An Automatic Classification of the Primary and the Corresponding Authors in Research Articles

Sukhwan Jung
Department of Computer Science
University of South Alabama
Mobile, USA
shjung@southalabame.edu

Rituparna Datta
Department of Computer Science
University of South Alabama
Mobile, USA
rdatta@southalabama.edu

Aviv Segev
Department of Computer Science
University of South Alabama
Mobile, USA
segev@southalabama.edu

Abstract— Researchers often rely on the byline order in a publication to estimate relative contributions made by its authors, an assumption on which existing author contribution measures are based. This byline-based approach is, however, incompatible with the alphabetical author ordering, a practice still employed by many research fields. Manually requesting authors to state their contributions can overcome the limitation of the existing methods. Such approaches, however, require resource-intensive data acquisition and preprocessing, rendering them ungeneralizable to existing bodies of bibliographic records. The present paper proposed a compromise by focusing on distinguishing the main contributors from the rest of the authors using machine learning algorithms, bypassing the limitation of both the byline-based numerical author contribution methods and ungeneralizable manual approaches. The experiment validated the proposed approach by successfully classifying both the primary and the corresponding authors shown as the first and the last author without utilizing byline orders. The Random Forest classifier showed the best performances, successfully classifying the first author, the last author, and both with the accuracy of 0.90, 0.89, and 0.76 respectively.

Keywords— byline analysis, machine learning, author credit measure, citation analysis, scientometrics

I. Introduction

Publications represent specific research findings in the field of knowledge. New research is built around a body of knowledge provided by existing publications, crediting them in the form of citations. The number of citations given to a publication, therefore, reflects its perceived scientific achievement and importance, which in turn indicate the level of recognition its authors received in the field of research. Distribution of the recognition to the multiple co-authors, however, is not straightforward, as co-authors often contribute to the common publication in varying roles and degrees. One author could have written more, while the other did more experiments.

This complicated nature of the author contributions resulted in the practice of contribution-based byline order, where the authors in the bylines are ordered by their relative contributions to the publication, with the exception of the corresponding author, who is often positioned last. The differences in the contribution between co-authors are still hard to numerically measure because there are various factors involved in author contributions such as writing, experimenting, data processing, method devising, validating, supervising, and so on. The number of possible author roles shows the complexity of factors required to ascertain the correct representation of author contributions, even with manual interpretations from the authors themselves.

There are a number of measures designed to distribute values among the authors in the byline using simple mathematical equations to automatically represent numerical author contributions. The most basic method is a straight counting [1], where only the first author is considered *cited*, where the complex problem of contribution distribution is removed from the equation. Such an approach does not match the scientific trend towards multi-authored publications and is quickly deemed unfit to represent author contributions [2]. The basic approach of equal contribution was unable to distinguish the main contributors from the rest of the authors, hence many of the measures were developed to base their calculations on the position of authors in the bylines [3]. These measures assign the highest values to the first authors based on the assumption that the bylines are ordered by their contributions, using various mathematical functions including proportions, geometric sequences, harmonic functions, and golden ratios. Some measures are also conscious of the corresponding authors, allocating special care to the last authors in bylines.

Regardless of the approaches, the aforementioned author contribution measures based on the byline orders share two main limitations. First, they treat every *i*th co-author in publications with *n* total authors to have exactly the same degree of contributions. This is an incorrect assertion of the author contributions, as author contribution differs for all publications even with the identical group of co-authors; the contribution of each author in each publication is different. Second, they cease to function properly where alphabetically-ordered bylines are concerned, which is still a common practice for many fields of research such as mathematics or business and finance, where more than two thirds the publications are intentionally ordered alphabetically by author names [4]. Manually requesting the authors for author contribution is a resource-intensive task which cannot be used on existing data. The complexity of the problem and lack of an effective and affordable solution led the researchers to use the basic approach of crediting all authors the same with full values; many of the major bibliographic search engines such as

Google Scholar or Scopus employ such an approach, increasing the citation count of all the authors when the publication is cited. The obvious discrepancies between the primary authors, the corresponding authors, and the rest are ignored to preserve generalizability and scalability.

The proposed approach is a compromise to overcome the problem of generalizability with the alphabetically-ordered byline dataset; instead of assigning numerical contribution values to the authors, the proposed approach focuses on distinguishing main contributors such as the primary and the corresponding authors from the rest. Ideally, the experiment result is compared against the golden answer set where the identity of the primary and the corresponding authors are known in the alphabetically ordered bylines, but ascertaining such a set in a large-scale dataset requires resource-intensive data collection and preprocessing. The proposed approach instead aims to identify the first and the last authors from the rest in a predominantly contribution-ordered byline dataset. The first and the last authors in such a dataset respectively represent the primary and the corresponding authors, effectively retaining the practical functionality of the proposed approach. It is also generalizable to a name-ordered byline dataset if the byline order is not considered during the calculation. This renders the research problem to classifying the main contributing – either the first or the last – authors from the author pool using author-related features, which are binary classification problems solvable with various machine learning (ML) algorithms. It is assumed that the main contributors are more active, more collaborative, and more recognized in the research fields than non-main contributors and the features related to the author activities are extracted from the bibliographic networks to train such classifiers. Six ML algorithms were implemented in the experiment: Decision Tree, Random Forest, K-Nearest Neighbors, Logistic Regression, Linear Discriminant Analysis, and Gaussian Naïve Bayes. Their classification results are compared to show the effect of different classification approaches on the performance of the proposed approach.

