Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (2) 160–174
http://journals.rudn.ru/miph

Research article

UDC 519.7
PACS 07.05.Tp
DOI: 10.22363/2658-4670-2026-34-2-160-174 EDN: JHMSMG

Modeling authorship attribution for short texts under training corpus degradation

Viktor M. Khvostenko[1], Andrey P. Kovalenko[2], Sergey Yu. Melnikov[1,3], Roman V. Meshcheryakov[1,4]

[1] *HSE University, 11 Pokrovsky Bulvar, Moscow, 109028, Russian Federation*

[2] *Academy of Cryptography of the Russian Federation, post office box 100, Moscow, 119331, Russian Federation*

[3] *RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation*

[4] *Institute of Control Science of RAS, ICS RAS, 65 building 2 Profsoyuznaya St, Moscow, 117342, Russian Federation*

(received: January 12, 2026; revised: February 10, 2026; accepted: February 20, 2026)

Abstract. Attributing authorship of short texts is a complex task, and its accuracy heavily depends on the quality and quantity of training data. In many practical scenarios, however, this data is subject to degradation. This study examines how two specific types of degradation – textual corruption and reduction in corpus size – affect the performance of six standard classifiers for authorship attribution of English-language tweets. We model textual corruption as random and independent character-level distortions applied to the texts in the training corpora. The test set (the texts whose authorship is to be attributed) remains unchanged. The experimental setup involves 50 authors, with 200 test texts. The volume of training texts per author varies from 800 to 50 (across 16 incremental steps), and the distortion level ranges from 0 to 20% (6 gradations). For each combination of corpus size and distortion level, we train and test six classifiers: Support Vector Machines (SVM), Logistic Regression, Naive Bayes, Random Forest, Decision Tree, and k-Nearest Neighbors (kNN). Authorial style is represented using Bag of Words (BOW), TF-IDF, word token N-grams, and their combinations. The results show that for most classifiers, degradation in the training sets leads to comparable reductions in attribution accuracy. The highest overall performance was achieved by the Support Vector Machine with TF-IDF features. It proved to be the most robust method both when the number of training texts was reduced and under high distortion levels. In the absence of distortions, SVM with TF-IDF achieved an accuracy of 0.92; at distortion levels of 5 and 20%, its accuracy reached 0.89 and 0.75, respectively. Logistic Regression with Bag of Words ranked second, delivering accuracies of 0.9, 0.87, and 0.74 under the same conditions. The kNN method demonstrated the lowest accuracy among the classifiers examined.

Key words and phrases: Authorship Attribution, Short Texts, Training Corpus Degradation, Textual Corruption, Classifier Performance

For citation: V. M. Khvostenko, A. P. Kovalenko, S. Y. Melnikov, R. V. Meshcheryakov "Modeling authorship attribution for short texts under training corpus degradation," *Discrete and Continuous Models and Applied Computational Science*, vol. 34, no. 2, pp. 160–174, 2026. DOI: 10.22363/2658-4670-2026-34-2-160-174. EDN: JHMSMG.

1. Introduction

Research into automatic authorship attribution dates back to the mid-20th century. Over this period, a wide range of methods has been developed – from statistical author profiles to modern large language models. Today, the task remains highly relevant due to the proliferation of digital communication channels (messengers, email, forums, blogs), which often lack robust user verification and are susceptible to identity forgery.

Contemporary studies in authorship attribution increasingly focus on more challenging variants of the classical problem. These include the analysis of short texts [1], attribution under conditions of textual distortion [2], and scenarios with a large number of candidate authors [3]. Each of these complications has been shown to significantly reduce attribution accuracy.

However, the combined effect of two simultaneous challenges—reduced training corpus size and textual distortions within those corpora—remains underexplored. This paper addresses this gap by investigating authorship attribution for short texts under such conditions of training data degradation.

We focus on random character-level distortions (specifically, symbol substitutions) and consider six standard classifiers: SVM, Logistic Regression, Naive Bayes, Random Forest, Decision Tree, and kNN. As stylistic features, we employ token N-grams ($N = 1, 2$) and TF-IDF values computed on various combinations of N-grams—features commonly used in authorship attribution studies.

The aim of this work is to analyze how the accuracy of these classifiers degrades when the training data is subject to random distortions and/or when the volume of the training corpus is reduced.

1.1. Negative effects arising from text distortions

Text distortions can arise from various sources, including typos, optical character recognition (OCR) errors [4], speech recognition errors [5], machine translation inaccuracies [6], and imperfections from decrypting certain ciphers [7]. Such distortions can significantly impair both the human readability of a text and the performance of automated processing systems.

A review by [8] examines how traditional text analysis methods—such as information retrieval, topic identification, summarization, and named entity extraction—are affected by distorted input. More recently, the impact of distortions on large language model (LLM) performance has been explored in [9].

The specific challenge of summarizing documents containing OCR errors was addressed in [10]. The study demonstrated that while the "Cut-and-Paste Text Summarization" method [11] performs well on clean text, its effectiveness drops sharply when applied to distorted text. The quality of the generated summary degrades rapidly with even low levels of noise, with tokenization errors and incorrect sentence boundary detection causing the most significant harm.

Errors introduced at the character level can also propagate through processing pipelines. In [4], the authors trace how OCR errors (character replacements, insertions, and deletions) lead to incorrect tokenization. This, in turn, causes faulty sentence segmentation and ultimately results in errors during part-of-speech tagging. Thus, in a pipeline-based approach, even minor initial distortions can amplify at each subsequent stage.

Interestingly, the impact of distortions varies by task. While [12] reports that OCR errors and synthetic character-level distortions have only a minor effect on document clustering accuracy, they significantly reduce the performance of topic modeling with Latent Dirichlet Allocation.

