The Million Authors Corpus: A Cross-Lingual and Cross-Domain Wikipedia Dataset for Authorship Verification

Abraham Israeli[*1], **Shuai Liu**[*2], **Jonathan May**[2] and **David Jurgens**[1]
[1]University of Michigan
[2]Information Sciences Institute, University of Southern California
{isabrah, jurgens}@umich.edu, {liushuai, jonmay}@isi.edu

Abstract

Authorship verification (AV) is a crucial task for applications like identity verification, plagiarism detection, and AI-generated text identification. However, datasets for training and evaluating AV models are primarily in English and primarily in a single domain. This precludes analysis of AV techniques for generalizability and can cause seemingly valid AV solutions to, in fact, rely on topic-based features rather than actual authorship features. To address this limitation, we introduce the Million Authors Corpus (*MAC*), a novel dataset encompassing contributions from dozens of languages on Wikipedia. It includes only long and contiguous textual chunks taken from Wikipedia edits and links those texts to their authors. *MAC* includes 60.08M textual chunks, contributed by 1.29M Wikipedia authors. It enables broad-scale cross-lingual and cross-domain AV evaluation to ensure accurate analysis of model capabilities that are not overly optimistic. We provide baseline evaluations using state-of-the-art AV models as well as information retrieval models that are not AV-specific, in order to demonstrate *MAC*'s unique cross-lingual and cross-domain ablation capabilities.

1 Introduction

Authorship verification (AV) aims to determine whether two or more texts are written by the same author. AV has attracted significant attention from researchers because of its practical and diverse applications, such as identity verification, account-linking, historical linguistics, forensic analysis, and AI-generated text detection. Despite various approaches having been explored for AV, the state-of-the-art (SOTA) models mainly focus on a data-driven representation learning approach (Rivera-Soto et al., 2021; Zhu and Jurgens, 2021; Wegmann et al., 2022; Fincke and Boschee, 2024). Although these representation models demonstrate promising performance on large-scale AV on various datasets (Klimt and Yang, 2004; Schler et al., 2006; Seroussi et al., 2014; Ni et al., 2019; Bevendorff et al., 2020; Khan et al., 2021), they are limited mainly to AV on English texts in a single domain because most dataset studies are single-domain and English-only. These constraints on domain and language further hinder the study of cross-lingual and cross-domain AV, which requires texts written by the same author but in different languages and/or domains.

The objective of this study is to research and address these limitations. To do so, we introduce the *Million Authors Corpus* (*MAC*)—a large dataset of content from Wikipedia that covers 60 languages.[1] The potential of Wikipedia as a corpus for linking textual content to its contributing users in multiple languages and domains has remained underexplored. Unlike existing datasets, *MAC* contains texts in dozens of languages and four different domains, as well as authors writing in multiple languages and domains. *MAC* enables a more comprehensive evaluation of AV models, which is impossible with existing datasets. In addition, *MAC* is suitable for training and evaluating other NLP tasks such as text-style transfer (Hallinan et al., 2023; Liu and May, 2024), and semantic-shift (Hamilton et al., 2016; Kutuzov et al., 2018).

We choose to base *MAC* on Wikipedia since: (i) Wikipedia is one of the largest public textual repositories in the world (Medelyan et al., 2009; Mesgari et al., 2015); (ii) The association between an editor and a page-edit is transparent, verified, and maintained by the Wikimedia Foundation (Ayers et al., 2008); and (iii) Wikipedia allows its users multiple ways to contribute and communicate, such as editing *article pages*[2] or discussing various topics over

[*] denotes equal contribution

[1] *MAC* includes 60 languages, two of which are English. One of these is "simple English," a simplified variant of standard English used as a separate Wikipedia project.

[2] Article pages are the content pages people normally think of as Wikipedia.

Findings of the Association for Computational Linguistics: ACL 2025, pages 25997–26017
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

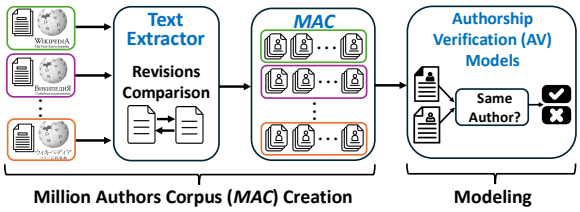


Figure 1: *MAC* creation and usage. We process 60 languages from Wikipedia to create *MAC*. It contains 60.08M text chunks contributed by 1.29M users. Using *MAC*, we train authorship verification models and compare their performance against existing baselines in novel cross-lingual and cross-domain settings.

talk pages. Users' writing styles and topics widely differ between these page types. Wikipedians are also welcome to contribute to multiple Wikipedia projects (i.e., languages). We make use of this and collect data from the different Wikipedia page types and languages to enable *cross-domain* and *cross-language* evaluation. Therefore, *MAC* consists of users that contribute to multiple Wikipedia namespaces (which we refer to as *domains*) as well as multilingual users.

We develop a novel pipeline to extract *substantive* contributions from Wikipedia in any language. These contributions, which we refer to as *text chunks*, must be long and contiguous new textual edits. We associate each text chunk with its corresponding author (i.e., a Wikipedian), we henceforth refer to as an *author*.

Due to Wikipedia's diversity, these chunks represent multiple genres, including encyclopedic text, social discussions, and debates, all across multiple topics and languages. Including only long enough contributions makes *MAC* practical and useful for AV models and makes the conclusions from AV experiments meaningful. Figure 1 illustrates the creation-evaluation flow of *MAC*.

Overall, *MAC* contains more than 60.08M text chunks from 60 Wikipedia languages. 1.29M contributers made at least two contributions to be included in *MAC*. It contains more than 560K authors who have substantial contributions to more than a single domain and more than 250K authors who have substantial contributions in more than a single language. These cross-domain and cross-lingual characteristics of *MAC* makes it valuable for various NLP tasks, primarily AV.

To demonstrate the value of *MAC*, we conduct an exploratory study and show that *MAC* could help answer five fundamental research questions (RQs) in AV. **RQ1**: Does an AV model trained on one language and domain perform well on unseen authors in the same language and domain? **RQ2** (respectively, **RQ3**): Does an AV model trained on one language (respectively, domain) perform well on unseen authors on other languages (respectively, domains)? **RQ4** (respectively, **RQ5**): Given a text written by an author, could an AV model find another text written by the same author in another language (respectively, domain)? The first three RQs are answerable using existing datasets, while the latter two are only answerable with *MAC*. Notably, our results on four AV models show that their performance differences are generally consistent across the first two RQs but starkly different across the latter three RQs, including the two only assessable by *MAC*, confirming the necessity of the cross-lingual and cross-domain evaluation powered by *MAC*.

To summarize, in this study, we make the following contributions:

1. We introduce *MAC*—a novel corpus that relies on Wikipedia data. Its uniqueness is the extraction of long text chunks from Wikipedia and its linkage to the contributing author. *MAC* is cross-lingual and covers 60 languages. It is also cross-domain as we collect data from different Wikipedia page types (e.g., article and talk pages). We make *MAC* available for the research community. The full dataset is available in the Zeondo data repository.[3] The concise data we use to train the presented models in the manuscript are also available in the Hugging-Face repository.[4]

2. Based on *MAC*, we conduct an exploratory experiment using four AV models to examine five fundamental RQs, two of which are unanswerable with existing datasets regarding cross-lingual and cross-domain AV.

3. The experimental results show that the performance of the models exhibits significantly different patterns on the two RQs only answerable by *MAC* compared to both the first two RQs and the third RQ. By this, we underlie the value of the cross-lingual and cross-domain characteristics of *MAC*.

2 *MAC* Creation

Here, we detail how *MAC* was built, illustrated by the three leftmost blocks in Figure 1.

[3] https://zenodo.org/records/15538126

[4] https://huggingface.co/datasets/Blablablab/MAC

NS	Description	# T.Chunks	%
0	Article page	30.75M	51.18%
1	Article talk page	5.60M	9.33%
2	User page	1.08M	1.81%
3	User talk page	22.65M	37.69%

Table 1: Wikipedia namespaces (NSs) that we include in *MAC*, which we refer to as 'domains' in this study. The two rightmost columns are the number and percentage of text chunks included in *MAC*.

2.1 Wikipedia Raw Data

Wikipedia maintains its content in multiple languages, each functioning as an independent project. In this study, we process the 60 languages with most of the Wikipedia content to get sufficient coverage. Appendix Table 3 contains a full list of the languages we process. The English Wikipedia is the largest and most active project, with over 6M articles and 48M registered editors. However, many other languages (e.g., French and German) have shown significant growth in recent years in both content and registered editors.[5]

Content in Wikipedia comes in multiple forms. Apart from collaboratively written article pages, each user has their own publicly readable 'user page'. Further, for each article and user page, there is a 'talk page'—a virtual space for the Wikipedia community to discuss, comment, and ask questions about relevant topics of the main article or user. Each of these four types of pages contains different content and writing styles. Wikimedia maintains different namespaces (NSs) for each, which we use throughout our study and call them *domains*. We use the Wikipedia numbering notation, which ranges from 0 to 3. While building *MAC*, we process all four domains, summarized in Table 1.