II. Background Research

Measuring the different contributions between co-authors is a problem with a long history. There are a number of measures in the field of informetrics assigning different contributions to authors according to their position in the byline [3], which are listed below using following notations: the author contribution *w(k)* for an author is calculated with *k* and A where A is the total number of authors and the author position in the byline is k = 1,…, A.

Standard counting is the most common practice where all authors receive full credit regardless of their position in the byline. This is the simplest method of all and is widely used for most bibliographic databases for its simplicity, but this method causes overweight problems, where the total weight of a publication increases proportional to the number of its authors, penalizing publications with fewer authors.

$$w(k) = 1$$

Fractional counting or uniform counting is introduced to deal with the overweight problem by normalizing the weight between the authors in a single byline [5], resulting in total weights of all papers to be one.

$$w(k) = \frac{1}{A}$$

While computationally simple, both the commonly used standard counting and the normalized fractional counting are limited in that they give equal credit to all of the authors for each publication, contrary to the widely accepted belief that authors contribute differently to any publication. Proportional counting is proposed to assign author weights relative to the order of authors in a byline, giving an inverse of their position as the weight [6], which is further normalized to solve the overweight problem.

$$w(k) = \frac{1}{k}, \qquad w(k) = \frac{2}{A} \times \left(1 - \frac{k}{A+1}\right)$$

Geometric counting is a similar approach for order-based author weight, where the ratio of author weights for consecutive authors is equal [7].

$$\frac{w(2)}{w(1)} = \frac{w(3)}{w(2)} = \cdots = \frac{w(A)}{w(A-1)} = \lambda$$

A simplified method is proposed with $\lambda = 0.5$ across all byline sizes [8].

$$w(k) = \frac{2^{A-k}}{2^{A} - 1}$$

Harmonic counting originated from the harmonic function [9] and is defined as

$$w(k) = \frac{1/k}{\sum_{k=1}^{A}(\frac{1}{k})}$$

Golden counting [7] assigns author weight based on the golden number $\varphi = 0.6180$ where $1 - \varphi = \varphi^{2}$, and is defined as

$$\begin{cases} w(k) = 1 & k = 1 & A = 1 \\ w(k) = \varphi^{2k-1} & k = 1, \dots, A-1 & A \geq 2 \\ w(k) = \varphi^{2k-2} & k = A & A \geq 2 \end{cases}$$

Counting measures based on byline orders are capable of capturing the diminishing degree of author contributions depicted in the byline, but fails to consider the corresponding authors, who are often placed as the last author in contribution-ordered bylines. Noblesse oblige counting bases its algorithm in the long tradition of having a corresponding author at the end of the byline, deeming the last author the most important among the list of authors by giving half of the total credit to the last author [10].

$$\begin{cases} w(k) = 0.5 & k = A \\ w(k) = \dfrac{1}{2 \cdot (A-1)} & k = 1, \dots, A-1 \end{cases}$$

The Noblesse oblige applies a fractional approach to all non-corresponding authors, and first/last counting is introduced to enable differential credit assignment to the non-corresponding authors; both the first author and the corresponding author receive full weight, while the authors in between receives relatively diminishing credit based on their position [11].

$$\begin{cases} w(k) = 1 & k = 1 \; or \; A = 1 \\ w(k) = 0.7 & k = 2 \; and \; A = 3 \\ w(k) = \dfrac{2 \cdot (A - k + 1)}{(A + 1) \cdot (A - 2)} & A - 1 \geq k \geq 2 \; and \; A \geq 4 \end{cases}$$

Although they are the most sophisticated approach, such methods are rarely practiced to measure the author contribution in practice. Standard counting, the most straightforward with the least consideration of the differences in the author credit, is the most used method. This is due to the limitation of the alternatives; various research domains continue practicing alphabetical ordering as an alternative author listing method, when the ratio of intentionally alphabetically-ordered authors can reach as high as 73.3% in the field of mathematics and 68.3% in business and finance [4]. One of the attempts to overcome such limitations is the CrediT[1] taxonomy, aiding a manual recording of author contribution by providing information on the author contribution for fourteen different categories. The manual approaches, however, are resource-intensive while being ungeneralizable to the other existing dataset due to the necessity of methods-specific data.

Most of the existing methods require author bylines ordered by the author contributions and hence are incapable of assigning different contribution values for authors in different bylines with the same relative position while being incompatible with the research domains where alphabetical ordering is practiced. The present paper proposes a compromise to author contribution, measuring where the exact numerical representation of author contribution is omitted; only the main contributors such as the primary and the corresponding authors are distinguished from the other authors. This is a binary classification problem which Machine Learning methods can solve. A binary classification is a form of supervised learning, classifying the labels of given data records into two distinct groups. The machine learning classifiers are trained on a dataset with known labels to build a statistical model classifying – predicting – the labels of new data. There are diverse applications of binary classifications ranging from text processing to medical diagnosis [12]–[19]; however, binary classifications have not been used in the context of the present work.

III. Dataset and Models

A. Dataset

The paper utilizes six ML binary classifiers to distinguish main contributors such as the primary and the corresponding authors from the other authors. A binary classification is a form of supervised learning, which requires the existence of a training set with known labels to function properly. There are no readily-available datasets where the identities of the primary and the corresponding authors are known in the alphabetically ordered bylines, however. Manually retrieving author contributions from the actual authors can take months for relatively small-scale data [20], hence the paper used the position in the contribution-ordered byline as the training labels instead. The first and the last authors in a predominantly contribution-ordered dataset respectively represent the primary and the corresponding authors, effectively retaining the practical functionality of the proposed approach. Byline orders are not used during the training and are only retained for validation purposes.

