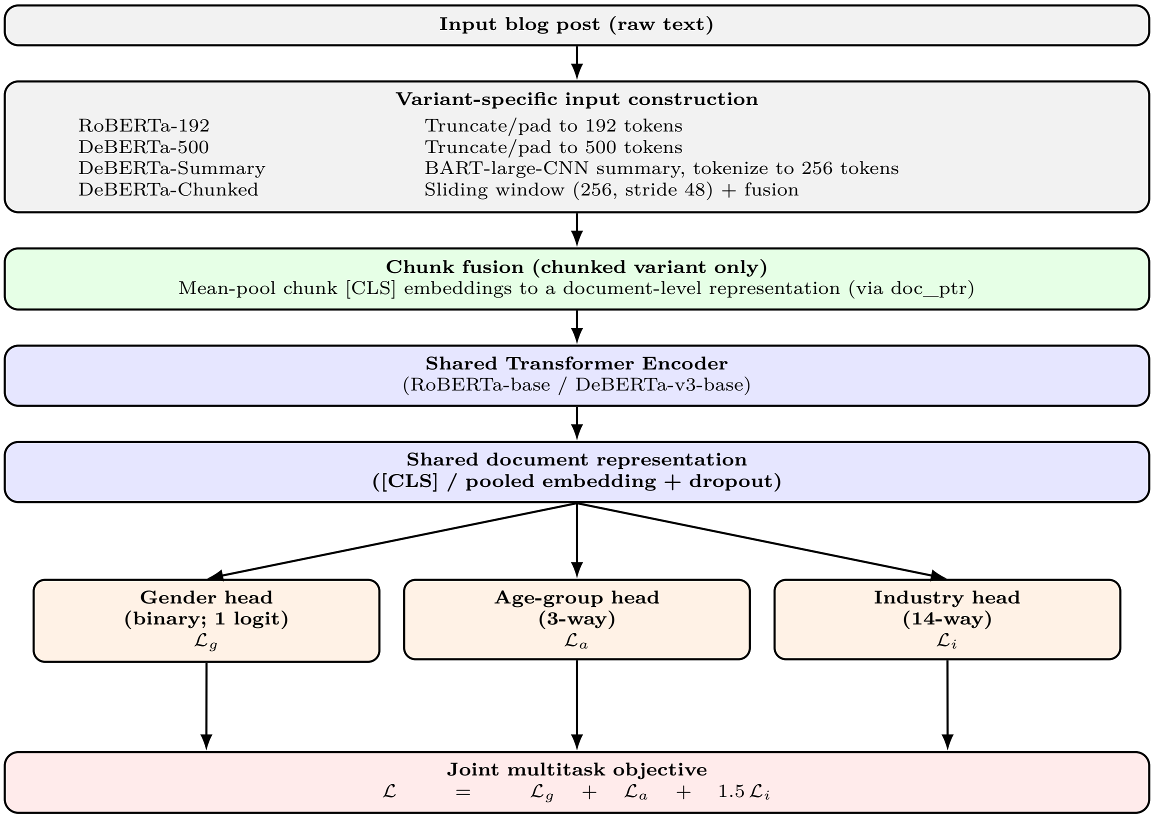

Input blog post (raw text)
Variant-specific input construction
RoBERTa-192
Truncate/pad to 192 tokens
DeBERTa-500
Truncate/pad to 500 tokens
DeBERTa-Summary
BART-large-CNN summary, tokenize to 256 tokens
DeBERTa-Chunked
Sliding window (256, stride 48) + fusion
Chunk fusion (chunked variant only)
Mean-pool chunk [CLS] embeddings to a document-level representation (via doc__ptr)
Shared Transformer Encoder
(RoBERTa-base / DeBERTa-v3-base)
Shared document representation
([CLS] / pooled embedding + dropout)
Gender head
(binary; 1 logit)
$\mathcal{L}_g$
Age-group head
(3-way)
$\mathcal{L}_a$
Industry head
(14-way)
$\mathcal{L}_i$
Joint multitask objective
$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_a + 1.5\,\mathcal{L}_i$

Figure 2: MTL architecture with *parallel* task heads and hard parameter sharing. Model variants differ only in input construction (truncation, long-context encoding, summarization, or chunking with mean pooling), while all heads are trained jointly with a weighted loss.

