CROSSNEWS: A Cross-Genre Authorship Verification and Attribution Benchmark

Marcus Ma[1], Duong Minh Le[1], Junmo Kang[1], Yao Dou[1],
John Cadigan[2], Dayne Freitag[2], Alan Ritter[1], Wei Xu[1]
[1]Georgia Institute of Technology
[2]SRI International
{marcus.ma, dminh6, junmo.kang, douy}@gatech.edu
{john.cadigan, daynefreitag}@sri.com
{wei.xu, alan.ritter}@cc.gatech.edu

Abstract

Authorship models have historically generalized poorly to new domains because of the wide distribution of author-identifying signals across domains. In particular, the effects of topic and genre are highly domain-dependent and impact authorship analysis performance greatly. This paper addresses the existing data gap in authorship for these resources by introducing CROSSNEWS, a novel cross-genre dataset that connects formal journalistic articles and casual social media posts. CROSSNEWS is the largest authorship dataset of its kind for supporting both verification and attribution tasks, with comprehensive topic and genre annotations. We use CROSSNEWS to demonstrate that current models exhibit poor performance in genre transfer scenarios, underscoring the need for authorship models robust to genre-specific effects. We also explore SELMA, a new LLM embedding approach for large-scale authorship setups that outperforms existing models in both same-genre and cross-genre settings.

Introduction

Accurately identifying the author of a document - also known as authorship analysis - plays a critical role in applications spanning forensic linguistics (Yang and Chow 2014), digital security, and content verification, all of which depend on identifying an author's unique and invariant writing style markers. Existing authorship models are able to achieve near perfect performance on popular authorship benchmarks such as Enron (Klimt and Yang 2004) and IMDb (Seroussi, Zukerman, and Bohnert 2014), but these datasets are confined to a single text genre.[1] Simply identifying the topic of a document and using n-gram features, for example, appears to be a reliable approximation of authorship on these datasets (Rosen-Zvi et al. 2004; Tyo, Dhingra, and Lipton 2022). This oversimplified approach limits the scope and scale of authorship models in real-world scenarios, where text documents are heterogeneous and one person may write multiple different types of texts on varied topics. Prior work studying genre transferability (Rivera-Soto et al. 2021; Sousa-Silva 2018) also assumes documents in different genres are written by separate sets of authors, such that any conclusions based on the impact of changing genre is also confounded by the change of the author pool.

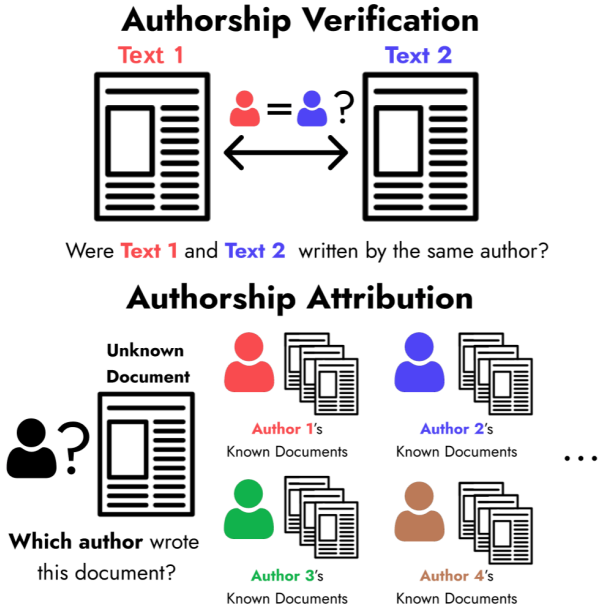

Figure 1: Authorship Verification asks if two arbitrary texts are written by the same person. Authorship Attribution is the task of choosing the author of a document from a set of known authors. CROSSNEWS features a cross-genre setup, where texts from the same author span different genres, such as formal news articles and informal social media posts, reflecting more challenging real-world scenarios.

To address these limitations, we introduce CROSSNEWS, a dataset that connects two different genres of texts written *by the same author*: (i) formal, long news articles, and (ii) informal, short Twitter/X posts. CROSSNEWS consists of a silver set and a gold set. The silver set is formed by matching news articles to Wikidata entities of writers with Twitter/X accounts, while the gold set is created manually from three news agencies with document topic labels. Although previous authorship research has investigated formal writing (Zhang et al. 2018) and social media (Barbon, Igawa, and Bogaz Zarpelão 2017; Boenninghoff et al. 2019) separately,

[1]We use "genre" in this paper to differentiate between social media posts, news articles, emails, and movie reviews. See Lee (2001) for more nuanced and detailed discussions about the differences between the terms "genre," "domain," and "medium."

	Dataset	Genre(s)	Cross-Topic	Cross-Genre	Train Authors	Train Docs	Test Authors	Test Docs
Verification	Blogs	Online Blogs	×	×	5000	70k	5000	70k
	MUD	Reddit Comments	×	×	100k	300k	20k	60k
	Amazon	Online Reviews	×	×	100k	1M	35k	350k
	PAN 2021	Fanfiction	✓	×	60k	120k	30k	60k
	PAN 2022/23	Emails, Essays, etc.	✓	✓	56	8k	56	8k
	CROSSNEWS	**News Articles, Tweets**	✓	✓	**2236**	**100k**	**500**	**20k**
Attribution	IMDb62	Movie Reviews	×	×	62	20k	62	5k
	Blogs50	Online Blogs	×	×	50	50k	50	16k
	CMCC	Interviews, Essays, etc.	×	✓	50	4k	50	1k
	Guardian	Book Reviews	×	✓	13	333	13	111
	CROSSNEWS	**News Articles, Tweets**	✓	✓	**500**	**15k**	**500**	**7.5k**

Table 1: Overview of existing authorship verification (top) and attribution (bottom) datasets.

our work is the first to examine how stylistic properties persist across both genres. Furthermore, as shown in Table 1, CROSSNEWS is an order of magnitude larger than any existing dataset for cross-genre verification and attribution.