Microsoft Academic Graph (MAG) is a heterogeneous bibliographic dataset [21] containing over 210 million publications and 254 million authors. While created by Microsoft in recent years, it was deemed competitive with major bibliographic search engines such as Google Scholar or Scopus [22]. The bulk download is available on Azure, which is updated weekly, and the bulk dataset as of January 30, 2018 used in the previous research [23] was downloaded for the experiment.

The dataset is filtered to a specific domain of *human computer interaction*, where bylines are predominantly ordered by contributions [4], when this domain is relatively small compared to other larger research domains. Publications with *HCI* and *Human-Computer Interaction* with and without the dash character in their keywords, titles, and fields of study (FOS) were used; the field of study is a categorization of publication records automatically built from an iterative graph link analysis and entity filtering by Microsoft, which are confirmed by scanning through their meta-information such as title, keywords, and abstracts. Any invalid entries missing id, year, authors, and references in their properties were also removed. The resulting dataset is a bibliographic graph with 170,060 authors, 712,228 publications, and 1,935,659 citations. For the purpose of the experiment, publications before the year 1998 were excluded, resulting in a final dataset with 140,494 authoring relationships.

The size of co-authors affect the relative ratio of the first and the last author in a byline, hence *minAuthorCount* was used to filter out the publications in the dataset with too small byline lengths. Three thresholds 1, 3, and 5 were used in the experiment; for each iteration, bylines with lengths smaller than *n* were excluded. An increase in *minAuthorCount*, as shown in Fig. 1, results in the smaller number of authoring relationships in the dataset with a lower ratio of *true* labels in the dataset. The total number of authoring relationships decreased from 140,494 to 48,607 with increasing *minAuthorCount*, while the ratio of the first and the last authors also decreased; the ratio of *isEither declines* from 59.35% to 34.93%, showing the labels become more imbalanced with smaller dataset size. Such skewness could affect the training results while the ratio did not reach extreme levels, and minority class was up-scaled for each experiment iteration to balance the number of *true* and *false* labels during the training and classifying process. To remove the possibility of classifiers trained with the order of authors given, ten-fold cross-validation is used to create ten training/test sets with randomized orders; the results for the experiment iteration are generated by averaging the ten outcomes.

[1] https://casrai.org/credit/

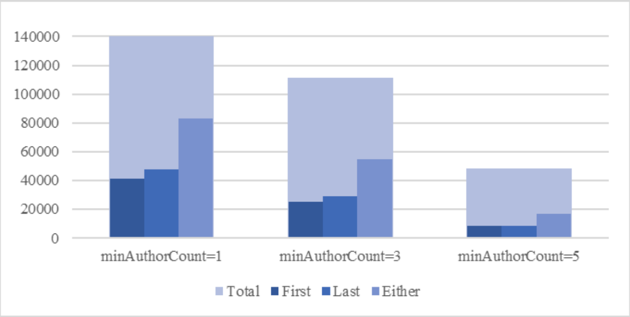


Fig. 1. Changes in the relative ratio of the first and last authors to the number of total authors in the dataset with different *minAuthorCount* threshold.

B. *The Main Contributor Classification*

Decision Tree is a supervised learning method that does not depend on any parameter tuning. It can be used for both classification and regression. Decision Tree works by creating a tree-like structural model to predict target variable value which learns rules to make an effective decision from input data. Decision Tree has internal nodes and leaves. Other than the root node, all internal nodes have one parent. Classification is performed recursively with a top-down manner from the top node to a leaf. Random Forest is a very well-known classification technique that can also be used for regression [24]. It was proposed in 2001 by Breiman [25]. It integrates several weak Decision Tree classifiers, and the final decision is made based on the highest number of votes; as a result, it is an ensemble learning approach. It has already been applied to exploit nonlinear relationships and multi-class classification and in the problems where the number of variables is more than the number of observations. K-Nearest Neighbors (KNN) is also a supervised machine learning technique that can be used for both classification and regression. Its working principle is very simple and easy to implement. As the name suggests, an object is classified based on the nearest training sample in the input variable space.

Logistic Regression is a mathematical model based on the sigmoid function for a binary regression matching several given inputs with a single binary output [26]. Use of the sigmoid function results in the easy conversion of regression estimation to the probability value ranging from 0 to 1. The existence of low and high plateaus at both ends acts as buffers to extreme input variables while remaining sensitive to the range of variables in the middle of the s-shaped curve [27].

Linear Discriminant Analysis (LDA), or the Fisher's linear discriminant [28], expresses one categorical outcome variable with the linear combination of multiple continuous input variables similar to Logistic Regression and can be viewed as the opposite of analysis of variance (ANOVA) where the input is discrete and the output is continuous [29]. A number of assumptions are made to the input data for multi-class LDA, but it is robust enough to remain effective when some of the assumptions were violated by the input data [30]. Gaussian Naïve Bayes is one of the simple classifiers based on mathematical probabilities, but its simplicity and scalability make it one of the widely used machine learning classifiers to date [31], [32]. As its name suggests, Gaussian Naïve Bayes is a probabilistic classifier for continuous input with Gaussian distributions, which are naively assumed to have strong independence between each other, using the Bayes' Theorem [33] to calculate the outcome's conditional probabilities [34].

