The Myth of Double-Blind Review Revisited: ACL vs. EMNLP

Cornelia Caragea[1], **Ana Sabina Uban**[2], **and Liviu P. Dinu**[2]
[1]Computer Science, University of Illinois at Chicago
Chicago, Illinois
[2]Faculty of Mathematics and Computer Science, University of Bucharest
Bucharest, Romania
cornelia@uic.edu, ana.uban@gmail.com, ldinu@fmi.unibuc.ro

Abstract

The review and selection process for scientific paper publication is essential for the quality of scholarly publications in a scientific field. The double-blind review system, which enforces author anonymity during the review period, is widely used by prestigious conferences and journals to ensure the integrity of this process. Although the notion of anonymity in the double-blind review has been questioned before, the availability of full text paper collections brings new opportunities for exploring the question: *Is the double-blind review process really double-blind?* We study this question on the ACL and EMNLP paper collections and present an analysis on how well deep learning techniques can infer the authors of a paper. Specifically, we explore Convolutional Neural Networks trained on various aspects of a paper, e.g., content, style features, and references, to understand the extent to which we can infer the authors of a paper and what aspects contribute the most. Our results show that the authors of a paper can be inferred with accuracy as high as 87% on ACL and 78% on EMNLP for the top 100 most prolific authors.

1 Introduction

The scientific peer-review process is indispensable for the dissemination of high-quality information (Hojat et al., 2003). However, one of the major problems with this process is bias (Williamson, 2003; Tomkins et al., 2017). For example, Tomkins et al. (2017) performed an experiment during the ACM Web Search and Data Mining conference 2017 to understand the potential bias in favoring authors from prestigious institutions and found that, indeed, when reviewers have access to the authors identities, they more often tend to favor well known authors from prestigious institutions. The double-blind review process is generally employed by top scientific journals and conferences in order to guarantee *fairness* of the paper selection, and thus, plays an essential role in how scientific quality is eventually measured (Meadows, 1998). It is designed to reduce the risk of bias in paper reviews, ensuring that all papers are judged solely based on their content and intrinsic quality and that any author has a fair chance of having a paper accepted, regardless of their prestige or previous work. The double-blind review process implies that the submitted papers have to be anonymized, i.e., the authors' names are not explicitly available with the papers, and any direct or indirect indications of who the authors might be (for example, referring to self-citations in the first person) are forbidden. Reviewers have access only to the papers' content, and the authors in turn do not know who their assigned reviewers are. Despite these strict considerations, the notion of anonymity in the double-blind review has still been questioned. Notably, Hill and Provost (2003) showed that the authors of a scientific paper can be inferred with fairly high accuracy using only the papers it references. Specifically, using the vector-space representations of the list of references (or citations) in a paper and measuring similarity between these representations and the pattern of citations of an author, they are able to infer the authors of a paper with accuracy up to 60% for the top-10% most prolific authors, and show that self-citations are an important predictive factor.

We are interested in further understanding how predictable the authorship of a paper is, and specifically what part of the paper gives it away. For this purpose, we include in our analysis additional text features to reflect various aspects of the text, as well as references, and make use of a more complex machine learning model, based on deep learning, for predicting authors based on these features. We focus on ACL and EMNLP, the top conferences in computational linguistics, which use the

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2317–2327,
Hong Kong, China, November 3–7, 2019.

double-blind review system to decide whether to accept papers for publication.

Our contributions are as follows: we train deep learning models on papers published in the ACL and EMNLP conferences, using features extracted from each paper's body of text as well as its references, and show that these models are able to predict authors with accuracy of about 87% on ACL and about 78% on EMNLP. We additionally perform an ablation study, for an in-depth analysis of the predictive value of each feature. We finally also show how the number of authors considered for analysis can affect performance.

The rest of the paper is organized as follows: In the next section we present related work. Then in Section 3, we describe our datasets. Section 4 deals with the methodology of our experiments, including our baseline algorithm and the details of the model we propose. In Subsections 4.2 and 4.3 we also discuss the features we used in our model and the details of their extraction and preprocessing steps. Section 5 describes the setup of our deep learning experiments, including the metrics we use for measuring performance. Finally, in Subsection 5.3 we report and discuss our results, and in Section 6 we present our conclusions.

2 Related Work

There have been several studies approaching the question of the integrity of the double-blind review process. An early study on blind review published in a journal of psychology (Ceci and Peters, 1984) shows that authors of anonymous papers could be identified by surveyed reviewers using the combination of the paper's references and the referee's personal background knowledge.

