A Multi-dimensional Constrained Framework for Derived Text Formats

Kei Du, Christof Schöch

University of Trier
Universitätsring 15, 54296 Trier
{duk, schoech}@uni-trier.de

Abstract

Derived Text Formats (DTFs) have been proposed as a solution to enable text and data mining while avoiding copyright infringement. Building on a review of recent empirical studies of DTFs on topic modeling, authorship classification, and sentiment analysis, this paper argues that DTFs should not be treated as static formats, but as variable and task-dependent representations shaped by multiple interacting factors. In response, we propose a multi-dimensional framework that conceptualizes DTFs as configurations within a structured space defined by both internal representation parameters and external constraints. The framework includes four internal representation dimensions—feature level, degree of reduction, transformation strategy, and aggregation level—as well as two external constraining forces: legal requirements and task-specific information needs. By emphasizing the interdependence of these dimensions, the proposed framework provides a systematic way to describe, compare, and design DTFs across different analytical contexts. Therefore, this paper contributes to a more theoretically grounded understanding of DTFs and offers guidance for their responsible and effective use in text and data mining in Digital Humanities.

Keywords: derived text formats, token-based DTFs, framework

1. Introduction

The practice of transforming texts into derived representations is well established in computational text analysis. However, the notion of derived text formats (DTFs), also known as extracted features, extends this practice by embedding it within a legal and infrastructural context, thereby shifting the focus from purely methodological considerations to the interplay between analytical utility and copyright compliance (Jett et al. 2020, Schöch et al. 2020). In recent years, driven by the principles of open science, an increasing number of researchers have been making their research data publicly available alongside their publications to ensure the reproducibility and falsifiability of their work. As a result, DTFs are getting increasing attention due to the increasing needs of the storage, publication, and reuse of research data built from in-copyright texts.

The definition of DTFs is constrained by conflicting requirements arising from the tension between legal compliance and text analysis. Legal requirements call for the reduction of textual information, while methodological needs rely on retaining such information; the tension between these two has not yet been systematically conceptualized or explained. Therefore, this paper aims to explore how to systematically describe and design DTFs within this context. In the following, we first explore the previous studies on DTFs in order to understand how different DTFs were applied in text and data mining. Then we share our observations and reflections on the use of DTFs and propose a multi-dimensional constrained framework to support the use of DTFs.

2. Previous evaluations of DTFs on text and data mining tasks

In recent years, a series of studies have evaluated DTFs' performance across various text and data mining tasks such as topic modeling, authorship classification and sentiment analysis.

Kocula (2022) systematically evaluated the performance of the Latent Dirichlet Allocation (LDA) algorithm across three DTFs: Term-Document Matrices (TDM), Segment-wise Abolished Sequence information (SAS), and Selectively Reduced Token information (TKN). This study conducted experiments using a corpus of 19th and 20th-century English novels and employed topic coherence (via the Palmetto library) and statistical significance tests to measure the quality of the generated topics. The results demonstrate that DTFs, particularly SAS and TKN, achieve topic coherence scores that are remarkably similar—and in some instances superior—to those of the original full text, suggesting that DTFs provide a robust, scientifically valid, and legally safe method for distributing research data.

Du (2023) investigated the impact of DTFs on authorship classification, with a particular focus on how information loss affects performance in stylometric analysis. The study examined three types of token-based DTFs, especially focusing on selectively replacing words in texts with their corresponding part-of-speech (POS) tags. To evaluate their effectiveness, experiments have been conducted using three corpora in three different languages. The experiment involves gradually replacing a certain proportion of the vocabulary in each text (from 0% to 100%) and measuring the resulting author classification

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 16–19
12 May 2026.

performance. The results show that moderate levels of information loss (replacing or removing up to 40% of words in texts) have relatively little impact on classification accuracy. Interestingly, replacing words with POS tags does not lead to better results than simply removing them, suggesting that POS information contributes little to authorship discrimination in this context.

Du & Schöch (2024) investigated how much information loss a BERT-based model can tolerate before its sentiment classification performance significantly degrades. The authors conducted experiments using both non-literary and literary datasets and tested two DTFs: DTF-1 (randomized word order) and DTF-2 (POS Replacement). The results show that BERT is remarkably robust at picking up semantic "signals" despite the loss of syntax in scrambled text. Also, when 40% of words are replaced by their corresponding POS tags, the text becomes almost impossible for a human to read or recognize (satisfying legal safety), yet the sentiment classification accuracy remains nearly identical to the original text.