1.2. On the attribution of authorship of distorted texts

A number of studies have investigated how distortions affect authorship attribution under various conditions. In [13], the problem is considered in a scenario where the training and test texts differ in genre and subject matter—a difference that negatively impacts identification accuracy. The study demonstrates that stylistic features based on character N-grams (particularly for $N = 3$) are more robust to such cross-genre and cross-topic variations than word-based features.

The robustness of character N-grams to textual noise itself was explored in [14] using the "Burrows' delta" method [15]. The authors employed frequency vectors of the most frequent words (MFW) and character N-grams as stylistic features, introducing random character-level distortions at noise levels ranging from 1 to 100% (in 1% increments). Experiments on relatively long historical texts in several European languages revealed a trade-off: higher attribution accuracy at low noise levels comes at the cost of reduced noise immunity, and vice versa. Specifically, increasing the length of MFW vectors improves accuracy when noise is minimal, while shorter vectors perform better as distortion levels rise. The study also found that vectors of the most frequent character 3-grams serve as reliable stylistic markers even under heavy distortion—in some corpora, authorial signal remained traceable up to a 60% noise level.

A different approach to distortion—optical character recognition (OCR) of noisy images—is examined in a series of studies [16–18]. The authors constructed a corpus of 25 English texts (approximately 850 words each) written by five authors. Printed pages were saved as high-resolution JPG images, then artificially degraded using MATLAB's "Salt and Pepper noise" filter [19] at levels from 1 to 7% (in 1% increments). The noisy images were subsequently processed by an OCR system, and authorship attribution experiments were conducted on the resulting texts. Four classifiers were tested, using words and character N-grams ($N = 1$–4) as stylistic features. Unsurprisingly, attribution accuracy decreased as noise increased, with classifiers based on 4-gram characters achieving the highest overall performance.

Finally, a real-world application of authorship attribution under distortion is presented in [2], which analyzes letters by the Brothers Grimm—Jacob (1785–1863) and Wilhelm (1786–1859). The corpus of 85 letters was digitized using the Transkribus system [20] for printed and handwritten text recognition. The proportion of words correctly recognized (matching expert manual transcription) was 88% for printed texts and 81% for handwritten ones. The authors argue that this level of accuracy is sufficient for practical authorship attribution, at least in this binary classification task. Using a support vector machine with a linear kernel (implemented in Scikit-learn [21]), they achieved better-than-chance attribution accuracy (above 50%) on texts where at least 20% of characters were undistorted.

2. Description of experiments

2.1. Corpus Description

The experiments were conducted on the English-language Twitter corpus introduced in [22]. This dataset has been widely used for benchmarking authorship attribution models (see, e.g., [23, 24]). It comprises tweets collected in 2009–2010 from 7,000 globally popular accounts; each account is considered a distinct author. From this pool of 7,000 authors, 50 were randomly selected for the present study, and 1,000 messages were retained for each.

Representative examples from the corpus are shown as follows. The first three texts are from the same author, texts 4 and 5 from another, texts 6 and 7 from a third, and texts 8, 9, and 10 each have different authors.

1. Now playing: Modern Talking - Angie's Heart. Tune in: http://stream.laut.fm/eurodance.m3u.
2. Now playing: A-ha - Take on me. Tune in: http://stream.laut.fm/eurodance.m3u.
3. Now playing: Mozzart - Malice And Vice. Tune in: http://stream.laut.fm/eurodance.m3u.
4. Build your Sales Marketing for Free. Utilize this permission emailing tool to successfully communicate with clients. http://ow.ly/pROT.
5. Build your corporate brand or a new brand through a Free marketing and sending permission based email solution. http://ow.ly/zZopC.
6. New blog post: Stop Smoking by Hypnosis at Self Hypnosis Stop Smoking http://bit.ly/2bRcbm.
7. New blog post: Stop Smoking Habitrol Gum | Little Green Papoose http://bit.ly/8a1JmE.
8. Obama says to reduce role of U.S. nuclear weapons (Reuters) http://bit.ly/9XIGY8.
9. The time in Hollywood, CA, is 6:05:03 am.
10. @julio_zin *smirks and smiles slyly* are you now?

2.2. Data Preparation and Experimental Design

The corpus was split into training and test sets using an 80/20 ratio. For each of the 50 authors, 800 messages were allocated as potential training data and 200 messages as test data.

All texts underwent preliminary preprocessing using the NLTK and Scikit-learn libraries, which included removal of hyperlinks and hashtags, word tokenization, stopword removal, and lemmatization.

The experimental workflow comprised eight stages, illustrated in Figure 1. After preprocessing, features were extracted from the original (clean) test texts. These features served as the reference for evaluation.

For the training data, two types of degradation were introduced. First, synthetic random distortions of the "symbol substitution" type were applied using the method described in [25]. For a given distortion probability of $p \in (0, 1)$, the text was scanned character by character. Each character was independently subjected to replacement with probability p; if replaced, it was substituted by a randomly selected Latin character. The distortion parameter p was chosen such that 0, 1, 3, 5, 10, and 20% of the text characters were distorted. Thus, for each original training text, six versions with different distortion levels were generated. The second type of degradation involved varying the number of training texts per author from 50 to 800 in increments of 50, yielding 16 distinct training set sizes.

These two factors were combined factorially, producing a set of training set variants. For each such variant, stylistic features were extracted. The following feature types were used:

- Bag of Words (BOW)—token counts;
- N-grams—word tokens and bigrams;
- TF-IDF—calculated on five different N-gram combinations: tokens; tokens + bigrams; tokens + bigrams + trigrams; bigrams + trigrams; and tokens + bigrams + trigrams + 4-grams.

Six classifiers from the Scikit-learn library were employed (six classifier variants): Support Vector Machines (SVM), Logistic Regression, Naive Bayes, Random Forest, Decision Tree, and k-Nearest Neighbors (kNN). Default parameters were used for all classifiers, except kNN where $k = 1$ was set.

