DIEq: Dynamic Identity Equilibrium for Author Disambiguation in KDD Cup 2024 WhoIsWho-IND Challenge

Zhixiang Lu*
University of Liverpool
Liverpool, UK
leodickbig2000@gmail.com

Hansheng Zeng
University of Hong Kong
Hong Kong, China

Yuqi Li
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China

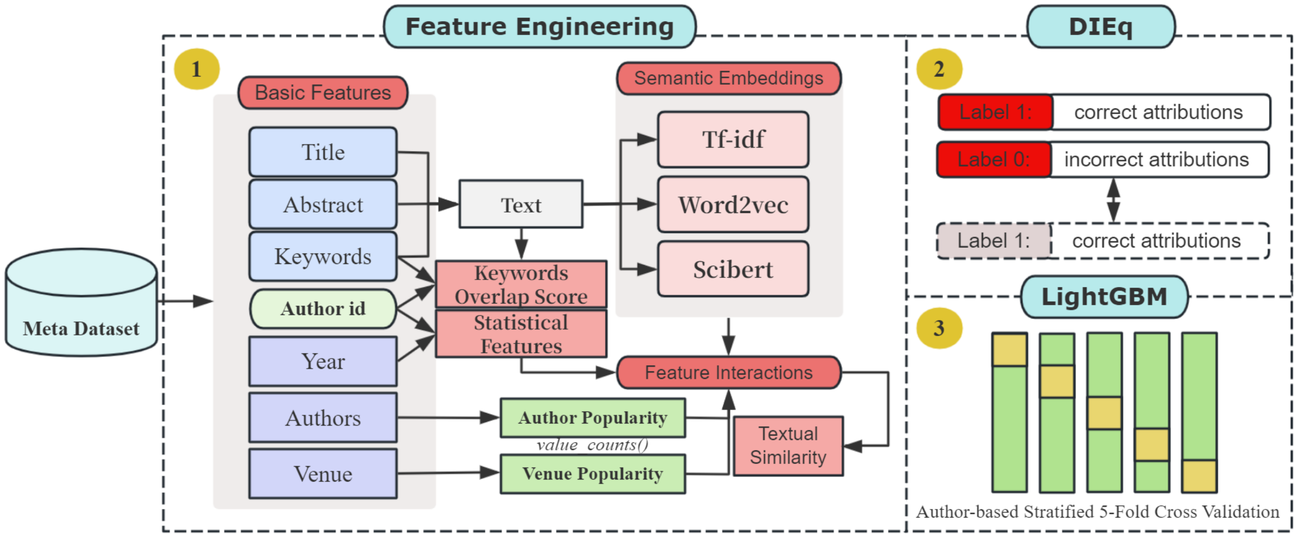

Figure 1: Our Solution Pipeline in WhoIsWho-IND task of KDD Cup 2024

Abstract

We propose Dynamic Identity Equilibrium (DIEq), a novel data preprocessing technique for author disambiguation. DIEq addresses dataset imbalance by simultaneously interpreting a subset of negative samples as both negative and positive, creating 'academic identity phantoms' that enrich the feature space. This approach not only exploits data imbalance but also accelerates convergence and enhances model prediction accuracy. Ranking in the top 10 of the WhoIsWho-IND task at KDD Cup 2024, our approach combines over 3,000 hand-crafted features with a 5-fold LGBM model, achieving a 0.5-1.2% increase in wAUC on test data, contributing to more precise academic impact assessment and knowledge discovery.

CCS Concepts

• **Computing methodologies → Information extraction**.

