Originally published in: Barth, Florian/Du, Keli/Calvo Tello, José/Genêt, Philippe/Lendvai, Piroska/Schöch, Christof/Trippel, Thorsten (Eds.): Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) – Paris: ELRA Language Resources Association, 2026. Pp. 34-43. (LREC 2026)

Multi-Label Text Classification of Derived Text Formats with DistilBERT

Jennifer Ecker, Roman Schneider
Leibniz Institute for the German Language
R5 6-13, D-68161 Mannheim
{ecker, schneider}@ids-mannheim.de

Abstract

Derived Text Formats enable the distribution of copyrighted texts by systematically perturbing linguistic information to reduce reconstructibility. However, the extent to which such information loss affects downstream text classification remains unclear. We investigate how controlled perturbations affect learning dynamics in transformer-based classification using two datasets and two strategies: POS-consistent replacement of 30%, 40%, and 50% of tokens, and random word-order shuffling. On Wikipedia data, POS replacement increases loss by 4–9% and reduces micro-F1 by 3–8%, depending on the replacement rate, while shuffling raises loss by 5% and lowers micro-F1 by 4%. Performance degrades monotonically with higher replacement rates, and shuffling yields results between the 30% and 40% conditions, indicating that DistilBERT relies more on lexical semantics than on word order. Experiments on specialist-domain data show the same pattern, demonstrating robustness across domains. To test cross-representation generalization, we train classifiers on both clean and perturbed texts and evaluate them on the respective alternate representation. Models trained on DTF data generalize better to clean text than vice versa, suggesting that perturbation-based training promotes more robust representations. Our findings position DTF as a promising strategy for reproducible, legally compliant, and robust NLP research.

Keywords: DTF, Multi-label classification, Text perturbations, Transformer robustness, Learning dynamics

1. Introduction

Reproducibility underpins rigorous empirical NLP research, yet copyright restrictions on large text corpora preclude their public dissemination, compromising scientific transparency and replication. Derived Text Formats (DTF) address this challenge by systematically perturbing texts—through techniques such as word embeddings, Part-of-Speech (POS) tag replacement, or word-order shuffling—to reduce reconstructibility while retaining utility for downstream tasks like classification (Rehm et al., 2007). However, Kugler et al. (2024) demonstrate high reconstructibility of contextualized word embeddings produced by transformer-encoders, confirming that such DTFs enable analysis but still risk copyright violation.

Multi-label classification assigns multiple labels to instances simultaneously, prevalent in hierarchical taxonomies like Wikipedia main topic categories (Tsoumakas and Katakis, 2007; Tarekegn et al., 2021). Early approaches relied on rule-based category matching, while recent methods fine-tune BERT (Bidirectional Encoder Representations from Transformers) variants on imbalanced datasets (Sanh et al., 2019). Perturbation robustness studies reveal BERT's vulnerability to adversarial attacks (e.g., TextFooler synonym swaps) and shuffling, underscoring lexical-semantic dependence over syntax (Jin et al., 2019; Hauser et al., 2021; Reimers and Gurevych, 2019). DTF-specific work has enabled privacy-preserving analysis in computational literary studies, yet – to our knowledge – no systematic evaluation exists of transformer-based text classification performance under controlled DTF degradation.

This study aims to fill that gap. We investigate the impact of selected DTF on transformer-based language models for large-scale classification tasks. Our experiments draw on two complementary corpus sources:

1. Wikipedia articles: This text type represents general-domain language use and provides a heterogeneous and widely used benchmark corpus. Its open-data status enables the parallel publication of original and derived versions, allowing explicit quantification of the textual alterations introduced by DTF.

2. Specialized scientific texts: Although likewise publicly accessible, these texts differ substantially in register and structure. They are domain-specific, terminologically dense, and frequently include illustrative example sentences (i.e., object-language material). Their fine-grained linguistic annotations impose distinct challenges for classification models.

The combination of these two corpora seems methodologically advantageous: Together, they facilitate a systematic assessment of the robustness and transferability of multi-label classifiers across registers and levels of annotation granularity, supporting the broader goal of transferring models to

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 34–43
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026
Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-138577

proprietary corpora.

This paper is structured as follows: Section 2 describes the data and methodology, Section 3 presents experimental results on robustness and train–test mismatch, and Section 4 discusses implications for privacy-preserving NLP.

2. Classification task

DTFs have already been tested for author classification (Du, 2023) and sentiment classification (Du and Schöch, 2024). In this work, we extend their use to a multi-label text classification setting. It is particularly interesting to see what effects different perturbations have when training this type of text classifier. In general, difficulties arise from imbalanced class distributions, whereby some classes are overrepresented, while others contain fewer examples. Furthermore, assigning multiple labels to one text expands the problem, creating many possibilities for combining the classes.