For each combination of distortion level, corpus size, feature type, and classifier, the model was trained on the degraded training data and subsequently tested on the clean test set. Accuracy was used as the primary evaluation metric, defined as the ratio of correctly attributed authors to the total number of predictions.

Testing the authorship attribution models was carried out in eight stages, as shown in Figure 1.

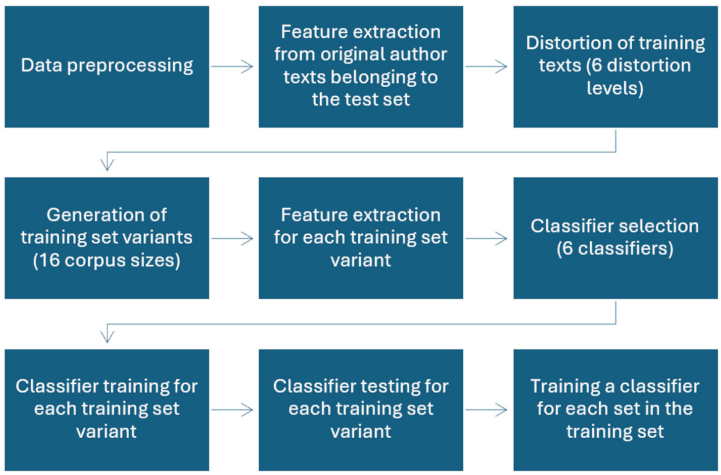

Figure 1. Stages of authorship attribution model testing

3. Results

The experimental results are presented in Figure 2 to 13. The figures show attribution accuracy for six different feature sets, color-coded as follows:

- Blue: Bag of Words (BOW);
- Orange: TF-IDF on tokens;
- Green: *N*-grams (tokens and token bigrams);
- Red: TF-IDF on tokens, bigrams, and trigrams;
- Purple: TF-IDF on bigrams and trigrams;
- Brown: TF-IDF on tokens, bigrams, trigrams, and 4-grams.

For each classifier, two figures are provided: one showing performance on clean training data as a function of training set size, and another showing performance under increasing distortion levels (1, 3, 5, 10, and 20%).

All classifiers were implemented using the Scikit-learn library with default parameters, unless otherwise specified. The kNN classifier required additional parameter tuning, described in Section 3.6.

3.1. Support vector machine (linear kernel)

The linear SVM classifier was implemented using the LinearSVC package from Scikit-learn with default parameters.

3.2. Logistic regression

The logistic regression classifier was implemented using the LogisticRegression package from Scikit-learn with default parameters.

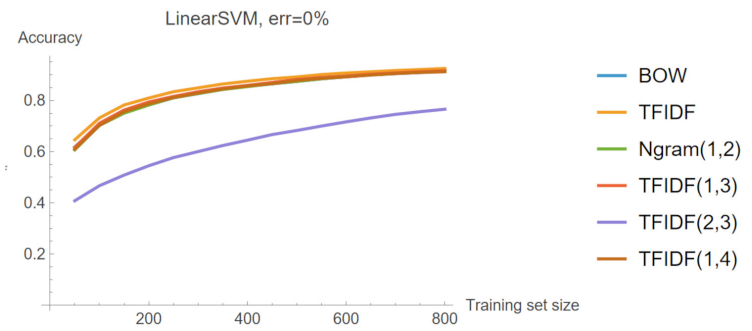

Figure 2. Accuracy of the linear SVM classifier as a function of training set size (clean training data)

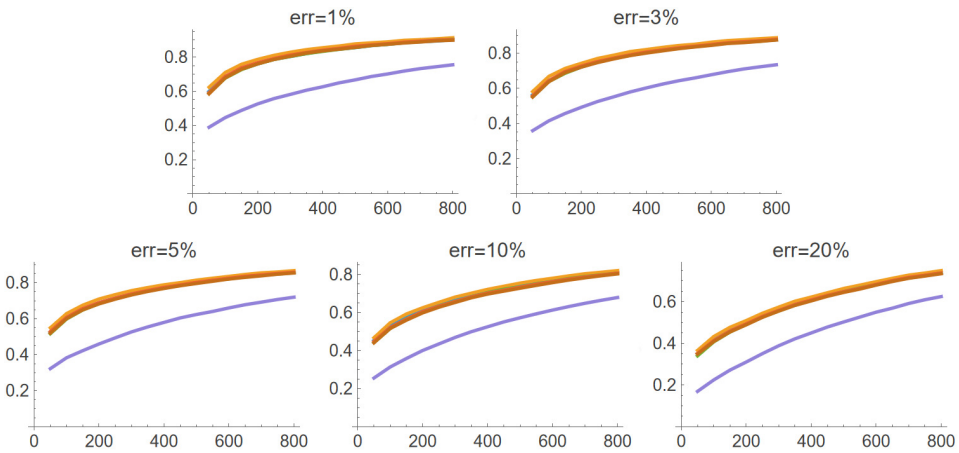

Figure 3. Accuracy of the linear SVM classifier as a function of training set size under different distortion levels (1, 3, 5, 10, and 20%)

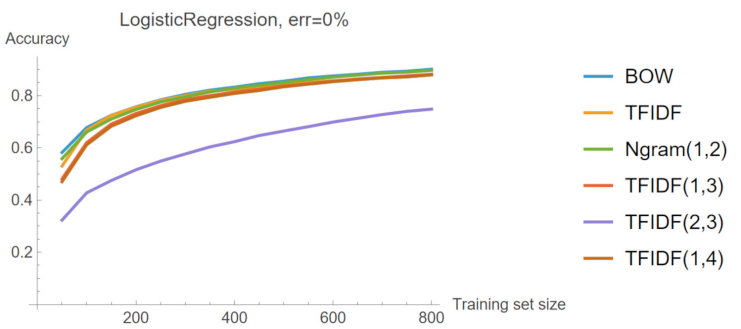

Figure 4. Accuracy of the logistic regression classifier as a function of training set size (clean training data)