Statistical studies on the difference between single-blind and double-blind peer review have more recently demonstrated that unveiling the identity of the authors to the reviewers leads to biased reviews, favoring more prestigious authors and institutions. Tomkins et al. (2017) performed a controlled experiment on scientific articles submitted to the 10th ACM Conference on Web Search and Data Mining, where for every article half of the reviewers had access to author information, while the other half did not. They found that single-blind reviewers are more likely to recommend famous authors for acceptance by a factor of 1.58. A few studies have previously proposed automatic approaches for author prediction for scientific articles. For example, Hill and Provost (2003) successfully predicted the authors of scientific articles published as part of the KDD Cup 2003 competition, using only information from the articles' references lists.

In a related task to ours, several studies have looked at authorship attribution on scientific articles, or predicting authors of scientific articles from an exclusively stylistic point of view. Althoff et al. (2013) studied authorship attribution on scientific articles specifically in the multi-author setting, using various text-based features (including word n-grams and various stylistic features) and models based on logistic regression and expectation maximization. Hitschler et al. (2017) performed experiments for predicting authors of ACL articles, restricting their data to only single-author articles. Their study focused on the style level, using only POS tag sequences, and showed that limiting the number of words considered as features can have a beneficial effect on the predictor's performance. Seroussi et al. (2012) proposed the use of an author-topic model (Rosen-Zvi et al., 2004) for the task of authorship attribution and showed promising results in a scenario with many authors. Rexha et al. (2015) analyzed the style of medical scientific articles and how the stylistic uniformity of an article varies with the number of co-authors.

Outside the world of scientific articles, a few previous studies showed the promise of using neural networks for authorship attribution. Bagnall (2015) successfully used a multi-headed Recurrent Neural Network for an author identification task at PAN 2015. The use of Convolutional Neural Networks (CNNs) for learning from text data was proposed by Kim (2014), where CNNs are successfully applied to several sentence classification tasks. Rhodes (2015) trained a Convolutional Neural Network on word embeddings for predicting authors of medium-sized texts, and Shrestha et al. (2017) used CNNs in an authorship attribution task on tweets. Luyckx and Daelemans (2008) studied the effects of having many authors as classes and of limited training data on author attribution - which are realistic, but difficult scenarios, common to our problem as well.

As far as we are aware, no other study has dealt with analyzing the authorship of articles published at ACL or EMNLP (or a comparably prestigious conference) without restricting the scenario to only a subtask (for example, focusing only on a subset of the data), or limiting the analysis to one

aspect of the text (for example, focusing on the stylistic level). While previous studies support the hypothesis that authors of a scientific article are possible to predict from an anonymized paper, we attempt to provide a fuller picture regarding what exactly it is about an anonymous article that can give away its authors.

3 Datasets

For evaluation, we used two datasets of articles from the computational linguistics conferences ACL and EMNLP, published on or before 2014 (Bird et al., 2008; Anderson et al., 2012). The ACL dataset contains $4,412$ articles authored by a total of $6,565$ unique authors, whereas the EMNLP dataset is comprised of $1,027$ articles written by $1,861$ unique authors in total.[1] Note that the size of the EMNLP dataset is much smaller than the size of the ACL dataset since EMNLP is a much newer conference compared to ACL. From each dataset, we normalized the author names to consist of the initial of the first name and the full last name and removed the authors with less than three articles (to ensure enough data for training and evaluation), leaving us with 922 authors for the ACL dataset and 262 authors for the EMNLP dataset (which represent our classes).

As illustrated in Figure 1, which plots the class distribution in each dataset (i.e., the number of articles per author in decreasing order), we can see that the distribution is very skewed, with the more prolific authors being responsible for many of the articles in each dataset and with many authors contributing only a few articles.

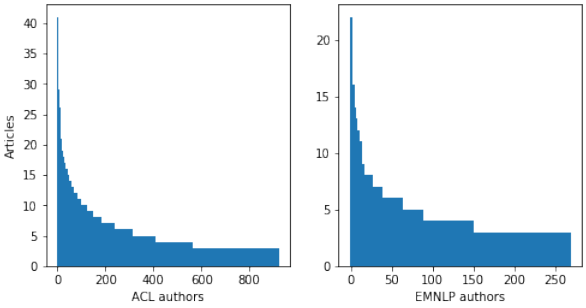

Figure 1: Class (author) distribution.

A similar, if not more pronounced, imbalance can be observed at the level of cited authors. For the purpose of our experiments, we also extracted and analyzed the references (or citation) lists of each article in our datasets, and looked at the distribution of citations across all cited authors. This distribution is illustrated in Figure 2, which plots the number of times each author is cited (in decreasing order). The values on the y axis in this figure are plotted using a logarithmic scale, since the distribution is very skewed.

Tables 1 and 2 show the top-15 most prolific authors as well as the top-15 most cited authors found in the citations lists in our datasets, for both ACL and EMNLP, respectively.

[1] Code and data available upon request.