We collect the data of all Wikipedia pages using the Wikipedia dump files, which are released monthly by the Wikimedia Foundation.[6] The Wikipedia dump files allow us to process all existing information on each page. This way, we can track and collect information about each revision of both the content pages and the talk pages. The dump files contain detailed information per revision, including time, editor name, textual content, etc. We use the April 2024 release of Wikipedia dump files for each of the 60 languages we process.

[5]List of Wikipedias: https://tinyurl.com/yckxuup4

[6]Wikipedia dumps: https://dumps.wikimedia.org

2.2 Raw Data Processing

A unique attribute of *MAC* is that it contains only *substantial*, long and contiguous contributions (i.e., text chunks). By doing so, we avoid the long tail of minor contributions, which are very common both in Wikipedia and in existing AV datasets and which, when included, call into question the validity of AV. In this section, we describe in detail how we extract these text chunks.

We process each Wikipedia project (i.e., language) independently. Our language-agnostic methodology makes it feasible to process Wikipedia pages in any language. To do that, we avoid using language-specific techniques that are not necessarily available across all languages (e.g., Arabic POS tagging). To process the textual content of each revision, we use existing Python packages built for Wikipedia (see Appendix Table 9).

In the remainder of the paper, we use the following notation. P is the set of Wikipedia pages that we process. For each page $p \in P$, we mark its N sequence of sorted revisions as $R_p^1, ..., R_p^i, ...R_p^N$. We mark R_p^* as the latest relevant revision of article p, which we compare each new revision against (i.e., to find added content).

Per page p, its sequence of N sorted revisions is taken as input one after the other while building the corpus. The Wikipedia dump files contain the page textual content for each $R_p^i, \forall i \in N$. They *do not* contain the textual added content of R_p^i compared to R_p^{i-1}, which forces us to compare revisions to identify new long textual contributions.

Before processing revision R_p^i, we check whether it is relevant. There are two scenarios where we skip R_p^i, such as a vandalism edit that deletes the page content. In Appendix Section A, we further explain those two scenarios. In cases when R_p^i is relevant, we follow two steps to extract its information and add to *MAC*:

Step 1: Revisions Comparison In this step, we extract the newly added textual content in R_p^i. We do so by comparing R_p^i with R_p^*.[7] By splitting the text into paragraphs and sentences, we can detect new or edited text at the sentence level. Sequences of new or edited text are concatenated into the same text chunk. Note that multiple text chunks can be added or edited within a single revision. Hence, R_p^i is associated with a *list* of new text chunks.

[7]As previously explained, some revisions are skipped, so the comparison is not always to R_p^{i-1}.

Step 2: Length Filtering Each text in *MAC* needs to be long enough to contain stylistic information for AV. However, many Wikipedia changes are minimal (e.g., typo correction), and therefore, we impose a length restriction where any edit must introduce at least α contiguous words. We set α=100 in English, and it is adjusted for each language due to morphological and typological differences (e.g., a language with shorter sentences than English but more morphologically rich words); Details of this are in Appendix Section B. This length presents a challenging setting on par with current datasets focused on social media and online reviews (e.g., Tyo et al., 2022).

We also impose a *maximum* length restriction. We exclude contributions that are longer than 5α words (e.g., 500 in English) to avoid edits, which are a recreation of existing articles, including translations and archiving of discussions in talk pages.

Wikipedia is a rich collaborative platform that allows users to contribute in different ways. Not all such contributions are relevant to our needs. Hence, we apply three data-cleaning processes to ensure we do not include undesired contributions. Those are: removal of tables, bots detection, and mixed languages content exclusion.[8] We further explain these cleaning processes in Appendix Section C.

For each text chunk that we save, we also record a list of essential attributes such as the author details (name and ID), page details (title, ID, and namespace), and edit time. Row-level examples of *MAC* are presented in Appendix Table 4.

2.3 *MAC* Analysis

In Table 2, we present basic statistics for the ten most dominant languages in *MAC*. The complete statistics list with each language in the corpus is found in Appendix Table 3. Here, *MAC* is used to train and evaluate AV models. Therefore, we monitor and report the number of authors with at least two contributions, as those authors are helpful for training AV models. The two rightmost columns in Table 2 highlight two uniquenesses of *MAC*—cross-domain and cross-language contributions. Since each text chunk is associated with a specific Wikipedian (i.e., author), we are able to track authors even when they contribute in different domains and languages. This tracking ability is thanks to Wikipedia's unified login mechanism.[9]

[8]We only remove mixed languages content when the primary language is non-English 'mixed' with English text.

[9]Unified login: https://tinyurl.com/4j6nfe8n

		Author Statistics		
Language	# T.Chunks	≥2 Contrib.	≥2 Domains	≥2 Langs.
English (en)	22.95M	618.76K	294.83K	93.47K
Gernam (de)	5.52M	109.89K	59.55K	23.77K
French (fr)	6.77M	85.88K	39.28K	20.82K
Spanish (es)	3.16M	73.64K	28.49K	17.61K
Russian (ru)	2.48M	49.32K	17.32K	12.90K
Italian (it)	2.21M	46.73K	19.38K	9.62K
Portuguese (pt)	1.86M	35.09K	11.22K	7.62K
Polish (pl)	1.66M	29.81K	16.03K	4.98K
Dutch (nl)	1.15M	20.48K	8.76K	6.26K
Ukrainian (uk)	883.37K	15.05K	4.41K	5.80K
All (60 Langs.)	60.08M	1.29M	568.30K	253.37K

Table 2: *MAC* statistics. We present the top ten dominant languages in *MAC* and the total (i.e., sum) numbers over all languages in the corpus. 'T.Chunks' and 'Contrib.' refer to Text Chunks and Contributions, respectively.

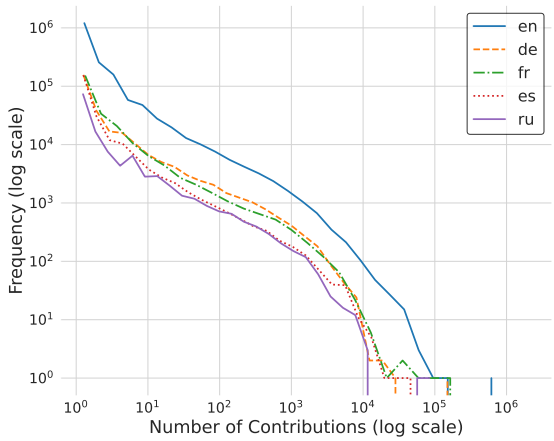
This way, users keep their username while editing different page types (i.e., domains) and languages.

English (en) is the most dominant language in the corpus, aligned with the pattern in other Wikipedia corpora (Guo et al., 2020; Perez-Beltrachini and Lapata, 2021). 48% of the authors in *MAC* are English-speaking.

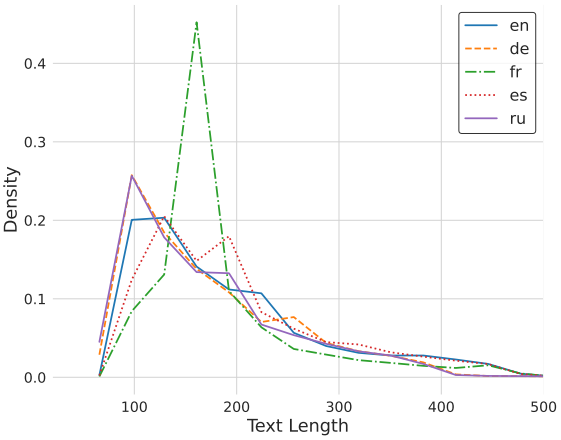
***MAC* analysis** *MAC* comprises both authors (i.e., Wikipedians) and their textual contributions. Accordingly, we examine the distribution of authors across languages, as well as the length of each textual instance in *MAC*. Figure2 presents these two distributions: a log-log plot of author contributions and a histogram of text lengths. As expected, the distribution of author contributions closely follows a power-law pattern, while the distribution of text lengths is left-skewed, reflecting the dataset's minimum length requirement of approximately 100 words per text.

Authors Overlap A practical way to analyze *MAC* is to calculate the overlap of authors between language pairs. We expect 'close' languages, such as Ukrainian and Belarusian, to have a high authors' overlap as both are East Slavic languages, with a high lexical similarity, commonly spoken by many in the post-Soviet states. By applying this approach across all language pairs, we construct a network in which nodes represent languages, and weighted edges denote connections based on the extent of author overlap.

Appendix Figure 5 illustrates this network. As expected, the most edited Wikipedia languages (e.g., English, German) dominate the network. Closely related languages such as Ukrainian-Russian-Belarusian and Arabic-Egyptian-Farsi



(a) Authors distribution



(b) Text length distribution

Figure 2: *MAC* characteristics. Subfigure (a) demonstrates the power-law distribution of authors' contributions. Subfigure (b) displays the left-skewed distribution of text lengths, with French being a notable exception. For clarity, we present only the five most prevalent languages in *MAC*.

maintain tight connections in the network.

3 Authorship Verification Experiments

Here, we introduce the AV tasks and experiments.

