WikiFirst: A Genre-Fixed, Content-controlled Corpus for Evaluating Content Effects in Authorship Analysis

Dung Tuan Nguyen, G. Çağatay Sat, Evgeny Pyshkin, and John Blake
School of Computer Science and Engineering
University of Aizu
Aizuwakamatsu
Japan
s1312004,s1312006,pyshe,jblake@u-aizu.ac.jp

Abstract

This paper presents the design and construction of *WikiFirst*, a corpus for investigating the impact of content variation on authorship similarity under a fixed genre. Prior work has investigated individual authorial style and impact of genre. However, the role of content has remained underexplored due to the lack of suitable data. We address this gap by constructing a Wikipedia-based corpus consisting exclusively of first revisions authored by non-anonymous editors, thereby ensuring high authorship certainty while maintaining a stable encyclopaedic genre.

1 Introduction

Authorship analysis aims to determine whether texts share a common author based on stylistic evidence, yet attribution performance is influenced by multiple interacting factors, notably individual style, genre, and content. While progress has been made in modelling authorial style (Sat et al., 2025) and mitigating genre effects, content variation remains a confounding factor that negatively impacts the accuracy of authorship attribution. A key limitation is the scarcity of corpora that allow systematic comparison of texts written by the same author on different topics while holding genre constant. In this work, we address this limitation by introducing *WikiFirst*, a corpus specifically designed to isolate content effects in authorship analysis.

The design of *WikiFirst* is motivated by the need for a reusable, methodologically transparent resource that enables researchers to explicitly evaluate and control for content variation in authorship analysis. By providing genre-fixed texts authored by the same individuals across multiple, well-defined content domains, *WikiFirst* supports systematic investigation of content effects.

This paper makes two primary contributions. First, we introduce *WikiFirst*, a new corpus for authorship analysis that provides high authorship certainty and controlled content variation under a fixed genre. Second, we propose a principled content taxonomy to categorise Wikipedia articles into eight broad content domains. This corpus allows researchers to investigate the impact of subject variation within a single genre on authorship features.

2 Related Work

Authorship analysis is a research field that spans forensic linguistics (Coulthard and Johnson, 2000; Nini et al., 2023), literary studies (Eder and Górski, 2023), and computational text analysis (e.g., Blake et al., 2025), concerned with identifying stylistic regularities that distinguish authorial style. Alongside methodological advances, a range of benchmark corpora has been developed, including literary texts (Güngör, 2015), emails (Hussain, 2020), forum posts, and social media data. While these resources have enabled steady progress, they frequently conflate multiple sources of variation, making it difficult to isolate the influence of individual factors such as genre and topic.

Genre effects, in particular, have been shown to substantially alter feature distributions, exacerbating the difficulty of accurate authorship attribution (Kestemont et al., 2020, 2021). In contrast, content variation has received comparatively limited direct attention, despite long-standing recognition that topical vocabulary can impact stylistic markers. The absence of corpora explicitly designed to control content while holding genre constant has impeded investigation of this issue.

Heini and Kredens (2024) created the 100 Idiolect corpus comprising different genres of writing written by 112 authors, enabling comparison of authorial style across multiple genres. More recently, (Ma et al., 2025) created the much larger CROSSNEWS corpus, allowing machine learning models to compare differences between genres, and

Proceedings of SIGHUM 2026, pages 323–327

to some extent across content areas. However, the content of 97.5% of the authors is limited to a single domain, making cross-content not feasible. Thus, this research aims to fill that niche by creating a corpus to investigate authorial feature stability and variation across different content domains.

2.1 Authorship Analysis and Confounds

Authorship attribution is influenced by interrelated confounding factors, including genre, topic, register, time period and medium (e.g., spoken vs. written). Among these, there is a growing body of work that shows the difficulty of cross-genre attribution (Kestemont et al., 2012), particularly for smaller datasets found in forensic contexts (Neal et al., 2017).

In addition, content effects are often treated implicitly (arguably by conflating topic with genre) rather than explicitly. Common strategies include removing content-bearing words or emphasizing function words under the assumption that they are less topic-sensitive. Such approaches may ameliorate content bias, but throw little insight into how content variation itself affects authorial stylistic similarity.