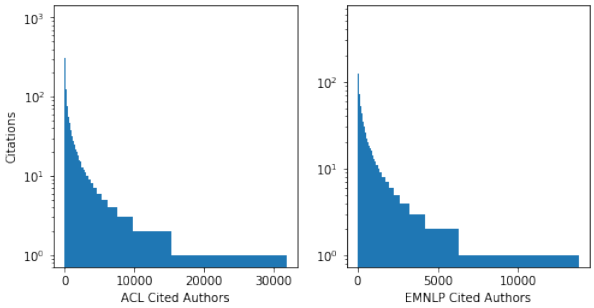

Figure 2: Number of citations per author.

Author	#Papers	Author	#Citations
D Klein	41	C Manning	656
M Zhou	39	D Klein	562
M Johnson	32	F Pereira	561
I Dagan	30	M Collins	539
C Manning	29	D Marcu	459
K Knight	28	P Coehn	441
M Zhang	27	S Roukos	414
Y Liu	27	E Charniak	400
Q Liu	27	A McCallum	393
R Barzilay	26	K Knight	388
N Smith	25	F Och	387
E Hovy	24	M Marcus	377
G Satta	23	M Johnson	355
H Li	23	D Jurafsky	349
Y Matsumoto	21	H Ney	348

Table 1: Top 15 most prolific authors and most cited authors in our ACL dataset.

Author	#Papers	Author	#Citations
D Klein	22	C Manning	333
N Smith	16	D Klein	308
H Ng	16	M Collins	246
C Manning	16	F Pereira	238
G Zhou	15	A McCallum	222
M Lapata	14	P Koehn	204
J Eisner	13	D Marcu	191
D Roth	13	F Och	188
Q Liu	12	R McDonald	179
M Collins	12	S Roukos	174
K Torisawa	11	D Jurafsky	173
M Zhang	11	K Knight	167
Y Liu	11	M Marcus	155
M Zhou	9	M Johnson	154
S Petrov	9	A Ng	143

Table 2: Top 15 most prolific authors and most cited authors in our EMNLP dataset.

Author Name Normalization. It is worth mentioning that our method for normalizing author names can produce collisions, and hence, ambiguities. However, we chose to normalize the names because the noise resulting from not doing so (i.e., having the same author encoded with multiple different ways of writing their name, especially prevalent in references) might be even more detrimental to learning than the possible ambiguities.

In order to understand the level of collisions in our classes (i.e., the author names in the headers), we show in Figure 3 the number of author names that result in three collisions, two collisions, or no collisions at all after normalization, for both ACL and EMNLP. As can be seen from the figure, the number of author names with collisions in each dataset is small. Among the names with collisions, 13 occur within the top most prolific 100 authors in ACL, and only 3 occur in the top 100 for EMNLP. Note that a similar analysis for the author names in the references lists is difficult since many names appear already in the normalized form (first name initial last name). For this reason, normalizing author names is also necessary for computing our baseline, which matches the names of article authors with names of cited authors.

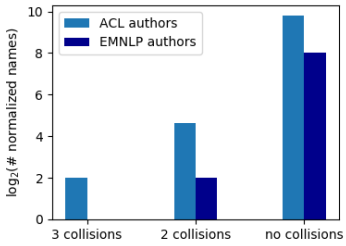

Figure 3: Collision analysis for author name normalization (shown on a $\log_2$ scale).

Author name normalization is in some cases useful, e.g., in the case of authors with middle names, which are sometimes explicit and other times omitted (there are 12 cases in the ACL dataset of authors with middle names whose names occur differently in different articles), or in the case of authors whose first names can have different spellings such as *Dan/Daniel Jurafsky*.

4 Methodology

4.1 Baseline

For our baseline we chose to focus solely on the citations of each article. As Hill and Provost (2003) have shown before, citations alone can be a strong indicator of the authors of an article, with self-citations being especially telling. Specifically, our baseline algorithm consists of simply ranking the authors cited in an article in reverse order of how frequently they were cited overall in the article's references list, and outputs as predicted authors the top (10) most cited authors in this ranking.

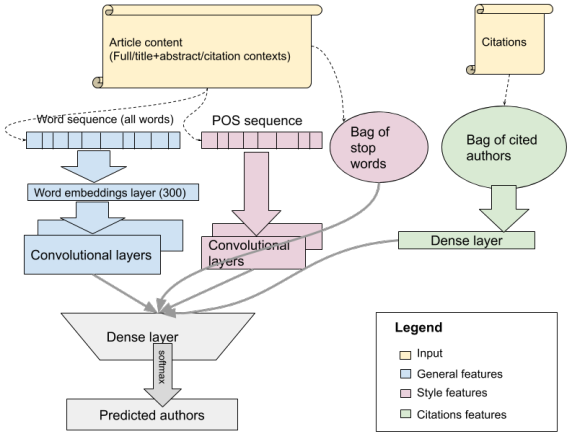

Figure 4: Network architecture

4.2 Proposed Model

For the purpose of our machine learning experiments, we formulate the problem as a supervised classification task, where each article is labelled with one or multiple authors, and a machine learning model learns to predict the set of correct labels (authors) for each data point (article). The order of the authors is not taken into consideration.