3.1 Task Definition

AV aims to identify whether two texts are written by the same author. Following recent works (Rivera-Soto et al., 2021; Zhu and Jurgens, 2021; Wegmann et al., 2022; Fincke and Boschee, 2024), we adopt a similarity-based verification approach in which a model is trained to compute similarity scores between a pair of texts and assign a high score to texts written by the same author. Formally, given a set of candidate *texts* $\mathcal{T} = \{t_1, \cdots, t_n\}$, and an *authorship* function $\mathrm{A}(t)$, where $\mathrm{A}(t_i) = \mathrm{A}(t_j)$ iff t_i and t_j are written by the same author, the goal of AV is to train a model M such that $\mathrm{M}(t_i, t_j) > \mathrm{M}(t_i, t_k)\ \forall\ i, j, k \in [n]$ where $\mathrm{A}(t_i) = \mathrm{A}(t_j) \neq \mathrm{A}(t_k)$. In practice, since $\mathcal{T}$ may be too large to exhaustively evaluate M for all pairs, and an operator of an AV model may wish to obtain authorship matches for a small set of candidate texts (e.g. for plagiarism detection), AV is formulated as an information retrieval (IR) task, where a small query set of texts is used to probe a much larger candidate set. Success of M is demonstrated (via standard IR metrics; see Section 3.2.4) by its ability to correctly assign scores such that a candidate text matching a query's author is found.

3.2 Experimental Setup

3.2.1 Data Post-processing

To evaluate AV models in an IR manner, we restructure *MAC* as an IR task and thus denote query-candidate pairs for training, validation, and test. Specifically, for each author in each language, we extract one pair of texts. To maximally reduce the effect of content in AV (Wegmann et al., 2022), we select the hard positive pair with the lowest SBERT (Reimers and Gurevych, 2019) cosine similarity if we have more than two texts to choose from. Then, per language, we split all extracted text pairs into training, validation, and test sets in a 7:1:2 ratio.

We only keep text pairs from domain 0 in the training and validation sets to hold out the other domains for evaluating out-of-domain generalizability. We also filter out texts of more than 300 words to avoid the risk of including translated texts. The text pair selection (and length filtering) substantially reduce the amount of data used in the AV experiments, thereby leading to the different number of authors in Table 2 ($\geq$2 Contrib.) and Table 8. For convenience, we denote the set of query and candidate texts as $\mathcal{Q} = \{q_0, \cdots, q_n\}$ and $\mathcal{C} = \{c_0, \cdots, c_n\}$, respectively.

3.2.2 AV Models

We consider two categories of models: off-the-shelf IR models and fine-tuned authorship representation models. The IR models serve as baselines for assessing the performance gain from fine-tuning.

IR Model We select two representative IR models: **BM25** (Robertson et al., 1994) and **SBERT** (Reimers and Gurevych, 2019). BM25 is one of the most widely used information retrieval (IR) models. Given a pair of texts, a relevance score is calculated based on token matching. We use this relevance score as the AV similarity score sim_{ij} for each pair

of texts q_i and c_j. Unlike BM25, SBERT is a neural representation model that transforms texts into embedding vectors, tuned explicitly for semantic similarity. For two texts q_i and c_j, we use the cosine similarity of their embedding vectors, $\mathbf{v}_{q_i}$ and $\mathbf{v}_{c_j}$ as their AV similarity score sim_{ij}, which can be formally defined as $\text{sim}_{ij} = \frac{\mathbf{v}_{q_i} \cdot \mathbf{v}_{c_j}}{\|\mathbf{v}_{q_i}\| \|\mathbf{v}_{c_j}\|}$.

Authorship Representation Model We evaluate two models directly fine-tuned on *MAC*. One is trained with the Sentence Transformer codebase (Reimers and Gurevych, 2019) using the multiple negatives ranking loss. For convenience, we refer to this model as **SBERT$_{AV}$** to distinguish it from the semantic SBERT in the previous category. The other is the state-of-the-art AV system **SADIRI** (Fincke and Boschee, 2024) that uses hard positives and in-batch hard negatives, which are negative examples whose authorship is difficult to distinguish from the anchor text, to improve the model performance in more challenging situations. Note that, in this work, we extract the hard positive pairs during data post-processing, so the main difference between SBERT$_{AV}$ and SADIRI is the hard negative batching technique.

We use the same equation for SBERT to calculate the AV similarity score for both SBERT$_{AV}$ and SADIRI. In this study, we only focus on AV models trained in a single language. Hence, we train an SBERT$_{AV}$ and a SADIRI model for each language in the training set so that we can hold out other languages and domains to test the generalizability of the models in RQs 2 and 3.

3.2.3 Evaluation Setup

We first train an SBERT$_{AV}$ and a SADIRI model on domain 0 data for each of the top 10 languages in *MAC* according to the size of the domain 0 data. We then evaluate these trained models, as well as the BM25 and SBERT baselines, on held-out test sets from *MAC* to examine the five RQs introduced in Section 1. Please see Appendix Section D.3 for training and test set statistics.

For RQs 1 to 3, we create a single-language single-domain test set for each language L and domain D by extracting a text pair with two texts in D from the test split of L for each author. Then, in **RQ1**, we examine the **in-language in-domain** performance of models using in-language domain 0 test sets; in **RQ2**, we examine the **out-of-language generalizability** of models using out-of-language in-domain[10] test sets. In particular, we evaluate the models on domain 0 single-language test sets in all unseen languages; in **RQ3**, we examine the **out-of-domain generalizability** using in-language[10] out-of-domain test sets. In particular, we evaluate the models on in-language single-domain test sets for all unseen domains (domains 1 to 3) and take the average score across all evaluation domains for each AV approach.

In **RQ4**, we examine the **cross-lingual verification** performance using a cross-lingual in-domain test set. This is created by gathering all authors in the test split for all languages, and for each of those who have contributions in domain 0 in at least two languages, constructing a single text pair using two texts in domain 0 randomly chosen from two different languages. Also, we always choose two texts from *different articles* to ensure they are not translations of each other. The same test set is used for all models; in **RQ5**, we examine the **cross-domain verification** performance using in-language cross-domain test sets. In particular, for each language, we construct a single text pair by randomly matching texts in different domains for each author who has contributions in at least two domains in the test split.

Note that, RQ2 (respectively, RQ3) is similar to RQ4 (respectively, RQ5) because they both involve multiple languages (respectively, domains). However, they differ fundamentally in how the document pairs are constructed. In RQ2 (respectively, RQ3), the two documents for each author are in the same language (respectively, domain), which is different from the language (respectively, domain) of the training data, whereas in RQ4 (respectively, RQ5), the two documents for each author are in different languages (respectively, domains).

3.2.4 Evaluation Metrics

We evaluate the AV models using Success@k, which is one of the most commonly used metrics in the AV literature (Khan et al., 2021; Rivera-Soto et al., 2021; Fincke and Boschee, 2024).[11] For-

[10] We use "in-domain" (respectively, "in-language") to indicate that the domain (respectively, language) for evaluation is the same as that used for training the fine-tuned models. For the two off-the-shelf IR baselines, we use the same model for all domains and languages.

[11] Some of them use Recall@k which is equivalent to Success@k in this work because we have exactly one candidate for each query in the evaluation sets.

mally, given two sets of texts $\mathcal{C}$ and $\mathcal{Q}$,

$$\text{Success@}k = \sum_{i=1}^{|\mathcal{Q}|} \mathbb{I}\left(\bigvee_{r=1}^{k} \text{A}(c_{r_i}) = \text{A}(q_i)\right) / |\mathcal{Q}| \quad (1)$$

where c_{r_i} refers to the text in $\mathcal{C}$ that has the r^{th} highest AV similarity to q_i w.r.t. an AV model. For each experiment, we form $\mathcal{Q}$ (respectively, $\mathcal{C}$) from the query (respectively, candidate) side of each pair of a validation or test subset, as described in Section 3.2.1. Validation subsets are used to choose the final model in each experiment, and we report results on the test subsets only. We use $k = 1$ as the main metric in this work for a strict evaluation. However, we also present results while using $k = 8$ in Appendix Section E.

3.2.5 Implementation Details

We use OkapiBM25Model from the Gensim Python library (Řehůřek and Sojka, 2010) for BM25 and the Sentence Transformer Python library (Reimers and Gurevych, 2019) for SBERT models and training. We use an implementation of the SADIRI system based on its description in the original paper (Fincke and Boschee, 2024). We use *only multilingual* language models in our experiments. We use paraphrase-multilingual-mpnet-base-v2 as the base model for semantic SBERT and the xlm-roberta-base (Conneau et al., 2020) as the base model for all fine-tuned authorship representation models. Please see Appendix Section D.1 for all hyperparameters.

4 Results

In this section, we present the evaluation results for all AV models on all five RQs. Due to space limitations, we mainly discuss the results for AV models trained on English data. Please see Appendix Section E for results over the rest of the languages. The results for RQs 1 to 5 are shown in bar groups 1 to 5 in Figure 3, respectively.

RQ1: Training on in-language in-domain data improves AV model performance, as shown in the first bar group in Figure 3; here, the two fine-tuned models (SBERT$_{AV}$ and SADIRI) perform better than both IR baselines (BM25 and SBERT), and between the two fine-tuned models, SADIRI outperforms SBERT$_{AV}$, which aligns with our expectation because we expect the hard negative batching of SADIRI benefits the model when evaluated on the same distribution as training data.

