Source Attribution of News Articles

Omid Bodaghi*, Tahera Fahimi*, Homa Habashi*, Allan Lyons†
CPSC 630 Group 1
*MSc
†PhD

***Abstract*—In this project, we aim to use news headlines to determine the reporter of news articles. We developed a project that can accurately distinguish the author based on the title of an article using a dataset of thousands of reporters and their accompanying article headlines and bodies. In this project, we addressed critical issues with machine learning-based approaches and proposed a solution based on information theoretic measurements. We use relative entropy to identify the author of the article and calculate the distance of the article from other authors. Our results show that our model is effective for binary-class author identification, particularly when the authors write in different areas. In addition, we demonstrate that our model outperforms one of the most well-known machine learning algorithms, the random forest algorithm, in terms of accuracy when dealing with these very short texts.**

I. Introduction

The question of authorship attribution and verification has been a question from the time that books began to be gathered in libraries [7]. Questions surrounding authorship, both to confirm authenticity [8] as well as to uncover spurious writings were matters of debate and outright fraud was at times also a concern. In *Poetics*, Aristotle used arguments based on the structure of the text to argue the authenticity of various works attributed to Homer [7]. In his work *De architectura*, Marcus Vitruvius Pollio, the first century BCE Roman architect and engineer, relates a story whereby document comparison was used to uncover plagiarism in a poetry competition.[1]

Over the last few years, we have started to see the impact of social media on our society that has in part been brought about by the wide adoption of smartphones which has allowed users to obtain and spread information faster than ever before. There have been some benefits to society as information has become more widely available; however, these same systems have also enabled the spread of so-called fake news including both misinformation as well as deliberate disinformation. One of the ways that fake news can be manifest is through misattribution by reporting that a certain document or story is from one source when it is really from another or by trying to deny that a document is from the attributed source. One potential antidote to this type of fake news is to uncover the truth and flag this misattribution. Although the psychology behind why individuals are more or less likely to be misled by false information is far beyond the scope of this paper, the fact that the misidentification is flagged at least gives the reader a chance at converging on the truth.

Many stylometric measurements such as sentence length, vocabulary diversity, and the ratio of *hapax legomena*,[2] as well as arguments such as those based on the discourse structure of a document require longer documents and often a larger corpus to compare with. Short documents or other small pieces of text resist these methods and so in this paper we restrict our work to the examination of the *titles* of news articles. We explore various techniques for authorship authentication and evaluate their effectiveness in accurately identifying the true author of these very short texts.

II. Background and Motivation

In its simplest form, source attribution is a generalized type of authorship attribution which in turn is the problem of identifying the author of the document in question. More specifically, authorship

[1] http://www.perseus.tufts.edu/hopper/text?doc=urn:cts:latinLit:phi1056.phi001.perseus-eng1:7

[2] A hapax legomenon is a word that only appears once in a given corpus. It is also used when describing a word that, although perhaps used several times, is only used in a single, indistinguishable context.

attribution can be described as correctly assigning the author of a document from within a group of suspected authors. Similarly, authorship verification refers to confirming that an alleged author is or is not the author of a document under consideration. Essentially, the task is to match the style and linguistic patterns of the anonymous text to those of the known authors to identify the most likely source of the text [1, 8].

In modern times, the field of authorship attribution has primarily relied on machine learning techniques in recent years, but there has also been interest in exploring information theoretic approaches. In contrast to machine learning methods that require a large amount of training data and complex algorithms, information theoretic techniques are typically simpler and more efficient.

Authorship attribution can be divided into three basic types of attribution problems: multi-class, two-class, and one-class attribution. Two-class, or binary, attribution is the case where all of the documents in question were written by a pair of authors and the problem is to properly assign unattributed documents to the correct author. Multi-class attribution is similar but is extended to involve multiple authors. One-class attribution can also be referred to as authorship verification [2]. In this case, the problem is to properly determine whether an unattributed document, which may have been written by any number of other authors, either belongs or does not belong to a set of documents composed by a single, known author.