As our model, we choose a neural network with several subcomponents corresponding to various types of features, as detailed below.

4.2.1 Features

In order to capture as many of the aspects of a scientific article as possible in our model, we extract and use various features, corresponding to different levels at which characteristics of the author could manifest. We categorize these into three main types of features:

- Content level: word sequences (consisting of 100-word sequences in the article's title, abstract and body).
- Style level: stopwords bag-of-words, part-of-speech sequences.
- Citation level: bag-of-words of cited authors.

Figure 4 shows a high-level view of the network architecture and its various components.

The network is designed to learn from each separate feature using dedicated subcomponents. At the content level, we use convolutional layers to learn from word sequences. CNNs have been shown to be successful in text classification tasks by Kim (2014). We use similar settings to the ones reccommended in this study - passing the word sequences through a single convolutional layer with 300 filters, and a kernel size of 9, followed by a max pooling layer. Before going through the convolutions, the word sequences are passed through a word embedding layer of 300 dimensions, initialized with the pre-trained word2vec embeddings available from Google, trained using the skip-gram objective (Mikolov et al., 2013). Using embeddings that are already pre-trained on a large dataset should benefit our task (our dataset being itself not very large), but since we use general-purpose pre-trained embeddings, we choose not to fix the embedding weights, but rather let the network further update them by learning from our data to tune them to our task and domain.

A separate convolutional layer is dedicated to learning from part-of-speech sequences. As in the case of word sequences, we consider the order of parts of speech in a text segment to be relevant, assuming certain types of syntactical constructs can be specific to certain authors. Thus, after tagging a text segment with parts of speech, we encode the result as part-of-speech sequences and pass them through a convolutional layer of 50 filters and kernel size 4, followed by a max pooling layer. The POS tags are given as one-hot vectors. Stopwords are extracted from each article segment and encoded as bag-of-words, keeping their frequencies, but not their order in the text. We used the stopword list available from the NLTK package. Stopwords frequencies are traditionally used in stylometry, being one of the most indicative features of an author's style (Koppel et al., 2009).

To extract knowledge from citations, we focus on cited authors, encoding for each article the authors cited in its references section, along with the citation frequencies for each author. The total number of cited authors in each of our datasets is much larger than the number of authors that contribute directly to one of the articles, e.g., over $22,000$ unique authors are cited in our ACL dataset. This makes the one-hot encoding that we use for cited authors to be very high-dimensional, so we pass the extracted feature through an additional lower-dimensional fully connected layer.

In the final layers of our network, we collect all of the output from each subcomponent dealing with the various features, and pass them through a dense component consisting of a fully connected layer and a Softmax layer that produces the network's predicted probabilities for each class.

4.3 Preprocessing and Feature Extraction

We extracted the text from PDFs using Grobid.[2] Several preprocessing steps were necessary before using the articles' text as features in our model.

For our text-related features we consider the title, abstract and body of the articles, and exclude references from the article's text, by removing them both from the references section and from the citations within the article text (so as to isolate text features from citation features). After normalizing and tokenizing the resulted text according to usual practice in natural language processing applications (including lowercasing every word, discarding numbers and punctuation, resolving end-of-line hyphenation), we construct a list of vocabulary words consisting of the most frequent 50,000 words in all texts. Our choice of vocabulary size was informed by a previous study looking at authorship on ACL data (Hitschler et al., 2017), which showed that 50,000 words is an optimal vocabulary size for authorship tasks on this dataset. For EMNLP, which is a smaller dataset, we restrict the minimum word frequency to 5 occurrences, leaving us with a vocabulary of approximately 23,000 words. Considering only the words in each vocabulary (and replacing all other words with an "unknown" token), we encode the text as word sequences of 100 words, padding the sequences with zeros if they are shorter. Further, our training examples consist of these word segments, rather than full articles. Before extracting content features, we discard outliers, ignoring articles consisting of either zero or more than $20,000$ words.

In addition, we also extract the context around citation mentions within the content of articles, by selecting a window of 100 characters around the citation (and excluding the citation itself), then applying the same text preprocessing steps as above only on this window. This is used as a separate feature, as described in Section 5.2.