2.2 Wikipedia-Based Corpora and Revision Data

Wikipedia has been widely used in natural language processing research due to its scale, accessibility, and rich revision metadata in a variety of areas, including low-resource language training (e.g. Hungarian BERT (huBERT), Nemeskey, 2020) and term-analysis for fields like natural sciences (Wulff, 2023). The encyclopedic writing style provides a relatively stable genre, while the revision history offers detailed information about contributors. However, to the best of our knowledge there is no Wikipedia corpus of first revisions, where authorship is attributable to a single editor. As a result, existing Wikipedia-based datasets are ill-suited for investigating stylistic consistency of authors across content domains.

3 Corpus Design

The design of *WikiFirst* is guided by the goal of isolating content variation as a variable in authorship analysis. To this end, the corpus is constructed from Wikipedia articles written in a single, stable genre, while deliberately varying topic across documents authored by the same individuals. Authorship certainty is prioritised by restricting inclusion to first revisions produced by non-anonymous editors, thereby avoiding ambiguity and potential style transfer introduced by collaborative editing.

By leveraging the structured metadata and revision history in Wikipedia, *WikiFirst* balances scale, control, and reproducibility. The corpus is intended to support both methodological analysis and empirical evaluation, enabling researchers to assess how content variation affects stylistic similarity independently of genre effects.

Four core design principles underpin the construction of *WikiFirst*. First, authorship certainty is ensured by selecting only the first revision authored by the page creators who are registered editors, providing a clear link between text and author. Second, genre control is achieved by limiting the corpus to encyclopaedic articles. Third, topic variation is introduced systematically by identifying prolific contributors who have created articles across multiple, well-defined content domains, enabling controlled comparison across topics. Finally, reproducibility is facilitated through transparent selection criteria and reliance on publicly available data, allowing the corpus to be reconstructed or extended in future work.

To operationalise the design principles of authorship certainty, genre control, topic variation, and reproducibility, the corpus was constructed using the following five inclusion criteria: (1) editors must be non-anonymous and (2) human who (3) have created initial articles in (4) five or more distinct content domains (5) prior to the OpenAI release of ChatGPT on 30 November 2022. Accounts identified as bots or otherwise automated were excluded.

4 Content Taxonomy

For authorship analysis, an effective taxonomy should capture content differences that plausibly affect lexical and discourse patterns, while remaining analytically simple. Content domains are treated as mutually exclusive categories, with each article assigned to a single domain to avoid ambiguity and facilitate controlled pairwise and cross-domain comparisons.

Content is divided into two main categories, namely (1) sciences and (2) humanities and cultural domains, each of which may be represented by 2×2 matrices, giving a total of 8 subcategories. Sciences may be divided broadly in two dimensions, namely (1) based on the degree to which laws govern the subject, and (2) whether the fo-

cus is more theoretical or practical. This may be represented as a 2×2 matrix as shown in Table 1

	Pure	Applied
Hard	e.g. Physics	e.g. Economics
Soft	e.g. Computer science	e.g. Business

Table 1: Two-dimensional taxonomy of scientific content domains (adapted from Becher and Trowler (2001)).

For the humanities and cultural domains, a complementary low-dimensional analytical taxonomy is adopted drawing on established distinctions between descriptive and normative discourse, and between institutional and cultural forms of knowledge. This taxonomy shown in Table 2 captures the broad differences in communicative purpose and cultural embedding while remaining analytically tractable.

	Descriptive	Normative
Institutional	e.g. History	e.g. Politics
Cultural	e.g. Biography	e.g. Religion

Table 2: Two-dimensional taxonomy of humanities and cultural content domains

Wikipedia uses a category graph rather than a strict classification hierarchy, allowing overlapping categories that evolve through community editing, but this flexibility results in cycles, redundancy, and inconsistent granularity. The thirteen Wikipedia categories are, therefore, mapped to the eight target content domains.

To ensure sufficient topic coverage per author, a balancing strategy is adopted that trades off taxonomic granularity against author availability. This is achieved by subdividing selected high-level categories into finer-grained domains where necessary, increasing domain diversity without compromising the taxonomic structure.

5 Data Collection and Extraction

Potential target authors were identified from lists of prolific Wikipedia contributors (Wikipedia contributors, 2025b).

A tailormade data collection script (see Algorithm 1 for pseudocode) was created that retrieves the initial revision of newly created pages and extracts the author identifier, revision timestamp, page title, and all textual content entered by the editor.