3. Reflections on the use of DTFs

By examining the decisions regarding the selection and use of DTFs in the studies mentioned above, we can identify the following characteristics of DTFs:

First, DTFs are not static; instead, they are flexible and adaptable text representations designed for different text and data mining tasks. Different tasks may require different DTFs. For example, authorship attribution only needs lexical information in text, while sequential information is necessary for training language models. When defining a DTF for a text and data mining task, a balance must be reached between, on the one hand, the legal regulations concerning copyright protection and the sharing of research data, and, on the other hand, the textual information required for the specific task. In other words, a DTF must establish trade-offs between preservation and deletion of textual information, that is between text recognizability and reconstructability, on the one hand, and analytical performance on the other.

Second, when determining a DTF, it is necessary not only to decide which types of textual information to retain or to delete, but also to determine the degree to which they should be retained or deleted. For example, we could randomly reorder 50% of the words in each sentence within a text, or replace 20% of the content words in the entire text with their corresponding POS tags. Of course, the extent to which such modifications or transformations are applied depends on the trade-offs mentioned earlier.

Third, when different DTFs are applied to the same task, if the results based on Format A are better than those based on Format B, this only indicates that Format A is more suitable for that task; it does not mean that Format A is a better DTF than Format B. This is because, during the process of transforming the same text into different DTFs, different pieces of information are filtered out of the text. For example, a document-term matrix keeps word frequency while discarding word order and syntactic relations, while a POS-based representation removes lexical content but retains aspects of grammatical structure. These transformations change the feature space on which computational models operate, as well as the statistical distribution and internal organization of the data. As a result, applying different DTFs to the same task is not simply a matter of using “more” or “less” text data for text mining; rather, it involves using different data to solve the same task. Performance differences cannot be simply attributed to methodological superiority, as these differences may reflect variations in how the underlying data is structured. Therefore, any evaluation of DTFs must consider the transformations and the information loss they introduce, rather than assuming that all derived formats are comparable simplifications of the same text.

Finally, a one-size-fits-all DTF is highly unlikely to exist, as different text and data mining tasks rely on different types of textual information. To establish a DTF capable of supporting as many tasks as possible, it is necessary to balance various types of textual information rather than optimizing solely for a single objective. This means the format should preserve diverse textual information—such as lexical, syntactic, and sequence features—while allowing for controlled information reduction. Such an attempt to address every aspect is likely to result in suboptimal performance across all tasks. Therefore, rather than pursuing a single universal format, it makes more sense to define DTF more flexibly to suit different analytical needs. However, if the same text is converted into different DTFs and all of them are made publicly available as research data, this increases the risk that the original text could be reconstructed using NLP technologies (such as large language models).

4. A multi-dimensional constrained framework of DTFs

Based on the above analysis, we believe it is necessary to establish a framework for describing DTFs so that users can take every important aspect into account when using DTFs. We consider DTFs to be the result of textual transformations shaped by mutually constraining conditions and introduce our multi-dimensional framework for DTFs.

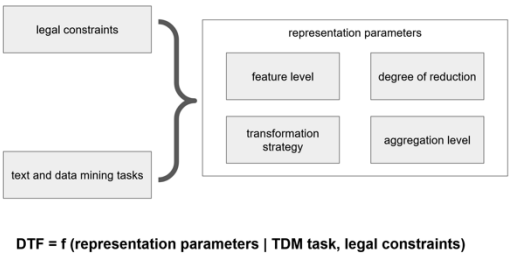

Figure 1. DTF as a multi-dimensional constrained framework

As presented in Figure 1, the graphical representation of the framework is at the top and the formula at the bottom defines DTF as a functional configuration of internal representation parameters that is dynamically shaped and restricted by two primary external forces: the legal constraints and the text and data mining tasks. The external forces serve as boundary conditions that restrict and guide the configuration of the representation parameters. The internal representation parameters have four components: feature level, degree of reduction, transformation strategy and aggregation level. Different internal parameters lead to different DTFs, each reflecting a specific balance between analytical utility and copyright compliance. Importantly, these dimensions are not independent of one another: choices made in one dimension affect the range of possible configurations in others. The following is a detailed explanation of each dimension.