Regardless of the specific attribution problem to be solved, document classification generally follows the same four basic phases: feature extraction, dimension reduction, classifier selection, and then finally evaluation [3]. Feature extraction encompasses cleaning and filtering the data, and then identifying the salient pieces of data that will be used for the latter analysis. Particularly with large documents or sets of documents, the amount of detailed data can become overwhelming and so to maintain performance, it is often necessary to use techniques to reduce the amount of information being fed to the algorithm. Once the data is extracted, it is necessary to choose an appropriate classifier that is as accurate as required while still being adequately performant and it should go without saying that the algorithm appropriate for one scenario might not be the best choice in another. Finally, the last step is to do the evaluation.

III. Related Work

Text classification has been widely studied for the purpose of authorship attribution, plagiarism detection, genre analysis, or data extraction and there are many existing text classification algorithms [3]. Although a bit older, Holmes approached stylometry from a statistical perspective. By counting words either in their raw or lemmatized form and comparing them with features such as sentence length, number of syllables, etc., he sought to use these statistics to discern a fingerprint of the stylistic traits of an author [1].

In modern times, the field of authorship attribution has primarily relied on machine learning techniques; however, there are also approaches that consider information theoretic approaches to this problem. In contrast to many machine learning methods that require a large amount of training data and complex algorithms, information theoretic techniques can be simpler and more efficient to implement.

For example, Zhao et al. [11] demonstrated the effectiveness of using relative entropy based on function-word distances for authorship attribution. This method is based on the idea that authors have distinctive patterns of using function words such as “the” or “of” which can be used to distinguish their writing from others. This approach has the advantage of being computationally less expensive than some advanced machine learning methods.

Le and Safavi-Naini [4] addressed the problem of authorship attribution for tweet-sized texts that are too short for traditional stylometry approaches which generally require larger documents. Due to the restricted length of a tweet, they correlated the entropy of character and word n-grams to identify patterns unique to different authors rather than measuring traditional features such as lexical richness or sentence length which are more appropriate for longer texts.

In addition to authorship attribution, information theoretic techniques have also been applied to keyword extraction. Le Thi and Safavi-Naini [6] proposed an information theoretic approach to identify the most salient keywords in a document for the purpose of keyword extraction. This approach has the advantage of being able to automatically extract keywords that are most relevant to the content of the document.

Finally, Le et al. [5] also expanded on unique word identification using an information theoretic approach. This method uses the concept of entropy to identify words that are highly informative and specific to a particular document or author. This approach has the potential to be useful in a range of applications, such as plagiarism detection or forensic analysis.

IV. Methodology

We present our research methods in three parts. First, we collected a large dataset. Second, we processed the data by employing cleaning and filtering techniques to produce an intermediate representation of our data. Finally, we analyzed this data using information theoretic techniques.

A. Data Acquisition

Before analyzing data, we first needed to collect an appropriate dataset. Rather than implementing our own crawler, use utilized an existing dataset gathered by Thompson [10] entitled "All the News 2.0" which is a significant expansion on an earlier version that is available on Kaggle [9]. This dataset is a collection of almost 2.7 million essays and news articles from 27 American publications. With articles spanning a period covering January 1, 2016 to April 2, 2020, the dataset is distributed as an 8.2 GB CSV file with fields for the date, year, month, day, author, title, the text of the article, the original source URL, publication name, and publication section (if any). The data schema is summarized in Table I.