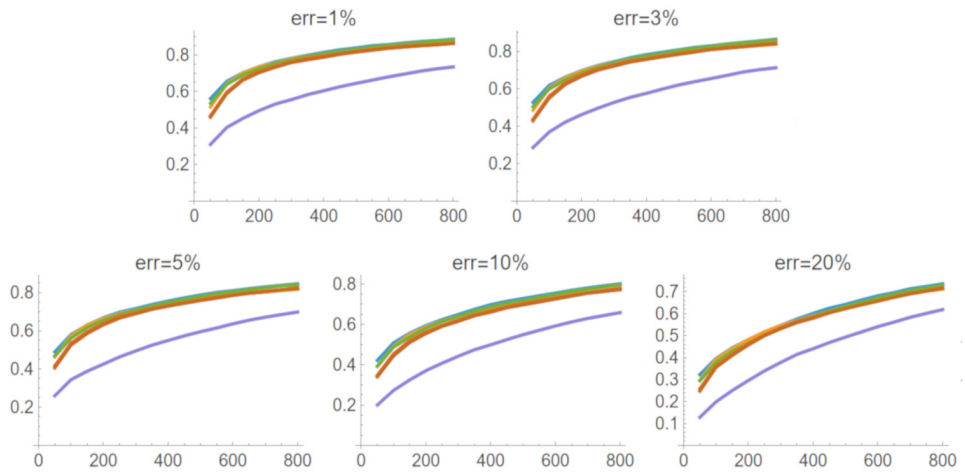

Figure 5. Accuracy of the logistic regression classifier as a function of training set size under different distortion levels (1, 3, 5, 10, and 20%)

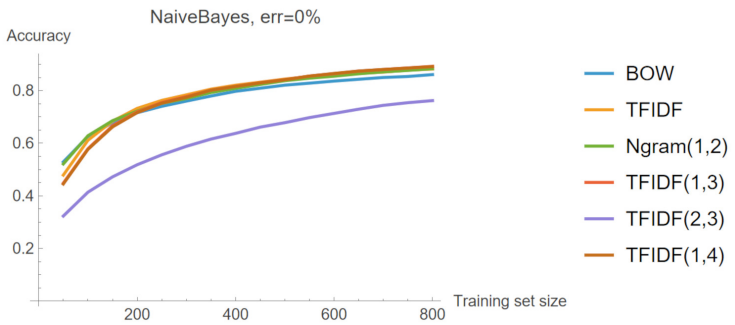

Figure 6. Accuracy of the Naive Bayes classifier as a function of training set size (clean training data)

3.3. Naive bayes

The Naive Bayes classifier was implemented using the MultinomialNB package from Scikit-learn with default parameters.

3.4. Random forest

The Random Forest classifier was implemented using the RandomForestClassifier package from Scikit-learn with default parameters.

3.5. Decision tree

The Decision Tree classifier was implemented using the DecisionTreeClassifier package from Scikit-learn with default parameters.

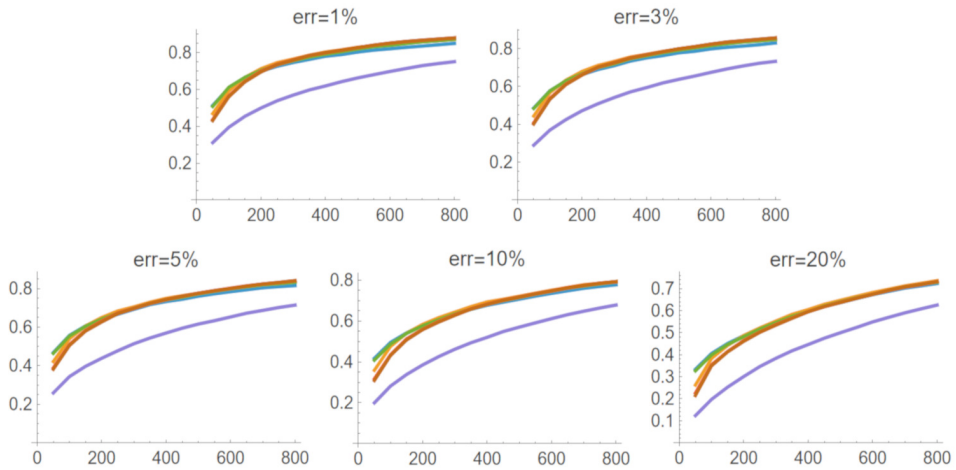

Figure 7. Accuracy of the Naive Bayes classifier as a function of training set size under different distortion levels (1, 3, 5, 10, and 20%)

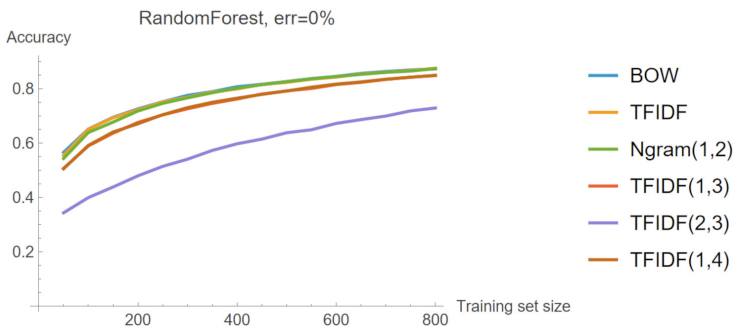

Figure 8. Accuracy of the Random Forest classifier as a function of training set size (clean training data)

3.6. k-Nearest Neighbors (kNN)

The kNN classifier was implemented using the KNeighborsClassifier package from Scikit-learn with the Euclidean distance metric. Since the performance of kNN depends critically on the choice of k, preliminary experiments were conducted to select an appropriate value.