Single-task baselines are trained using class-weighted cross-entropy loss for each task independently, ensuring that minority classes were fairly represented during optimization. Figure 2 summarizes the shared-encoder, multi-head setup and the input-handling variants. For the multitask (MTL) models, the total loss is the weighted sum of the individual task losses:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_a + 1.5\,\mathcal{L}_i$$

We apply a weight of 1.5 to the industry loss to compensate for its higher difficulty and cardinality. All models are optimized with AdamW (weight decay 0.01) using a linear learning-rate schedule with 10% warmup steps and learning rate 2e-5. The dataset is split into 85% training and 15% test sets, stratified by industry to preserve distributions. We use different batch sizes across models depending on their computational cost. To prevent data leakage, we use an author-level train–test split: all posts from each author are assigned exclusively to either the training set or the test set. This guarantees evaluation on entirely unseen writers. All models are evaluated using macro-averaged F1-score and accuracy, with MTL models using a composite F1-score (average across tasks). Training is performed using PyTorch on cloud-based GPUs provided by Kaggle and Google Colab (T4/A100).

3.2.1 Single-Task Machine Learning Baselines

For comparison, we trained classical machine learning baselines (Logistic Regression and SVM) using TF-IDF features, and deep learning models (BERT and LSTM) on raw text sequences under the same split. The TF-IDF vectorizer was configured with unigrams and bigrams to capture both individual word and short phrase patterns, excluding tokens appearing in fewer than two documents or in more than 90% of documents.

The BERT-base-uncased single-task baseline is fine-tuned for sequence classification with a maximum input length of 192 tokens, ensuring direct comparability with the RoBERTa-192 MTL variant. The bidirectional LSTM likewise operates on sequences truncated or padded to 192 tokens. Both models follow the same training protocol as the MTL transformers. The LSTM required more training epochs because we kept the learning rate consistent with the transformer models, which slows convergence but ensures a fair and controlled comparison.

3.2.2 RoBERTa-Based MTL with 192-Token Length

Our first MTL model uses RoBERTa-base to jointly predict gender, age group, and industry, leveraging shared linguistic features. The novel combination of these tasks, tailored for the cleaned Blog Authorship Corpus, uses a shared RoBERTa encoder (768-dimensional [CLS] token output) with dropout (p=0.1) and task specific linear heads for gender (R1), age group (R3), and industry (R14). Lg is binary cross-entropy for gender, and La and Li are cross-entropy losses with label smoothing (0.05) for age group and industry, with 1.5 weight on industry to prioritize its complexity. Texts are tokenized (max length: 192) and dynamically padded. The model was fine-tuned for 10 epochs (batch size: 16) with AdamW, a cosine scheduler (100 warmup steps), and gradient clipping (1.0).

3.2.3 DeBERTa-Based MTL with 500-Token Length

To capture longer contexts, we developed an MTL model using DeBERTa-V3-base, novel for its disentangled attention mechanism. It predicts the same tasks as the RoBERTa model, using a shared encoder (768-dimensional [CLS] output) with dropout (p=0.1) and identical task-specific heads. Texts are tokenized with a maximum length of 500 and dynamically padded. The model is fine-tuned for 16 epochs using the same architecture as the other variants. Model checkpointing saved the best and last states based on composite F1-score, enhancing training robustness.

3.2.4 MTL with Summarized Text and 256-Token Length

To explore concise inputs, we apply our DeBERTa-V3-based MTL model to summarized texts generated by facebook/bart-large-cnn, a novel approach to reduce text length while preserving key information. Blog posts are summarized using beam search (beam size: 5) with length constraints and repetition mitigation. Summaries are tokenized with a maximum length of 256 and dynamically padded. The model architecture mirrors the RoBERTa model, fine-tuned for 12 epochs with identical hyperparameters and checkpointing.

3.2.5 MTL with Chunked Text

To handle long posts, we develop a chunking-based variant of our DeBERTa-V3-base MTL model. While it retains the same underlying architecture and optimization setup as the original model, we introduce a sliding-window tokenizer (stride = 48, max length = 256) to segment extended posts into overlapping chunks. Chunk embeddings are mean-pooled (attention-masked), passed to task-specific heads, and the resulting logits are averaged per document to compute document-level losses and predictions. The model is fine-tuned for 19 epoches. This approach enhances feature extraction from longer texts while remaining compatible with the existing MTL framework.