TABLE I
"All the News" data schema

Field	Type	Description
date	str	Datetime of article publication
year	int	Year of article publication
month	float	Month of article publication
day	int	Day of article publication
author	str	Article author, if available. Multiple authors are separated by a comma
title	str	Article title
article	str	Artcile text, without paragraph breaks
url	str	Article URL
section	str	Section of the publication in which the article appeared, if applicable
publication	str	Name of the article publication

One of the challenges we had working with this data were the inconsistencies in some of the article metadata fields. For example, sometimes the author field contained just the author name while in other cases it contained the entire byline such as "By Barbara Feder Ostrov, Kaiser Health News" or in a format that did not specifically match the documented schema such as "By Steve Almasy and Tessa Carletta, CNN." Additionally, sometimes the name in that field did not identify the writer of the article, for example, "By CNN Staff with graphics from Joyce Tseng." Since the dataset is suffiently large, we were able to address the challenges of the data in the author field by being selective with the records we chose to analyse.

As far as it was available, we used the section field as a topic indicator. Although not a perfect indicator of the article semantics, and suffering from some of the same limitations as the author field, it was still useful since it does reveal that, for an example, that, according to some internal editorial decision, an article belonged in the "business" section as opposed to the "entertainment" section.

B. Data Cleaning and Selection

As described in Section II, all collected data needs some sort of cleaning and filtering to be useful. This is particularly the case with our dataset in order to address the challenges described above in Section IV-A. In addition, the large size of the dataset proved to be a challenge due to memory constraints on some of our computers.

To address these challenges, we wrote some utility programs in Python to extract the features that we intended to analyze and to filter the data down to a manageable size without having to hold the entire dataset in memory at once. Our main utility, `parse_file.py`, is designed to read an arbitrary CSV file, optionally filter each record based on an arbitrary regular expression, and then extract the desired field along with some metadata into a JSON structure that could either be saved or piped to the next analysis stage. This JSON data structure was designed to include the target data field processed into any number of formats that would be useful to the next stage of processing.

C. Data Analysis

To create our model and test it, we followed the following six steps:

1) As the dataset includes various authors with different numbers of articles, we filtered out authors with less than 4000 articles as well as all articles that indicated that they might have more than a single author. To accomplish this, we used `jq`, a command-line tool for processing JSON data. First, we extracted a list of the authors using our `parse_file.py` utility and then we sorted the JSON file based on the author of the headlines using the following command:

```
jq '.|=sort_by(.key)'
news_title.json >
news_title_sorted.json
```

Next, we used the following command to list the number of articles for each author:

```
jq '.|=group_by(.key)|map(author:
.[0].key, value: map(.callTimeMs)
| length) | map((.author):
.value)' news_title_sorted.json
> news_title_frequency.json
```

After that, we wrote a Python script named `authors_freq.py` to extract authors with more than 4000 articles.

2) We then used a Python script named `json_generator.py` to generate a separate JSON file for each author, including the list of headlines written by that author. In the script, we ran the following command for each prolific author:

```
python3 parse_file.py -f
./all-the-news-2-1.csv -k
author -F title -a null -r
'author=^ *author_name *$' -o
./author_name.json
```

Here, "`-f`" defines the original dataset, "`-k`" defines the keys (authors or publications), "`-F`" defines the parts of the data that we need (title, article, date, etc.), "`-a`" is the feature set (bigram, trigram, or raw text), "`-r`" is a filter over data (we used a regex to filter only records where the author equals the author_name), and "`-o`" specifies the output file.

Fig. 1. Kullback–Leibler Divergence Formula

$$D_{KL}(P||Q) = \Sigma_{x \in \mathcal{X}} P(x) log \frac{P(x)}{Q(x)}$$

3) We split the data into training and test data, similar to supervised learning approaches. We used 80% of the headlines of each author for training and 20% for testing. We did this using a split function in `model.py`.

4) With the training data, we converted raw headlines to n-grams. We tested bigrams and trigrams for characters and words and observed that character bigrams worked slightly better.

5) For each author, we calculated the occurrence probability for each bigram as $\Pr(x) = \frac{freq(x)}{total}$. Here, x represents a bigram, $freq(x)$ is the number of occurrences of x in the text, and $total$ is the total number of bigrams in the text. We repeated this process for each author, using the training data.