Table 2 presents the accuracy of the kNN classifier for different values of k (from 1 to 10) across all six distortion levels. These results were obtained using TF-IDF features on tokens, which proved to be the best-performing feature set for the maximum training corpus size (see Table 2 below). The highest accuracy for each distortion level is highlighted in bold.

Based on these results, $k = 1$ was selected for all subsequent experiments. Figure 12 and 13 show the accuracy of the kNN classifier (with $k = 1$) as a function of training set size for clean and distorted training data, respectively.

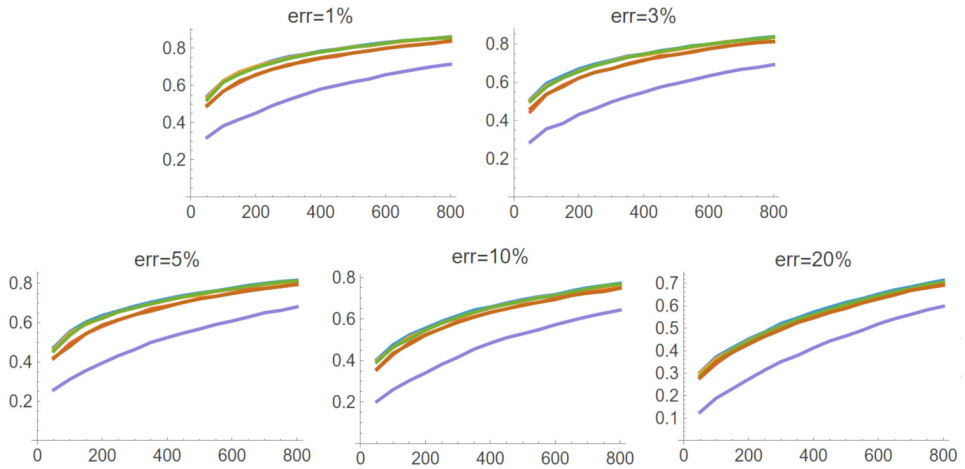

Figure 9. Accuracy of the Random Forest classifier as a function of training set size under different distortion levels (1, 3, 5, 10, and 20%)

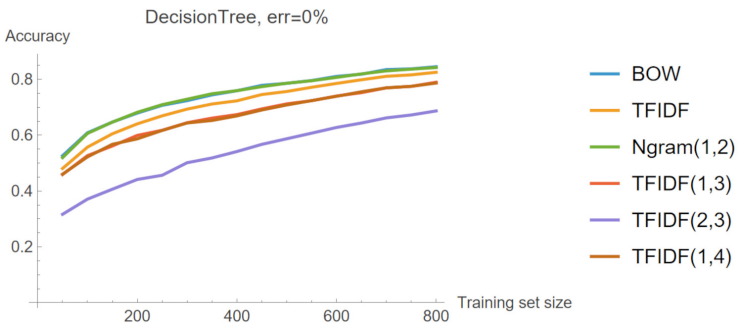

Figure 10. Accuracy of the Decision Tree classifier as a function of training set size (clean training data)

Table 1

Accuracy of the kNN classifier for the maximum training corpus size (800 texts per author) with $1 \leqslant k \leqslant 10$

k	1	2	3	4	5	6	7	8	9	10
Err=0%	**0.83**	0.76	0.74	0.73	0.72	0.71	0.71	0.71	0.72	0.72
Err=1%	**0.82**	0.74	0.72	0.70	0.70	0.69	0.69	0.69	0.69	0.69
Err=3%	**0.79**	0.71	0.68	0.66	0.65	0.64	0.64	0.64	0.64	0.64
Err=5%	**0.77**	0.68	0.65	0.63	0.61	0.6	0.6	0.6	0.6	0.6
Err=10%	**0.72**	0.62	0.59	0.56	0.54	0.52	0.52	0.52	0.51	0.51
Err=20%	**0.66**	0.54	0.49	0.45	0.42	0.41	0.4	0.39	0.39	0.39

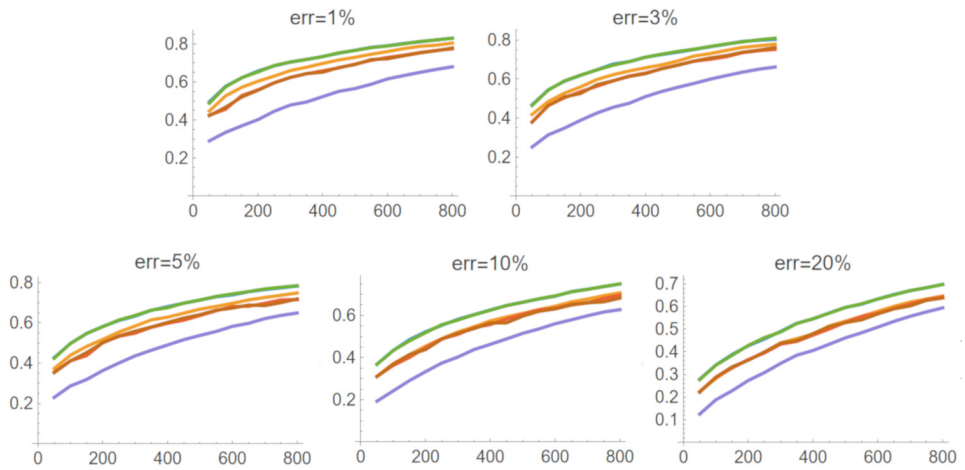

Figure 11. Accuracy of the Decision Tree classifier as a function of training set size under different distortion levels (1, 3, 5, 10, and 20%)