Keywords

Author Disambiguation, Data Imbalance Mitigation, Feature Engineering, Adversarial Learning, Text Representation

ACM Reference Format:
Zhixiang Lu, Hansheng Zeng, and Yuqi Li. 2024. DIEq: Dynamic Identity Equilibrium for Author Disambiguation in KDD Cup 2024 WhoIsWho-IND Challenge. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The rapid growth of academic publications has made author disambiguation an increasingly critical challenge in managing academic knowledge graphs [1, 20, 21]. This task is particularly crucial for accurately assessing research impact, facilitating collaboration, and enhancing the overall integrity of scientific databases [3, 13]. However, the inherent imbalance in author attribution datasets, where correctly attributed papers significantly outnumber misattributed ones, poses a substantial challenge to developing effective disambiguation models [12, 19].

Traditional approaches to author disambiguation have relied heavily on metadata analysis and string-matching techniques [9].

[*] First and corresponding author.

KDDCup'24, August 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

More recent methods have incorporated machine learning algorithms, leveraging features extracted from publication metadata, content analysis, and citation networks [22]. Despite these advancements, the problem of data imbalance continues to hinder the performance of many state-of-the-art models, often resulting in high false positive rates or reduced sensitivity to genuine cases of misattribution [16].

In this paper, we introduce Dynamic Identity Equilibrium (DIEq), a novel data preprocessing technique designed to leverage issues in author disambiguation tasks. DIEq operates by simultaneously interpreting a subset of negative samples (misattributed papers) as both negative and positive, effectively creating "academic identity phantoms". This approach strategically exploits dataset imbalance while enriching the feature space, compelling the model to learn more nuanced and discriminative representations of author identities.

Our method builds upon recent advancements in adversarial learning and data augmentation techniques, adapting these concepts to the specific challenges of academic data mining. By combining DIEq with a comprehensive set of over 3,000 hand-crafted features and employing a Light Gradient Boosting Machine (LGBM) model [15], we demonstrate significant improvements in both model convergence speed and prediction accuracy.

We evaluate our approach on the WhoIsWho-IND task of KDD Cup 2024, a large-scale author disambiguation challenge focused on detecting misattributed papers in author profiles. Our experiments show that DIEq not only strategically leverages the data imbalance issue but also leads to a 0.5-1.2% increase in Weighted Area Under the Curve (wAUC) on test data, outperforming several baseline methods.

The rest of this paper is organized as follows: Section 2 provides a comprehensive description of the dataset used in the KDD Cup 2024 WhoIsWho-IND task, along with our initial Exploratory Data Analysis (EDA) findings. Section 3 details our proposed DIEq method and feature engineering approach. Section 4 presents our experimental setup and results. Finally, Section 5 concludes the paper and discusses future directions.

2 WhoIsWho-IND KDD Cup 2024

2.1 Competition Task

The WhoIsWho-IND task of the OAG-Challenge, focuses on detecting incorrectly attributed papers in academic literature. The main task is to develop models to distinguish between correctly and incorrectly assigned papers using a dataset of 148,309 labelled samples, with a significant class imbalance (131,024 correct and 17,285 incorrect attributions). The challenge lies in leveraging text length distributions and additional paper attributes (e.g., title, abstract, authors, keywords) to overcome the substantial overlap between classes. This task aims to enhance author name disambiguation, crucial for improving the accuracy of academic knowledge graphs and supporting various applications in scientific research and policymaking.

2.2 Dataset

The training dataset for this task comprised 148,309 samples, with 131,024 instances of correctly assigned papers (Label 1) and 17,285 instances of incorrectly assigned papers (Label 0) [21], which implies that class imbalance (approximately 7.58:1 ratio) is a crucial factor that should be addressed in the model development strategies.

2.3 Text Length Distribution

The **Figure 3** in the appendix illustrates the distribution of text lengths in the WhoIsWho-IND task dataset from the OAG-Challenge, utilizing density curves to compare the text length distributions of different labels. The vast majority (99.7%) of the samples have text lengths below 4,096 characters, while the primary concentration of text lengths between 0 and 2,000 characters. Incorrectly assigned papers (Label 0) display a pronounced peak near short text length(<128 characters). This could indicate a higher likelihood of misattribution for extremely short texts.