The paper aims to classify the primary and the corresponding authors from others in a set of bylines. Given that the research domain in the dataset predominantly utilized contribution-based author ordering, it is effectively the same as classifying the first and the last authors without utilizing the byline orders in the classification process. The binary classification task for the experiment hence is *given a publication, classify the author types without accessing byline orders*. Each author in a byline is classified whether he/she is 1) the first author *isFirst*, 2) the last author *isLast*, or 3) either of them *isEither* using only the information unrelated to their byline position. Six ML algorithms were utilized to generate set of results for each author type. Sklearn (https://scikit-learn.org/) is a Python library based on SciPy to provide robust machine learning functions to the Python environment, supported by companies such as Nvidia, Microsoft, and Intel. The library is focused on the modeling phase of the process, allowing freely modification and filtration of the dataset as required during the experiment. Six widely used existing ML classifiers explained above are implemented in the Sklearn library, and were used in the experiment; *decision tree classifier, random forest classifier, k-nearest neighbors classifier, linear logistic regression, linear discriminant analysis, and Gaussian naïve Bayes*. Changes to the default function attributes were kept to a minimum to classify labels without tuning; only the random state for the *decision tree* and the *random forest* classifiers were set to a set value of zero to preserve the classification result over multiple iterations. Each classifier is trained on the training set filtered by different *minAuthorCount* with the selected set of features, then its accuracy is assessed on the test set. Accuracies of the six ML classifiers were compared against three baseline classifiers; a most frequent classifier (B_freq) that always predicts most frequently shown labels, a stratified classifier (B_strat) that predicts labels proportional to the labeling ratio in the training set, and a uniform classifier (B_unif) that uniformly predicts all labels.

The machine learning binary classifiers provide means of identifying the binary class of given data based on the series of inputs. TABLE I. shows the seven features used to train ML classifiers and descriptions for each feature.

TABLE I. FEATURES USED FOR THE ML CLASSIFIERS

Feature	Description
citeC	Citation count in the dataset.
pubC	Publication count in the dataset.
coauthC	Overlapping co-author count in the dataset.
citeCy	Citation count up to the year y.
pubCy	Publication count up to the year y.
coauthCy	Overlapping co-author count up to the year y.
yDiff	Number of years from y to the last year *lastYear* = 2018.

The underlying assumption in the proposed approach is that the main contributors are more active, more collaborative, and more recognized, and such author features were used to train the classifiers to distinguish the main contributors from other authors. The features used were limited to the basic features for the ease of computation and to show that the proposed experiment is generalizable; all features can be extracted from any bibliographic records. Citation count *citeCy*, the number of publications *pubCy*, and the number of co-authors *coauthCy* were calculated for authors $a_1,a_2,\ldots,a_n$ of any publication *p*, where *n* is the number of authors in *p*. *citeCy*, *pubCy*, and *coauthCy* are sensitive to the publication year *y* of *p*, which are the same author features calculated up to the year *y*. The features within the whole dataset and up to year *y* are differentiated to test if the differences in such numbers impact the classification result. An age of publication *yDiff* = *lastYear* – *y* is added as a publication-related feature, leading to a total of seven features.

24 feature combinations in TABLE II. were selected out of 63 possible combinations to analyze the impact of different criteria used during the feature selection: six individual features TABLE II. (a), three sets grouped by author activity TABLE II. (b), two sets grouped by year sensitivity TABLE II. (c), and one set with all six features TABLE II. (d). Twelve feature combinations with and without *yDiff* result in a total of 24 feature combinations, and a total of 2,160 experiment iterations.

TABLE II. TWELVE COMBINATIONS OF AUTHOR FEATURES USED IN THE EXPERIMENT.

<table>
<tr><td>(a)</td><td colspan="6"><b>Individual features</b></td></tr>
<tr><td></td><td>citeC</td><td>citeCy</td><td>coauthC</td><td>coauthCy</td><td>pubC</td><td>pubCy</td></tr>
<tr><td>(b)</td><td colspan="6"><b>Grouped by author activity</b></td></tr>
<tr><td></td><td colspan="2">citeC, citeCy</td><td colspan="2">coauthC, coauthCy</td><td colspan="2">pubC, pubCy</td></tr>
<tr><td>(c)</td><td colspan="6"><b>Grouped by year sensitivity</b></td></tr>
<tr><td></td><td colspan="3">citeC, coauthC, pubC</td><td colspan="3">citeCy, coauthCy, pubCy</td></tr>
<tr><td>(d)</td><td colspan="6"><b>All six features</b></td></tr>
<tr><td></td><td colspan="6">citeCy, coauthCy, pubCy, citeC, coauthC, pubC</td></tr>
</table>

C. *Evaluation*

Four types of results – True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) - were counted from the confusion matrices generated for each classification result. Accuracy, sensitivity, and specificity in TABLE III. are calculated from the confusion matrices, then averaged to evaluate classification results from a combination of input variables. The Receiver Operating Curve (ROC) is drawn for each method as well.

TABLE III. METRICS USED FOR BINARY CLASSIFICATION RESULT.