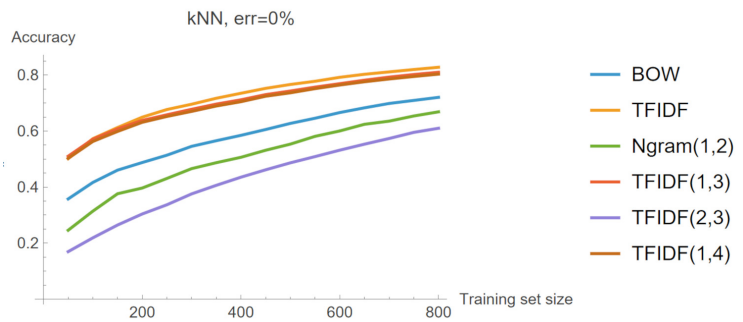

Figure 12. Accuracy of the kNN classifier ($k = 1$) as a function of training set size (clean training data)

4. Analysis and discussion of results

Table 2 summarizes the best attribution accuracy achieved by each classifier for three distortion levels (0%, 5%, and 20%) and five training set sizes (800, 600, 400, 200, and 50 texts per author). The color coding indicates which stylistic feature set yielded the best performance for each classifier under each condition. The feature sets are: Bag of Words (BOW), TF-IDF on tokens, N-grams (1,2), TF-IDF on tokens+bigrams+trigrams (denoted TF-IDF(1,3)), and TF-IDF on tokens+bigrams+trigrams+4-grams (denoted TF-IDF(1,4)).

For the baseline scenario (clean training data with full corpus size), our results closely match those reported in [24], confirming the validity of our experimental setup.

A notable pattern emerges from Figure 2–13: the TF-IDF(2,3) feature set—comprising only bigrams and trigrams of tokens—consistently underperforms compared to all other feature sets. This can be attributed to two factors. First, all other feature sets include unigrams (individual tokens), which appear to be highly informative for authorship attribution. Second, token N-grams are particularly

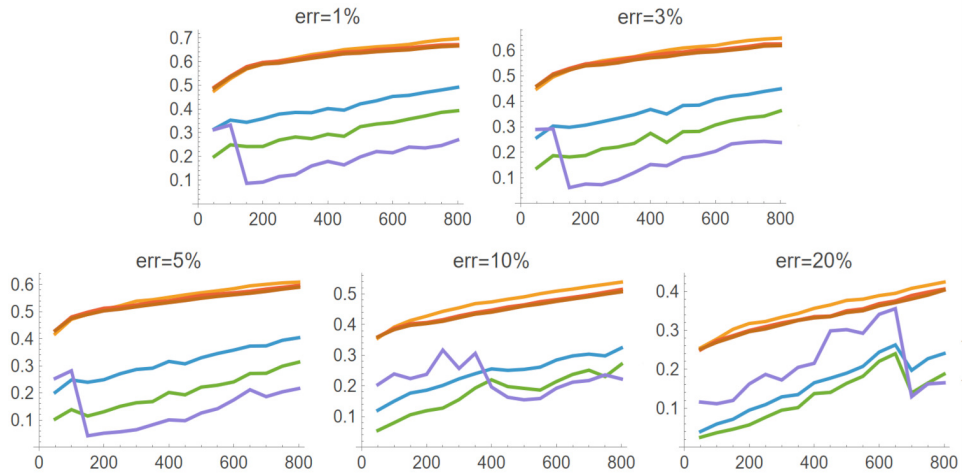

Figure 13. Accuracy of the kNN classifier ($k = 1$) as a function of training set size under different distortion levels (1, 3, 5, 10, and 20%)

vulnerable to character-level distortions, as random symbol substitutions can easily disrupt word boundaries and token sequences.

Support Vector Machine (SVM) demonstrated the highest overall accuracy across nearly all parameter combinations tested. This finding aligns with the well-established suitability of SVM for high-dimensional, sparse data typical of text classification tasks.

Logistic Regression consistently ranked second, delivering comparable performance to SVM in many scenarios, particularly when using Bag of Words features (as noted in Section 3.2).

Naive Bayes, Random Forest, and Decision Tree showed intermediate performance, with accuracy declining predictably as distortion levels increased and training set sizes decreased.

k-Nearest Neighbors (kNN) proved to be the least effective classifier in our experiments, exhibiting both low accuracy and high variability. This poor performance is likely due to the well-known sensitivity of distance-based methods to the high dimensionality of feature spaces—a phenomenon often referred to as the "curse of dimensionality."

An interesting observation is that $k = 1$ consistently yielded the best results for the kNN classifier (see Table 2). While this might appear counterintuitive—since smaller k values typically increase sensitivity to noise—the structure of the Twitter corpus provides an explanation. As illustrated in Table 1, many authors in the dataset exhibit highly formulaic writing patterns (e.g., automated "Now playing..." posts or templated promotional messages). This creates a situation where each author's texts have many close neighbors within their own collection, making the single nearest neighbor a surprisingly effective choice. This phenomenon bears some resemblance to the "overfitting" effects discussed in [26], though in this case the simplicity of the 1-*NN* rule proves advantageous given the repetitive nature of the data.