3 Method

3.1 Feature Engineering

We meticulously engineered over 3,000 features, encompassing a wide range of textual, semantic, and author-specific statistical attributes (see **Figure 1**). Our comprehensive feature set includes:

- `Popularity Features`: Authors and venue popularity are derived from frequency distributions, computed using MapReduce to capture citation patterns
- `Semantic Embeddings`: Tf-idf [10], Word2Vec [11]and Scibert based representations [14]
- `Textual Similarity Metrics`: Employing cosine similarity measures to quantify textual relationships
- `Keywords Overlap Score`: Quantify the overlap of keywords in each author's research domain in the corresponding article
- `Author-grouped Statistical Features`: Statistical features to capture author relationships and influence
- `Feature Interactions`: Generating composite features to capture complex relationships between authors' research domains

3.2 Definition of Keywords Overlap Score

For each author, we calculate the top 100 most frequent words in all texts of each author. Let W_{author} represent all texts for the author author, and Count(w, W_{author}) represent the occurrence count of word w in W_{author}. We define the set of top 100 most frequent words for the author author as:

$$T_{\text{author}} = \{w \mid w \in \text{Top100words}(W_{\text{author}})\} \tag{1}$$

For each sample i, we calculate the keywords overlap score between the sample's text y_i and the top 100 keywords set T_{x_i} of the sample's corresponding author x_i. Let y_i represent the text of sample i, and T_{x_i} represent the set of top 100 keywords for the author x_i. The overlap score overlap_score(y_i, T_{x_i}) is defined as:

$$KeywordsOverlapScore(y_i, T_{x_i}) = \frac{\left|\{w \mid w \in y_i \cap T_{x_i}\}\right|}{\left|T_{x_i}\right|} \tag{2}$$

3.3 Embedding Cosine Similarities

Given two embedding matrices, one representing the mean embedding of all articles related to an author and the other representing the embedding of the current sample article, we compute the cosine similarity between these embeddings for each sample.

Let $E = [e_1, e_2, \ldots, e_n]$ be the matrix of embeddings for all articles related to an author, and let $F = [f_1, f_2, \ldots, f_n]$ be the matrix of embeddings for the current sample articles. For each sample i, we calculate the following:

1. Compute the mean embedding of all articles related to the author of the i-th sample:

$$\mu_i = \frac{1}{m_i} \sum_{j=1}^{m_i} e_{ij} \tag{3}$$

where m_i is the number of articles related to the author of the i-th sample.

2. Calculate the difference between the mean embedding μ_i and the embedding of the i-th sample article f_i:

$$d_i = f_i - \mu_i \tag{4}$$

3. Calculate the cosine similarity between the mean embedding μ_i and the embedding of the i-th sample article f_i:

$$CosineSims(d_i, f_i) = \frac{d_i \cdot f_i}{\|d_i\|\|f_i\|} \tag{5}$$

Thus, the process can be summarized as follows:

(1) Convert text columns *Text1*(all related articles) and *Text2*(current articles) to embedding matrices E and F, respectively.
(2) Initialize an empty list *CosineSims*.
(3) For each i from 1 to n:
  (a) Compute the mean embedding μ_i of all articles related to the author of the i-th sample.
  (b) Calculate the difference $d_i = f_i - \mu_i$.
  (c) Compute $CosineSims(d_i, f_i)$.
  (d) Append the result to *CosineSims*.
(4) Return the list *CosineSims*.

3.4 Dynamic Identity Equilibrium

Author disambiguation is a critical task in bibliometrics and digital libraries, where accurately identifying the authors of academic papers is essential. Traditional methods often struggle with dataset imbalance, leading to suboptimal performance. In addressing the challenges of author disambiguation, we introduce Dynamic Identity Equilibrium (DIEq), an innovative data preprocessing technique inspired by concepts from quantum mechanics and adversarial learning that revolutionizes the approach to dataset imbalance. DIEq operates on the principle of quantum superposition in the academic identity space, allowing a strategic subset of negative samples to exist simultaneously as both negative and positive entities.