3.3 Evaluation Metrics

Let $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ be the test set for a given task with C classes, where $y_n \in \{1, \dots, C\}$ is the gold label and $\hat{y}_n$ is the predicted label. We report *Accuracy* and *Macro-F1* for each task, and a *Composite Macro-F1* that aggregates performance across the three tasks (gender, age group, industry).

Accuracy. Accuracy is the fraction of correct predictions:

$$\text{Acc} \;=\; \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[\hat{y}_n = y_n]\,, \tag{1}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Per-class Precision, Recall, and F1. For class c, define true positives, false positives, and false negatives as

$$\text{TP}_c = \sum_{n=1}^{N} \mathbb{I}[y_n = c \wedge \hat{y}_n = c], \quad \text{FP}_c = \sum_{n=1}^{N} \mathbb{I}[y_n \neq c \wedge \hat{y}_n = c], \quad \text{FN}_c = \sum_{n=1}^{N} \mathbb{I}[y_n = c \wedge \hat{y}_n \neq c]. \tag{2}$$

Precision and recall for class c are:

$$P_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \qquad R_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, \tag{3}$$

and the class-wise F1-score is:

$$F1_c = \frac{2P_cR_c}{P_c + R_c}. \tag{4}$$

(When a denominator is zero, we follow the standard convention of setting the corresponding quantity to zero.)

Macro-F1. Macro-F1 averages the class-wise F1-scores uniformly, giving equal weight to each class:

$$\text{MacroF1} \;=\; \frac{1}{C} \sum_{c=1}^{C} F1_c. \tag{5}$$

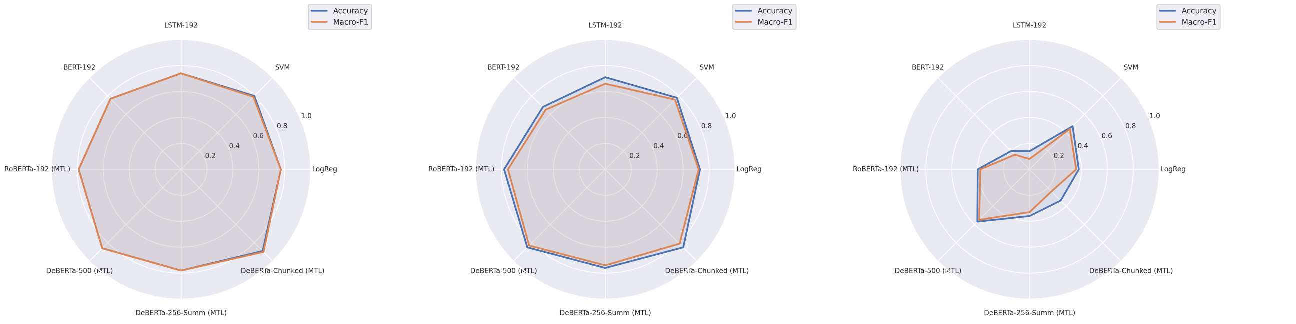
Composite Macro-F1 across tasks. Let MacroF1_g, MacroF1_a, and MacroF1_i denote macro-F1 for gender, age group, and industry, respectively. We define the composite score as:

$$\text{CompositeF1} = \frac{\text{MacroF1}_g + \text{MacroF1}_a + \text{MacroF1}_i}{3}. \tag{6}$$

This metric summarizes overall multitask performance while treating the three tasks equally.

4 Results of Experiments

Table 2 reports accuracy and macro-F1 scores on the held-out test set for all baselines and multitask models. Single-task baselines (Logistic Regression, SVM with TF-IDF, LSTM, and BERT) reach 0.74–0.79 macro-F1 on gender, 0.65–0.76 on age group, and 0.08–0.44 on the 14-class industry task. Classical single-task



(a) Gender (Binary)

(b) Age Group (3 classes)

(c) Industry (14 classes)