We evaluate a wide range of authorship models on both tasks using CROSSNEWS. Our experiments show that, while prior work finds statistical learning models, such as N-gram method (Koppel and Schler 2004), achieve the state-of-the-art performance in the single-genre settings (Tyo, Dhingra, and Lipton 2022), these models generalize poorly in CROSSNEWS' cross-genre settings. Also, while LLMs perform well in small-scale attribution setups (Hung et al. 2023), existing methods generalize poorly to large-scale attribution experiments. We demonstrate a new LLM embedding approach, SELMA, which achieves state-of-the-art accuracy for authorship verification and attribution without requiring additional training. Furthermore, while there is a general consensus that verification and attribution are closely linked, little work has been done to compare the tasks side by side. We find that topic impacts verification and attribution in contrasting ways. In summary, our main contributions are:

- The creation of CROSSNEWS, the largest cross-domain authorship dataset for both verification and attribution.
- A zero-shot LLM-based method for verification and attribution, SELMA, that utilizes instruction-tuned embeddings with task-specific prompts.
- A combined analysis of experiments on both the verification and attribution tasks, which have been frequently studied individually but rarely in conjunction.
- Authorship experiments of the effects of genre and topic.
- Publicly accessible code to support future research.[2]

Related Work

Authorship analysis aims to identify unique signals of individual authors.[3] There are two main tasks: (i) verification, where models determine if two documents are by the same or different authors, and (ii) attribution, where models assign documents to one author among a set of known authors (see Figure 1). For authorship verification, the PAN organization[4] provides the most widely used benchmarks. The verification datasets from both PAN 2020 (Kestemont et al. 2020) and 2021 (Kestemont et al. 2021) contain texts crawled from FanFiction.[5] While the original 2021 test set was designed to be harder than the 2020 dataset, top-performing models from the PAN competition actually see better results on the 2021 test set. However, Brad et al. (2022) recreated the PAN 2021 setup with different splits and yielded the opposite result, underscoring that model performance tends to be dataset-specific and not generalizable. In addition, Rivera-Soto et al. (2021) adapted the Amazon review dataset (Ni, Li, and McAuley 2019) and Reddit Million User Dataset (MUD) (Khan et al. 2021), which were not originally verification datasets, to the verification task by sampling text pairs from the most active users. Standard authorship attribution benchmarks include the IMDb62 (Seroussi, Zukerman, and Bohnert 2011) and Blogs50 (Schler et al. 2006) dataset, which contain documents from a small number of prolific online writers. SOTA models perform very well on IMDb62 (98%) and Blogs50 (75%) (Tyo, Dhingra, and Lipton 2022), highlighting the lack of challenging attribution datasets.

Authorship datasets typically consist of a single genre due to the difficulty of linking authors across genres, as unique identifiers in one data source do not align with author labels in other sources. As a result, cross-genre authorship datasets are scarce and very limited in size. For attribution, the CMCC dataset, (Goldstein-Stewart et al. 2008) consisting of a collection of interviews, written essays, and emails from 21 authors on six sociopolitical topics, and the Guardian opinion dataset (Stamatatos 2013) contain 756 and 444 documents, respectively. For verification, the PAN 2022 and 2023 verification tasks (Stamatatos et al. 2022, 2023) use documents written by 112 individuals in four different genres (i.e., essays, emails, texts, and memos) from the Aston 100-Idiolects dataset (Heini, Kredens, and Pezik 2021). However, these verification and attribution datasets are too small to leverage modern techniques of authorship analysis (e.g., embedding-based methods). To address this need, we introduce the large-scale CROSSNEWS dataset.

[2] https://github.com/mamarcus64/CrossNews

[3] See also the survey by Tyo, Dhingra, and Lipton (2022).

[4] https://pan.webis.de/

[5] https://www.fanfiction.net/

The CROSSNEWS Dataset

Cross-genre authorship datasets are particularly hard to construct because authors often change their pen names, and usernames on social media may not match real names. Because of this difficulty, we collect two separate sets of data - a large, silver set from automated Wikidata-Twitter linkages, and a smaller, gold set collected manually. Combined, they make CROSSNEWS the largest cross-genre authorship dataset to our knowledge (See Table 1).

Silver Set Construction

Data Collection We utilized Wikidata (Vrandečić and Krötzsch 2014), a knowledge base derived from Wikipedia. We queried Wikidata for all entities who are English-speaking journalists, have Twitter/X accounts and have written for free public news websites. We linked entities to the articles they authored from these websites. Then, we extracted the author names from the articles and matched their names to possible Wikidata entities. We define the two names as matching if they have a Jaro-Winkler similarity (Winkler 1990) of above 0.95. After retrieving article texts for these authors, we removed duplicate articles via an LSH filter (Indyk and Motwani 1998) on the body text. For each Wikidata entity that had at least one linked author, we pulled their tweets via the Twitter/X API[6]. We capped the number of tweets written per author to 600 and the number of articles written per author to 200. This process yielded 2,260 journalists with a total of 65,589 articles (articles per author: mean 29.1, median 6) and 1,083,221 tweets (tweets per author: mean 479.3, median 599), which constitutes the silver data portion of CROSSNEWS.

Manual Verification To verify the accuracy of the above approach, we hired three annotators to manually link 300 journalists and their articles to their corresponding Wikidata entry, if it existed. All of the annotators had a college-level education. Annotators compared the journalist's name and news articles side by side with a list of potential Wikidata entries and selected the entry that corresponded to the journalist, or `None` or `Unsure`. The three annotators had a Fleiss' Kappa (Fleiss 1971) of 0.878. From the 300-journalist sample we estimate the automatic linkage method has a linkage accuracy of 93.6%.

Gold Set Construction

We constructed a manual gold set for evaluation by selecting authors from the New York Times (NYT), the Guardian, and the Times of India, three news organizations not in the silver set. We randomly chose authors who have written at least 100 articles, manually identified their Twitter/X handle, and collected their tweets via the Twitter/X API. We collected 175 NYT, 155 Guardian, and 170 Times of India journalists. For each of the 500 total authors, we collected 100 tweets and 100 articles for a total of 100,000 gold documents. For text processing, we removed location tags and author bylines. Each article is labeled as one of five topics (politics, economy, sports, culture, other) based on the news organization's categories. Additionally, we manually annotated 2000 articles with more fine-grained topic labels within each topic category (see Figure 2). The most frequent article topic is politics with 60.7% of articles. The vast majority (97.6%) of authors primarily write articles on one topic.