For the Wikipedia texts, we assign categories based on the Wikipedia main topic classification taxonomy[1] as of October 22, 2024. Because this taxonomy evolves over time, categories may be merged, removed, or newly introduced. In this study, we use the following top-level classes: Academic disciplines, Business, Communication, Concepts, Culture, Economy, Education, Energy, Engineering, Entertainment, Entities, Food and drink, Geography, Government, Health, History, Human behavior, Humanities, Information, Internet, Knowledge, Language, Law, Life, Lists, Mass media, Mathematics, Military, Nature, People, Philosophy, Politics, Religion, Science, Society, Sports, Technology, Time, and Universe.

For the specialist texts, we employ fine-grained categories derived from a dedicated domain-specific ontology (Lang et al., 2018). Further details are provided in the data section below.

2.1. Data

2.1.1. Wikipedia articles

The Wikipedia text data are extracted from XML sources (Kupietz et al., 2019) containing articles included in the German Reference Corpus *DeReKo* (Kupietz et al., 2018). We build the training corpus using a rule-based string-matching method that leverages the category links stored in the XML metadata. Each article contains a reference to its corresponding Wikipedia category in the ‹classCode› element. Using these references, we extract all terms corresponding to the main topic classifications. A lexicon of main categories (e.g., *Geography*) is created, and each lexicon entry is matched against the category names in the XML. Categories in the XML are often deep in the Wikipedia hierarchy (e.g., *Geography of Saxony-Anhalt*), where the parent main category could theoretically be determined by traversing the hierarchy upwards. However, to maintain practical and straightforward labeling, this traversal is not performed, and categories are assigned directly from the XML tag (e.g., *Geography of Saxony-Anhalt* remains as-is).

Using this approach, the resulting Wikipedia dataset comprises 430,767 texts totaling 421,659,916 tokens for classifier training. The assigned categories serve as thematic classes, with each text associated with an average of 1.24 class labels. Measured in characters, document length ranges from 1 to 466,945 characters, with a mean of 4,341 (median: 2,420; standard deviation: 7,765). When measured in words, texts range from 1 to 66,034 words, with a mean of 595 (median: 335; standard deviation: 1060). The observed minimum document length of a single character can be attributed to artifacts introduced during the automated data extraction process (e.g., formatting remnants or parsing inconsistencies). These extremely short texts represent rare outliers, as reflected by the substantially higher median (2,420 characters) and mean (4,341 characters) document lengths. Due to their negligible semantic content, such instances are unlikely to have a meaningful impact on classifier training.

2.1.2. Specialist texts

The specialist texts are linguistics research texts written by expert linguists and published in the online grammar portal *grammis* (Schneider and Lang, 2022). We use 1,649 documents from this source, which are part of COrLiCo (Corpus for the Oral–Literate Continuum) (Schneider, 2026). Each document has been manually annotated with 593 fine-grained thematic categories, averaging 2.03 labels per text. The category distribution is highly skewed, with some frequent labels (e.g., *Wortstellung, Wortbildung, Komposition*) and around 300 singleton labels (e.g., *Adjunktorgruppe, Diktumsgraduierung, Gattungsname*). For comparability with the Wikipedia sub-corpus, we retain only the 39 most frequent linguistic categories for classification.

In terms of length, the *grammis* texts exhibit substantial variation. Measured in characters, document length ranges from 100 to 36,194 characters, with a mean of 2,826 (median: 1,995; standard deviation: 2,956). When measured in words, texts range from 10 to 4,875 words, with a mean of 358 (median: 244; standard deviation: 392). This indicates a highly heterogeneous dataset with a long-tailed distribution of document lengths.

[1] https://en.wikipedia.org/wiki/Category:Main_topic_classifications

We expand this specialist text dataset using controlled paraphrastic augmentation to increase the amount of training data. Each of the 1,649 original texts is paraphrased ten times with Cohere's Command A, a 111B-parameter large language model optimized for multilingual processing of long documents (Cohere, 2025), which we access locally via an Ollama runtime. The model is prompted as a German linguistics expert and rewrites each text to preserve the original content and specialized terminology while varying the surface form and phrasing. To ensure diversity and quality, we discard near-duplicate paraphrases that exceed a cosine similarity threshold of 0.80. Additionally, we manually inspect a random sample of 100 outputs to verify that the paraphrases are grammatical, semantically faithful, and suitable as additional training instances.

The generated texts differ slightly in their length distribution from the original *grammis* texts. In characters, they range from 82 to 159,724, with a mean of 1,996 (median: 1,807; standard deviation: 1,869). In terms of words, lengths vary between 9 and 14,918 words, with a mean of 245 (median: 222; standard deviation: 209). Compared to the original texts, the paraphrases are on average shorter and less variable, although a small number of extreme outliers remain.