The extraction of the part-of-speech features is done by applying the Stanford POS tagger (Toutanova et al., 2003) to the word sequences, re-

[2] https://github.com/kermitt2/grobid

sulting in part-of-speech sequences corresponding to each article segment. Stopwords are encoded as bag-of-words for each article segment.

Citations extracted from the "References" section of each article are encoded as bag-of-authors - unordered sets of citation frequencies corresponding to each cited author. Recall that author names (when they occur either as authors of the target article, or as authors of a cited paper in a references list) are normalized to consist of the initial of the first name and the full last name (see Section 3).

5 Experimental Setup and Results

The nature of our dataset requires special attention to the setup of the training experiments, one of the main particularities of the data being the skewness of the label distribution. We split the dataset in three subsets: one for actual training, one for validation (used for tuning hyperparameters), and the third for testing performance. At this stage, we ensure that each of the three sets contains at least one article from each author in our labeled set. This also implies that we exclude any authors with less than three articles, obtaining 922 authors for the ACL dataset and 262 for EMNLP (which are the different classes in our supervised learning problem). Given our datapoints, as explained in the previous section, consisting of article segments (word sequences of 100 words) rather than full articles, we also ensure that all segments extracted from one given scientific article are appointed to the same set, and not split between training, validation and test. We take this precaution to make sure that anything our network learns is not an artifact of the particular article, but rather of its author.

Lastly, to reduce the impact of the label skewness on our trained model, we use weighted sampling for generating the training examples, making sure the probability of generating a training example from any class is approximately the same across classes. For this to be possible, a final adjustment had to be made to our training examples. Our datasets, comprising of scientific articles, with each article having been written either by a single author or in co-authorship between several authors, essentially consists of multi-label examples. For training, we transform the training examples from multi-label examples to single-label examples, by generating several copies of the same datapoint, each labelled with only one of its authors, whenever a text was written by more than one author. This allows us to perform weighted sampling, as well as use a simple softmax layer as the final layer in the network, which generates one predicted label for any training example, and cross-entropy loss as our loss function. At the evaluation stage, we use the original multi-labeled examples, to be able to correctly measure the model's performance using our metrics.

Dataset	Nr authors	Training	Valid	Test
ACL	Top 100	73,261	12,260	12,348
ACL	Top 200	124,460	18,756	19,967
ACL	Mid 200	25,746	12,476	13,618
ACL	Bottom 200	13,673	12,378	12,074
ACL	Top 500	185,451	29,498	30,232
ACL	All 922	157,427	41,978	44,427
EMNLP	Top 100	39,852	8,432	7,971
EMNLP	Mid 100	13,486	7,557	7,950
EMNLP	Bottom 100	8,403	6,801	7,769
EMNLP	All 262	49,743	17,518	17,671

Table 3: Data size (in number of article segments).

Dataset	Nr authors	Training	Valid	Test
ACL	Top 100	350	185	197
ACL	Top 200	1,315	258	263
ACL	Mid 200	343	190	199
ACL	Bottom 200	180	181	176
ACL	Top 500	1,758	413	428
ACL	All 922	1,604	697	710
EMNLP	Top 100	325	90	89
EMNLP	Mid 100	130	83	88
EMNLP	Bottom 100	86	78	86
EMNLP	All 262	345	191	86

Table 4: Data size (in number of articles).

Tables 3 and 4 show the number of article segments and the number of articles we end up with after extracting the features and splitting the ACL and EMNLP datasets, respectively, into the training, validation, and test sets, in each of the experimental settings.

5.1 Metrics

Depending on whether we see the list of authors that contributed to an article as a sorted or unsorted list, a machine learning model that can predict this set of authors can be designed either as a multi-label classifier or as a ranking model. We use the former, and disregard the order of the authors of an article, assuming it is not always relevant to the quantity of each author's contribution to the article text, thus representing the authors for an article as an unordered set of labels.

We do, however, consider the order of the predicted classes in the model's output. For evaluating our model, we use both the performance

metrics that are suited for multi-class classification (where we treat the model's predictions as unordered sets) and metrics suited to ranking problems, which are generally used in information retrieval (where we see the list of model predictions as a sorted list, ranked according to the Softmax probabilities in the model's output). These performance metrics are as follows:

- Accuracy@k - computed as the number of articles for which at least one true author was in the top k predicted authors. We use $k = 10$ as this number was shown to perform well in other search and retrieval tasks (Spink and Jansen, 2004).
- Mean Average Precision (MAP)
$$MAP = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{r} \sum_{k=1}^{r} P@k$$
where $\mathcal{A}$ is the set of articles, and precision at rank k $P@k$ is the number of correct authors within the first k predicted authors relative to the number of predictions (k). Here too we use a maximum rank of $r = 10$.
- Mean Average Recall (MAR)
$$MAR = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{r} \sum_{k=1}^{r} R@k$$
where recall at rank k $R@k$ is defined as the number of correct authors within the first k predicted authors relative to the total number of true authors ($r = 10$ as before).
- Mean Reciprocal Rank (MRR)
$$MRR = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{r_a}$$
where r_a is the rank at which the first correct predicted author was found for article $a \in \mathcal{A}$.