Metric	**Formula**
Accuracy	The ratio of all True classifications, (TP+TN)/All
Sensitivity	The ratio of correct Positive classifications, TP/P
Specificity	The ratio of correct Negative classifications, TN/N

IV. RESULTS

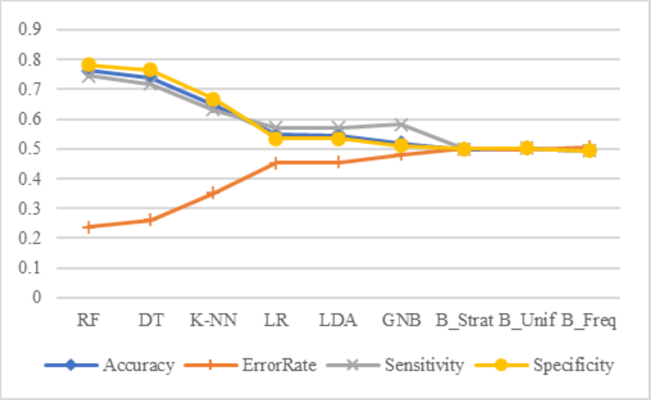
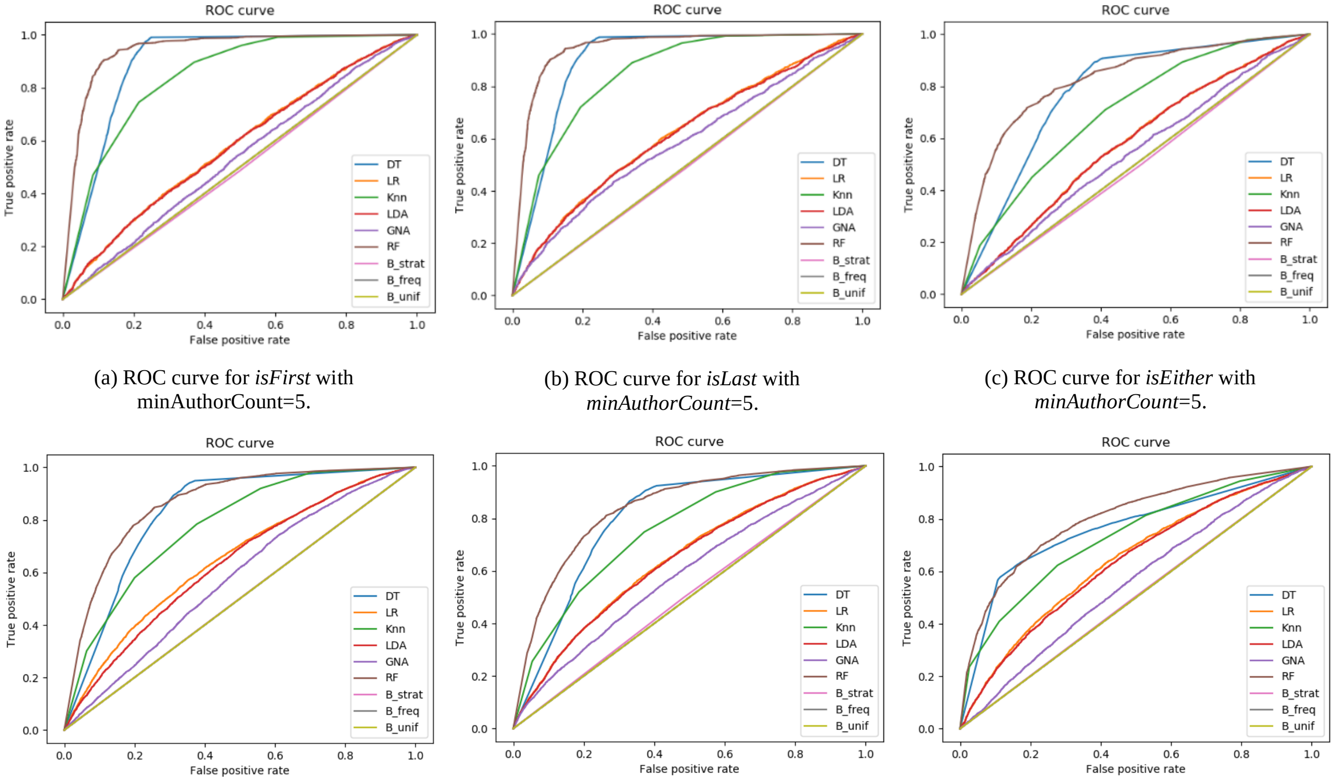


Fig. 2. Classification Metrics for Classification Results for *isEither* with Seven Features and *minAuthorCount*=5.

The experiment results showed that all six ML algorithms were able to classify the main contributors better than all three baseline methods B_freq, B_strat, and B_unif for all three labels with different *minAuthorCount*, with classification accuracy reaching up to 0.89 on specific combinations of ML algorithms and target labels. Three labels *isFirst*, *isLast*, and *isEither* were classified over the dataset filtered with a different *minAuthorCount*=1, 3, 5 using 24 feature combinations, resulting in a total of 216 classification accuracy results for each of the nine classifiers. Fig. 2 visualizes a sample result for *isEither* classification comparing nine methods trained with all features on the dataset with *minAuthorCount*=1.

B_unif and B_freq baseline classifiers do not perform well in the negatively-skewed data used in the experiment, with the former returning lower negatives and the latter returning no positives. Upscaling the labels resulted in all baseline methods becoming binary guessing in practice, leading to roughly 0.5 accuracy for all metrics. With only one label classified for each iteration, the result for B_freq contains either no positives or negatives depending on the training set's label ratio; sensitivity or specificity cannot be calculated in such cases, which are ignored in the graph.