This process generates up to 16,490 paraphrase candidates, resulting in a final specialist dataset of 18,139 linguistic texts, totalling 4,628,838 tokens.

A critical point concerns the extensive use of paraphrastic augmentation. While this procedure substantially increases the amount of available training data, it also introduces a strong dependency on model-generated text: roughly 90% of the resulting specialist dataset consists of LLM-produced paraphrases rather than naturally occurring expert texts. This raises the question to what extent observed patterns reflect genuine properties of the underlying corpus as opposed to artefacts of the generative model.

The paraphrases are generated without explicitly specified decoding parameters (e.g., temperature or nucleus sampling), meaning that the degree of variation is not systematically controlled and may depend on implicit model defaults. At the same time, large language models tend to regularize stylistic variation, smooth rare constructions, and favor preferred lexical and syntactic patterns. Even in the absence of deterministic decoding, this can induce subtle distributional shifts, for instance a bias toward more prototypical or "canonical" formulations. Although we filter near-duplicates and manually verify a subset of outputs, such measures cannot fully eliminate these effects.

Consequently, results obtained on this dataset should be interpreted with caution. Performance gains may partly reflect a model's ability to learn and exploit the stylistic and structural regularities of the generating LLM, rather than the full diversity of authentic specialist writing. Future work could address this limitation by (i) explicitly reporting and controlling decoding parameters to improve reproducibility, (ii) comparing against non-augmented baselines, and (iii) systematically analyzing differences between human- and model-generated texts, for example with respect to lexical diversity and syntactic variation.

2.1.3. Generating Derived Text Formats

For both Wikipedia and linguistic specialist texts, we generated multiple DTF versions using the derived text formatter provided by MONAPipe (Dönicke et al., 2022). We systematically generate four DTF variants to probe distinct information bottlenecks:

- **Word replacement**: Replace X% of content words per sentence with their POS tags (X $\in$ {30, 40, 50}), preserving syntactic structure while eliminating lexical-semantic information for substituted tokens.
- **Word-order randomization**: Fully shuffle tokens at document level, eliminating sequential and syntactic relations while preserving the full lexicon.

These perturbations create orthogonal degradation axes: word replacement ablates *semantics* at controlled rates, while word-order randomization ablates *syntax/sequence*, enabling dose-response analysis of DistilBERT's representational reliance on each signal type. Specifically, the graduated POS replacement rates quantify lexical contribution incrementally – each 10% increase in preserved content words should yield measurable F1-micro gains if semantics dominate. Conversely, randomization tests positional encoding utility, hypothesizing lesser degradation since transformer self-attention is permutation-invariant in theory (though task-specific patterns may emerge). This controlled experimental design isolates DistilBERT's feature learning priorities – lexical vs. sequential – while mimicking real-world DTF privacy constraints, providing actionable guidance for copyright-compliant corpus sharing.

2.2. Classifier

We fine-tune a pre-trained DistilBERT model[2] from HuggingFace to perform multi-label classification using PyTorch Lightning. DistilBERT (Sanh et al.,

[2] https://huggingface.co/distilbert/distilbert-base-german-cased

2019) was chosen for its compact size and fast processing speed. It is a knowledge-distilled version of BERT, a transformer-based language model that produces contextualized word representations by considering both left and right context. Through distillation, a smaller "student" model learns to approximate the behavior of a larger pre-trained BERT, retaining much of its semantic and syntactic understanding while being computationally more efficient. This makes DistilBERT particularly well suited for large-scale and resource-sensitive classification tasks.

All models were trained for up to 25 epochs for the Wikipedia data and for up to 40 epochs for the specialist data. For the Wikipedia data, the baseline model and one DTF model were trained five times to quantify residual stochasticity. All experiments used identical data splits and random seeds for NumPy, PyTorch, and CUDA. Although the hyper-parameters had to be tuned separately for each data set to achieve optimal performance, the resulting configurations remained stable across all DTF versions within a given data set. All code will be released in a public repository for the replication of the classification models.

For training, we use Binary Cross-Entropy with Logits Loss, computed independently for each class and averaged across all classes. Each label is thus treated as a separate binary classification task. The loss measures the discrepancy between predicted probabilities and true labels and serves as the differentiable objective optimized during training. We report both training and validation loss to monitor optimization dynamics and generalization. While decreasing training loss indicates successful fitting, divergence between training and validation loss can reveal overfitting or poor generalization.