Phantom Identity Generation: In this task, DIEq transforms all negative samples into positive ones, inspired by quantum superposition [17]. The transformation is governed by:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{6}$$

where $|0\rangle$ and $|1\rangle$ represent negative and positive states respectively, and $|\alpha|^2 + |\beta|^2 = 1$. In our case, $\beta = 1$ and $\alpha = 0$ for all samples, analogous to a quantum measurement collapsing the superposition [23]. The phantom strength P(x) is defined as:

$$P(x) = \sigma(w \cdot f(x) + b) \tag{7}$$

where σ is the sigmoid function, w is a learned weight vector, and $f(x)$ is a non-linear transformation. P(x) modulates the impact of each transformed sample on the model, similar to the amplitude amplification in Grover's algorithm [8]. This approach strategically leverages class imbalance while preserving information density, akin to quantum error correction techniques [7]. The resulting enriched feature space facilitates more nuanced author disambiguation, as demonstrated in recent work on quantum-inspired machine learning [2, 23].

Adversarial Learning: In adversarial learning, perturbations are introduced to input samples to challenge and improve model robustness [6]. Similarly, DIEq creates 'adversarial' samples by transforming negative instances into positive ones, challenging the model to learn more nuanced decision boundaries. This process is analogous to the generator in Generative Adversarial Networks (GANs) [5], where the generator creates samples to deceive the discriminator.

The phantom strength P(x) in DIEq serves a role similar to the discriminator in GANs, modulating the impact of transformed samples. This creates a dynamic equilibrium between the original data distribution and the transformed one, reminiscent of the Nash equilibrium sought in adversarial training [18].

Furthermore, the quantum-inspired superposition in DIEq aligns with recent work on quantum adversarial learning [4], where quantum states are used to represent and manipulate adversarial examples. This quantum perspective offers a unique approach to enhancing model generalization in author disambiguation tasks. By incorporating these adversarial learning concepts, DIEq utilized class imbalance and potentially improved the model's robustness and generalization capabilities in author disambiguation scenarios.

4 Results

By applying DIEq to the WhoIsWho-IND task of KDD Cup 2024, we observed a significant improvement in model performance as shown in **Table 1**. The introduction of academic identity phantoms not only improved model performance but also facilitated faster convergence during training. Our method achieved a remarkable 0.5-1.2% increase in wAUC on test data, pushing the boundaries of author disambiguation accuracy.

Table 1: Model Performance Comparison

Model*	wAUC†	TT (s)‡
LGBM+Basic Features	0.612	22.4
LGBM+Scibert Embeddings	0.654	37.9
LGBM+Basic Features+Scibert Embeddings	0.679	42.1
LGBM+Feature Engineering	0.715	334.8
LGBM+Feature Engineering+DIEq	0.722	67.5

* Model: **Single-Fold** with Author-based Stratified Sampling
† wAUC: Weighted Area Under the Curve on **Validation Leaderboard**
‡ TT (s): Training Time (seconds)

Our DIEq method demonstrated significant improvements in model performance on the WhoIsWho-IND task of KDD Cup 2024. The LGBM+Feature Engineering+DIEq model achieved a wAUC of 0.722, representing a 0.5-1.2% increase over baseline models on the test set. This improvement, while numerically small, is substantial in the context of author disambiguation tasks. The introduction of academic identity phantoms effectively utilized the inherent class imbalance, a persistent challenge in scholarly databases. Notably, DIEq facilitated faster convergence during training, reducing the computational time from 334.8s to 67.5s compared to feature engineering alone, marking a 79.8% reduction in training time.