6) To test the model, we used Kullback–Leibler Divergence (KLD), also known as relative entropy, for identifying the author of a news headline. The Kullback-Leibler Divergence is a statistical distance used to quantify the dissimilarity between two probability distributions, denoted by P and Q. It is calculated as shown in Figure 1.

For each test text, we calculated its distance from each author using KLD distance. The smallest distance indicated the closest author, and we selected that author as the calculated author for the test document. We made two slight changes to the original KLD distance formula to fix potential problems. Mathematically, $\log(0)$ and $\frac{x}{0}$ are undefined and should not be in our calculations. However, if there is a new bigram in the test dataset that did not exist in the train dataset, we would have $\frac{x}{0}$ in the KLD formula. Furthermore, if a bigram is present in the training dataset but not in the test dataset, then the KLD distance formula would

Fig. 2. Calculating Accuracy

$$accuracy = \frac{\text{Number of correct prediction}}{\text{Total number of predictions}}$$

yield a $\log(0)$, which is undefined. To address these issues, we only considered bigrams that are present in both the training and test datasets. This ensures that the probability distribution used in the KLD distance calculation is well-defined and prevents any undefined values from arising.

The system effectiveness can be evaluated using various measures such as accuracy or F1 score.[3] We will focus exclusively on accuracy in this study since accuracy represents the proportion of correct predictions, which is an important factor in assessing the system's efficiency. Accuracy, as shown in Figure 2, is defined as the ratio of correct predictions compared to the total number of predictions.

D. Comparison to an AI Model

To enhance our comprehension of our model's efficacy, we compare our results using KLD with the results obtained using a machine-learning algorithm, specifically the Random Forest (RF) algorithm, as it is frequently implemented for classification purposes and is known for its expediency. In order to have comparable results, we worked with the same data we used for our 2-class experiments.

V. Results

In this section we report our results in three parts that relate to our experiments. First we show the results we obtained using a two-class model, then we show similar work in a multi-class context. Finally, we compare our two-class results with similar results obtained using a machine learning Random Forest model.

A. Two-class

For the present study, we selected writers with more than 4000 attributed single-author articles and calculated the accuracy. Table II displays the two-class result. Based on the input of authors in the row and column fields, each table cell represents the system's accuracy in identifying the correct author. For example, the first row shows the data of *Max Bowden*, and the accuracy of detecting the test data of *Daive Quinn* in the presence of *Max Green* is shown in row 1, column 5. It is evident that the system's accuracy in identifying the author *Karen Mizoguchi* is generally high, as most values in Table II are close to one. However, there is one instance where the system failed to identify the author correctly, resulting in 60% accuracy. This outcome can be explained by considering each author's distinct writing criteria. For instance, *Karen* primarily covers topics related to the startup and music business, while *Daive* focuses on entertainment and lifestyle. Due to the close connection between these two criteria, there may be several news titles with identical wording, making it difficult for the system to determine the author accurately.

B. Multi-class

As part of our analysis, we computed multi-class results for certain authors. As an example, the multi-class results for classes one through three indicate that the model was trained using the data from these classes and then evaluated for accuracy. The following shows the test classes:

1 For *Max Greenwood*, *Rebecca Savransky*, and *Dave Quinn*, the corresponding accuracies are $63\%, 60\%,$ and 98%, respectively.

2 The corresponding accuracies for *Max Greenwood*, *Rebecca Savransky*, *Dave Quinn*, *Karen Mizoguchi* are $63\%, 61\%, 59\%,$ and 68% respectively.

3 For *Max Greenwood*, *Rebecca Savransky*, *Dave Quinn*, *Karen Mizoguchi*, *Stephanie Petit*, accuracies are as follow: $63\%, 61\%, 40\%, 48\%,$ and 55%.