The results indicate that RF (Random Forest), DT (Decision Tree), and KNN are best at classifying either the first or the last authors with high accuracy, sensitivity, and specificity compared to the Logistic Regression (LR), LDA, and Gaussian Naïve Bayes (GNB). RF outperformed the other two methods, reaching an overall accuracy of 0.90 for *isFirst*. DT showed a marginally lower performance of 0.87 while KNN had a significantly lower accuracy of 0.77. RF showed the best performance in sensitivity and specificity as well, stating its superiority over the others. LR, LDA, and GNB are outperformed by the other ML methods while still outperforming the three baseline methods. The same is true for *isFirst* and *isLast* as visualized by the ROC curves in Fig. 3(a, b), where RF, DT, and KNN have noticeably higher area under the curve (AUC) value, reaching over 0.94, 0.89, and 0.84 respectively, while LR, LDA, and GNB all failed to reach 0.62. The likely cause of such performance disparity is the multicollinearity nature of the used author features, reducing the predictive accuracy of the latter three linear algorithms.



(a) ROC curve for *isFirst* with minAuthorCount=5.

(b) ROC curve for *isLast* with *minAuthorCount*=5.

(c) ROC curve for *isEither* with *minAuthorCount*=5.

(d) ROC curve for *isFirst* with *minAuthorCount*=1.

(e) ROC curve for *isLast* with *minAuthorCount*=1.

(f) ROC curve for *isEither* with *minAuthorCount*=1.

Fig. 3. ROC Curve for Six ML algorithms and Three Baselines with Seven Features.

This proves that classifying the main contributors with author-based features and ML algorithms can result in a meaningful degree of precision, while non-linear classifiers show superior performances. Larger disparities between ROC curves in Fig. 3(a), Fig. 3(b) compared to Fig. 3(c) also prove that the ML-based approach is capable of distinguishing the difference between the primary and the corresponding authors more than the sum of them, even without algorithm tuning.

Relative performance differences persist with different *minAuthorCount* as shown in Fig. 3(d-f); RF, DT, and KNN show AUC over 0.83, 0.79, and 0.75 on average. The classification metrics for different thresholds can be seen in TABLE IV. ; three non-performing linear classification algorithms were removed from the analysis as they showed relatively smaller performance increases. Accuracy, sensitivity, and specificity of RF and DT increased with higher thresholds classifying all three labels, with *isEither* showing an exception

TABLE IV. CHANGES IN CLASSIFICATION ACCURACIES, SENSITIVITIES, AND SPECIFICITIES FOR RANDOM FORST, DECISION TREE, AND K-NEAREST NEIGHBORS WITH VARYING *MINAUTHORCOUNT*, TRAINED BY ALL SEVEN FEATURES.

		Accuracy			**Sensitivity**			**Specificity**		
	minAuthor Count	***isFirst***	***isLast***	***isEither***	***isFirst***	***isLast***	***isEither***	***isFirst***	***isLast***	***isEither***
Random Forest	1	0.7961	0.7720	0.7299	0.7695	0.7524	0.7588	0.8286	0.7948	0.7069
	3	0.8350	0.8236	*0.6850*	0.7980	0.7966	*0.6763*	0.8824	0.8560	*0.6948*
	5	**0.8957**	**0.8940**	**0.7630**	**0.8566**	**0.8598**	**0.7457**	**0.9442**	**0.9352**	**0.7830**
Decision Tree	1	0.7787	0.7545	0.7333	0.7545	0.7373	0.7908	0.8080	0.7744	0.6948
	3	0.8134	0.7994	*0.6696*	0.7744	0.7693	*0.6659*	0.8654	0.8371	*0.6734*
	5	**0.8680**	**0.8645**	**0.7388**	**0.8174**	**0.8183**	**0.7173**	**0.9378**	**0.9267**	**0.7652**
KNN	1	0.7100	0.6876	**0.6656**	0.6827	0.6678	**0.6839**	0.7470	0.7127	0.6506
	3	0.7279	0.7130	*0.6193*	0.6894	0.6833	*0.6114*	0.7859	0.7543	*0.6284*
	5	**0.7659**	**0.7684**	0.6475	**0.7107**	**0.7180**	0.6315	**0.8604**	**0.8491**	**0.6677**

showing a temporary performance drop at *minAuthorCount* = 3. This is due to the fact that single-authored publications are utilized with the minimum threshold. These authors are both the first and the last author at the same time, providing bridges between otherwise distant first and last authors for other publications with common feature profiles between the two. KNN exhibits a more pronounced effect of single-authored publications, where the accuracy and sensitivity both decrease with a larger threshold for *isEither*. This indicates that distance-based ML algorithms can be used to identify the main contributors in a field with more single-authors, while the algorithms based on Decision Tree are more suitable for fields with heavy co-authoring behaviors.

The effect of different feature sets used in the training is observed in TABLE V. , where the classification accuracies of the Random Forest method are shown in the ascending order of accuracy for *isEither*. The table shows that the ML method outperformed the baseline results even when trained with a single author feature such as the number of citations or the number of publications. *yDiff* increases the classification accuracy by 7.04% on average, but there was no single feature responsible for the high performance reached with all seven features; higher performances in *isFirst*, *isLast*, and *isEither* all have high correlations with the number of features used in the training. The identification of the publications' main contributors relies on the combination of author activities, which reflects the diverse considerations required for the problem. Such correlations between the number of features used and the classification accuracy are less pronounced in the three less performing ML algorithms. TABLE VI. showed *citeC* was the single most important input for the least performed Gaussian Naïve Bayes method, followed by *citeC* used with *yDiff*. The likely cause for such a pattern is the overall performance of the method being too close to the random baselines, rendering the ranking of feature sets insignificant.