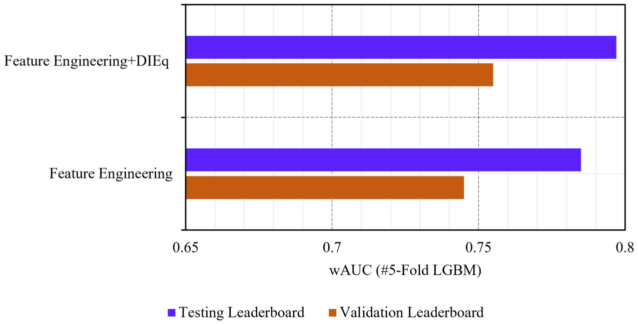

Figure 2: DIEq Performance Comparison

Performance gains were consistent across both validation and testing leaderboards, with **Figure 2** illustrating superior results for Feature Engineering+DIEq on both metrics. The progressive improvement across model configurations, from LGBM+Basic Features (wAUC: 0.612) to our final model, underscores the cumulative benefits of each methodological enhancement. Moreover, the approach demonstrated robustness to potential over-fitting, maintaining its performance advantage in transitioning from validation to testing phases.

5 Conclusion

This study presents DIEq as a novel and effective approach for enhancing author disambiguation in imbalanced academic datasets. By leveraging academic identity phantoms, our method significantly accelerates training convergence, addressing two critical challenges in large-scale bibliometric analysis. The observed 0.5-1.2% improvement in wAUC, while seemingly modest, represents a significant advancement in the highly competitive domain of author disambiguation, where marginal gains often translate to substantial real-world impact. Our approach's consistent performance across validation and testing phases underscores its generalizability and reliability, which are crucial factors for deployment in production environments. The substantial reduction in training time without compromising accuracy highlights the method's efficiency, making it particularly valuable for processing large-scale scholarly databases. These findings contribute to the broader field of entity resolution in academic literature and offer promising directions for addressing imbalanced datasets in related domains. Furthermore, the success of DIEq in this context suggests potential applications in other areas of natural language processing and information retrieval where data imbalance is prevalent. Future work could explore the adaptability of this method to other entity disambiguation tasks and its integration with emerging deep learning architectures for even greater performance gains.

References

[1] Diego R. Amancio, Osvaldo N. Oliveira Jr, and Luciano da F. Costa. 2015. On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks. *Europhysics Letters* 109, 6 (2015), 68001. https://doi.org/10.1209/0295-5075/109/68001
[2] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. 2017. Quantum machine learning. *Nature* 549, 7671 (2017), 195–202.
[3] Ricardo G. Cota, André A. Ferreira, Cristiano Nascimento, Marcos André Gonçalves, and Alberto H. F. Laender. 2017. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology* 68, 3 (2017), 984–999. https://doi.org/10.1002/asi.23736
[4] Pierre-Luc Dallaire-Demers and Nathan Killoran. 2018. Quantum generative adversarial networks. *Physical Review A* 98, 1 (2018), 012324.
[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
[7] Daniel Gottesman. 1997. Stabilizer codes and quantum error correction. *arXiv preprint quant-ph/9705052* (1997).
[8] Lov K Grover. 1996. A fast quantum mechanical algorithm for database search. *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* (1996), 212–219.
[9] Hui Han, Lee Giles, Hongyuan Zha, Cheng Li, and Kostas Tsioutsiouliklis. 2004. Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*. IEEE, 296–305.
[10] Thorsten Joachims and Baroper Str. 1997. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. *Springer US* (1997).
[11] Joseph Lilleberg, Yun Zhu, and Yanqing Zhang. 2015. Support vector machines and Word2vec for text classification with semantic features. In *IEEE International Conference on Cognitive Informatics Cognitive Computing*.
[12] Xiao Liu, Da Yin, Xingjian Zhang, Kai Su, Kan Wu, Hongxia Yang, and Jie Tang. 2021. OAG-BERT: Pre-train Heterogeneous Entity-augmented Academic Language Model. (2021).
[13] Yong Liu, Wenyi Li, Zhiqing Huang, and Qingyong Fang. 2015. A fast method based on multiple clustering for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology* 66, 3 (2015), 634–644.
[14] Polina Lobanova, Pavel Bakhtin, and Yaroslav Sergienko. [n. d.]. Identifying and Visualizing Trends in Science, Technology, and Innovation Using SciBERT. *IEEE Transactions on Engineering Management* PP ([n. d.]).
[15] Qi Meng. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Neural Information Processing Systems*.
[16] Sahar Momeni and Philipp Mayr. 2016. Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names. In *Research and Advanced Technology for Digital Libraries*. Springer, 386–391.
[17] Michael A Nielsen and Isaac L Chuang. 2010. *Quantum computation and quantum information*. Cambridge university press.
[18] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*. 2234–2242.
[19] Anderson F Santana, Marcos A Gonçalves, Alberto HF Laender, and Anderson A Ferreira. 2017. Incremental Author Name Disambiguation by Exploiting Domain-Specific Heuristics. *Journal of the Association for Information Science and Technology* 68, 4 (2017), 931–945.
[20] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, and Zhong Su. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[21] Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, Tianyi Han, Yuwei An, Dan Zhang, Weng Lam Tam, Kun Cao, Yunhe Pang, Xinyu Guan, Huihui Yuan, Jian Song, Xiaoyan Li, Yuxiao Dong, and Jie Tang. 2024. OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
[22] Yutao Zhang, Fan Zhang, Ping Yao, and Jie Tang. 2018. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1002–1011.
[23] Wojciech Hubert Zurek. 2003. Decoherence, einselection, and the quantum origins of the classical. *Reviews of modern physics* 75, 3 (2003), 715.