Since our datapoints are article segments rather than full articles, the measured performance on the model's output will be with respect to these segments: for example, an accuracy of 10% denotes that 10% of the *segments* were correctly classified.

We additionally adapt these metrics to be able to also output the performance at *article-level*, which allows us to properly compare it to our baseline's performance (which is measured on full articles). We accomplish this by grouping the article segments in our test set according to the article they were extracted from, and for each article, order all probabilites in the network's output for the given article, and consider the top predicted classes across this global ranking as the model's predictions for the target article.

5.2 Experiment Settings

In order to understand the contribution of each feature for predicting the authors of a paper, we perform an ablation study - isolating and using in turn only certain features and combinations of features and deliberately not using others.

In a first experiment, we compare our model's performance on both the entire set of authors in each dataset, and only a subset of the authors, e.g., selecting only the top most prolific ones (top-100, top-200, and so on) and ignoring the rest. In this experiment, we only consider articles authored by these selected authors, both for training and test. We expect performance should be higher on this subset where rare authors are discarded, since the pronounced skewedness of the data makes so that for many authors in the tail of the distribution there are very few data points to train on.

In a second experiment, we group our features according to the level of the text that they represent: content level (word sequences), style level (stopwords and POS sequences) or citation level (cited authors). We run our model using the features in each of the groups in turn, and ignoring the rest. The measured performance in each separate setting should be an indicator of the importance of the specific feature (or feature combination) and of the aspect of the article that it captures.

Finally, we also experiment separately with our text features, in order to understand which part of the scientific text is more specific to its author, and how much of the text is really useful in predicting the author. We look separately first at the entire body of the article, secondly at only the title and abstract, and thirdly only at the context of the citations that occur in the text - assuming these might be useful by giving away something about the author's citation pattern that cannot be inferred from the references list alone. Citation contexts have been shown to capture useful information in previous studies focusing on text summarization (Qazvinian et al., 2010; Abu-Jbara and Radev, 2011; Qazvinian and Radev, 2008), document indexing (Ritchie et al., 2008), keyphrase extraction (Gollapalli and Caragea, 2014; Caragea et al., 2014), and author influence in digital libraries (Kataria et al., 2011).

These last experiments aiming to analyze feature importance are performed in the small-scale setting, where only the top 100 authors are considered. We report all results both at segment level

Features	Acc	MAP	MAR	MRR
All features (top100cls)	**84.65**	9.87	46.22	25.28
All features (top200cls)	70.00	**18.07**	54.41	48.96
All features (top500cls)	63.77	16.82	43.34	41.60
All features (mid200cls)	62.38	14.02	49.44	40.78
All features (last200cls)	54.18	11.38	43.22	35.21
All features (922cls)	52.41	13.78	31.29	33.52
Content + Ref (100cls)	75.38	16.74	**58.42**	**49.46**
Content (100cls)	49.87	8.87	33.23	25.34
Style ft.	25.21	3.22	13.81	9.20
Content (100cls)	49.87	8.87	33.23	25.34
Title+abstr.	16.19	2.71	9.52	6.22
Ref. contexts	35.53	5.07	21.27	13.54

Table 5: Results on ACL dataset (segment level).

Features	Acc	MAP	MAR	MRR
All features (top100cls)	**87.88**	**24.71**	**78.37**	**72.33**
All features (top200cls)	66.78	16.93	51.10	46.20
All features (top500cls)	60.74	15.36	39.71	38.68
All features (mid200cls)	54.82	11.14	40.38	32.99
All features (last200cls)	50.00	8.61	36.24	25.53
All features (922cls)	47.59	11.74	27.04	28.77
Content + Ref (100cls)	76.13	16.37	57.50	48.73
Content (100cls)	48.29	10.51	37.16	31.44
Style ft.	26.70	3.18	14.25	9.26
References	86.67	24.35	78.31	71.37
Content (100cls)	48.29	10.51	37.16	31.44
Title+abstr.	18.08	2.56	10.23	6.34
Ref. contexts	30.95	4.15	11.20	18.16
Baseline (100cls)	54.86	11.57	42.76	34.37
Baseline (200cls)	55.70	12.08	40.82	34.48
Baseline (500cls)	56.17	12.73	38.01	34.65
Baseline (922cls)	57.25	14.79	28.25	35.67

Table 6: Results on ACL dataset (article level).

and at article level, with the exception of the experiment where the only features used are references, which are article-level features.