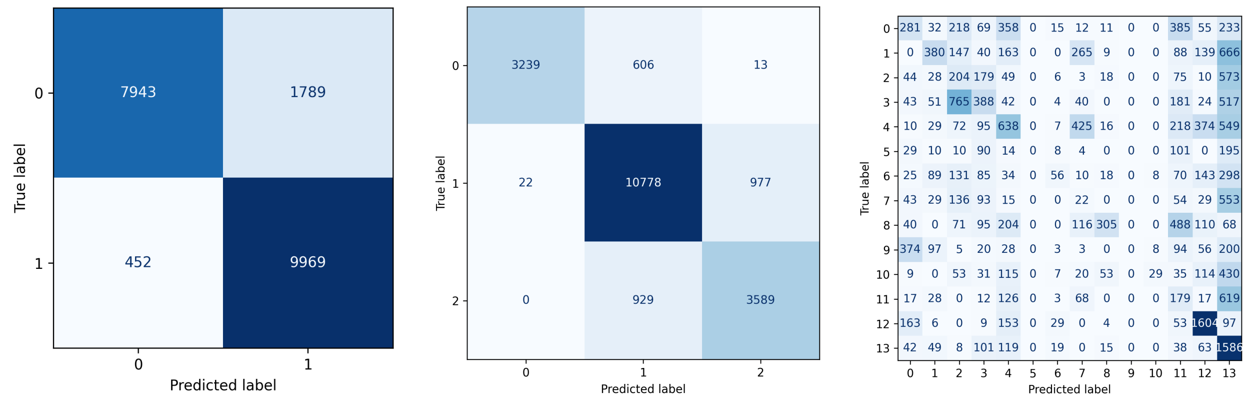
Figure 3: Radar plots compare Accuracy and Macro-F1 across single-task and multitask models.

models perform reasonably on gender and age but struggle dramatically on the 14-class industry task, with the strongest baseline (SVM) reaching only 0.47 accuracy and 0.44 macro-F1. This confirms that industry prediction requires richer contextual representations than those captured by bag-of-n-grams or short-context models. All four multitask transformer variants substantially outperform the single-task baselines.

Joint multitask training improves performance on all three tasks compared to training separate single-task models on the same architecture. The DeBERTa-V3-base model with a 500-token context is the best-performing configuration, achieving the highest composite macro-F1 (0.75) across all setups, along with strong per-task gains: gender F1 of 0.86, age-group F1 of 0.83, and, most notably, industry F1 of 0.55 (+11 points over the best single-task model). Extending the context length from 192 tokens (RoBERTa) to 500 tokens (DeBERTa) further increases scores on every task. Replacing the contiguous 500-token input with either BART summarization or sliding-window mean pooling preserves most of the performance on gender and age group while reducing industry macro-F1 substantially. Figure 3 summarizes these trends using radar plots, which compare Accuracy and macro-F1 across models for each task. The plots highlight the consistent gains from multitask learning and make clear that improvements are largest and most uneven for industry prediction, while gender and age exhibit more stable performance across modeling choices.Although the sliding-window variant achieves the strongest gender macro-F1 (0.90) among the multitask models, its very low industry macro-F1 (0.24) substantially reduces its overall composite performance. This contrast shows that gender and age cues remain fully recoverable when the input document is split into short independent chunks, whereas industry prediction requires coherent long-range context that is disrupted by chunking.

Table 2: Performance of Single-Task ML and Multitask Learning Models on Gender, Age Group, and Industry Prediction Tasks

Model	**Epochs**	**Gender** **Acc / F1**	**Age Group** **Acc / F1**	**Industry** **Acc / F1**	**Composite** **F1**
Single-Task ML Baselines					
Logistic Regression		0.77 / 0.77	0.73 / 0.72	0.38 / 0.36	–
SVM		0.80 / 0.79	0.78 / 0.76	0.47 / 0.44	–
LSTM (192 tokens)	30 / 66 / 26	0.74 / 0.74	0.71 / 0.66	0.14 / 0.08	–
BERT (192 tokens)	4 / 4 / 2	0.77 / 0.77	0.68 / 0.65	0.20 / 0.16	–
Multitask Learning Models					
RoBERTa (192 tokens)	10	0.79 / 0.79	0.78 / 0.75	0.40 / 0.38	0.64
DeBERTa (500 tokens)	16	0.86 / 0.86	**0.85 / 0.83**	**0.57 / 0.55**	**0.75**
DeBERTa (256 tokens, Summ.)	12	0.78 / 0.78	0.76 / 0.74	0.36 / 0.33	0.62
DeBERTa (chunked)	19	**0.89 / 0.90**	**0.85** / 0.81	0.34 / 0.24	0.65