In multi-label settings with class imbalance, the loss can be dominated by the large number of negative label predictions per instance (Lin et al., 2017). Because it is averaged across all classes, correctly predicting negatives can substantially reduce the overall loss. Consequently, low loss values do not necessarily imply strong performance on positive or minority labels. As noted by Terven et al. (2025), the loss should therefore primarily be interpreted as an optimization objective rather than a comprehensive performance metric.

To evaluate predictive quality, we report the F1-micro score. Unlike the loss, the F1-score operates on discrete predictions and combines precision and recall into a single measure. The micro-averaged variant aggregates decisions across all labels, making it well-suited for imbalanced multi-label scenarios and aligned with our focus on overall predictive behavior. We report both training and validation F1-micro to assess classification quality on seen and unseen data. In contrast, F1-macro assigns equal weight to each class and would mainly capture performance changes in minority labels rather than reflecting overall model behavior.

3. Experiments and Results

We divide our study into two complementary experiments: Experiment A evaluates classifier robustness under controlled text perturbations, using both the Wikipedia and specialist texts corpora to assess domain-general degradation patterns across general-domain and terminologically dense registers. Experiment B investigates representational transfer under train-test mismatch by cross-evaluating models trained on clean vs. DTF-perturbed texts; this analysis is restricted to the Wikipedia corpus, as its larger size and balanced class distribution enable more reliable estimation of generalization gaps, whereas the smaller specialist corpus risks inflated variance from stochastic effects in fewer runs.

3.1. Experiment A: Robustness to Controlled Text Perturbations

3.1.1. Comprehensive Evaluation on Wikipedia Dataset

These experiments were conducted with two different types of data manipulation: (i) replacing X% of the words in the training sentences with their corresponding part-of-speech (POS) tags and (ii) completely randomizing the word order in the training texts. For both manipulations, models with identical hyper-parameters and initializations (seeds) were trained to ensure a fair comparison.

Figure 1 presents the training and validation loss over 20 epochs for the baseline model and two DTF models of the Wikipedia data. The baseline model achieved the best performance, with training loss converging to approximately 0.30 and validation loss stabilizing at the same level, indicating effective learning without significant overfitting. The POS replacement strategy (DTF: POS 50%) resulted in substantially degraded performance, with both training and validation losses plateauing around 0.39. Notably, the training and validation curves remained closely aligned throughout training, suggesting that while overfitting was avoided, the model's capacity to learn discriminative features was severely limited by the reduced lexical information. The randomization strategy (DTF: Randomize) yielded intermediate results, with validation loss converging to approximately 0.34. Although this represents a significant performance degradation compared to the baseline, the model performed better than the POS replacement condition, likely because the original lexical items remained accessible despite the disrupted word order.

To further investigate the impact of lexical information loss, we conducted an ablation study varying the POS replacement rate at 50%, 40%, and 30%. Figure 2 depicts the loss curves for these three configurations. The results demonstrate a clear dose-response relationship between POS replacement rate and model performance. The 30% replacement condition achieved the best performance among the three variants, with validation loss converging to approximately 0.34. The 40% replacement condition showed intermediate performance at around 0.36 validation loss, while the 50% replacement condition exhibited the poorest performance, plateauing at approximately 0.39.

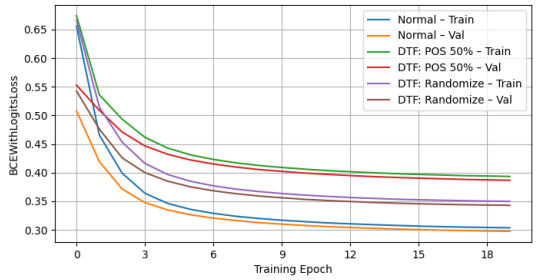

Figure 1: Training and validation loss across epochs for Wikipedia baseline and types of DTF.

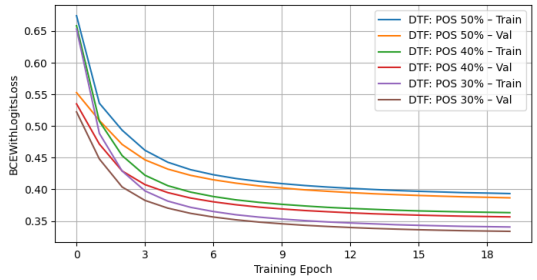

Figure 2: Training and validation loss across epochs for Wikipedia DTF POS types.