[6]https://developer.twitter.com/en/docs/twitter-api

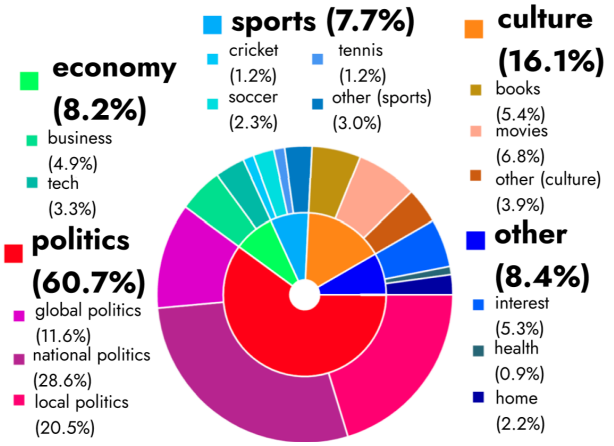

Figure 2: Topic distribution of CROSSNEWS gold articles.

Authorship Models

We benchmark three types of state-of-the-art authorship models on the CROSSNEWS dataset: (i) non-Transformer models, (ii) embedding methods, and (iii) LLM prompting.

Non-Transformer Methods

Unlike other areas of NLP where Transformer models dominate, authorship methods that do not utilize neural networks are still very competitive (Tyo, Dhingra, and Lipton 2022). We consider N-gram, PPM, and O2D2, all three of which have been top-performing models on previous authorship benchmarks (Tyo, Dhingra, and Lipton 2022; Neal et al. 2017; Kestemont et al. 2021).

N-gram (Koppel and Schler 2004) This method constructs character, part of speech, and n-grams features to train a logistic classifier, and is reported to outperform modern verification models (Tyo, Dhingra, and Lipton 2022). In the authorship verification task, N-gram creates a single model that runs binary logistic regression on the difference in features between the two texts, while in the attribution task, one model is made per author and the prediction is selected via the max of these models.

PPM (Teahan and Harper 2003) **P**rediction by **P**artial **M**atching (PPM) predicts future words based on the context of previously seen words via hierarchical word probability graphs. For verification, PPM creates a compression model for a single text and applies it to the second, calculating cross-entropy between the prediction and the true text. For attribution, a compression model is calculated for each author based on their known texts, and test documents are labeled as the author whose model produces the lowest cross-entropy.

O2D2 (Boenninghoff, Nickel, and Kolossa 2021) **O**ut-**O**f-**D**istribution **D**etector model was the highest-performing model in the PAN 2021 verification task event (Kestemont et al. 2021). This method first produces authorship embeddings from a trained LSTM model and modifies the final result based on uncertainty modeling and Bayes factors from the underlying training text distribution.

Embedding Methods

Authorship embedding models use contrastive learning (Goldberger et al. 2004) to encode representations of documents such that the similarity between two documents composed by the same author in a vector space is maximized. Embedding models can be used for both verification and attribution. For verification, embedding methods classify pairs based on a specific threshold of the cosine similarity between the two documents, where the similarity threshold is calculated as the value that classifies the most number of correct labels on the validation set. For attribution, embedding models create author embeddings by averaging all of the known document embeddings of a given author, then match an unknown document embedding to the closest author embedding via cosine similarity. We consider the PPM, LUAR, and STEL embedding models. For all embedding models, we use a pre-trained RoBERTa (Liu et al. 2019) encoder and train further on CROSSNEWS' silver set.

PART (Huertas-Tato et al. 2022) **P**re-trained **A**uthorship **R**epresentation **T**ransformer (PART) uses a Transformer encoder to initially embed the text document into a sequence of semantic word embeddings, then trains an LSTM to create a style embedding from the word embeddings via a contrastive loss function. Notably, the Transformer encoder is frozen, so the LSTM itself is the only trainable part of the architecture.

LUAR (Rivera-Soto et al. 2021) **L**earning **U**niversal **A**uthorship **R**epresentations (LUAR) also contrastively learns authorship embeddings by finetuning a Transformer encoder. LUAR samples many windows of 32-token excerpts across a single document (for verification) or multiple documents (for attribution) to encode, then applies self-attention to these window embeddings to produce a single embedding. LUAR is trained to create authorship embeddings from a variable number of windows to improve model generalization for arbitrary document numbers and lengths.

STEL (Wegmann, Schraagen, and Nguyen 2022) **ST**yle **E**va**l**uation framework model (STEL) fine-tunes the Transformer encoder with a contrastive loss, but designs training data to create data pairs that discuss the same topic with the aim to disentangle writing style from topic association.

Zero-shot LLM Methods

Given its effectiveness, LLM prompting has been recently applied to authorship tasks. These recent works, including PromptAV (Hung et al. 2023) and LIP (Huang, Chen, and Shu 2024), have achieved considerable success on the verification task; however, prompting LLM for the attribution task has the key limitation that performance degrades quickly once a single prompt contains too many documents. While Huang, Chen, and Shu (2024) reported competitive performance by LLM prompting for attribution, their experiments contained no more than 20 authors with a single document per author. Applying their methods to CROSSNEWS' 500-author attribution setup barely outperforms a random baseline. To handle an arbitrarily large number of documents and authors, we explore a new method that combines LLM embeddings with prompting, namely **S**tyle **E**mbeddings from **L**anuage **M**odels for **A**uthorship (SELMA). We describe three prompting methods (for verification) and SELMA (for verification and attribution) below.

Task Description Only (TaskOnly) Prompt This baseline prompting method involves only providing the task description in the input prompt and querying the model for the result. We adopt the task description from Huang, Chen, and Shu (2024) and construct the input prompt for the verification task as follows:

> *Verify if two input texts were written by the same author. Provide your answer simply with True or False.*
> *Input Text 1: (text 1)*
> *Input Text 2: (text 2)*
> *Answer:*

PromptAV (Hung et al. 2023) Besides the task description, this method includes the eight most relevant stylistic variables for the model to attend to in the input prompt, such as special characters and punctuation style and prompts the model to use Chain-of-Thought prompting (Wei et al. 2023).