Require: User CSV file F_{in}, Output Directory D_{out}, Batch Size $B \leftarrow 50$
Ensure: JSON files containing classified text and token counts for each user

Metadata processing
$Users \leftarrow$ ReadCSV(F_{in})
for all $user \in Users$ **do**
  $Creations \leftarrow \emptyset$
  $SeenTitles \leftarrow \emptyset$
  $AllArticles \leftarrow$ GetContribs($user$)
  for all $article \in AllArticles$ **do**
    if $article.title \notin SeenTitles$ **then**
      $SeenTitles \leftarrow SeenTitles \cup \{article.title\}$
      $Creations \leftarrow Creations \cup \{article\}$
  SaveJSON($Creations, D_{out} + "/" + user$)
Content fetching and domain classification
for all $file \in$ ListFiles(D_{out}) **do**
  $Entries \leftarrow$ LoadJSON($file$)
  $Pending \leftarrow Entries$
  $Batches \leftarrow$ ChunkList($Pending, B$)
  for all $batch \in Batches$ **do**
    $RevIDs \leftarrow \{e.revid \mid e \in batch\}$
    $Titles \leftarrow \{e.title \mid e \in batch\}$
    $TextMap \leftarrow$ FetchText($RevIDs$)
    $CatMap \leftarrow$ FetchCat($Titles$)
    for all $e \in batch$ **do**
      $T \leftarrow TextMap[e.revid]$
      if "#REDIRECT" in T **then**
        $e.dom \leftarrow$ "System/Redirect"
        $e.numToken \leftarrow 0$
        $e.text \leftarrow$ ""
      else
        $e.text \leftarrow T$
        $e.numToken \leftarrow$ length(T)
        $Cat \leftarrow CatMap[e.title]$
        if $Cat \neq \emptyset$ **then**
          $e.dom \leftarrow getDom(Cat)$
        else
          $e.dom \leftarrow$ "NA"
    SaveJSON($Entries, file$)
    Sleep(1.0) ▷ Avoid 429 error code

Algorithm 1: Wikipedia User Contribution Fetch Pipeline

The original content fetched from the Wikipedia API is in Wikitext format, a markup language that intersperses content with formatting syntax. The text was cleaned and preprocessed using the WikiText cleaning script (Pryzant, n.d.). This script utilizes the `mwparserfromhell` library to iteratively strip Wikitext markup, converting it into plain text. A multi-pass approach is employed, parsing and stripping the code three times to handle complex nested structures. Additionally, custom regular expressions are used to aggressively remove citations, HTML artifacts, URLs, table formatting characters, and non-ASCII characters. Finally, after all filtration is done, authors who had less then 10 documents are removed

from the corpus. The code used can be found at https://github.com/himynameiszim/wikifirst.

6 Corpus Statistics

Statistical analysis of the corpus was done using R programming language (R Core Team, 2025), Quanteda (Benoit et al., 2018) package, and Tidyverse (Wickham et al., 2019) library.

WikiFirst consists of 100 unique authors contributing a total of 226,098 documents across 13 categories. The corpus contains approximately 42,710,363 tokens in total, with a mean document length of 188.902 tokens (SD = 641.68). Each author contributes an average of 2260.98 documents, spanning a mean of 12.16 categories per author.

Table 3 provides a high-level overview of the corpus composition.

Item	Number
Number of authors	100
Number of documents	226,098
Number of content domains	13
Mean domains per author	12.16
Mean documents per author	2260.98
Total word tokens	42,710,363
Mean tokens per doc	188.902
Standard dev. (tokens per doc.)	641.68

Table 3: Summary statistics for *WikiFirst*

Table 4 shows the number of texts in each of the Wikipedia categories.

Category	Texts
Culture and the arts	7,735
General reference	1,810
Geography and places	9,531
Health and fitness	564
History and events	2,798
Human activities	15,524
Mathematics and logic	9,639
Natural and physical sciences	1,991
People and self	30,888
Philosophy and thinking	8,392
Religion and belief systems	2,161
Society and social sciences	27,254
Technology and applied sciences	107,811

Table 4: Document distribution across Wikipedia content categories

The Wikipedia categories are highly imbalanced with only four out of 13 categories comprising over 10,000 texts, and four containing fewer than 2,000. These categories were mapped to the proposed context taxonomy, resulting in the distribution shown in Table 5.