4 The corresponding accuracies for *Dave Quinn*, *Karen Mizoguchi*, *Stephanie Petit* are $40\%, 48\%,$ and 55%.

5 For *Alexia Fernandez*, *Brett Samuels*, *Stephanie Petit*, the accuracies are : $65\%, 96\%,$ and 65%.

To justify the outcomes, we use Table III, which shows the criteria each author writes in. When attempting to match authors with their respective specialty criteria, the model's performance is less

[3] An F1 score is a metric used to evaluate machine learning models that measures both the precision as well as the recall of a model.

TABLE II
ACCURACY FOR 2-CLASS AUTHORS

		Comparison									
		Max Greenwood	**John Bowden**	**Rebecca Savransky**	**Julia Manchester**	**Dave Quinn**	**Brett Samuels**	**Alexia Fernandez**	**Jordain Carney**	**Karen Mizoguchi**	**Stephanie Petit**
Training	**Max Greenwood**		0.56	0.60	0.57	0.99	0.63	1.00	0.84	0.99	1.00
	John Bowden	0.63		0.69	0.59	0.99	0.59	0.99	0.84	0.99	1.00
	Rebecca Savransky	0.61	0.63		0.65	0.99	0.66	1.00	0.84	0.99	1.00
	Julia Manchester	0.69	0.65	0.69		0.99	0.66	0.99	0.86	0.99	1.00
	Dave Quinn	0.97	0.97	0.98	0.98		0.97	0.72	0.99	0.73	0.68
	Brett Samuels	0.60	0.56	0.65	0.60	0.98		0.99	0.84	0.99	0.99
	Alexia Fernandez	1.00	1.00	0.98	1.00	0.55	0.99		1.00	0.60	0.67
	Jordain Carney	0.74	0.72	0.77	0.76	0.98	0.78	0.99		0.98	0.98
	Karen Mizoguchi	0.97	0.95	0.97	0.96	0.57	0.97	0.62	0.99		0.65
	Stephanie Petit	0.97	0.96	0.96	0.97	0.61	0.96	0.68	0.99	0.65	

TABLE III
AUTHORS AND THEIR SPECIALTIES

Author	**Speciality**
Max Greenwood	Political
John Bowden	Coverage Congress and campaigns
Rebecca Savransky	Education
Julia Manchester	political
Dave Quinn	Entertainment, Lifestyle
Brett Samuels	White House and the Trump campaign
Alexia Fernandez	workers' rights for the Center for Public Integrity
Jordain Carney	politics
Karen Mizoguchi	music business
Stephanie Petit	celebrity news and captivating human interest stories

than optimal for authors working in similar fields. For instance, the multi-class outcomes for *Max Greenwood* and *Rebecca Savransky* are relatively poor, likely due to their similar areas of expertise. However, the performance of the multi-class model for *Dave Quinn* aligns with our expectations, given his distinct specialty criteria.

C. Comparison to Random Forest Model

We evaluated our work by comparing it with the results obtained using a Random Forest (RF) machine learning model. In the same manner as our other experiments, the RF model was trained using 80% of the data and then evaluated using the remaining 20% of the data during the testing phase. Table IV provides the accuracy values for the RF model in correctly identifying the author in the 2-class model.

Based on a comparison of the data in Tables II and IV, it is evident that our model is able to identify the author with greater accuracy than the machine learning approach. We suspect that the RF model encounters difficulty assigning the correct label to authors with similar topics due to overlapping vocabulary, which renders the model incapable of effectively distinguishing these very brief texts.

VI. FUTURE WORK

This section considers future work and possible methods to improve this study.

Within the scope of our experiments, we relied solely on common bigrams to calculate the KLD between the test data and each training data. However, to further enhance this system, we propose the utilization of all bigrams instead of restricting to bigrams common to both sets being compared. For instances in which certain bigrams do not occur in the training data, we propose using a normal distribution to select the corresponding bigram for the training data.