LIP (Huang, Chen, and Shu 2024) Linguistically Informed Prompting (LIP) describes the task in the input prompt, and explicitly provides examples of stylistic markers for the model to attend to, while asking the model to describe the different writing styles of the two authors.

SELMA (this work) In our SELMA method, we utilize an instruction-tuned LLM specifically designed for text embeddings to encode unknown and known documents into a shared embedding space to measure stylistic similarity. In particular, we use `e5-mistral-7b-instruct` (Wang et al. 2024), which was instruction-tuned on the sentence similarity embedding task by synthetically generating text retrieval data by querying LLMs. SELMA compares document embeddings in pairs, where one document has an instruction concatenated before the text and the other document does not. An `[EOS]` token is also appended to this input to be fed into the Transformer-based LLM, and its corresponding embedding is extracted from the final layer's `[EOS]` position. We use the following instruction:

> *Instruct: Retrieve stylistically similar text.*
> *Query: (text)*

For each verification task pair, one document is chosen randomly and embedded with the instruction prepended, and the other is embedded without instruction. The documents are classified via cosine similarity following the same procedure as the other embedding methods. For attribution, the test query document is embedded with the instruction prepended and each of the known reference documents are

	Model	Train Genre Pair Type	Test Genre Pair Type: Article-Article Acc ↑	Article-Article F1 ↑	Tweet-Tweet Acc ↑	Tweet-Tweet F1 ↑	Article-Tweet Acc ↑	Article-Tweet F1 ↑
	Random Baseline	-	50.0	50.0	50.0	50.0	50.0	50.0
Non-Transformer Methods	N-Gram (Tyo et al. 2022)	Article-Article	69.5$_{\pm 3.10}$	73.1$_{\pm 1.59}$	49.5$_{\pm 1.64}$	66.3$_{\pm 0.35}$	52.4$_{\pm 2.01}$	60.1$_{\pm 4.75}$
		Tweet-Tweet	51.1$_{\pm 0.04}$	66.6$_{\pm 0.02}$	69.4$_{\pm 2.58}$	71.9$_{\pm 2.04}$	52.5$_{\pm 0.28}$	51.2$_{\pm 2.22}$
		Article-Tweet	58.3$_{\pm 1.70}$	69.5$_{\pm 0.56}$	55.9$_{\pm 1.99}$	67.6$_{\pm 0.38}$	55.8$_{\pm 1.49}$	62.2$_{\pm 1.85}$
	PPM (Teahan and Harper 2003)	Article- Article	67.5$_{\pm 1.03}$	62.8$_{\pm 0.19}$	61.2$_{\pm 0.13}$	38.4$_{\pm 1.35}$	53.4$_{\pm 0.27}$	17.7$_{\pm 0.18}$
		Tweet-Tweet	48.1$_{\pm 1.47}$	66.9$_{\pm 0.37}$	70.2$_{\pm 0.44}$	68.7$_{\pm 0.17}$	52.5$_{\pm 1.42}$	54.9$_{\pm 0.25}$
		Article-Tweet	50.2$_{\pm 0.70}$	66.9$_{\pm 1.12}$	68.9$_{\pm 1.23}$	69.2$_{\pm 0.01}$	54.7$_{\pm 1.22}$	55.1$_{\pm 0.12}$
	O2D2 (Boenninghoff et al. 2021)	Article-Article	69.8$_{\pm 0.86}$	73.9$_{\pm 0.25}$	53.9$_{\pm 1.36}$	63.9$_{\pm 0.13}$	57.0$_{\pm 1.05}$	55.3$_{\pm 2.29}$
		Tweet-Tweet	69.8$_{\pm 0.75}$	75.8$_{\pm 0.48}$	61.0$_{\pm 0.48}$	70.0$_{\pm 0.12}$	63.1$_{\pm 0.02}$	63.4$_{\pm 0.72}$
		Article-Tweet	65.6$_{\pm 0.82}$	73.4$_{\pm 0.19}$	62.6$_{\pm 0.44}$	68.6$_{\pm 0.12}$	61.9$_{\pm 0.46}$	68.3$_{\pm 0.27}$
Embedding Methods	PART (Huertas-Tato et al. 2022)	Article-Article	70.0$_{\pm 0.55}$	69.5$_{\pm 0.52}$	66.9$_{\pm 0.82}$	60.5$_{\pm 0.74}$	58.9$_{\pm 0.72}$	40.0$_{\pm 0.55}$
		Tweet-Tweet	76.8$_{\pm 0.09}$	78.4$_{\pm 0.83}$	67.1$_{\pm 0.00}$	66.0$_{\pm 0.83}$	61.1$_{\pm 1.45}$	55.0$_{\pm 0.52}$
		Article-Tweet	72.9$_{\pm 1.16}$	78.5$_{\pm 0.74}$	59.0$_{\pm 1.03}$	69.6$_{\pm 1.41}$	63.7$_{\pm 1.16}$	69.8$_{\pm 0.91}$
	LUAR (Rivera-Soto et al. 2021)	Article-Article	73.7$_{\pm 1.84}$	70.8$_{\pm 3.38}$	61.1$_{\pm 2.98}$	66.2$_{\pm 1.93}$	59.8$_{\pm 1.15}$	58.8$_{\pm 5.19}$
		Tweet-Tweet	68.6$_{\pm 1.17}$	70.9$_{\pm 5.28}$	72.9$_{\pm 4.50}$	70.8$_{\pm 6.34}$	63.7$_{\pm 3.72}$	61.1$_{\pm 9.56}$
		Article-Tweet	75.52$_{\pm 3.32}$	74.7$_{\pm 2.78}$	69.61$_{\pm 3.47}$	68.47$_{\pm 2.74}$	67.8$_{\pm 3.02}$	66.8$_{\pm 3.64}$
	STEL (Wegmann et al. 2022)	Article-Article	54.2$_{\pm 6.02}$	63.3$_{\pm 1.91}$	49.4$_{\pm 2.06}$	65.4$_{\pm 0.78}$	48.2$_{\pm 1.14}$	58.0$_{\pm 7.89}$
		Tweet-Tweet	52.8$_{\pm 4.01}$	63.7$_{\pm 0.36}$	50.3$_{\pm 3.47}$	62.6$_{\pm 3.90}$	50.6$_{\pm 1.66}$	61.6$_{\pm 6.21}$
		Article-Tweet	51.0$_{\pm 4.42}$	21.3$_{\pm 30.77}$	50.7$_{\pm 1.76}$	20.8$_{\pm 30.03}$	48.6$_{\pm 1.00}$	19.6$_{\pm 2.76}$
LLM Prompting LLaMA-3 70B	Task Description Only (TaskOnly)	–	58.6$_{\pm 2.65}$	38.1$_{\pm 1.28}$	73.1$_{\pm 0.99}$	57.5$_{\pm 0.96}$	48.8$_{\pm 0.48}$	6.12$_{\pm 2.33}$
	PromptAV (Hung et al. 2023)	–	76.1$_{\pm 1.18}$	77.2$_{\pm 0.89}$	84.9$_{\pm 0.93}$	79.1$_{\pm 0.48}$	58.9$_{\pm 0.93}$	36.9$_{\pm 0.64}$
	LIP (Huang, Chen, and Shu 2024)	–	73.7$_{\pm 1.13}$	77.1$_{\pm 0.09}$	82.0$_{\pm 1.93}$	78.4$_{\pm 0.95}$	64.0$_{\pm 1.08}$	59.3$_{\pm 1.70}$
SELMA Mistral-7B	No Prompt	–	77.1$_{\pm 0.28}$	80.9$_{\pm 0.38}$	53.3$_{\pm 0.64}$	68.4$_{\pm 0.55}$	64.3$_{\pm 0.25}$	73.3$_{\pm 0.38}$
	Task Description Only (TaskOnly)	–	85.8$_{\pm 0.27}$	86.0$_{\pm 0.18}$	66.1$_{\pm 0.34}$	74.6$_{\pm 0.2}$	78.5$_{\pm 0.12}$	79.8$_{\pm 0.23}$
	PromptAV (Hung et al. 2023)	–	85.8$_{\pm 0.18}$	85.8$_{\pm 0.18}$	65.0$_{\pm 0.24}$	73.8$_{\pm 0.16}$	78.1$_{\pm 0.29}$	79.3$_{\pm 0.23}$
	LIP (Huang, Chen, and Shu 2024)	–	82.5$_{\pm 0.21}$	84.2$_{\pm 0.16}$	58.8$_{\pm 0.38}$	70.6$_{\pm 0.05}$	72.4$_{\pm 0.36}$	77.0$_{\pm 0.24}$