TABLE V. CLASSIFICATION ACCURACIES FOR RANDOM FOREST METHOD WITH MINAUTHORCOUNT=5, IN ACESNDING ORDER.

Feature Set	isFirst	isLast	isEither
coauthCy	0.5439	0.5667	0.5173
coauthC	0.5471	0.5728	0.5350
pubCy	0.5505	0.5789	0.5389
citeCy	0.5430	0.5733	0.5395
pubC	0.5518	0.5748	0.5440
…	…	…	…
coauthCy_coauthC_yDiff	0.7648	0.7683	0.6677
citeCy_coauthCy_pubCy_yDiff	0.8260	0.8269	0.7126
citeC_coauthC_pubC_yDiff	0.8746	0.8734	0.7479
citeCy_coauthCy_pubCy_citeC_coauthC_pubC	0.8727	0.8725	0.7480
citeCy_coauthCy_pubCy_citeC_coauthC_pubC_yDiff	0.8957	0.8940	0.7630

TABLE VI. CLASSIFICATION ACCURACIES FOR GAUSSIAN NAÏVE BAYES METHOD WITH MINAUTHORCOUNT=5, IN ACESNDING ORDER.

Feature Set	*isFirst*	*isLast*	*isEither*
citeC	0.5171	0.5287	0.5110
citeC_yDiff	0.5175	0.5295	0.5113
citeC_citeCTotal	0.5143	0.5270	0.5126
citeC_citeCTotal_yDiff	0.5147	0.5271	0.5128
citeCTotal	0.5098	0.5222	0.5132
…	…	…	…
publishC_publishCTotal	0.5235	0.5526	0.5211
publishC_publishCTotal_yDiff	0.5239	0.5526	0.5211
citeCTotal_coauthCTotal_publishCTotal	0.5159	0.5432	0.5212
publishCTotal_yDiff	0.5191	0.5522	0.5219
publishCTotal	0.5171	0.5496	0.5227

The experiment showed that the first and the last authors in a predominantly contribution-ordered byline dataset can be classified with high degree of accuracy. Different classification accuracies for ML algorithms indicated some are more preferable than others for the task, but all six showed accuracy above three baseline classifiers for classifying the primary authors, the corresponding authors, and the main contributors. Basic features based on four sets of information – year of publication, publication count, citation count, number of co-authors – resulted in high classification accuracy reaching up to 89% for both the first and the last author classifications, indicating performance improvement is possible through more elaborate feature selection and engineering. The general applicability of binary machine learning classifiers and the basic author features used in the proposed approach allowed the proposed approach be generalizable to any existing bibliographic datasets.

V. CONCLUSION

The main contributors to multi-authored publications are often distinguished by their position in the byline under the assumption that the primary authors are placed first and the corresponding authors last. Bylines alphabetically ordered by the author names invalidate the assumption, making existing methods not applicable in a number of research domains. The present paper proposed a ML-based author classification approach using author features excluding their byline orders. Alphabetically-ordered byline datasets lack a readily-available golden answer set on the identity of the primary and the corresponding authors, and therefore the first and the last authors in a predominantly contribution-ordered byline dataset were used in the experiment.

The proposed approach successfully identified the main contributors with different ML binary classifiers, while also showing the capability to distinguish the difference between the primary and the corresponding authors to a high degree even without algorithm tuning. Exclusion of byline orders in the process guarantees identical results if the bylines were ordered alphabetically, showing the generalizability of the proposed approach. Six widely used ML algorithms were employed without any tuning to compare the effect of different basic ML approaches. All of the ML algorithms perform better than the baseline even when trained with a single author feature such as the number of publications the author has; this result supports the validity of the author classification based on the author features extracted from bibliographic graphs. Decision Tree and Random Forest performed best, reaching accuracies over 89% for classifying the first and the last authors, while linear regression methods performed worse due to the feature multicollinearity; this proves that classifying the main contributors with author-based features and ML algorithms can result in a meaningful degree of precision, while non-linear classifiers are better than others. Performance analysis over different byline length thresholds also revealed that the distance-based ML algorithm can be used to identify the main contributors in a field with more single-authored publications, while the algorithms based on decision trees are more suitable for fields with heavy co-authoring.

The compromise between accurate author contribution measure and generalizability resulted in a successful author classification based on their research activities. The validity of the proposed approach is shown by the experiment, and future works will be focused on performance improvement, training additional ML algorithms with tuning with more author features as suggested by the experiment result. Theoretically shown generalizability will be empirically proved by testing it on the set of alphabetically ordered publications with known primary and corresponding authors.

References

[1] J. R. Cole and S. Cole, “Social Stratification in Science,” *American Journal of Sociology*, vol. 83, no. 2, pp. 491–492, Sep. 1977, doi: 10.1086/226571.

[2] D. Lindsey, “Production and Citation Measures in the Sociology of Science: The Problem of Multiple Authorship,” *Social Studies of Science*, vol. 10, no. 2, pp. 145–162, May 1980, doi: 10.1177/030631278001000202.

[3] R. Todeschini and A. Baccini, *Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research*. John Wiley & Sons, 2016.