TABLE IV
RANDOM FOREST MODEL ACCURACY FOR 2-CLASS AUTHORS

		Comparison									
		Max Greenwood	**John Bowden**	**Rebecca Savransky**	**Julia Manchester**	**Dave Quinn**	**Brett Samuels**	**Alexia Fernandez**	**Jordain Carney**	**Karen Mizoguchi**	**Stephanie Petit**
Training	**Max Greenwood**		0.49	0.49	0.64	0.96	0.58	0.89	0.82	0.96	1.0
	John Bowden	0.64		0.66	0.66	0.93	0.66	0.96	0.82	0.94	0.99
	Rebecca Savransky	0.6	0.6		0.7	0.97	0.67	0.97	0.8	0.98	0.96
	Julia Manchester	0.58	0.53	0.55		0.89	0.56	1.0	0.84	0.96	0.97
	Dave Quinn	0.95	0.95	0.9	0.93		0.98	0.66	0.95	0.71	0.68
	Brett Samuels	0.59	0.49	0.51	0.57	0.94		0.99	0.8	0.92	0.89
	Alexia Fernandez	0.94	0.94	0.88	0.99	0.57	0.94		0.97	0.58	0.7
	Jordain Carney	0.82	0.74	0.8	0.8	1.0	0.77	0.94		1.0	0.94
	Karen Mizoguchi	0.92	0.94	0.9	0.94	0.48	0.99	0.6	0.92		0.57
	Stephanie Petit	0.89	0.92	0.88	0.9	0.57	0.9	0.65	0.93	0.62	

Throughout our investigation, we evaluated bigram and trigram for characters, as well as bigram for words, ultimately selecting the approach that yielded the most favorable results. Love [7] proposes a more effective method of author detection that incorporates weights for each metric and subsequently creates a weighted combination of these metrics. For future work, we also propose investigating whether the syntactic structure of these very short texts provides enough information to distinguish authors or whether any variance is more related to the topic as opposed to the writer.

VII. PLANNING VS EXECUTION

In our original, internally conceptualized plan, our intent was to distribute the project tasks equitably among our group members to best utilize our various skills and talents. The project can be divided into a few main steps along with some parallel tasks.

1) Defining the project topic: The initial step involved researching various works related to information theory to identify potential project topics. To start with, **Allan** and **Tahera** researched machine-learning-based information theory, while **Homa** and **Omid** studied game-based randomness generation. After analyzing various papers and consulting with Dr. Rei, we decided to focus on authorship attribution.

2) Surveying existing approaches: After identifying the main topic, we conducted a survey of existing information theoretic approaches to gain a better understanding of the field. **Allan, Tahera, Homa, and Omid** were all involved in this step.

3) Obtaining a data set: **Tahera** was responsible for identifying and obtaining the data set that we used in our project.

4) Writing the proposal: Once our project was organized, we collaborated on writing the proposal. **Allan, Tahera, and Omid** contributed to this step.

5) Continuing research: Once the proposal was approved, we continued reading additional related papers, including artificial intelligence related approaches, to gain further insights and knowledge. **Allan, Tahera, Homa, and Omid** all contributed.

6) The central part of the project involved implementing and testing various approaches to authorship attribution, which we divided into two main tasks: a) Dataset processing: **Allan** was responsible for cleaning and organizing the dataset to ensure that it was structured uniformly. As part of this step **Allan** was responsible for the overall design of the software and data flow. He and **Homa** worked on preparing a set of functionalities to flexibly process the data in different modes. b) Model building and testing: **Tahera and Omid** were responsible for

building and testing the model, as explained in the data analysis subsection.

7) As the native English speaker, **Allan** was responsible for for the final editing of the written components as well as the LaTeX formatting.