Category	Texts
Pure hard science	11,630
Applied hard science	108,375
Pure soft science	11,341
Applied soft science	15,524
Descriptive institutional content	2,798
Normative institutional content	27,254
Descriptive cultural content	38,623
Normative cultural content	10,553

Table 5: Mapped document distribution categories

Overall, 73% of all authors have at least 10 documents in 7 out of 8 domains, allowing the corpus to be used in a robust, content-focused model training environment. The proposed eight-domain taxonomy produces a substantially more balanced distribution, with every domain containing at least 2,700 texts and seven of the eight domains exceeding 10,000 texts. This improved balance supports more reliable cross-domain comparison and reduces the risk that observed differences in authorship similarity are driven by data sparsity rather than genuine content effects.

7 Conclusion

In conclusion, our dataset contains more than 200,000 documents across 8 content domains, which is sufficiently diverse for content modeling in authorship attribution tasks. It should be noted that our dataset is subject to the licensing terms of Wikipedia, namely Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA).

As the information contained is crowd-sourced, its truth value and acceptability is dependent on the contributing editors (Wikipedia contributors, 2025a). However, as this dataset is created for authorship attribution model training purposes, authorship style and not the objective truth is paramount. The dataset can be accessed from doi:10.34740/kaggle/dsv/14797130.

Limitations

A primary limitation of the final corpus is the authorial signature is affected by the editorial rules of Wikipedia. The adherence to these relatively strict genre rules, however, enhances the generic integrity of the corpus. In addition, all the authors in the dataset are top contributors in Wikipedia, making the texts obtained closer to the norm of Wikipedia, lowering authorial diversity.

Finally, the corpus consists of articles that were written since the creation of Wikipedia, which may

introduce minor temporal variation in terminology or individual writing habits. However, this variation is relatively insignificant compared to the timescales typically associated with substantial language change, and all authors have texts ranging from years ago reducing the risk of time-related bias in the dataset.

References

Tony Becher and Paul Trowler. 2001. *Academic tribes and territories*. McGraw-Hill Education.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

John Blake, Abu Saleh Musa Miah, Krzysztof Kredens, and Jungpil Shin. 2025. Detection of AI-Generated texts: A Bi-LSTM and attention-based approach. *IEEE Access*, 13:71563–71576.

Malcolm Coulthard and Alison Johnson. 2000. *Forensic Linguistics: An Introduction to Language in the Justice System*. Routledge.

Maciej Eder and Rafał L Górski. 2023. Stylistic fingerprints, pos-tags, and inflected languages: A case study in polish. *Journal of Quantitative Linguistics*, 30(1):86–103.

Abdulmecit Güngör. 2015. Benchmarking authorship attribution techniques using over a thousand books by fifty victorian era novelists. Master's thesis, Purdue University, West Lafayette, IN, USA.

Annina Heini and Krzysztof Kredens. 2024. Remote data collection in sociolinguistics: lessons from the COVID-19 pandemic. *International Journal of Social Research Methodology*, 27(6):747–759.

Javed Hussain. 2020. Enron email dataset.

Mike Kestemont, Kim Luyckx, Walter Daelemans, and Thomas Crombez. 2012. Cross-genre authorship verification using unmasking. *English Studies*, 93:153–160.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Janek Bevendorff, Martin Potthast, and Benno Stein. 2021. Overview of the cross-domain authorship verification task at PAN 2021. In *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*.

Marcus Ma, Duong Minh Le, Junmo Kang, Yao Dou, John Cadigan, Dayne Freitag, Alan Ritter, and Wei Xu. 2025. CROSSNEWS: A cross-genre authorship verification and attribution benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24777–24785.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36.

Dávid Márk Nemeskey. 2020. *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Budapest University of Technology and Economics.

Andrea Nini, Peter Burnap, Matthew L. Williams, and Kevin Knight. 2023. Register variation in malicious forensic texts: An exploratory analysis. In *Proceedings of the Corpus Linguistics Conference*.

Reid Pryzant. n.d. wiki text cleaner. `https://gist.github.com/rpryzant/561cc1b4d372cce7479fd14290eacbc3`. GitHub Gist. Accessed: 2026-01-06.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

G. Çağatay Sat, John Blake, and Evgeny Pyshkin. 2025. Modelling the relative contributions of stylistic features in forensic authorship attribution. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 1066–1073, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wikipedia contributors. 2025a. Reliability of Wikipedia. [Online; accessed 5-Jan-2026].

Wikipedia contributors. 2025b. Wikipedia: List of Wikipedians by number of edits. [Online; accessed 27-Dec-2025].

Peter Wulff. 2023. Network analysis of terms in the natural sciences insights from wikipedia through natural language processing and network analysis. *Education and Information Technologies*, 28(11):14325–14346.