4.1. Conclusion

This study investigated authorship attribution for short texts under training data degradation, combining two challenging conditions: random character-level distortions and reduction in corpus size. Using a Twitter corpus of 50 authors with 1,000 texts each, training data were varied from 800 to

Table 2

Single division error with leading monomial weight of 95%

	800	600	400	200	50
			Err=0%		
SVM	0.92	0.91	0.88	0.83	0.65
LR	0.9	0.88	0.85	0.78	0.58
NB	0.89	0.87	0.83	0.76	0.53
RF	0.88	0.86	0.82	0.75	0.57
DT	0.84	0.82	0.78	0.71	0.53
kNN	0.72	0.7	0.67	0.63	0.51
			Err=5%		
SVM	0.89	0.87	0.83	0.77	0.58
LR	0.87	0.84	0.8	0.72	0.53
NB	0.86	0.84	0.79	0.71	0.49
RF	0.84	0.81	0.76	0.7	0.51
DT	0.81	0.78	0.73	0.65	0.47
kNN	0.65	0.63	0.6	0.56	0.46
			Err=20%		
SVM	0.75	0.71	0.64	0.54	0.37
LR	0.74	0.7	0.62	0.52	0.32
NB	0.74	0.7	0.63	0.52	0.33
RF	0.71	0.67	0.59	0.48	0.3
DT	0.7	0.65	0.57	0.46	0.28
kNN	0.42	0.39	0.37	0.32	0.25
	BOW	TFIDF	Ngram (1,2)	TFIDF (1,3)	TFIDF (1,4)

50 texts per author across six distortion levels (0–20%), while test texts remained unchanged. Six classifiers—SVM, Logistic Regression, Naive Bayes, Random Forest, Decision Tree, and kNN—were evaluated using BOW and TF-IDF features on various N-gram combinations. With the exception of kNN, all classifiers exhibited comparable accuracy degradation patterns, and those performing well on complete data maintained their advantage under degraded conditions. SVM with TF-IDF features proved most robust across all training set sizes and distortion levels, achieving accuracies of 0.92, 0.89, and 0.75 at 0, 5, and 20% distortion respectively, followed closely by Logistic Regression with BOW (0.90, 0.87, and 0.74). Decision Tree and Random Forest showed greater vulnerability to data reduction and distortion, while kNN yielded the poorest and most unstable results, likely due to the high dimensionality of the feature space. Feature sets including unigrams consistently outperformed those based solely on higher-order N-grams, underscoring the importance of token-level information

for attribution of short, noisy texts. These findings quantify the robustness of common attribution methods and identify SVM with TF-IDF as the most reliable choice when training data quality cannot be guaranteed.

Author Contributions: Conceptualization, Melnikov S.Yu.; methodology, Melnikov S.Yu.; software, Khvostenko V.M.; validation, Kovalenko A.P.; formal analysis, Meshcheryakov R.V.; research, Khvostenko V.M.; resources, Kovalenko A.P.; writing—initial draft preparation, Khvostenko V.M.; writing—review and editing, Melnikov S.Yu.; visualization, Khvostenko V.M.; supervision, Meshcheryakov R.V.; project administration, Melnikov S.Yu.; funding acquisition, Meshcheryakov R.V. All authors have read and agreed with the published version of the manuscript.

Funding: The study was supported by the Russian Science Foundation grant No. 24-11-00340 (https://rscf.ru/project/24-11-00340/).

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Declaration on Generative AI: The author has not employed any generative AI tools.

References

[1] M. Eder, "Does size matter? Authorship attribution, small samples, big problem," *Digital Scholarship in the Humanities*, vol. 30, pp. 167–182, 2015. DOI: 10.1093/llc/fqt065

[2] G. Franzini, M. Kestemont, G. Rotari, M. Jander, J. Ochab, J. Byszuk, and M. Eder, "Attributing authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm," *Frontiers in Digital Humanities*, vol. 5, p. 4, 2018. DOI: 10.3389/fdigh.2018.00004

[3] K. Luyckx and W. Daelemans, "Authorship attribution and verification with many authors and limited data," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, 2008, pp. 513–520. DOI: 10.3115/1599081.1599147

[4] D. Lopresti, "Optical character recognition errors and their effects on natural language processing," in *Proceedings of the 3rd Workshop on Analytics for Noisy Unstructured Text Data*, 2009, pp. 115–122. DOI: 10.1145/1568293.1568312

[5] M. Kubis, P. Szymański, M. Ziółko, and K. Wróbel, "Open challenge for correcting errors of speech recognition systems," in *Language and Technology Conference*, 2019, pp. 322–337. DOI: 10.1007/978-3-030-33241-5_27

[6] A. A. Goncharov, N. V. Buntman, and V. A. Nuriyev, "Errors in machine translation: Classification problems," *Sistemy i Sredstva Informatiki*, vol. 29, no. 3, pp. 92–103, 2019, (in Russian). DOI: 10.14357/08696527190308

[7] E. Dawson and L. Nielsen, "Automated cryptanalysis of XOR plaintext strings," *Cryptologia*, vol. 20, no. 2, pp. 165–181, 1996. DOI: 10.1080/0161-119691884871

[8] L. V. Subramaniam, S. Roy, T. A. Faruquie, and S. Negi, "A survey of types of text noise and techniques to handle noisy text," in *Proceedings of the 3rd Workshop on Analytics for Noisy Unstructured Text Data*, 2009, pp. 115–122. DOI: 10.1145/1568293.1568312

[9] G. Sperduti and A. Moreo, *Misspellings in natural language processing: A survey*, 2025. arXiv: 2501.16836.

[10] H. Jing, D. Lopresti, and C. Shih, "Summarizing noisy documents," in *Proceedings of the Symposium on Document Image Understanding Technology*, 2003, pp. 111–119.

[11] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 2017, pp. 1–6. DOI: 10.1109/ICCCSP.2017.7944061

[12] D. D. Walker, W. B. Lund, and E. K. Ringger, "Evaluating models of latent document semantics in the presence of OCR errors," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 1–10.

[13] E. Stamatatos, "On the robustness of authorship attribution based on character n-gram features," *Journal of Law and Policy*, vol. 21, p. 7, 2013.

[14] M. Eder, "Mind your corpus: Systematic errors in authorship attribution," *Literary and Linguistic Computing*, vol. 28, no. 4, pp. 603–614, 2013. DOI: 10.1093/llc/fqt039

[15] N. Mamaev, X. Piotrowska, M. Marusenko, and A. Ronzhin, "Burrows's delta for authorship attribution of Russian literary texts," in *CEUR Workshop Proceedings*, vol. 2233, 2017.