RQ2: Trained AV models can generalize to unseen languages, as seen in the second bar group in Figure 3. The trend also closely resembles the patterns observed in RQ1, but the performance gain over BM25 is slightly less than that in RQ1, which indicates that while the fine-tuned models are generalizable to other languages, the improvement in out-of-language performance is relatively smaller. The fine-grained results on the top five languages in Figure 4 also indicate that for each evaluation language, the models trained in the same language perform the best or just worse than the model trained in English data in most cases, which falls within our expectation because the English subset of *MAC* contains the most data.

RQ3: In contrast to RQ2, fine-tuned models do not generalize to new domains, as seen in the third bar group in Figure 3. Neither fine-tuned model is better than the strongest baseline, BM25, and SBERT$_{AV}$ even exhibits a worse performance than BM25, which suggests that out-of-domain generalization is harder than out-of-language generalization for AV, and training in single domain data may even harm the performance in unseen domains. Note that, although the task in RQ3 is harder than RQ1, the scores in RQ3 are higher than RQ1. The main reason is that the size of the test set for RQ3 is much smaller than RQ1, which makes the retrieval in RQ3 easier.

RQ4: Unlike the observations in RQs 2 and 3, fine-tuned AV models may or may not generalize to identifying authors across language, depending on the training algorithm used, as shown in the 4$^{\text{th}}$ bar group in Figure 3. This result exhibits a markedly different pattern where SBERT also outperforms BM25, and SBERT$_{AV}$ outperforms both BM25 and SADIRI. BM25 performs the worst mainly because it relies on exact token matching, which does not work across languages with different vocabularies, while a potential reason for the underperformance of SADIRI is that its training strategy overemphasizes in-language in-domain hard negative optimization which harms cross-lingual performance.

RQ5: Fine-tuned models do not generalize to identifying authors across domains, as shown in the 5$^{\text{th}}$ bar group in Figure 3. Both trained models are worse than the strong baseline BM25 and marginally better than the weak baseline SBERT. The worse performance of SADIRI differs from the observation in Fincke and Boschee (2024) that training on hard positives and in-batch hard negatives dramatically improves the performance on cross-genre AV. One possible reason for the in-

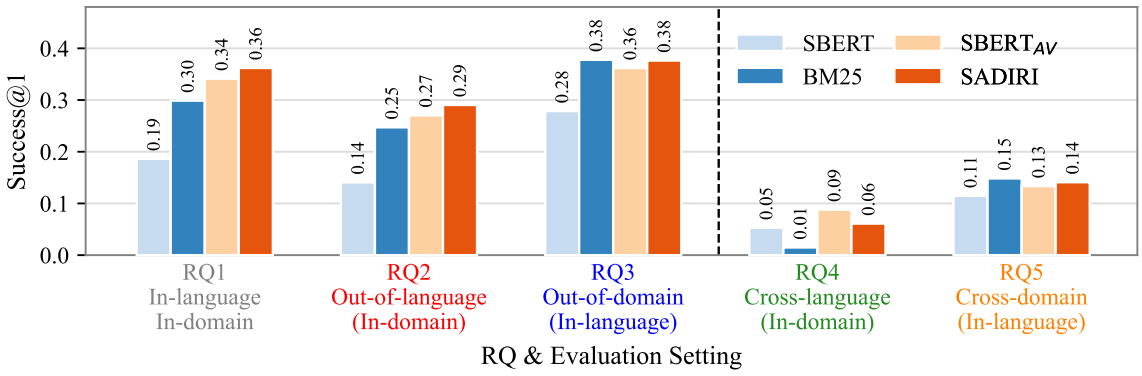


Figure 3: Success@1 scores for the two baselines (SBERT and BM25) and the two fine-tuned models (SBERT$_{AV}$ and SADIRI) trained on English data. The scores for RQ2 are the averages across only unseen evaluation languages, and the scores for RQ3 are the averages across only unseen domains (domains 1 to 3). In general, the performance of the four models exhibits similar patterns on RQs 1 and 2, whereas the patterns on the other RQs, especially RQs 4 and 5, are markedly different. All differences in this figure are statistically significant based on paired t-tests ($p < 0.05$), except for the difference between BM25 and IR-AA in RQ3 and the difference between IR-AA and SBERT in RQ4.

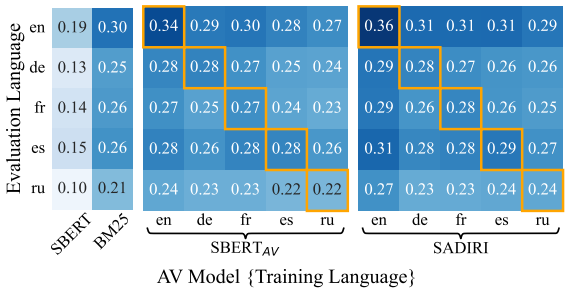


Figure 4: The Success@1 scores for out-of-language generalization for the top five languages. The scores highlighted in the orange boxes are the in-language scores for which the models are trained and evaluated in the same language. Boxes in this figure are auto-shaded such that darkness correlates with Success scores across the entire table.

consistency is that we only trained the models on single-domain data, which significantly degraded the quality of the hard positives, thereby constraining the models' cross-domain generalizability. Comparing with the training data in Fincke and Boschee (2024), it reveals that training on multi-domain data or data from various sources may be the key to stronger cross-lingual generalizability.

5 Discussion

Overall, as seen in Figure 3, *MAC* enables the comparison of AV models under conventional scenarios such as those on the left (RQs 1 to 3). The results indicate that fine-tuned AV models perform better than IR baselines for AV in the same language and domain as training. Fine-tuned models also generalize well to AV in the same domain but unseen languages, but do not generalize to unseen domains at all (RQ3, 3rd bar group in Figure 3). Over the other 2 RQs, which are only assessable using *MAC*, the results are starkly different on RQ4, whereas the pattern on RQ5 closely resembles that in RQ3. In general, the results indicate that AV models' performances are not consistent across different evaluation conditions, so to have a more comprehensive understanding of the capability of AV models, testing them under different conditions is necessary. Existing datasets can cover some conditions, but *MAC* enables the evaluation of other situations like cross-lingual AV and cross-domain AV in different languages.

Moreover, the poor performance of the fine-tuned models in RQs 3 to 5 reveals challenges in AV involving multiple domains and multiple languages, which may serve as new directions for future research in this area, and *MAC*, as it contains high volumes of data from multiple domains and languages, enables researchers to build models to tackle these specific aspects of AV.

While building and evaluating *MAC*, we paid extra attention to excluding translation cases. Furthermore, a practical way to avoid the over-representation of accounts that are shared by multiple people is to limit the number of samples we use per author (while running experiments).[12] We use this approach in our research and extract only one pair of texts for each author in each language.

[12]This assumes that most accounts are not shared by multiple people.

6 Related Work

We first summarize AV-related research and then discuss existing Wikipedia corpora used in NLP.

AV Models With the development of NLP in the past few decades, the techniques in AV have evolved through three main stages. In the pre-neural era, AV works mainly focused on feature-based approaches (Koppel and Schler, 2004; Stamatatos, 2009; Stolerman et al., 2014). Later, with the widespread use of neural networks (NN), especially (pre-trained) transformers, NN-based representation models have become mainstream and demonstrated superior performance on large-scale AV (Rivera-Soto et al., 2021; Zhu and Jurgens, 2021; Wegmann et al., 2022; Fincke and Boschee, 2024). Recently, leveraging the strong zero- and few-shot capability of LLMs, some studies showed that LLMs could achieve a decent performance on AV without further downstream fine-tuning, but the findings are limited to small-scale AV (Hung et al., 2023; Huang et al., 2024). In this work, we concentrate on large-scale AV due to its higher practicality and, therefore, primarily adopt NN-based representation models in our experiments.

AV Datasets Datasets also play a crucial role in AV research for both model training and evaluation. AV requires corpora with authorship labels which could be collected from various sources such as email (Klimt and Yang, 2004), newswire agency (Lewis et al., 2004), newspaper (Stamatatos, 2013). Data from these sources often contain a relatively small number of authors. With the increasing number of internet users, later works started to collect data from public websites such as Amazon (Ni et al., 2019; Keung et al., 2020) and Reddit (Rivera-Soto et al., 2021). Datasets from these sources are substantially larger in scale and serve as the foundation of the SOTA AV models (Rivera-Soto et al., 2021; Fincke and Boschee, 2024). These datasets contain up to one million authors (Rivera-Soto et al., 2021) and texts in six languages (Keung et al., 2020), while *MAC* is 10 times larger in terms of both the numbers of authors and languages. Moreover, *MAC* also contains texts from four domains and authors writing in multiple languages and domains, enabling a more comprehensive evaluation of the AV models presented in Section 3.2.3.

Wikipedia Corpora Wikipedia serves the research community as a textual source in various domains and NLP tasks (Pan et al., 2017; Kaffee et al., 2018; Zhu et al., 2019; Sathe et al., 2020; Guo et al., 2020; Ta et al., 2023). A series of studies have used information about Wikipedia editors to create data sets that associate text with authors (Yang et al., 2017; Maki et al., 2017; Jaidka et al., 2021) and others have leveraged edit histories for a variety of purposes such as simplification (Yatskar et al., 2010) or grammatical error correction (Faruqui et al., 2018). These and other Wikipedia-related corpora have been released under the 'CC BY-SA' Wikipedia license terms, which we also rely on while releasing *MAC* to the research community.