Figure 3 presents the F1-micro scores over 20 epochs for the baseline model and two DTF models. The baseline model achieved the best performance, with F1-micro scores converging to approximately 0.48 on the validation set and 0.47 on the training set by epoch 20, demonstrating robust classification capability. While this absolute performance may appear modest, it reflects the inherent difficulty of the task, which involves multi-label classification over 39 categories. Importantly, the focus of this study is on the relative impact of perturbation strategies rather than absolute performance, and the baseline therefore serves as a consistent reference point for comparison. The POS replacement strategy (DTF: POS 50%) resulted in substantially degraded performance, with F1-micro scores plateauing at approximately 0.40 for validation and 0.39 for training. The closely aligned training and validation curves throughout the learning process suggest that the model reached its learning capacity early, limited by the reduced semantic information available when half of the tokens were replaced with generic part-of-speech tags. The randomization strategy (DTF: Randomize) yielded intermediate results, with F1-micro scores converging to approximately 0.43-0.44 on both training and validation sets. This represents a moderate performance degradation of roughly 4-5% compared to the baseline. Notably, the randomization condition outperformed the POS replacement strategy by 3-4%, confirming that preserving lexical semantics even when word order is disrupted provides more discriminative information for multi-label classification than maintaining syntactic structure alone.

Figure 4 depicts the F1-micro score curves for POS replacement rate at 50%, 40%, 30%. The 30% replacement condition achieved the best performance among the three variants, with F1-micro scores converging to approximately 0.45 on validation and 0.44 on training. The 40% replacement condition showed intermediate performance, reaching around 0.43 on validation and 0.42 on training, while the 50% replacement condition exhibited the poorest performance, plateauing at approximately 0.40 for validation and 0.39 for training. Notably, all three conditions showed remarkably similar learning dynamics, with rapid improvement during the first 3-5 epochs followed by gradual convergence. The consistent gap of approximately 2-4% F1-micro points between successive replacement rates suggests that each additional 10% of lexical information preserved contributes meaningfully to classification performance.

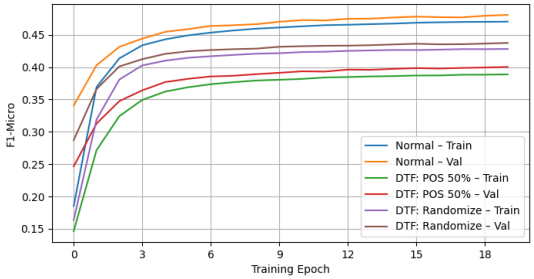

Figure 3: Training and validation F1-micro score across epochs for Wikipedia baseline and types of DTF.

The difference between the loss and the F1-score for the different DTF models from the baseline model is shown in Table 1 displaying the exact differences to the baseline model indicated above.

Model	Δ Train-Loss	Δ Val-Loss	Δ Train-F1	Δ Val-F1
Baseline	0.0000	0.0000	0.0000	0.0000
DTF: POS 50%	0.0870	0.0866	−0.0803	−0.0788
DTF: POS 40%	0.0583	0.0581	−0.0507	−0.0505
DTF: POS 30%	0.0356	0.0353	−0.0296	−0.0293
DTF: Randomize	0.0450	0.0444	−0.0416	−0.0430

Table 1: Difference in loss and F1-score from the baseline model of the Wikipedia data for all DTF models.

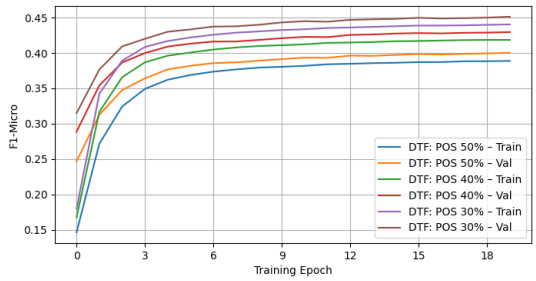

Figure 4: Training and validation F1-micro score across epochs for Wikipedia DTF POS types.

To determine whether the differences arise from the training data or the model itself, we trained the baseline configuration and one DTF configuration (DTF: POS 50%) five times with different random seeds. Due to a limitation of GPU computing resources only these two models were trained more than once. The resulting variance and standard deviation across runs was negligible (see Table 2) indicating that the chosen hyper-parameters are stable and that the model architecture is robust.

3.1.2. Targeted Evaluation on Specialist Dataset

We extend the robustness analysis to the specialist corpus, quantifying degradation in loss and F1-micro relative to the baseline across DTF conditions. The 39-class model achieves a validation F1-micro score of 0.53 (validation loss: 0.28), indicating solid multi-label performance given the fine-grained categories and terminological density. Figures 5 and 6 reveal patterns analogous to those observed for the Wikipedia data.