Table 2: Accuracy and F1 of non-Transformer and Embedding-based models (all trained on the silver data in CrossNews), as well as zero-shot LLM prompting and SELMA, on the three test pair types in CROSSNEWS for authorship verification. Darker cell colors indicate better performance and bold fonts indicate the best within a column.

embedded without instruction. An author embedding is created by averaging the embeddings of the known documents, and the test embedding is compared to each candidate author embedding via cosine similarity to rank and identify the most similar author.

We also conduct experiments of SELMA with different instructions in the prompt. We consider no instruction (SELMA + No Prompt), with only the task description (SELMA + TaskOnly), and by adopting the prompts that were originally designed for authorship verification in PromptAV (SELMA + PromptAV) and LIP (SELMA + LIP).

Authorship Verification Experiments

We formulate the authorship verification task as a binary classification problem: given two documents, output 1 if they share the same author and 0 otherwise.

Experiment Setup

To prepare data for the task, we sample document pairs from CROSSNEWS to form positive (if two documents have the same author) and negative (if two documents have different authors) examples. To balance between data efficiency and diversity, following (Stamatatos et al. 2022, 2023; Hu et al. 2023), we ensure each document is present in exactly one negative and one positive pair, and the number of positive and negative pairs are equal. We create three different train and test sets based on the genres of the two documents in a verification pair: (1) Article-Article: both documents are news articles, (2) Tweet-Tweet: both documents are tweets, and (3) Article-Tweet: one document is an article and the other one is a tweet. For each genre pair type, we sample 100,000 document pairs from CROSSNEWS' silver set, with an average of 67 pairs per train author, and 20,000 document pairs from CROSSNEWS gold set, with an average of 60 pairs per test author. The document pairs from the silver set are then used to form train and validation sets, with the ratio of 8:2, while the pairs from the gold set are used to construct the test set. For evaluation, we report accuracy and F1, as this combination informs the overall predictive power and bias towards predicting one class. Given that authorship performance is highly correlated to input text length (Koppel, Schler, and Argamon 2011; Eder 2013), following Stamatatos et al. (2022); Embarcadero-Ruiz et al. (2022), we concatenate tweets to a minimum length of 500 characters.

For all experiments in this paper, models are run on a single NVIDIA A40 GPU, except for LLaMA-3-70B for prompting, which is run on six A40's, with a total compute time of approximately 400 hours to train and evaluate all verification and attribution models sequentially. Experiment results are averaged over five runs with different random seeds.