Most Wikipedia-related studies focus on a single language, usually the most popular ones, such as English or German. While some others (Pan et al., 2017) leverage Wikipedia's multilingual capability, none tackle the broad combination of languages and domains as we do in *MAC* by extracting long textual content from edits done by Wikipedians.

7 Summary

Our novel *Million Authors Corpus* (*MAC*) spans 60 languages and four domains, linking long textual contributions to their respective authors while excluding minor edits. *MAC* consists of more than 60M cleaned text passages from contributions by Wikipedian authors. Here, we study AV models using the 1.29M authors who made 2+ contributions in *MAC* across the 60 languages; more than 560K authors contributed in different domains, such as an article page and a talk page. In addition, *MAC* includes multilingual authors who contributed to at least two languages.

MAC enables new studies of cross-lingual and cross-domain AV evaluations; i.e.,, given a text written by an author, finding another text by the same author written in a different language or domain. Our study demonstrates that in conventional AV scenarios, a selection of fine-tuned models surpass IR baselines on RQ1 and RQ2, but not on RQ3. However, we show that in new novel and challenging settings of cross-lingual and cross-domain AV (RQ4 and RQ5), we observe starkly different patterns and models, as a whole, perform substantially lower. Our results highlighting the value of the comprehensive AV model evaluation powered by *MAC* and point to the future research directions of AV in cross-domain and cross-lingual modeling.

8 Limitations

We process 60 languages out of more than 300 languages that exist in Wikipedia. Unfortunately, many languages have a relatively small user base and content, which is irrelevant to our study. We select the 60 major Wikipedia languages in terms of the number of pages and user base. Notably, we *do not* include the Cebuano Wikipedia[13] in *MAC*, although it is the second largest Wikipedia project, articles-wise. The reason is the extremely low number of contributors to this project—153 active users. We also limit our study to the four main page types in Wikipedia (i.e., namespaces 0 to 3) out of the 26 page types Wikipedia maintains.[14]

Another limitation of the new corpus is the association between text and users. Page edits on Wikipedia are automatically recorded by those who edit the page. However, we do not control the way people generate their textual content and whether those are 'copied-paste' or translated from other sources. This limitation is mostly relevant to article pages (namespace 0) rather than user or talk pages.

However, we handle translations to some extent using two methods: (i) We filer out from *MAC* too long textual contributions (see Step 2 in Section 2.2), which are likely to be an article translation; and (ii) While evaluating the models for **RQ4**, we do not include the same article for the same author in two different languages (see Section 3.2.3). Most of the Wikipedia content, and so of *MAC* has been created before the GenAI era, where both humans and LLMs create textual content.

In our study, we make sure to cover different domains within Wikipedia by including data from different namespaces. However, future research is needed to establish the degree to which authorship verification in the Wikipedia domains is portable to *other domains* and vice versa. For example, literary authorship style may not be captured here, nor may authorship identity over many years of an author's lifetime.

9 Ethical Considerations

This research relies exclusively on publicly available data from Wikipedia, adhering strictly to ethical standards for data collection and analysis. All data used in this study are accessible to the public through Wikipedia's platform, ensuring transparency and compliance with open data principles.

We adhere to the terms and conditions specified by Wikipedia's licensing framework, including the Creative Commons Attribution-Share license (CC BY-SA). All derivative works, analyses, and datasets generated from Wikipedia data comply with these licensing requirements, ensuring proper attribution and alignment with open knowledge practices.

Finally, we use the data solely for research and modeling purposes, advancing the research of AV models in the NLP domain while respecting the rights and intentions of the Wikipedia community. By following these principles, we aim to balance the pursuit of scholarly insights with the ethical responsibility of protecting individuals and maintaining data integrity.

Authorship verification is a dual-use technology with societal benefits and risks. While AV—and forensic linguistics in general—has been used to help law enforcement identify individuals involved in crimes such as human trafficking (e.g., Olsson, 2009; Saxena et al., 2023) and has uses in historical and cultural applications (e.g., Gurney and Gurney, 1998; Juola et al., 2008), such technology could also be used maliciously, such as to de-anonymize individuals posting under pseudonyms. We publish *MAC* to facilitate the research in AV models that can be useful in various areas. Such malicious usages have the potential to threaten internet users' privacy and other unnecessary effects. Therefore, we will release *MAC* with guidance to forbid such malicious use cases and intend it solely for research purposes.

Acknowledgments

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

[13] Cebuano Wikipedia: https://tinyurl.com/mr3d6z8h

[14] Wikipedia namespaces: https://tinyurl.com/46jw4u7u

References

Phoebe Ayers, Charles Matthews, and Ben Yates. 2008. *How Wikipedia works: And how you can be a part of it*. No Starch Press.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Overview of pan 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–383, Cham. Springer International Publishing.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A multilingual corpus of Wikipedia edits for modeling language and discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium. Association for Computational Linguistics.

Steven Fincke and Elizabeth Boschee. 2024. Separating style from substance: Enhancing cross-genre authorship attribution through data selection and presentation. *arXiv preprint arXiv:2408.05192*.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Penelope J Gurney and Lyman W Gurney. 1998. Authorship attribution of the scriptores historiae augustae. *Literary and Linguistic Computing*, 13(3):119–131.

Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. 2023. STEER: Unified style transfer with expert reinforcement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7546–7562, Singapore. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Martin Haspelmath. 2005. *The world atlas of language structures*. Oxford university press.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.

Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Lee. 2023. Who wrote it and why? prompting large-language models for authorship verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14078–14084, Singapore. Association for Computational Linguistics.

Kokil Jaidka, Andrea Ceolin, Iknoor Singh, Niyati Chhaya, and Lyle Ungar. 2021. WikiTalkEdit: A dataset for modeling editors' behaviors on Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2191–2200, Online. Association for Computational Linguistics.

Patrick Juola et al. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018. Learning to generate Wikipedia summaries for underserved languages from Wikidata. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 640–645, New Orleans, Louisiana. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. A deep metric learning approach to account linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5275–5287, Online. Association for Computational Linguistics.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

Shuai Liu and Jonathan May. 2024. Style transfer with multi-iteration preference optimization. *ArXiv preprint*, abs/2406.11581.

Keith Maki, Michael Yoder, Yohan Jo, and Carolyn Rosé. 2017. Roles and success in Wikipedia talk pages: Identifying latent patterns of behavior. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1026–1035, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Olena Medelyan, David Milne, Catherine Legg, and Ian H Witten. 2009. Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. "the sum of all human knowledge": A systematic review of scholarly research on the content of w ikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245.

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

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

John Olsson. 2009. *Wordcrime: Solving crime through forensic linguistics*. A&C Black.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from Wikipedia. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6874–6882, Marseille, France. European Language Resources Association.

Vageesh Saxena, Benjamin Ashpole, Gijs van Dijck, and Gerasimos Spanakis. 2023. IDTraffickers: An authorship attribution dataset to link and connect potential human-trafficking operations on text escort advertisements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8444–8464, Singapore. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Engelson Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Assoc. Inf. Sci. Technol.*, 60:538–556.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n -gram features. *Journal of law and policy*, 21:7.

Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. 2014. Breaking the closed-world assumption in stylometric authorship attribution. In *Advances in Digital Forensics X*, pages 185–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hoang Thang Ta, Abu Bakar Siddiqur Rahman, Navonil Majumder, Amir Hussain, Lotfollah Najjar, Newton Howard, Soujanya Poria, and Alexander Gelbukh. 2023. Wikides: A wikipedia-based dataset for generating short descriptions from paragraphs. *Information Fusion*, 90:265–282.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *ArXiv preprint*, abs/2209.06869.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368, Los Angeles, California. Association for Computational Linguistics.

Lei Zheng, Christopher M Albano, Neev M Vora, Feng Mai, and Jeffrey V Nickerson. 2019. The roles bots play in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qi Zhu, Xiang Ren, Jingbo Shang, Yu Zhang, Ahmed El-Kishky, and Jiawei Han. 2019. Integrating local context and global cohesiveness for open information extraction. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 42–50. ACM.

A Revisions Filtering

As we highlight in Section 2, we filter some of the revisions (i.e., edits). In this Section, we further explain the two criteria we apply to filter some of the revisions.

Not all edits of a Wikipedia page should be considered while building *MAC*. We consider two scenarios where we omit R_p^i. The first is when R_p^i is later deleted by a Wikipedia admin. In such a case, the textual content of the deleted revision appears as empty text in the dump file. It makes no sense for this to be considered a valid edit. The second case is when a user makes a sequence of consecutive edits to a page. If so, we 'aggregate' their contributions and consider only the last revisions out of the sequence. These two special cases can explain why R_p^* is not always the latest revision we processed but rather the last relevant revisions which R_p^i should be compared against.