[16] H. Sayoud et al., "Automatic authorship attribution of noisy documents," in *Proceedings of the FLAIRS Conference*, 2017, pp. 202–205.

[17] H. Benzerroug and S. Khennouf, "Author identification of corrupted OCR-based texts," *HDSKD Journal*, vol. 3, pp. 91–99, 2017.

[18] Z. Hamadache and H. Sayoud, "Authorship attribution of noisy text data with a comparative study of clustering methods," *International Journal of Knowledge and Systems Science*, vol. 9, no. 2, pp. 45–69, 2018. DOI: 10.4018/IJKSS.2018040103

[19] N. Bindal, P. Singh, V. Singh, and D. Gupta, "A systematic review of state-of-the-art noise removal techniques in digital images," *Multimedia Tools and Applications*, vol. 81, pp. 31 529–31 552, 2022. DOI: 10.1007/s11042-022-12980-7

[20] S. Colutto, P. Kahle, S. Guenter, and G. Muehlberger, "Transkribus: A platform for automated text recognition and searching of historical documents," in *15th International Conference on eScience (eScience)*, 2019, pp. 463–466. DOI: 10.1109/eScience.2019.00073

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel, "Authorship attribution of micro-messages," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1880–1891.

[23] C. Suman, S. Saha, and P. Bhattacharyya, "Authorship attribution of microtext using capsule networks," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 4, pp. 1038–1047, 2021. DOI: 10.1109/TCSS.2021.3073690

[24] T. Alsanoosy, B. Shalbi, and A. Noor, "Authorship attribution for English short texts," *Engineering Technology and Applied Science Research*, vol. 14, no. 5, pp. 16 419–16 426, 2024. DOI: 10.48084/etasr.7794

[25] V. M. Khvostenko et al., "Two-parameter model of synthetic distortions in the problem of assessing the readability of distorted texts," *The European Physical Journal Special Topics*, vol. 234, pp. 3865–3870, 2025. DOI: 10.1140/epjs/s11734-025-01567-8

[26] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

Information about the authors

Khvostenko, Viktor M.—Researcher of Cybersecurity CPS Department of HSE University (e-mail: vkhvostenko@hse.ru, ORCID: 0009-0000-1985-1401)

Kovalenko, Andrey P.—Professor, Doctor of Technical Sciences, Full member of Academy of Cryptography of the Russian (e-mail: kovalenko.ap@cryptoacademy.gov.ru, ORCID: 0009-0007-8777-8622)

Melnikov, Sergey Yu.—Doctor of Physical and Mathematical Sciences, Professor of Department of Probability Theory and Cyber Security of RUDN University; Chief Researcher of Cybersecurity CPS Department of HSE University (e-mail: melnikov-syu@rudn.ru, ORCID: 0000-0002-9023-9896)

Meshcheryakov, Roman V.—Professor, Doctor of Technical Sciences, Chief Researcher of Cybersecurity CPS Department of HSE University; Head of the Laboratory, Institute of Control Science of RAS (e-mail: mrv@ieee.org, ORCID: 0000-0002-1129-8434)

RUSSIAN METADATA

УДК 519.7
PACS 07.05.Tp
DOI: 10.22363/2658-4670-2026-34-2-160-174 EDN: JHMSMG

Моделирование атрибуции авторства коротких текстов в условиях деградации обучающего корпуса

В. М. Хвостенко[1], А. П. Коваленко[2], С. Ю. Мельников[1,3], Р. В. Мещеряков[1,4]

[1] *Национальный исследовательский университет «Высшая школа экономики», Покровский бульвар, д. 11, Москва, 109028, Российская Федерация*

[2] *Академия криптографии Российской Федерации, а/я 100, Москва, 119331, Российская Федерация*

[3] *Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация*

[4] *Институт проблем управления им. В. А. Трапезникова Российской академии наук, ул. Профсоюзная, д. 65, стр. 2, Москва, 117342, Российская Федерация*

Аннотация. Точность атрибуции авторства текстов критически зависит от качества и объема обучающих данных. Изучается, как два типа деградации обучающих авторских коллекций, искажение текста и сокращение размера корпуса, влияют на точность шести стандартных классификаторов при атрибуции авторства англоязычных твитов. Рассматриваются случайные независимые искажения типа «замена символов», применяемые к текстам в обучающих корпусах. Тестовая выборка остается неизменной. Эксперименты проводились на текстах 50 авторов (по 200 тестовых текстов). Количество обучающих текстов одного автора варьировалось от 800 до 50 (16 промежуточных шагов), а уровень искажения – от 0 до 20% (6 градаций). Для каждой комбинации размера корпуса и уровня искажения обучались и тестировались шесть классификаторов: метод опорных векторов (SVM), логистическую регрессию, наивный байесовский классификатор, случайный лес, дерево решений и метод k-ближайших соседей (kNN). В качестве признаков авторского стиля использовались мешок слов (BoW) и TF-IDF для токенов и N-грамм. Установлено, что для большинства классификаторов деградация обучающих данных приводит к сопоставимому снижению точности атрибуции. Наилучшие результаты показал метод опорных векторов с TF-IDF. Он оказался наиболее устойчивым как при сокращении количества обучающих текстов, так и при высоких уровнях искажений. В отсутствие искажений метод SVM с TF-IDF достиг точности 0.92; при уровнях искажений 5 и 20% его точность составила 0.89 и 0.75 соответственно. Логистическая регрессия с с мешком слов заняла второе место, показав точность 0.9, 0.87 и 0.74 в тех же условиях. Метод kNN продемонстрировал наименьшую точность среди рассмотренных классификаторов.

Ключевые слова: атрибуция авторства, короткие тексты, ухудшение качества обучающего корпуса, искажение текста, производительность классификатора