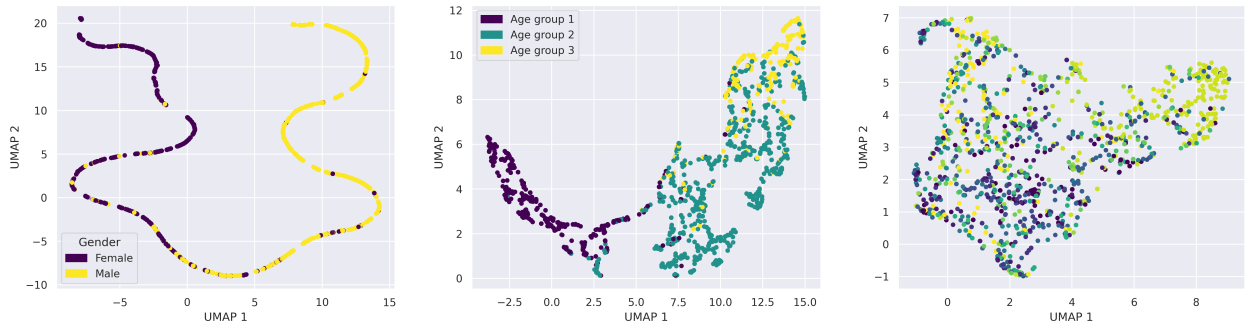


(a) Gender

(b) Age Group

(c) Industry

Figure 4: Confusion matrices (absolute counts) for the model overall (DeBERTa (Chunked)).



(a) Colored by gender

(b) Colored by age group

(c) Colored by industry

Figure 5: UMAP 2D projections of the final embeddings produced by the DeBERTa (chunked) model, colored by different task labels.

These results confirm that long contiguous context is the decisive factor for strong overall performance, especially on the challenging industry task.

Figure 4 presents the confusion matrices computed using the chunking model, which offers multi-chunk representations suitable for fine-grained error analysis while also reducing the computational cost of generating these visualizations. Figure 4a shows that gender prediction is highly reliable, with errors affecting only about 8% of the test set. Age-group classification(Figure 4b) exhibits clear diagonal dominance across all three categories. The 18–29 group achieves the highest recall, while the 13–17 group shows the least cross-class confusion. Most remaining errors arise between the two adult brackets (18–29 and 30–48), with confusion distributed roughly symmetrically in both directions. Industry performance(Figure 4c) varies substantially by class frequency. Smaller industries, particularly Finance & Property and Non-Profit, which together account for only about 7% of the dataset, exhibit negligible diagonal mass and are never predicted correctly. In contrast, high-support categories such as Student, Technology, and Education (collectively 32% of the data) exhibit strong within-class accuracy. These patterns reveal strong performance on well-represented classes but persistent challenges on low-support or stylistically overlapping categories.

Figure 5 shows UMAP projections of the [CLS] representations learned by our model. When colored by gender (Figure 5a), the embedding space splits into two remarkably clean, compact clusters with virtually no overlap, consistent with the observed gender F1 of 0.90 and confirming that robust stylistic cues are captured even in a multitask setting. Age groups (Figure 5b) form three well-defined regions: the dominant 18–29 cohort occupies a dense central area, while the younger (13–17) and older (30–48) groups appear as distinct arms with only minor mixing at the adult boundary. The 14-class industry projection (Figure 5c) reveals a more complex yet highly structured manifold: major industries (e.g., Education, Technology, Arts) correspond to large, coherent clusters, whereas rarer categories are more scattered and partially merge with semantically similar groups. This pattern directly explains the moderate macro-F1 of 0.24 on the industry task and underscores the benefit of longer contiguous context for disambiguating subtle topical and professional signals.

5 Discussion and Conclusion

Our study shows that multitask learning improves author-profiling performance across gender, age, and industry, including an 11% macro-F1 gain for industry over strong single-task baselines. Models with broader input context perform best overall, suggesting that long-form blog posts benefit from representations that capture extended stylistic and topical cues. Our comparison of long-document strategies indicates that different approaches highlight distinct aspects of the text and thus offer complementary advantages. Error patterns and embedding visualizations further reveal the relative difficulty of age and industry prediction due to overlapping semantic cues and boundary ambiguities. Overall, the results demonstrate that robust demographic inference on long-form blog text requires effective multitask representations and careful handling of document length.

Multitask learning succeeds because gender, age, and industry share overlapping lexical and topical cues, enabling richer representations than single-task training. Longer context helps for the same reason: diagnostic signals are sparse and distributed across posts. In contrast, chunking fragments the document, disrupting global structure and raising computational cost, while summarization compresses the text, often removing fine-grained cues and subtle demographic signals needed for accurate prediction.