B Edit Filtering

Each edit in MAC is intended to be a sufficiently long piece of text to encoder some aspect of an author's style. While stylometric work has generally used used longer text on the order of several hundred word (Neal et al., 2017), shorter text still contain rich style information and, given their higher frequency in data, are valuable for learning to recognize style. Here, we adopt a threshold-based approach for selecting edits, where any contiguous edit of at least α words is counted. We adopt α=100 for English for three reasons: (1) this length is in the same range as more recent and challenging AV datasets (e.g., Tyo et al., 2022), (2) the shorter length allows models trained on MAC to potentially generalize to other common short-document domains such as reviews and social media, which are more common in practice, and (3) recent shared research programs such as IARPA HIATUS have adopted a 100 word document length in their own metrics.[15]

Setting α=100 for all languages will not work well, however. Languages vary in their information density per word (Comrie, 1989), with some languages having much shorter or longer sentences than English due to morphology (Haspelmath, 2005). There is no standard approach to determine how to rescale such a threshold for all languages. Therefore, we adopt a data-driven approach to set α for each language based on its relative morphological density. Here, for each language ℓ, we first calculate the mean number of characters per whitespace-separated token, $\overline{w}_\ell$. Note that we intentionally avoid using an LLMs tokenizer to decide on intra-token units (i.e., from subwords), as these tokenizers' outputs vary significantly in our data based on whether the tokenizer was originally fit to language in that data so subword units for different languages would not be equivalently scoped.

We then rescale α for each language as $\frac{100\overline{w}_{\text{English}}}{\overline{w}_\ell}$, rounding to the nearest whole character. For example, Russian had 5.66 characters per tokens, compared with 4.80 for English, roughly correlating with the more complex morphology of Russian; as a result α=85 for Russian where shorter texts are allowed to account for richer intra-token stylistic variation. By rescaling α using in-domain data, we aim to directly match authors' writing behaviors in the data.

C Data Cleaning – Further Information

We execute the cleaning process of *MAC* while maintaining a "precision-biased" approach – prioritizing methods that ensure the cleanliness of the corpus at the expense of potentially excluding some relevant textual content. We apply the following three filters:

1. **Tables**. Wikipedia contains information not only in the format of textual sentences/paragraphs but also in tables. To avoid minor contributions from users who add/edit big tables (which do not contain much text), we remove all tables from all page revisions.

2. **Bots**. Bots operate in Wikipedia (Zheng et al., 2019). Systematically, most of their 'contributions' are not included in *MAC* as we only capture long textual contributions, while bots are less likely to do so. However, we apply a simple rule in the English Wikipedia to detect bots. We filter out contributions of usernames that start/end with the 'bot' regex.[16]

3. **Mixed languages**. Naturally, the dominant language per project is the project's language (e.g., Italian in the Italian Wikipedia). However,

[15] https://www.iarpa.gov/images/PropsersDayPDFs/HIATUS/IARPA-BAA-22-01_HIATUS_Amend001_C.pdf

[16] We apply this regex rule as case insensitive.

Wikipedians are not limited to writing in a specific language. It is most common for users to use English on Talk pages in non-English Wikipedia projects. To avoid having a lot of English content in non-English projects, we apply a language detection method[17] and remove cases of English text used in non-English Wikipedia projects.

D More Experimental Details

D.1 Hyperparameters

For a more controlled comparison between SBERT$_{AV}$ and SADIRI, we train them using the same set of hyperparameters listed in Table 5. Because the sizes of the subsets vary substantially by language, to ensure a similar number of gradient updates in one epoch for each language, we use different batch sizes for different languages. For BM25, we use the default hyperparameters for OkapiBM25Model in Gensim, which are listed in Table 6.

D.2 Hardware & Runtime

The *MAC* creation process over 60 languages took 16.5 days while using 100 Intel Xeon Gold 6138 CPU cores. All AV models in this work are trained using a single NVIDIA A40-48GB GPU. The AV models training times for SBERT$_{AV}$ and SADIRI are listed in subsection D.2.

D.3 Training & Test Sets Statistics

We list the number of query-candidate pairs for each subset in each language in Table 8. The numbers also represent author counts because we keep exactly one pair of query and candidate texts for each author in all training, validation, and test subsets. For RQ4, we only have a single cross-lingual test set with 1,635 text pairs (authors).

E More Experimental Results

We present the out-of-language generalization scores for all top ten languages in Figure 6 and the scores for RQs 1 to 5 for the two baselines and the two fine-tuned models trained on each of the top ten languages in Figure 7 and Figure 8.

In general, the two fine-tuned models trained in the top five languages consistently outperform the two IR baselines for RQ1, but for RQ2, only the models trained in the top three languages are consistently better than the IR baselines; for other training languages, the results are mixed for RQs 1 and 2. This observation suggests that training an AV model with decent in-language in-domain performance requires a certain amount of data, and this requirement on data size is larger for in-domain out-of-language generalization.

In contrast, the observations for all top ten languages on the other three RQs are generally consistent with the results for English in Figure 3. Specifically, for RQs 3 and 5, both trained models are worse than the strongest baseline BM25 in most cases; otherwise, their performances are the same; for RQ 4, both trained models consistently outperform the two IR baselines with only two exceptions (SADIRI model trained on German and Ukrainian).

We use Success@1 in the main result analysis because some of our evaluation sets a small and Success@1 can provide stricter results on small dataset. However, we also present Success@8 results in Figure 6, Figure 7, and Figure 8 for reference. The patterns of Success@8 results are generally consistent with those observed in Success@1 results. The only main difference is that SBERT$_{AV}$ outperforms SADIRI in most cases.

F Artifacts

F.1 Created Artifacts

In this study, the main artifact we make is *MAC*—a dataset of Wikipedia textual content associated with its author. The new corpus will be released for research use under the general Wikipedia licensing framework, including the Creative Commons Attribution-Share license (CC BY-SA). We further detail this in Section 9.

F.2 Use of Existing Artifacts

Table 9 outlines the models and code libraries employed in this work. Our use of these artifacts is consistent with their intended use.

G Use of AI Assistants

We use Grammarly and ChatGPT to check grammar and polish our manuscript, but only for grammar corrections and minor edits. We take full responsibility for all the content in our manuscript.

[17] Pycld2 package: https://tinyurl.com/4ewazbkp

	General Statistics			Author Statistics		
Language	# T.Chunks	# Chars	# Words	≥2 Contrib.	≥2 Domains	≥2 Langs.
Africaans (af)	41942	54591376	7848343	446	190	476
Amharic (am)	2929	4066547	713146	81	24	79
Arabic (ar)	2611219	2281733767	357439206	20988	7447	4472
Egyptian Arabic (arz)	307873	242898738	35042567	341	111	534
Belarusian (be)	118570	146621963	17866988	939	251	915
Bulgarian (bg)	265511	349182870	48966887	4823	1507	1405
Bengali (bn)	27561	50031224	5839917	178	43	132
Chechen (ce)	439	562391	68339	17	10	38
Cherokee (chr)	59	78955	11556	6	1	6
Czech (cs)	646080	849184442	114027574	12370	4310	2383
Gernam (de)	5525478	7257565213	934707414	109890	59549	23767
Greek (el)	295937	396833046	53634195	5869	2199	1959
English (en)	22953734	28310105431	4357302494	618758	294826	93465
Spanish (es)	3164341	4130539248	629555918	73644	28496	17605
Persian (fa)	1146456	1207151742	204651153	20381	7182	2944
Fula (ff)	96	128576	17060	6	0	14
Finnish (fi)	239296	428914597	44023582	2794	1323	1936
French (fr)	6766096	8576232169	1272364513	85884	39282	20821
Gujarati (gu)	8988	10578958	1547331	311	118	240
Hausa (ha)	23437	33542354	5385245	259	40	118
Hebrew (he)	476210	617562793	97213800	13278	6599	2147
Hindi (hi)	90692	120210890	21155221	4046	1449	2434
Hungarian (hu)	568839	709201304	85380427	10610	4450	1973
Armenian (hy)	198688	268542332	32330669	4269	2183	749
Indonesian (id)	404223	503307831	64705169	11180	2725	2763
Icelandic (is)	33321	37953918	5241063	828	294	384
Italian (it)	2211421	2842813803	413914078	46727	19382	9623
Japanese (ja)	599453	580791400	42052366	21560	5107	2497
Javanese (jv)	20529	24534947	3135389	434	119	328
Georgian (ka)	110869	128682863	13990251	1539	455	442
Kazakh (kk)	82239	97437119	10856557	1767	328	463
Malagasy (mg)	17806	22237969	2777122	48	12	79
Macedonian (mk)	83544	109737982	15301346	1447	284	528
Malayalam (ml)	254186	407844848	34966538	1635	578	798
Marathi (mr)	34176	38964783	5104805	1072	333	528
Burmese (my)	25050	32277716	2346444	466	130	192
Mazanderani (mzn)	1867	2184441	358170	39	18	72
Dutch (nl)	1149502	1455471618	205590939	20481	8757	6257
Punjabi (pa)	26763	38065954	6661363	509	90	228
Polish (pl)	1658171	2269438656	291628512	29805	16031	4987
Portuguese (pt)	1861950	2200190604	329245240	35091	11217	7615
Russian (ru)	2475746	3243405352	410405284	49317	17319	12900
Simple English (simple)	1042	1648009	207574	79	19	166
Serbian (sr)	311606	379683724	51996543	5357	2121	1535
Sundanese (su)	10853	12784831	1631605	223	51	180
Swedish (sv)	654611	798513490	108859092	13227	5407	4015
Swahili (sw)	19860	24114993	3289445	601	131	281
Telugu (te)	106079	155946105	17814054	800	295	397
Thai (th)	228603	209536522	13635803	6014	1799	947
Tagalog (tl)	33436	42459061	6052162	590	175	525
Turkish (tr)	658897	950765875	109644757	12602	3709	2812
Tatar (tt)	23609	33501616	3961305	281	95	197
Ukrainian (uk)	883369	1149425314	143511866	15053	4414	5800
Urdu (ur)	53131	72645622	14204222	990	360	562
Uzbek (uz)	67241	101189529	11410171	2103	536	390
Vietnamese (vi)	381356	576413011	113053125	7706	2372	1333
Wu (wuu)	325	441370	29505	32	6	29
Yoruba (yo)	3157	3879056	666500	142	11	89
Chinese (zh)	113722	131874811	9352539	7070	2018	2762
Zulu (zu)	937	1334536	138028	51	8	57
All (60 languages)	60083121	74727560205	10794832477	1287054	568296	253373

Table 3: Statistics for all 60 languages included in *MAC*. 'T.Chunks' and 'Contrib.' refer to Text Chunks and Contributions, respectively. In this table, we include data that comes from Wikipedia users with at least two contributions.