Results

The model performance for each combination of train/test genre pair types is reported in Table 2. Overall, models

Model	Known Genre-Unknown Genre											
	Article-Article			Article-Tweet			Tweet-Article			Tweet-Tweet		
	Acc ↑	R@8 ↑	Avg. Rank ↓	Acc ↑	R@8 ↑	Avg. Rank ↓	Acc ↑	R@8 ↑	Avg. Rank ↓	Acc ↑	R@8 ↑	Avg. Rank ↓
Random Baseline	0.2	1.6	250/250	0.2	1.6	250/250	0.2	1.6	250/250	0.2	1.6	250/250
N-gram	**61.4**	84.7	1/9	8.12	20.9	101/151	3.45	10.4	160/182	24.4	43.6	14/62
PPM	50.8	71.6	1/36	7.01	19.1	89/144	7.59	19.6	107/149	32.0	49.8	9/64
PART	26.0	61.4	5/17	2.81	12.6	88/133	7.61	28.4	28/50	28.0	55.9	6/31
LUAR	28.3	61.3	4/22	7.48	23.6	51/104	8.32	26.0	40/92	19.0	41.61	16/57
STEL	1.91	9.47	84/116	0.35	2.05	235/238	0.75	4.15	181/204	1.24	6.96	139/173
SELMA + No Prompt	52.8	80.2	1/7	15.8	38.6	19/70	18.7	47.8	9/40	31.0	55.9	6/40
SELMA + TaskOnly	56.9	**87.9**	**1/5**	**18.4**	**42.2**	**15/69**	20.1	50.1	8/40	**33.3**	**59.4**	**4/31**
SELMA + PromptAV	55.8	87.3	1/6	17.8	41.5	16/70	20.3	49.8	9/41	31.1	56.4	5/35
SELMA + LIP	55.6	86.1	1/6	15.7	37.8	20/72	**21.5**	**51.3**	**8/38**	31.2	56.0	5/36

Table 3: Results for the four Known-Unknown Genre combinations for the authorship attribution experiment on CROSSNEWS, consisting of all 500 authors in the gold set with 30 known documents per author. Avg. Rank is displayed in the format of (Median Rank)/(Mean Rank). Darker cell colors indicate better performance and bold fonts indicate the best within a column.

Article Features			Tweet Features		
Type	Feature	Weight	Type	Feature	Weight
Word 1-gram	said	0.215	Char 1-gram	#	0.532
Word 1-gram	trump	0.198	Char 1-gram	’	0.400
Char 3-gram	gam	0.187	Char 2-gram	..	0.374
POS 1-gram	:	0.168	Char 2-gram	.[space]	0.343
Char 1-gram	”	0.167	Char 1-gram	@	0.325
Char 1-gram	0	0.166	Char 1-gram	!	0.321
POS 1-gram	NNPS	0.165	Char 2-gram	,[space]	0.266
Word 1-gram	says	0.164	Char 1-gram	”	0.258
Char 1-gram	’	0.156	Char 3-gram	...	0.229
Char 2-gram	-[space]	0.151	Char 3-gram	tps	0.213

Table 4: Ten most important N-gram features for the Article-Article and Tweet-Tweet N-gram verification model, ordered by magnitude of feature weight.

perform best on the same-genre test pairs (Article-Article and Tweet-Tweet) and perform worst on Article-Tweet pairs. These results are consistent with previous findings that models have a hard time with cross-genre generalization (Rivera-Soto et al. 2021; Wang et al. 2023). Both N-gram and PPM, two of the non-Transformer methods, see large dropoffs in performance in cross-genre settings, marginally improving over the random baseline in terms of accuracy on the Article-Tweet test pairs. To explain this performance drop, we compare the important N-gram features between the Article-Article and Tweet-Tweet models (Table 4). We find the model attends to punctuation that is commonplace to specific genres, such as “- , :” for articles, and “#, @, !” for tweets, which limits the models’ transferability to other genres. O2D2, however, sees better performance on the Article-Tweet test pairs as it continuously re-samples its training data to prepare for domain shifts in testing. Meanwhile, the PART and LUAR embedding models perform better across more test and train setups. STEL, however, predicts almost every label as “different author”, yielding its close to or worse than random baseline accuracy. For LLM-based approaches, Prompting + LIP notably achieves state-of-art accuracy of 84.9 on the Tweet-Tweet pairs. SELMA achieves the best accuracy on the Article-Article and Article-Tweet setups, outperforming LLM Prompting despite the underlying LLM having ten times fewer parameters. This indicates that comparing documents in a latent embedding space may be a better approach to capturing stylometric features than directly prompting the LLM for a comparison.

Authorship Attribution Experiments

The Authorship Attribution task aims to determine the most likely author of a document from a predefined set of authors.

Experiment Setup

For this task, we use only the gold set from CROSSNEWS, as both the training text (known-authorship documents) and test text (unknown-authorship documents) must come from the same author pool. Our attribution setup contains all 500 gold authors with 30 known documents and 15 unknown documents per author. We create a set of known and unknown documents for both the Article and Tweet Genre, resulting in four separate evaluation setups for each combination of known-unknown document sets. We report accuracy, R@8 (the probability the correct author appears in the top 8 predicted authors), and the average rank of the true author (median and mean).

Results

Attribution results are presented in Table 3. Similar to verification, models perform much better in single-genre settings than in cross-genre settings. While N-gram performs the best in the Article-Article setup, it does not perform as well in cross-genre setups as grammar structures and vocabulary, the two main feature groups that N-gram utilizes, change across genres. Meanwhile, for embedding models, PART and LUAR perform better for both cross-genre setups, with R@8s of 28.4 and 26.0 for the Tweet-Article setup. Besides the accuracy of the Article-Article setup, SELMA outperforms all other models, particularly in cross-genre settings. For LLMs, we see that, as opposed to verification, prompt choice does not impact performance very much. This could be because the verification task only has 2 documents per

Article Setup	Verification		Attribution	
	All Topics	Single Topic	All Topics	Single Topic
N-gram	70.4	64.5 (-5.9%)	45.6	65.8 (+20.2%)
PPM	67.9	58.9 (-9.0%)	39.0	45.5 (+6.5%)
PART	70.3	64.9 (-5.4%)	32.6	39.5 (+6.9%)
LUAR	73.9	70.9 (-3.0%)	38.5	43.1 (+4.6%)
STEL	54.3	58.5 (+4.2%)	4.04	5.04 (+1.0%)
TaskOnly	73.0	70.4 (-2.6%)	64.4	52.2 (-12.2%)
PromptAV/AA	76.1	72.3 (-3.8%)	60.0	50.3 (-9.7%)
LIP	73.7	65.5 (-8.2%)	54.3	44.1 (-10.2%)

Table 5: Accuracy for verification and attribution on two Article-Article test sets, one containing documents from all topics and the other containing documents from a single topic (global politics). Change in accuracy between test sets (All Topics and Single Topic) is shown in color. The verification task uses LLM prompting and attribution uses SELMA.