5.3 Results

Table 5 shows the classification results measured on our data points consisting of article segments on ACL, whereas Table 6 contains the values of the metrics aggregated at article-level on the same ACL dataset. Tables 7 and 8 show similar results on the EMNLP dataset. Underlined scores are best within each group and bold scores are best overall. Figure 5 shows the accuracy of our best deep learning model on ACL and EMNLP as we increase the number of authors (classes), compared with the baseline model that considers only references and with a random model.

As can be seen from the tables and figure, results show that references are still the features that by far contribute the most to predicting the author(s) of an article. The importance of the references list also contributes to the strong performance of the baseline, which is able to correctly predict authors for as much as 54.86% of the articles on ACL and 57.80% on EMNLP, for the top 100 classes. Interestingly, the baseline's performance is comparable to the results reported by (Hill and Provost, 2003) who use a similar method, even on a different dataset. Feeding the extracted references into the deep network further boosts the predictive power of the references feature, reaching on its own an accuracy of 86.67% on ACL and 78.49% on EMNLP in an experiment looking at the top 100 most prolific authors in each dataset.

Features	Acc	MAP	MAR	MRR
All features (top100cls)	79.97	**19.98**	**60.14**	55.33
All features (mid100cls)	57.10	12.02	43.70	35.52
All features (last100cls)	62.96	12.02	46.10	36.08
All features (262cls)	59.47	15.65	37.42	37.90
Content + Ref (100cls)	**80.27**	19.93	59.88	**55.89**
Content (100cls)	48.59	8.28	28.66	22.15
Style ft.	25.98	3.96	12.95	10.29
Content (100cls)	48.59	8.28	28.66	22.15
Title+abstr.	15.48	2.14	7.38	6.15
Ref. contexts	21.47	2.73	9.20	7.39

Table 7: Results on EMNLP dataset (segment level).

Features	Acc	MAP	MAR	MRR
All features (top100cls)	**78.49**	17.70	56.09	48.74
All features (mid100cls)	51.68	11.09	39.49	33.91
All features (last100cls)	51.16	9.28	37.96	27.11
All features (262cls)	57.86	13.84	34.65	33.81
Content + Ref (100cls)	77.41	18.19	56.88	50.60
Content (100cls)	56.98	11.27	37.81	31.08
Style ft.	31.82	4.93	16.21	13.23
References	**78.49**	**18.79**	**58.04**	**52.26**
Content (100cls)	56.98	11.27	37.81	31.08
Title+abstr.	10.75	0.82	3.47	2.32
Ref. contexts	15.90	16.83	6.21	4.25
Baseline (100 cls)	57.80	12.59	42.31	36.01
Baseline (262 cls)	58.90	15.06	24.71	36.92

Table 8: Results on EMNLP dataset (article level).

The more general text features (content level) are the second best predictor, whereas the style-level features come last. Even if much less predictive than references, these text-based features are still far better than chance at predicting the true authors. For example, on ACL, in the setting with 100 possible classes, the expected accuracy of a random predictor (according to our definition of accuracy), would be around 10%, whereas when using all 922 classes, the chance accuracy is 1%.

With regard to the parts of the article text that seem to be most predictive, reference contexts seem to play a more important role than the title and abstract, even though using the full article content still gives the best results on both datasets.

Moreover, as we go from the top 100 most prolific authors to the last (rare) authors, the performance keeps decreasing. For example, the ac-

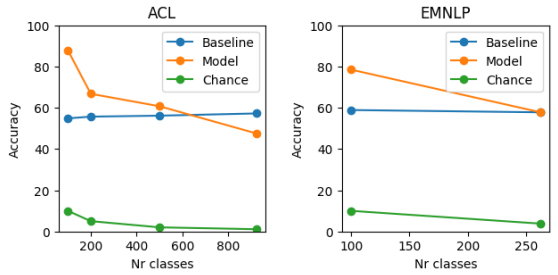

Figure 5: Accuracy with number of authors considered.

curacy on ACL at article level decreases from 87.88% (achieved for top 100 most prolific authors) to 50% (achieved for last 200 rare authors).

5.4 Error Analysis

In order to achieve a better understanding of the model's weaknesses and more generally of the difficulties of predicting authors of scientific papers, we examine the set of misclassified articles in the ACL test set, and compute the misclassification rate for an author as the number of their articles for which the model did not assign to them, divided by the total number of articles authored by them. An interesting finding is that the correlation between the rank of the author (in order of their number of written articles) and the misclassification rate is 0.35, showing that more prolific authors tend to be more accurately classified. One of the most misclassified authors in the top 5 most prolific authors is Christopher Manning (40% of articles are misclassifed, among which there are *Accurate Unlexicalized Parsing* and *Deep Learning for NLP (without Magic)*. For the first paper, some of the predicted authors are: Eugene Charniak, Mark Johnson, Lenhart K Schubert, Dan Klein, and Daniel Jurafsky, with Dan Klein being indeed one of the authors. Other 45 articles not authored by Christopher Manning were predicted as being written by this author, possibly due to a large number of his citations in the articles' references list and/or similar keywords with those of Christopher Manning.