	Wiki	Page	Revision	NS	Time	User	Text
1	en	Don Henrie	254495622	0	2008-11-27	Ursula darling	"His first feature film appearance was in the Aleister Crowley biopic Abbey ... performed a trio of tasks that gave no scientific proof to validate his condition and made extraordinary claims about himself."
2	en	Talk: Aua, American Samoa	332113525	1	2009-12-16	Tama. falealili	"Who ever wrote this article is so irresponsible ... Tuisamoa was also the founder of the Manuá islands and was known as the TuiManua, Tuifiti and Tuitoga."
3	en	Titledlive	152989039	2	2007-08-22	Titledlive	Hi, my name is Travis ... My professional motto: to create intuitive design and motion while remaining innovative and fresh.
4	en	Talk: Gamer9678	622240710	3	2014-08-21	Qed237	"Hi, exactly what edit of mine are you having any problem with? ... There is no reason for anyone to change that as it makes no change to the visual part of page, only that it is faster to build."

Table 4: *MAC* examples. The examples are presented across the four distinct namespaces (NS). For clarity and accessibility, we include only examples from the English Wikipedia to ensure that all readers can comprehend the textual content.

Parameter	Value
Batch Size	64 (en - ru) / 32 (it - uk)
# Epochs	6
Learning Rate	5e-5

Table 5: Training hyperparameters for both SBERT$_{AV}$ and SADIRI. We use different batch sizes for different languages due to large differences in data size among them. Please see Table 8 for the order of the languages.

Parameter	Value
k_1	1.5
b	0.75
ϵ	0.25

Table 6: BM25 hyperparameters

Model	Total Training Time (hr)
SBERT$_{AV}$	5.48
SADIRI	12.82

Table 7: Training times for SBERT$_{AV}$ and SADIRI on a single NVIDIA A40-48GB. The numbers are sums across the 10 languages used in the experiments. Please see Table 8 for the full list.

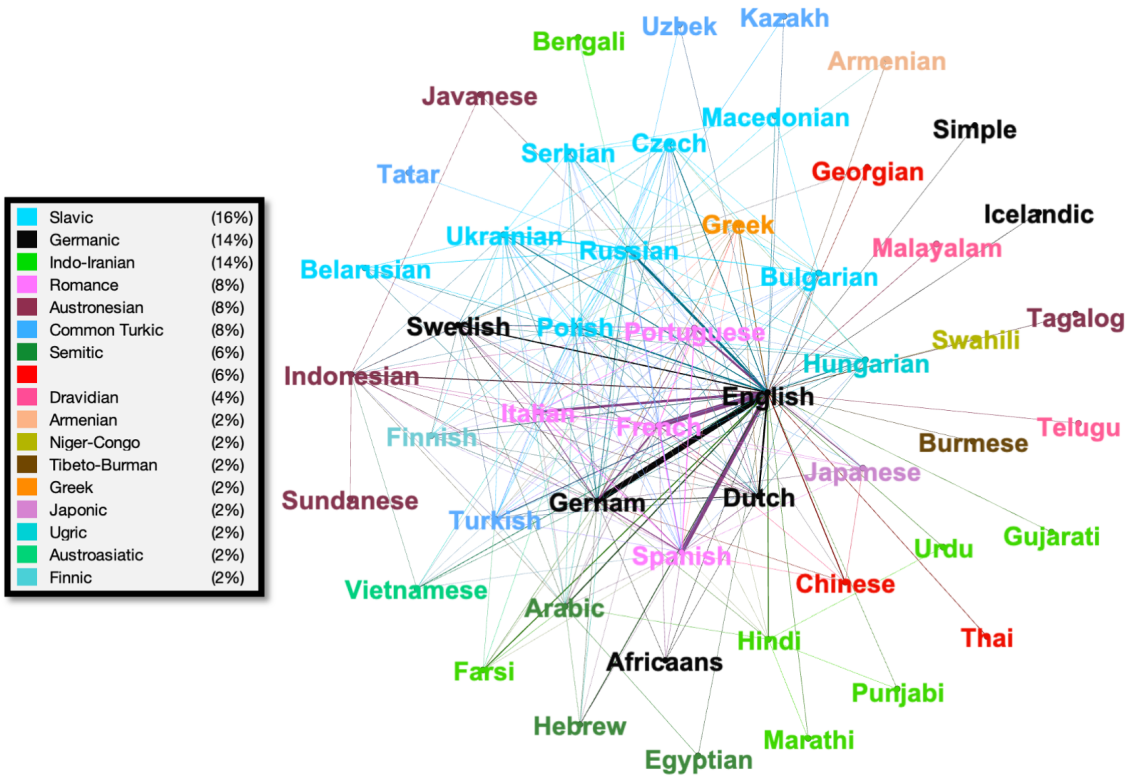


Figure 5: Authors overlap network in *MAC*. Each node is a Wikipedia project (i.e., language). Weighted edges are assigned according to the authors' overlap between node pairs. The most popular languages in Wikipeida are the most central nodes in the network. Closely related languages are likely to have a tight connection in the network. Node colors are based on the language family, which is detailed in the figure legend. The network was rendered using the Fruchterman-Reingold layout.

Language	Training		Test				
	Training	Validation	Domain 0	Domain 1	Domain 2	Domain 3	Cross-domain
English (en)	152,569	21,680	43,671	6,708	2,241	3,761	32,916
German (de)	31,326	4,431	9,029	1,659	219	579	6,915
French (fr)	24,512	3,385	6,936	482	299	407	3,743
Spanish (es)	21,490	3,021	6,074	293	92	404	2,602
Russian (ru)	20,234	2,916	5,738	284	42	218	1,731
Italian (it)	14,970	2,164	4,192	207	91	391	1,907
Portuguese (pt)	11,874	1,728	3,425	77	64	181	994
Polish (pl)	9,255	1,342	2,580	69	302	137	1,822
Dutch (nl)	7,118	1,029	2,027	79	18	102	825
Ukrainian (uk)	6,446	940	1,884	27	23	37	377

Table 8: The number of text pairs (authors) for training, validation, and single-language test sets.

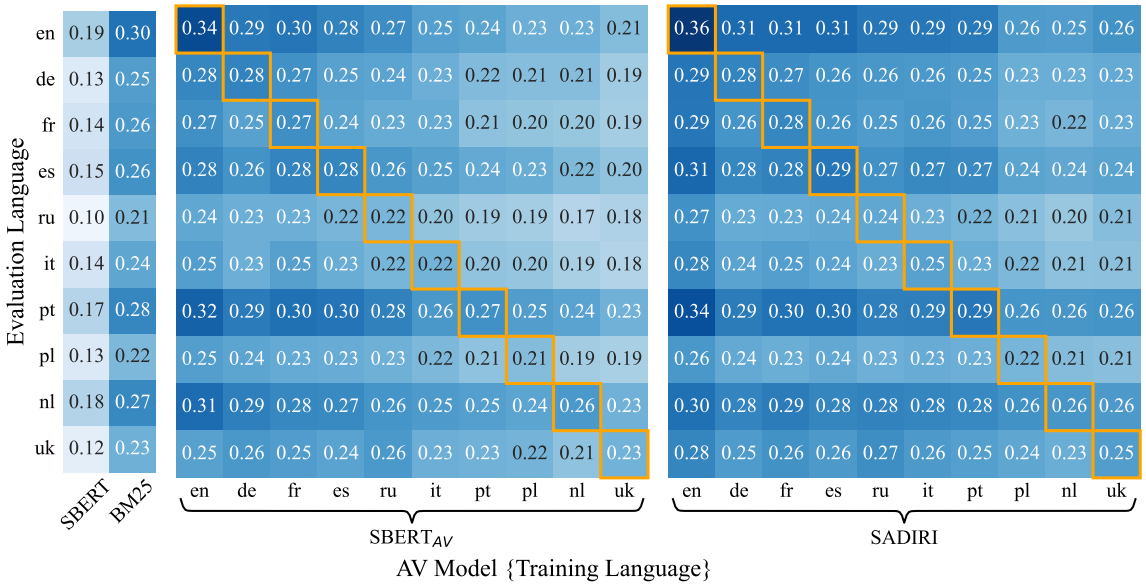


Figure 6: The Success@1 scores for out-of-language generalization for the top ten languages. The scores highlighted in the orange boxes are the in-language scores for which the models are trained and evaluated in the same language. Boxes in this figure are auto-shaded such that darkness correlates with Success scores across the entire table.