Tweet Setup	Verification		Attribution	
	All Topics	Single Topic	All Topics	Single Topic
N-gram	69.4	52.2 (-17.2%)	38.7	37.1 (-1.6%)
PPM	70.2	57.9 (-12.3%)	44.9	42.5 (-2.4%)
PART	67.1	64.4 (-2.7%)	48.9	48.2 (-0.6%)
LUAR	72.9	61.8 (-11.1%)	34.7	36.1 (+1.4%)
STEL	50.3	53.3 (+3.0%)	4.71	3.91 (-0.8%)
TaskOnly	82.8	67.3 (-15.5%)	52.4	49.7 (-2.7%)
PromptAV/AA	84.9	65.9 (-19.0%)	42.0	42.6 (+0.6%)
LIP	82.0	64.6 (-17.4%)	34.7	40.4 (+5.7%)

Table 6: Accuracy for verification and attribution on two Tweet-Tweet test sets, one containing documents from all topics and the other containing documents from a single topic (global politics). Change in accuracy between test sets (All Topics and Single Topic) is shown in color. The verification task uses LLM prompting and attribution uses SELMA.

prompt, so prompt engineering to mention specific stylistic aspects has a large influence on model predictions; attribution, on the other hand, has many more document references to disambiguate style without the need for explicit prompting. Finally, despite the dip in performance between the same-genre and cross-genre settings, all models still perform better than the random baseline. This observation indicates that markers of authorship are present across genres, and that strong identification of one genre has transferable properties to other unseen genres.

The Influence of Topic on Model Performance

In addition to genre, topic diversity significantly influences performance (Kestemont et al. 2021; Khan et al. 2021; Stamatatos 2018; Wang et al. 2023). To test this, we create two test sets of documents for both news articles and tweets - one containing 75 random authors from the CROSSNEWS gold set with documents from all topics, and one with 75 authors who wrote primarily in "Global Politics", with only the "Global Politics" documents selected. We present model accuracy in Tables 5 and 6. For verification, models tend to perform worse when all pairs contain the same topic, while in attribution, models tend to perform better. This could be because for verification, models have no previous context for any other documents written and have much less data to work with. As a result, these models tend to rely on topic first as a predictor of authorship, so models will over-classify matching topic as matching authorship. This is especially pronounced in the Tweet setup - models perform significantly worse when both tweets are on the same topic. Individual tweets contain much less information than articles, so models rely more on non-stylistic information like topic.

However, for attribution, non-LLM models perform much better in the same-topic context when compared to an all-topic dataset. Previous literature has assumed that authorship verification and attribution are very closely related (Tyo, Dhingra, and Lipton 2022; Koppel et al. 2012), which would imply that isolating the topic effect would produce the same change in performance for both attribution and verification. However, the increase in accuracy in attribution for the article experiment indicates that current models utilize topic differently between verification and attribution. While verification models only have access to two documents, attribution models are exposed to many more documents per author. Because topic diversity adds variability to many features these models use, such as vocabulary and frequently occurring phrases, when topics are isolated, attribution models are able to better distinguish between authors. The size of the all-topic document vocabulary was 53% larger than the single-topic vocabulary. This explains why the top-performing attribution model, N-gram, sees a 20.2% increase in accuracy between the all-topic and single-topic setup, as unique word n-grams that correlate highly with specific authors are much more likely to occur with a smaller vocabulary size. A notable exception is that LLMs perform worse in the Attribution setup for the Single Topic dataset. Looking to the Tweet setup, the relative accuracy difference between the All Topic dataset and Single Topic dataset is much better for the Attribution setup than the Verification setup. This provides further evidence that models are much more sensitive to a topic in verification setups as opposed to attribution setups.

Conclusion and Future Work

In this work, we present CROSSNEWS, the largest cross-genre authorship dataset. Evaluations of authorship models on CROSSNEWS show existing models perform poorly in genre transfer setups. Additionally, investigations into topic show that verification and attribution models process document topicality differently, a departure from existing authorship literature that suggests that verification and attribution models behave similarly to each other. Our findings show that future work should focus on building **generalizable** authorship models that explicitly avoid domain-specific signals. Moreover, different training approaches should be investigated that contain a wide range of train genre combinations. Finally, we demonstrate the feasibility of a new zero-shot LLM-based approach, SELMA, which outperforms all other models on CROSSNEWS. These promising results indicate that the future of authorship may lie with LLMs capable of robustly differentiating between genres and domains.

Acknowledgements

The authors would like to thank Suraj Mehrotra, Anantharaman Iyer, and Omansh Bainsla for their help in data annotation. This research is supported in part by ODNI and IARPA via the HIATUS program (contract 2022-22072200004). 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.

References

Barbon, S.; Igawa, R. A.; and Bogaz Zarpelão, B. 2017. Authorship Verification Applied to Detection of Compromised Accounts on Online Social Networks. *Multimedia Tools and Applications*, 76(3): 3213–3233.

Boenninghoff, B.; Hessler, S.; Kolossa, D.; and Nickel, R. M. 2019. Explainable Authorship Verification in Social Media via Attention-based Similarity Learning. In *2019 IEEE International Conference on Big Data (Big Data)*, 36–45.

Boenninghoff, B. T.; Nickel, R. M.; and Kolossa, D. 2021. O2D2: Out-Of-Distribution Detector to Capture Undecidable Trials in Authorship Verification. *CoRR*, abs/2106.15825.

Brad, F.; Manolache, A.; Burceanu, E.; Barbalau, A.; Ionescu, R. T.; and Popescu, M. 2022. Rethinking the Authorship Verification Experimental Setups. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5634–5643. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Eder, M. 2013. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2): 167–182.

Embarcadero-Ruiz, D.; Gómez-Adorno, H.; Embarcadero-Ruiz, A.; and Sierra, G. 2022. Graph-Based Siamese Network for Authorship Verification. *Mathematics*, 10(2).

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.

Goldberger, J.; Hinton, G. E.; Roweis, S.; and Salakhutdinov, R. R. 2004. Neighbourhood Components Analysis. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Goldstein-Stewart, J.; Goodwin, K.; Sabin, R.; and Winder, R. 2008. Creating and Using a Correlated Corpus to Glean Communicative Commonalities. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Heini, A.; Kredens, K.; and Pezik, P. 2021. Aston 100-Idiolects Project.