Our final project closely aligns with what we had initially anticipated in our proposal. Using the news dataset, preprocessing the data, using relative entropy techniques to distinguish the authors, and then comparing these results with a machine learning model all have been done in our project. However, there were a minor differences. Our original goal was to identify the news organization for an unknown news article instead of determining the authorship. We tested our model for organization identification, but it did not perform well for two main reasons: 1) The dataset we used was biased toward larger news organizations. Some organizations had more than a hundred thousand articles while others had only around a thousand. This led to a bias in the trained models towards larger organizations. 2) We anticipate that the main problem was the homogeneity of the news organizations. In contrast, with source similar to Twitter, each user might follow an specific style of writing, have ideosyncratic misspellings, or use a personal set of emojis and abbreviations. However, with news agencies, reporters usually are specialized writers and respect the customs of writing news headlines. Also note that the motivation of our work was to restrict our data news headlines, with lengths typically around 8 to 15 words, in order to test our approach in a restricted environment.

References

[1] David I. Holmes. "Authorship Attribution". In: *Computers and the Humanities* 28.2 (Apr. 1994), pp. 87–106. ISSN: 0010-4817, 1572-8412. DOI: 10.1007/BF01830689. URL: http://link.springer.com/10.1007/BF01830689 (visited on 2023-04-14).

[2] Moshe Koppel and Jonathan Schler. "Authorship Verification as a One-Class Classification Problem". In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 62. ISBN: 1-58113-838-5. DOI: 10.1145/1015330.1015448. URL: https://doi.org/10.1145/1015330.1015448.

[3] Kowsari et al. "Text Classification Algorithms: A Survey". In: *Information* 10.4 (Apr. 2019), p. 150. ISSN: 2078-2489. DOI: 10.3390/info10040150. URL: https://www.mdpi.com/2078-2489/10/4/150 (visited on 2023-02-04).

[4] Hoi Le and Reihaneh Safavi-Naini. "On De-anonymization of Single Tweet Messages". In: *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. Tempe AZ USA: ACM, Mar. 2018, pp. 8–14. ISBN: 978-1-4503-5634-3. DOI: 10.1145/3180445.3180451. URL: https://dl.acm.org/doi/10.1145/3180445.3180451 (visited on 2023-02-05).

[5] Hoi Le, Reihaneh Safavi-Naini, and Asadullah Galib. "Secure Obfuscation of Authoring Style". In: *Information Security Theory and Practice*. Ed. by Raja Naeem Akram and Sushil Jajodia. Vol. 9311. Cham: Springer International Publishing, 2015, pp. 88–103. ISBN: 978-3-319-24017-6. DOI: 10.1007/978-3-319-24018-3_6. URL: http://link.springer.com/10.1007/978-3-319-24018-3_6 (visited on 2023-02-06).

[6] Hoi Le Thi and Reihaneh Safavi-Naini. "An Information Theoretic Framework for Web Inference Detection". In: *Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence*. Raleigh North Carolina USA: ACM, Oct. 2012, pp. 25–36. ISBN: 978-1-4503-1664-4. DOI: 10.1145/2381896.2381902. URL: https://dl.acm.org/doi/10.1145/2381896.2381902 (visited on 2023-02-04).

[7] Harold Love. "Attributing Authorship: An Introduction". In: *Attributing Authorship*. Cambridge: Cambridge University Press, 2002. ISBN: 0-521-78948-6.

[8] Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship, The Federalist*. Addison-Wesley Series in Behavioral Science. Reading, Mass: Addison-Wesley, 1964.

[9] Andrew Thompson. *All the News*. URL: https://www.kaggle.com/datasets/snapcrack/all-the-news.

[10] Andrew Thompson. *All the News 2.0*. July 2022. URL: https://components.one/datasets/all-the-news-2-news-articles-dataset/.

[11] Ying Zhao, Justin Zobel, and Phil Vines. "Using Relative Entropy for Authorship Attribution". In: *Information Retrieval Technology*. Ed. by David Hutchison et al. Vol. 4182. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 92–105. ISBN: 978-3-540-45780-0. DOI: 10.1007/11880592_8. URL: http://link.springer.com/10.1007/11880592_8 (visited on 2023-02-04).