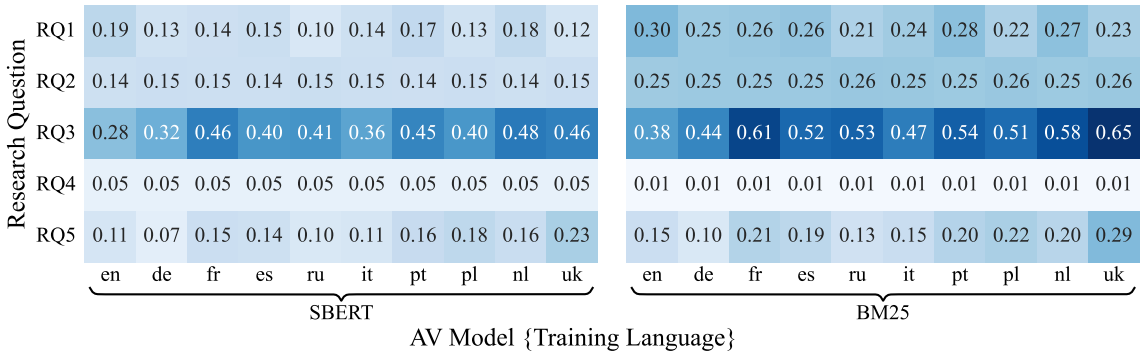


Figure 7: Success@1 scores for the two baseline models. The scores for RQ2 are the averages across only unseen evaluation languages, and the scores for RQ3 are the averages across only unseen domains (domains 1 to 3). Neither baseline model is trained on MAC. The training language indicates the evaluation set in each column is the same as the evaluation set for the fine-tuned models trained in that language.

Type	Name	License
Model	paraphrase-multilingual-mpnet-base-v2 (278M)	Apache-2.0
	xlm-roberta-base (278M)	MIT
Library	Gensim	LGPL-2.1
	Transformers	Apache-2.0
	Sentence Transformers	Apache-2.0
	SADIRI	N/A
	mwparserfromhell	MIT
	wikitextparser	GPL-3.0
	pycld2	Apache-2.0

Table 9: Existing artifacts used in this work as well as their corresponding licenses. The number of parameters for each model is provided in parentheses following the model name. The link to the documentation or the official website of each artifact is provided through the hyperlink embedded in the model name. For more details on SADIRI, please refer to the original work (Fincke and Boschee, 2024).

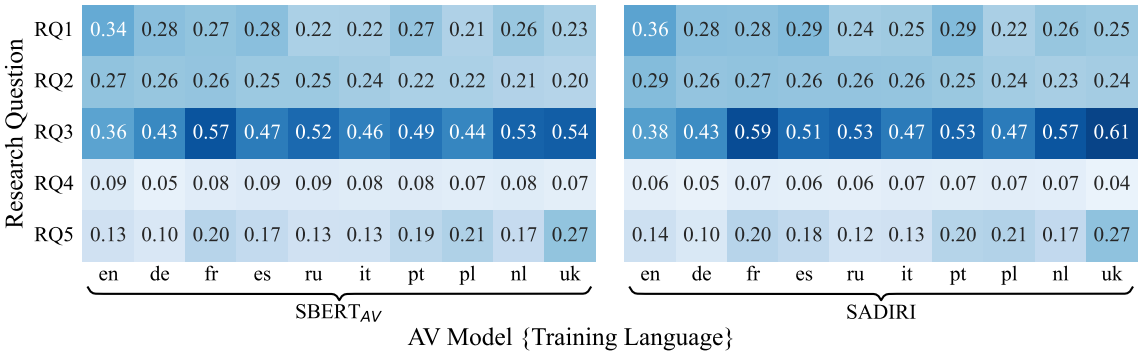


Figure 8: Success@1 scores for the two fine-tuned models. The scores for RQ2 are the averages across only unseen evaluation languages, and the scores for RQ3 are the averages across only unseen domains (domains 1 to 3).

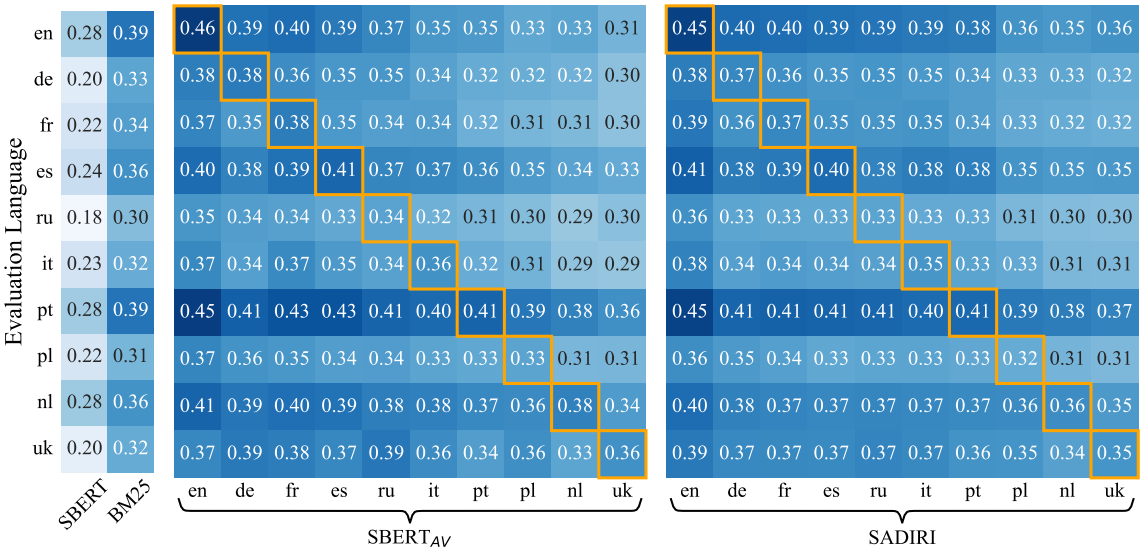


Figure 9: The Success@8 scores for out-of-language generalization for the top ten languages. The scores highlighted in the orange boxes are the in-language scores for which the models are trained and evaluated in the same language. Boxes in this figure are auto-shaded such that darkness correlates with Success scores across the entire table.

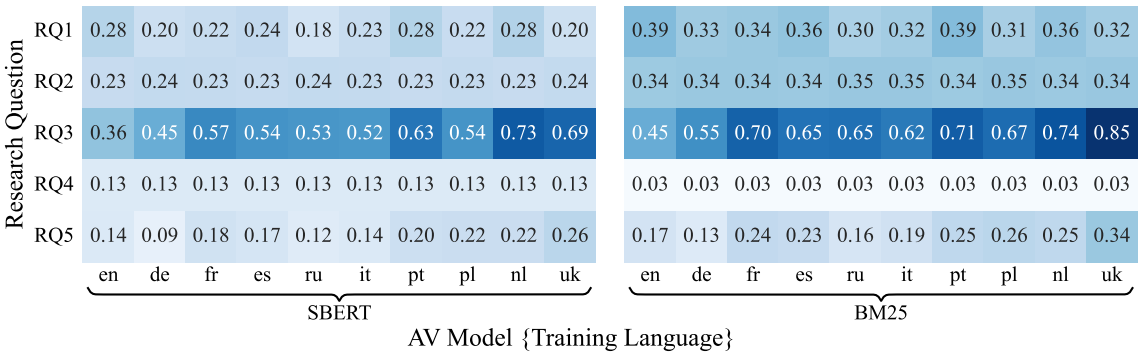


Figure 10: Success@8 scores for the two baseline models. The scores for RQ2 are the averages across only unseen evaluation languages, and the scores for RQ3 are the averages across only unseen domains (domains 1 to 3). Neither baseline model is trained on MAC. The training language indicates the evaluation set in each column is the same as the evaluation set for the fine-tuned models trained in that language.

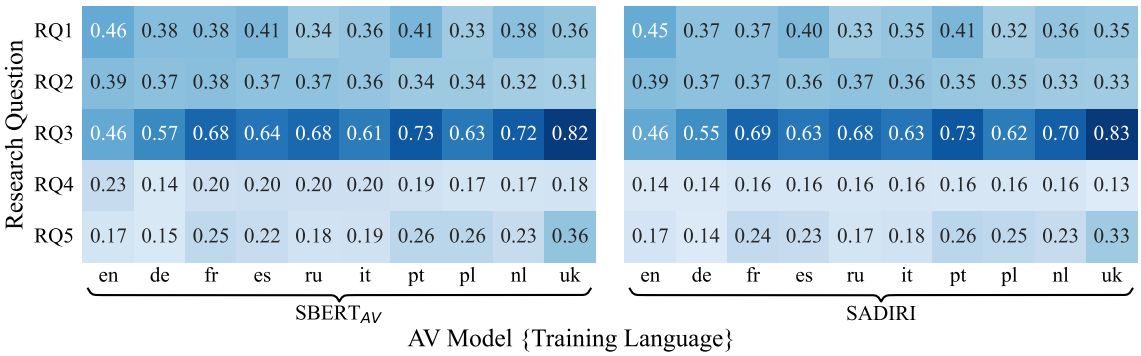


Figure 11: Success@8 scores for the two fine-tuned models. The scores for RQ2 are the averages across only unseen evaluation languages, and the scores for RQ3 are the averages across only unseen domains (domains 1 to 3).