Qualitative comparison against expert-assigned ground-truth keywords reveals linguistically plausible predictions that align reasonably well with the texts' core themes:

- Text "Wie flektieren entlehnte Adjektive?" (expert: *Adjektiv, Deklination, Flexion, Lehnwort*) predicts *Adjektiv* (0.6428) directly alongside related categories *Satzmodus* (0.5555) and *Satzadverbiale* (0.5567), plausibly extending the adjectival inflection focus.
- Text "Subjektkomplement im Vorfeld" (expert: *Vorfeld, Subjekt, Komplement*) assigns high probabilities to *Passiv* (0.9545), *Vorfeld* (0.7806), *Satzadverbiale* (0.7222), and *Supplement* (0.5436), capturing key aspects of clause-initial argument structure and word order.
- Text "Satz-Nomen- und Phrase-Nomen-Komposita" (expert: *Komposition, Nominalphrase, Phrase, Satz*) predicts *Wortstellung* (0.7111), *Mittelfeld* (0.6096), and *Wortart* (0.5850), reasonably associating compounding with phrasal and sentence-level syntactic phenomena.

These examples demonstrate face-validity: the model's top predictions are interpretable and thematically coherent with expert labels, focusing on grammatical core concepts (inflection, word order, argument structure) rather than random or implausible categories. This supports DistilBERT's capacity for meaningful thematic classification in specialized registers, though exact keyword recovery remains partial as expected in a reduced 39-class taxonomy. Moreover, text length and the usage of specialized terminology also play an important role in the predictions, as less terminology, shorter text length, and an overload of many example sentences illustrating grammatical phenomena lead to a more inaccurate assessment.

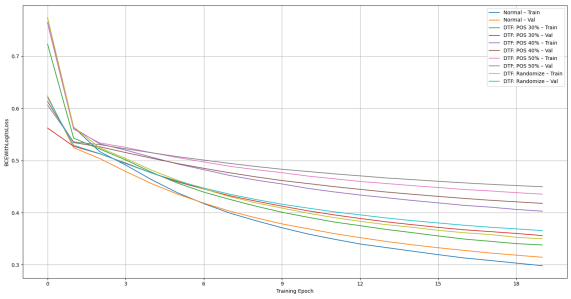

Figure 5: Training and validation loss across epochs for specialist dataset baseline and types of DTF.

Furthermore, Table 3 confirms the same consistent ranking observed for Wikipedia: POS replace-

Metric	Baseline		DTF: POS 50%	
	Var	Std-Dev	Var	Std-Dev
Train loss	1.1×10^{-5}	3.38×10^{-3}	2.5×10^{-6}	1.59×10^{-3}
Val loss	1.2×10^{-5}	3.50×10^{-3}	2.4×10^{-6}	1.54×10^{-3}
Train F1-Micro	4.0×10^{-6}	2.10×10^{-3}	7.0×10^{-7}	8.35×10^{-4}
Val F1-Micro	8.0×10^{-6}	2.82×10^{-3}	2.0×10^{-6}	1.43×10^{-3}

Table 2: Variance and standard deviation across five runs with different random seeds for the baseline and DTF (POS 50%) models.

Model	Δ Train-Loss	Δ Val-Loss	Δ Train-F1	Δ Val-F1
Baseline	0.0000	0.0000	0.0000	0.0000
DTF: POS 50%	0.1482	0.1457	−0.1904	−0.2136
DTF: POS 40%	0.1109	0.1092	−0.1312	−0.1418
DTF: POS 30%	0.0432	0.0469	−0.0651	−0.0745
DTF: Randomize	0.0570	0.0569	−0.0863	−0.0975

Table 3: Difference in loss and F1-score from the baseline model of the specialist data for all DTF models.

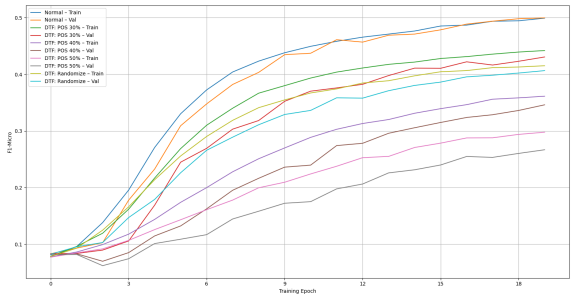

Figure 6: Training and validation F1-micro score across epochs for specialist dataset baseline and types of DTF.

ment at 50% causes the largest drops (e.g., validation F1-micro worsens by 0.2136 relative to baseline), with performance improving monotonically as replacement rates decrease to 40% (Δ=0.1418) and 30% (Δ=0.0745); word-order randomization falls between POS-40 and POS-30 (Δ=0.0975). Absolute degradation is markedly steeper, roughly 2–3× worse across metrics.