- **Legal constraints**: This dimension covers the legal requirements regulating the legitimacy of DTFs, particularly those related to recognizability, reconstructability, and copyright compliance. It specifies the extent to which the original text content can be retained or reconstructed.
- **Text and data mining tasks**: This dimension refers to the specific information requirements of particular text and data mining tasks, such as authorship attribution, topic modeling, text re-use, or sentiment analysis. It determines which textual features must be kept ensuring the validity of the analysis.
- **Feature level**: This dimension refers to the selection of linguistic features kept in the DTFs, such as word forms, lemmas, POS tags, or semantic and syntactic relationships. It determines which textual information are available for analysis.
- **Degree of reduction**: This dimension quantifies the proportion of information that must be deleted or transformed in the original text. It defines the overall degree of transformation, ranging from minor modifications to total loss of information.
- **Transformation strategy**: This dimension describes methods for keeping, replacing, or deleting textual elements, such as random substitution, POS-based filtering, or building static and contextual embeddings. It determines how information loss is distributed within each text in a corpus.
- **Aggregation level**: This dimension indicates the structural level at which DTFs are constructed, ranging from token-level and sentence-level transformations to document-level or corpus-level representations. It defines how textual information is organized and interpreted.

5. Limitations

While the proposed multi-dimensional framework provides a systematic approach to defining and designing DTFs, several limitations remain.

First, although legal requirements are identified as a core constraint, the framework does not yet account for the specific statutory variations across different international jurisdictions, such as the differences between European copyright exceptions and US Fair Use.

Second, the current framework lacks quantitative metrics for measuring reconstructability and recognizability; while it defines the degree of simplification from a qualitative perspective, it does not provide a technical threshold to ensure irreversibility when confronting large language models.

Finally, the empirical evidence supporting this framework is primarily based on traditional token-based analysis, and further research is needed to validate its applicability to more abstract representations, such as high-dimensional embeddings or large-scale generative AI workflows.

6. Conclusion

This paper suggests that derived text formats (DTFs) should not be understood as static representations, but rather as flexible yet constrained configurations shaped by both text analytical and legal considerations. By examining the existing empirical research across various text and data mining tasks, we argue that differences in DTFs design fundamentally alter the structure of the underlying data. Consequently, performance differences cannot be explained in isolation from the transformation processes that generate them.

To this end, we propose a multidimensional framework that conceptualizes DTFs as contextual configurations. By distinguishing

between internal representation parameters and external constraints, this framework provides a systematic approach for defining, describing and comparing DTFs. It also emphasizes the dynamic nature of DTF construction, viewing it as a process shaped by competing factors rather than a static choice of format.

This framework provides a more robust theoretical foundation for understanding DTFs and highlights the trade-offs involved in their application. Future research could further refine this framework, explore its applicability to a broader range of text and data mining tasks, and investigate the risks regarding publishing different DTFs of the same text, particularly considering the advances in NLP technology and their potential implications for text reconstruction.

7. Acknowledgments

This work was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

Author contributions:

Keli Du: Conceptualization, Methodology, Investigation, Visualization, Writing - original draft, Writing - review & editing.

Christof Schöch: Funding acquisition; Supervision, Writing - review & editing.

8. Bibliographical References

Du, K. (2023). Understanding the impact of three derived text formats on authorship classification with Delta. DHd 2023 Open Humanities Open Culture. 9. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2023), Trier, Luxemburg. https://doi.org/10.5281/zenodo.7715299.

Du, K., & Schöch, C. (2024). Shifting Sentiments? What happens to BERT-based Sentiment Classification when derived text formats are used for fine-tuning. Digital Humanities Conference 2024 (DH2024), Washington, DC. https://doi.org/10.5281/zenodo.18161643.

Jett, Jacob, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, and J. Stephen Downie (2020). The HathiTrust Research Center Extracted Features Dataset (2.0). DOI: http://doi.org/10.13012/R2TE-C227.

Kocula, M. (2021). Volltext vs. abgeleitetes Textformat: Systematische Evaluation der Performanz von Topic Modeling bei unterschiedlichen Textformaten mit Python. https://doi.org/10.5281/zenodo.5552487.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020). “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen”. In: Zeitschrift für digitale Geisteswissenschaften 5. DOI: http://doi.org/10.17175/2020_006.