A Reproducibility

Our solution's code is available at: https://github.com/Leo1998-Lu/KDD2024-WhoIsWho.

We describe in detail how to run the code in the readme files in that git repository. We also describe the execution environment. The top level directory provides instruction on how to run each directory code in which order. Our solution primarily relies on CPU-based computations, utilizing an i9-14900K processor and 128GB of RAM. We did not use any GPU for our model training or inference. Due to the CPU-intensive nature of our approach, the total runtime for feature extraction and model training varied depending on the dataset size and complexity. However, our method generally completed within a reasonable timeframe on the specified hardware. We provide detailed instructions in our repository on how to reproduce our results, including steps for feature generation and model training. Our code is optimized for RAM usage and does not require GPU resources.

B Online Resources

The pre-trained weights for **Scibert** that we used in the competition can be downloaded from the Huggingface Hub at the following link: https://huggingface.co/allenai/scibert_scivocab_uncased.

C Appendix

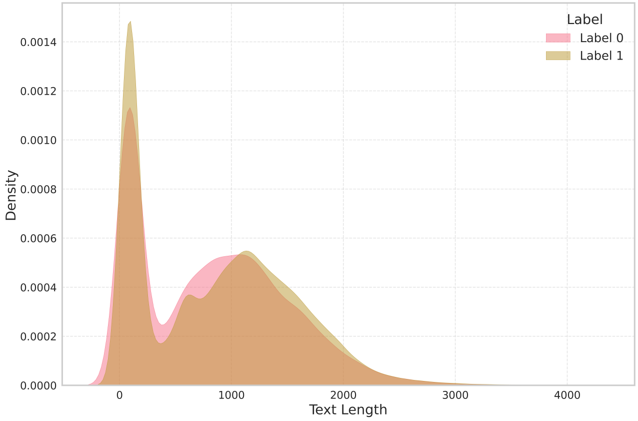

Figure 3: Text Length Distribution of Different Labels

Table 2: Impact of Text Length on Modeling

Model	Text Length	Validation wAUC
LGBM+Scibert Embeddings-64	64	0.611
LGBM+Scibert Embeddings-128	128	0.613
LGBM+Scibert Embeddings-256	256	0.628
LGBM+Scibert Embeddings-512	512	0.634
LGBM+Scibert Embeddings-1024	1024	0.646
LGBM+Scibert Embeddings-4096	4096	0.654