The results likely reflects the specialist corpus's extreme terminological density, where domain-specific lexical items — such as technical terms denoting fine-grained grammatical phenomena — carry disproportionate discriminative weight for thematic classification, rendering POS substitution particularly catastrophic compared to Wikipedia's more general-domain prose with broader semantic redundancy. In such narrow, technical registers, DistilBERT's reliance on precise lexical-semantic cues dominates even more markedly than positional or syntactic signals, amplifying overall sensitivity to DTF perturbations. Syntax preservation alone proves insufficient to sustain classification performance when irreplaceable terminology is systematically ablated. These findings underscore a register-dependent vulnerability: while general-domain models tolerate moderate degradation via contextual inference, specialist tasks demand verbatim lexical fidelity, posing steeper challenges for privacy-preserving text transformations in domain-specific NLP applications.

3.2. Experiment B: Train/Test Mismatch

The aim of this experiment is to answer the question: how strongly is the learned representation tied to a specific text form? Two models were trained on the Wikipedia data. The first model (B1) was trained on clean text (no alterations) and evaluated using DTF (50% POS replacement). The second model (B2) was trained on DTF with 50% POS replacement and evaluated on clean text. We chose the 50% POS replacement, because it produces the largest statistically reliable performance drop of approximately 8% micro-F1 loss on the validation set, while still allowing training to converge. This provides a balanced level of difficulty, enabling meaningful analysis of the influence of text form. For training both models, we used early stopping to avoid computational overload.

Figure 7 shows the training and validation loss under representational train–test mismatch. Model B1 (Train clean→Val DTF (50% POS)) fails to transfer from lexically grounded representations to representations where lexical semantics is partially removed and replaced by syntactic category information. The training loss (clean) decreases sharply and continuously and the validation loss (DTF) only decreases at the beginning and plateaus early at a significantly higher level. Overall, training on clean text leads to representations that rely heavily on lexical-semantic cues and exhibit

limited transfer to POS-based DTF. For model B2 (Train DTF (50%)→Val clean), the training loss (DTF) decreases more slowly and remains higher. The validation loss (clean) decreases continuously and reaches lower values than in model B1 (clean→DTF). Training on POS-based DTF induces representations that generalize better to clean text despite reduced optimization efficiency.

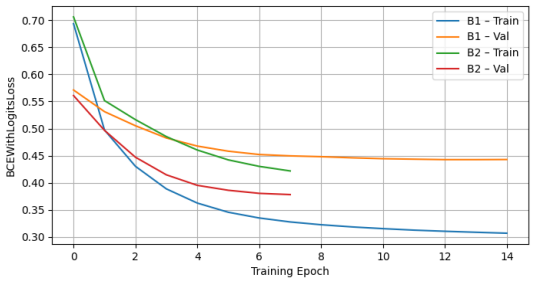

Figure 7: Training and validation loss under representational train–test mismatch. Models are trained either on clean text or on DTF (50% POS replacement) and evaluated on the opposite representation.

Figure 8 presents the F1-micro scores. The training F1-micro score of model B1 rises quickly and high (to 0.47). The validation F1-micro score rises significantly more slowly, reaching saturation at 0.33 and remaining clearly below training F1-micro score. The model learns very well on clean text. However, much of this knowledge cannot be transferred to POS-DTF due to missing lexical-semantic clues. High in-domain performance does not translate to robustness under representational degradation. The training F1-micro score of model B2 rises more slowly and remains below the clean training F1-micro score of model B1 (0.37). The validation F1-micro score rises steadily to reach 0.35, which is above the validation F1-micro score of model B1. According to the lower training F1-micro score, DTF training is more difficult. The model learns more robust features that are less dependent on lexical content. Overall, the generalization of DTF training to clean text is better than the reverse. F1-micro scores reveal an asymmetric transfer behavior: models trained on clean text achieve high in-domain performance but struggle to generalize to POS-based DTF, whereas models trained under POS-induced information limitations exhibit lower training performance, but improved generalization to clean text.

4. Conclusion

Using public-domain data from both Wikipedia and specialist linguistics corpora, this study demonstrates DTF's viability for transformer-based thematic classification by quantifying perturbation effects on learning dynamics and generalization.

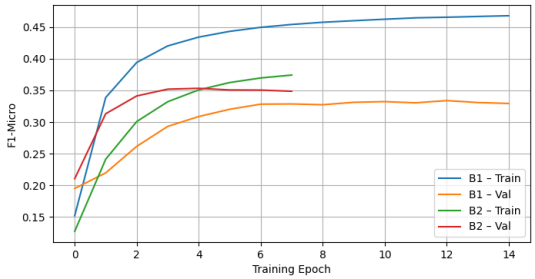

Figure 8: Training and validation F1-micro score under representational train–test mismatch. Models are trained either on clean text or on DTF (50% POS replacement) and evaluated on the opposite representation.