We introduce the first transformer-based multitask framework that jointly predicts gender, age group, and 14 industries from long-form blog text, provide the first systematic evaluation of input-length strategies for this setting, and offer new evidence on how demographic signals depend on context range, clarifying why gender is local, age is medium-range, and industry requires global topical coherence, supported by embedding-space analyses that reveal how shared encoders organize these signals.

Future work may deepen bias analysis by examining how stylistic cues align with demographic attributes and by modeling the latent stylistic dimensions that drive these associations. It would also be valuable to incorporate insights from recent fairness-aware multitask learning methods and bias-auditing frameworks, given the sensitive nature of demographic inference. A key limitation of current author, profiling research—including our own, is the assumption that writing style is static. In reality, individuals systematically shift their vocabulary, syntax, and topical focus across time, audience, and life stage, producing substantial intra-author variation. Our embedding analyses already hint at this dynamism: age-cohort clusters stretch and overlap in ways that reflect gradual linguistic evolution rather than fixed traits. Developing models that explicitly account for such personal evolution—whether through temporal modeling, continual learning, or diachronic multitask objectives—represents a critical and exciting direction for achieving truly robust and temporally stable profiling systems. The structure visible in our present data suggests that this goal is not only necessary but entirely feasible.

Acknowledgments

Anonymous acknowledgments

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. Docbert: Bert for document classification, 2019. URL https://arxiv.org/abs/1904.08398.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, learning hierarchical language structures. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=mPQKyzkA1K.

Anonymous. Preprocessing pipeline (anonymized for review), 2025a.

Anonymous. Cleaned blog authorship corpus (anonymized for review), 2025b.

Shijie Chen, Yu Zhang, and Qiang Yang. Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32, 2024.

Andrea Cimino, Felice Dell'Orletta, and Malvina Nissim. Tag-it @ evalita2020: Overview of the topic, age, and gender prediction task for italian. pp. 243–251, 01 2020. ISBN 9791280136329. doi: 10.4000/books.aaccademia.7262.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL `https://arxiv.org/abs/2006.03654`.

Zhile Jiang, Shuai Yu, Qiang Qu, Min Yang, Junyu Luo, and Juncheng Liu. Multi-task learning for author profiling with hierarchical features. In *WWW '18: Companion Proceedings of the The Web Conference 2018*, pp. 55–56, 04 2018. ISBN 9781450356404. doi: 10.1145/3184558.3186926.

Atharva Kulkarni, Lucio M. Dery, Amrith Setlur, Aditi Raghunathan, Ameet Talwalkar, and Graham Neubig. Multitask learning can improve worst-group outcomes. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=sPlhAIp6mk`.

Indira Lanza-Cruz, Rafael Berlanga, and María José Aramburu. Multidimensional author profiling for social business intelligence. *Information Systems Frontiers*, 26(1):195–215, February 2023. doi: 10.1007/s10796-023-10370-0. URL `https://doi.org/10.1007/s10796-023-10370-0`.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL `https://aclanthology.org/2020.acl-main.703/`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL `https://arxiv.org/abs/1907.11692`.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. Author profiling for hate speech detection, 2019. URL `https://arxiv.org/abs/1902.06734`.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593, September 2016. doi: 10.1162/COLI_a_00258. URL `https://aclanthology.org/J16-3007/`.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination, 2011. URL `https://arxiv.org/abs/1107.4557`.

Hyunji Park, Yogarshi Vyas, and Kashif Shah. Efficient classification of long documents using transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 702–709, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.79. URL `https://aclanthology.org/2022.acl-short.79/`.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Pan16 author profiling, September 2016. URL `https://doi.org/10.5281/zenodo.3745963`.

Francisco Rangel, Paolo Rosso, Manuel Montes y Gómez, Martin Potthast, and Benno Stein. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier (eds.), *Working Notes Papers of the CLEF 2018 Evaluation Labs*, volume 2125 of *CEUR Workshop Proceedings*, September 2018. URL `https://ceur-ws.org/Vol-2125/invited_paper_15.pdf`.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. Effects of age and gender on blogging., 01 2006a.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 199–205, 01 2006b.

Rachael Tatman. Blog authorship corpus, 2017. Dataset available at `https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus`.

Vishesh Thakur and Aneesh Tickoo. Text2gender: A deep learning architecture for analysis of blogger's age and gender, 05 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Hend Zouari. French axa insurance word embeddings: Effects of fine-tuning bert and camembert on axa france's data, 2020.