6 Conclusions and Discussion

We showed that the top most prolific authors of anonymized scientific articles can be predicted with high accuracy, with characteristics of the authors being apparent at all levels of the text: from content to style. Still, the most direct indicator of who the author of a paper might be comes from the papers that are referenced - both from the references themselves and from the citation context they occur in within the content of an article. Our work contributes to the debate about the double-blind reviewing process and aims at contributing toward the rapidly emerging field of *Fairness in AI*. Although we found that the most prolific authors can be inferred with accuracy as high as 87.88% on ACL and 78.49% on EMNLP, the authors with less papers are more and more difficult to infer, which enforces the benefits of the double-blind review in offering any author a fair chance of having their papers accepted in top venues.

The finding that authors of anonymized papers can be predicted with such high accuracy bears important consequences for the way scientific articles are reviewed and published. De-anonymizing articles means compromising the integrity of the review and selection process. The insights into how the authors of an article can be inferred are not only interesting, but could help guide a reconsidered approach of the way we write papers for submission to various venues. Still, the findings and conclusions of this paper should not be seen as a premise that the portions of an article which help a *neural network* identify authorship are the same as those which help a *human reviewer* identify authorship, and are not necessarily expected to inform how humans perform peer review.

In future experiments, more attention to the contribution of each author of the article might lead to further improvements in the prediction performance. In this article, we construct our training datasets as if all parts of an article were written by all authors, which is not accurate, and could even put an upper bound on the network's performance, by providing it with contradictory information during training. Techniques for segmenting the text according to their probable authorship could help improve the method.

Acknowledgments

The first two authors of this paper (CC and AU) contributed equally. We thank our anonymous reviewers for their constructive comments and feedback, which helped improve our paper. This research is supported by NSF CAREER award 1802358 and NSF CRI award 1823292 to Cornelia Caragea. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 500–509, Portland, Oregon, USA. Association for Computational Linguistics.

Tim Althoff, Denny Britz, and Zifei Shan. 2013. Authorship attribution in multi-author documents.

Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 2012. Towards a computational history of the ACL: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21, Jeju Island, Korea. Association for Computational Linguistics.

Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891*.

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan R. Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar. Association for Computational Linguistics.

Stephen J Ceci and Douglas Peters. 1984. How blind is blind review? *American Psychologist*, 39(12):1491.

Sujatha Das Gollapalli and Cornelia Caragea. 2014. Extracting keyphrases from research papers using citation networks. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1629–1635. AAAI Press.

Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review?: author identification using only citations. *Acm Sigkdd Explorations Newsletter*, 5(2):179–184.

Julian Hitschler, Esther van den Berg, and Ines Rehbein. 2017. Authorship attribution with convolutional neural networks and POS-eliding. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58, Copenhagen, Denmark. Association for Computational Linguistics.

Mohammadreza Hojat, Joseph S. Gonnella, and Addeane S. Caelleigh. 2003. Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process. *Advances in health sciences education : theory and practice*, 8 1:75–96.

Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea, and C. Lee Giles. 2011. Context sensitive topic models for author influence in document networks. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 513–520, Manchester, UK. Coling 2008 Organizing Committee.

A.J. Meadows. 1998. *Communicating Research*. Academic Press, San Diego.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 689–696, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vahed Qazvinian, Dragomir R. Radev, and Arzucan Özgür. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 895–903, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andi Rexha, Stefan Klampfl, Mark Kröll, and Roman Kern. 2015. Towards authorship attribution for bibliometrics using stylometric features. In *CLBib@ ISSI*, pages 44–49.

Dylan Rhodes. 2015. Author attribution with cnns. *Avaiable online: https://www. semanticscholar. org/paper/Author-Attribution-with-Cnn-s-Rhodes/0a904f9d6b47dfc574f681f4d3b41bd840871b6f/pdf (accessed on 22 August 2016)*.

Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 213–222, New York, NY, USA. ACM.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States. AUAI Press.

Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. 2012. Authorship attribution with author-aware topic models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–269, Jeju Island, Korea. Association for Computational Linguistics.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

Amanda Spink and Bernard J Jansen. 2004. *Web search: Public searching of the Web*, volume 6. Springer Science & Business Media.

Andrew Tomkins, Min Zhang, and William D Heavlin. 2017. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.

Alex Williamson. 2003. What will happen to peer review? *Learned Publishing*, 16:15–20.