Hu, X.; Ou, W.; Acharya, S.; Ding, S. H.; D'Gama, R.; and Yu, H. 2023. TDRLM: Stylometric Learning for Authorship Verification by Topic-Debiasing. *Expert Systems with Applications*, 233: 120745.

Huang, B.; Chen, C.; and Shu, K. 2024. Can Large Language Models Identify Authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 445–460. Miami, Florida, USA: Association for Computational Linguistics.

Huertas-Tato, J.; Huertas-Garcia, A.; Martin, A.; and Camacho, D. 2022. PART: Pre-trained Authorship Representation Transformer. arXiv:2209.15373.

Hung, C.-Y.; Hu, Z.; Hu, Y.; and Lee, R. 2023. Who Wrote it and Why? Prompting Large-Language Models for Authorship Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14078–14084. Singapore: Association for Computational Linguistics.

Indyk, P.; and Motwani, R. 1998. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, 604–613. New York, NY, USA: Association for Computing Machinery. ISBN 0897919629.

Kestemont, M.; Manjavacas, E.; Markov, I.; Bevendorff, J.; Wiegmann, M.; Stamatatos, E.; Potthast, M.; and Stein, B. 2020. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In *Conference and Labs of the Evaluation Forum*.

Kestemont, M.; Manjavacas, E.; Markov, I.; Bevendorff, J.; Wiegmann, M.; Stamatatos, E.; Potthast, M.; and Stein, B. 2021. Overview of the Cross-Domain Authorship Verification Task at PAN 2021. In *Conference and Labs of the Evaluation Forum*.

Khan, A.; Fleming, E.; Schofield, N.; Bishop, M.; and Andrews, N. 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*, 5275–5287. Online: Association for Computational Linguistics.

Klimt, B.; and Yang, Y. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *Machine Learning: ECML 2004*, 217–226. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-30115-8.

Koppel, M.; and Schler, J. 2004. Authorship Verification as a One-Class Classification Problem. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, 62. New York, NY, USA: Association for Computing Machinery. ISBN 1581138385.

Koppel, M.; Schler, J.; and Argamon, S. 2011. Authorship Attribution in the Wild. *Language Resources and Evaluation*, 45(1): 83–94.

Koppel, M.; Schler, J.; Argamon, S.; and Winter, Y. 2012. The "Fundamental Problem" of Authorship Attribution. *English Studies*, 93: 284–291.

Lee, D. 2001. Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the Bnc Jungle. *Language Learning and Technology*, 5.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V.

2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Neal, T.; Sundararajan, K.; Fatima, A.; Yan, Y.; Xiang, Y.; and Woodard, D. 2017. Surveying Stylometry Techniques and Applications. *ACM Comput. Surv.*, 50(6).

Ni, J.; Li, J.; and McAuley, J. 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)*, 188–197. Hong Kong, China: Association for Computational Linguistics.

Rivera-Soto, R. A.; Miano, O. E.; Ordonez, J.; Chen, B. Y.; Khan, A.; Bishop, M.; and Andrews, N. 2021. Learning Universal Authorship Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 913–919. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, 487–494. Arlington, Virginia, USA: AUAI Press. ISBN 0974903906.

Schler, J.; Koppel, M.; Argamon, S.; and Pennebaker, J. 2006. Effects of Age and Gender on Blogging. In *Papers from the 2006 AAAI Spring Symposium*, 199–205.

Seroussi, Y.; Zukerman, I.; and Bohnert, F. 2011. Authorship Attribution with Latent Dirichlet Allocation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, 181–189. USA: Association for Computational Linguistics. ISBN 9781932432923.

Seroussi, Y.; Zukerman, I.; and Bohnert, F. 2014. Authorship Attribution with Topic Models. *Computational Linguistics*, 40(2): 269–310.

Sousa-Silva, R. 2018. Computational Forensic Linguistics: An Overview of Computational Applications in Forensic Contexts. In *Language and Law*, volume 5, 118–143.

Stamatatos, E. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21: 421–439.

Stamatatos, E. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3): 461–473.

Stamatatos, E.; Kestemont, M.; Kredens, K.; Pezik, P.; Heini, A.; Bevendorff, J.; Stein, B.; and Potthast, M. 2022. Overview of the Authorship Verification Task at PAN 2022. *CEUR workshop proceedings*, 3180: 2301–2313.

Stamatatos, E.; Kredens, K.; Pezik, P.; Heini, A.; Bevendorff, J.; Stein, B.; and Potthast, M. 2023. Overview of the Authorship Verification Task at PAN 2023. *CEUR workshop proceedings*, 3180.

Teahan, W. J.; and Harper, D. J. 2003. *Using Compression-Based Language Models for Text Categorization*, 141–165. Dordrecht: Springer Netherlands. ISBN 978-94-017-0171-6.

Tyo, J.; Dhingra, B.; and Lipton, Z. 2022. On the State of the Art in Authorship Attribution and Authorship Verification. *IJCNLP 2023*.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10): 78–85.

Wang, A.; Aggazzotti, C.; Kotula, R.; Soto, R. R.; Bishop, M.; and Andrews, N. 2023. Can Authorship Representation Learning Capture Stylistic Features? *Transactions of the Association for Computational Linguistics*, 11: 1416–1431.

Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; and Wei, F. 2024. Improving Text Embeddings with Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11897–11916. Bangkok, Thailand: Association for Computational Linguistics.

Wegmann, A.; Schraagen, M.; and Nguyen, D. 2022. Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, 249–268. Dublin, Ireland: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Winkler, W. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*.

Yang, M.; and Chow, K.-P. 2014. Authorship Attribution for Forensic Investigation with Thousands of Authors. In *ICT Systems Security and Privacy Protection*, 339–350. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-55415-5.

Zhang, R.; Hu, Z.; Guo, H.; and Mao, Y. 2018. Syntax Encoding with Application in Authorship Attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2742–2753. Brussels, Belgium: Association for Computational Linguistics.