Experiment A yields several key insights: the baseline's superior performance confirms the joint importance of lexical-semantic and sequential information; POS replacement degrades performance more severely than randomization — especially in specialist texts due to terminological density — revealing lexical semantics' greater discriminative power, with performance scaling linearly such that each additional 10% of preserved tokens yields measurable gains. Randomization, in turn, demonstrates syntax's secondary but meaningful contribution.

Experiment B shows that models trained with DTF generalize more effectively to clean text than vice versa. While DTF training leads to slower optimization, it improves cross-representation generalization. In contrast, models trained exclusively on clean text exhibit pronounced sensitivity to representational shifts.

The results indicate that DistilBERT relies primarily on lexical-semantic information, as evidenced by the stronger performance degradation under POS replacement compared to randomization. The asymmetric transfer performance – DTF-trained models generalizing better to clean text than vice versa – demonstrates that training under controlled information loss induces more robust representations, less prone to overfitting on specific lexical cues. This aligns with findings in adversarial robustness studies, where BERT-like models trained on perturbed inputs develop features resilient to distributional shifts, akin to regularization effects observed in TextFooler attacks and adversarial training setups (Jin et al., 2019; Hauser et al., 2021). For DTF applications in computational literary studies, moderate POS replacement rates offer an optimal balance, given the linear scaling of performance

with preserved lexical content, while prioritizing semantics over syntax.

A central issue not addressed in this study is the reconstructibility of the generated DTF texts. Prior to publication, this should be assessed to mitigate potential data protection risks. In our case, however, this concern is moot, as all data are publicly available.

5. Acknowledgements

The work for this paper has been carried out within the SATEK project, funded by the German Research Foundation, grant number 531750631. One of the authors further acknowledges involvement in the Text+ project (grant number 460033370), which contributed to the development of this work.

The authors acknowledge the use of a large-language model (LLM) to draft the descriptive text for the figures in Chapter 3.1.1. The drafts were subsequently revised and edited by the authors, who retain full responsibility for the content.

6. Bibliographical References

Cohere. 2025. Command a: An enterprise-ready large language model. *10.48550/arXiv.2504.00698*.

Tillmann Dönicke, Florian Barth, Hanna Varachkina, and Caroline Sporleder. 2022. MONAPipe: Modes of narration and attribution pipeline for German computational literary studies and language analysis in spaCy. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS)*, pages 8–15, Potsdam.

Keli Du. 2023. Understanding the impact of three derived text formats on authorship classification with delta. In *DHd*, page 309.

Keli Du and Christof Schöch. 2024. Shifting sentiments? what happens to bert-based sentiment classification when derived text formats are used for fine-tuning.

Jonas Hauser, Zhao Meng, Damián Pascual, and Roger Wattenhofer. 2021. Bert is robust! a case against synonym-based adversarial examples in text classification. *arXiv preprint arXiv:2109.07403*.

Di Jin, Zhijing Shin, Junjie Kim, Jiaqi Duan, Xiangyu Tang, Tongche McCoy, Michael Ramezani, Hannaneh Levine, and Byron C Wallace. 2019. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*.

Kai Kugler, Simon Münker, Johannes Höhmann, and Achim Rettinger. 2024. Invbert: Reconstructing text from contextualized word embeddings by inverting the bert pipeline. *Journal of Computational Literary Studies*, 2:1–18.

Marc Kupietz, Harald Lüngen, Paweł Kamocki, and Andreas Witt. 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christian Lang, Roman Schneider, and Karolina Suchowolec. 2018. Extracting specialized terminology from linguistic corpora. In Eric Fuß, Marek Konopka, Beata Trawiński, and Ulrich H. Waßner, editors, *Grammar and Corpora 2016*, pages 425–434. Heidelberg University Publishing, Heidelberg.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007. Corpus masking: Legally bypassing licensing restrictions for the free distribution of text collections. In *Digital Humanities 2007*, pages 166–170.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Roman Schneider and Christian Lang. 2022. Das grammatische Informationssystem grammis – Inhalte, Anwendungen und Perspektiven. *Zeitschrift für germanistische Linguistik*, 50(2):407–427.

Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

Juan Terven, Diana-Margarita Cordova-Esparza, Julio-Alejandro Romero-González, Alfonso Ramírez-Pedraza, and Edgar A Chavez-Urbiola. 2025. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58(7):195.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

7. Language Resource References

Kupietz, Marc and others. 2019. *DeReKo - Deutsches Referenzkorpus: wpd19*. Institute for the German Language (IDS), Mannheim.

Schneider, Roman. 2026. *Abgeleitete Textformate zu gesprochener und geschriebener Sprache im Nähe-Distanz-Kontinuum*. Institute for the German Language (IDS), Mannheim.