[4] L. Waltman, “An Empirical Analysis of the Use of Alphabetical Authorship in Scientific Publishing,” *Journal of Informetrics*, vol. 6, no. 4, pp. 700–711, Oct. 2012, doi: 10.1016/j.joi.2012.07.008.

[5] Q. Burrell and R. Rousseau, “Fractional Counts for Authorship Attribution: A Numerical Study,” *Journal of the American Society for Information Science*, vol. 46, no. 2, pp. 97–102, 1995, doi: 10.1002/(SICI)1097-4571(199503)46:2<97::AID-ASI3>3.0.CO;2-L.

[6] G. V. Hooydonk, “Fractional Counting of Multiauthored Publications: Consequences for the Impact of Authors,” *Journal of the American Society for Information Science*, vol. 48, no. 10, pp. 944–945, 1997, doi: 10.1002/(SICI)1097-4571(199710)48:10<944::AID-ASI8>3.0.CO;2-1.

[7] N. Assimakis and M. Adam, “A New Author’s Productivity Index: P-Index,” *Scientometrics*, vol. 85, no. 2, pp. 415–427, Nov. 2010, doi: 10.1007/s11192-010-0255-z.

[8] L. Egghe, R. Rousseau, and G. V. Hooydonk, “Methods for Accrediting Publications to Authors or Countries: Consequences for Evaluation Studies,” *Journal of the American Society for Information Science*, vol. 51, no. 2, pp. 145–157, 2000, doi: 10.1002/(SICI)1097-4571(2000)51:2<145::AID-ASI6>3.0.CO;2-9.

[9] S. E. Hodge, D. A. Greenberg, and C. E. Challice, “Publication Credit,” *Science*, vol. 213, p. 950, Aug. 1981, doi: 10.1126/science.213.4511.950.

[10] H. A. Zuckerman, “Patterns of Name Ordering among Authors of Scientific Papers: A Study of Social Symbolism and Its Ambiguity,” *American Journal of Sociology*, Oct. 2015, doi: 10.1086/224641.

[11] C.-T. Zhang, “A Proposal for Calculating Weighted Citations based on Author Rank,” *EMBO Reports*, vol. 10, no. 5, pp. 416–417, May 2009, doi: 10.1038/embor.2009.74.

[12] Ł. Gadomer and Z. A. Sosnowski, “Knowledge Aggregation in Decision-Making Process with C-Fuzzy Random Forest using OWA Operators,” *Soft Comput*, vol. 23, no. 11, pp. 3741–3755, Jun. 2019, doi: 10.1007/s00500-018-3036-x.

[13] S. T. Selvi, P. Karthikeyan, A. Vincent, V. Abinaya, G. Neeraja, and R. Deepika, “Text Categorization using Rocchio Algorithm and Random Forest Algorithm,” in *2016 Eighth International Conference on Advanced Computing (ICoAC)*, Chennai, India, Jan. 2017, pp. 7–12, doi: 10.1109/ICoAC.2017.7951736.

[14] E. V. A. Sylvester *et al.*, “Applications of Random Forest Feature Selection for Fine-Scale Genetic Population Assignment,” *Evol Appl*, vol. 11, no. 2, pp. 153–165, Feb. 2018, doi: 10.1111/eva.12524.

[15] S. Mitra, K. M. Konwar, and S. K. Pal, “Fuzzy Decision Tree, Linguistic Rules and Fuzzy Knowledge-Based Network: Generation and Evaluation,” *IEEE Trans. Syst., Man, Cybern. C*, vol. 32, no. 4, pp. 328–339, Nov. 2002, doi: 10.1109/TSMCC.2002.806060.

[16] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, “The Comparation of Text Mining with Naive Bayes Classifier, Nearest Neighbor, and Decision Tree to Detect Indonesian Swear Words on Twitter,” in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Denpasar, Bali, Indonesia, Aug. 2017, pp. 1–5, doi: 10.1109/CITSM.2017.8089231.

[17] Y.-Y. Song and Y. Lu, “Decision Tree Methods: Applications for Classification and Prediction,” *Shanghai Arch Psychiatry*, vol. 27, no. 2, pp. 130–135, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.

[18] S. Tan, “Neighbor-Weighted K-Nearest Neighbor for Unbalanced Text Corpus,” *Expert Systems with Applications*, vol. 28, no. 4, pp. 667–671, May 2005, doi: 10.1016/j.eswa.2004.12.023.

[19] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” *Procedia Technology*, vol. 10, pp. 85–94, 2013, doi: 10.1016/j.protcy.2013.12.340.

[20] S. Boyer, T. Ikeda, M.-C. Lefort, J. Malumbres-Olarte, and J. M. Schmidt, “Percentage-based Author Contribution Index: a universal measure of author contribution to scientific articles,” *Res Integr Peer Rev*, vol. 2, no. 1, p. 18, Dec. 2017, doi: 10.1186/s41073-017-0042-y.

[21] D. Herrmannova and P. Knoth, “An Analysis of the Microsoft Academic Graph.” http://mirror.dlib.org/dlib/september16/herrmannova/09herrmannova.html.

[22] S. E. Hug, M. Ochsner, and M. P. Brändle, “Citation Analysis with Microsoft Academic,” *Scientometrics*, vol. 111, no. 1, pp. 371–378, Apr. 2017, doi: 10.1007/s11192-017-2247-8.

[23] S. Jung and W. C. Yoon, “Citation-based Author Contribution Measure for Byline-Independency,” Presented at 2019 IEEE International Conference on Big Data (IEEE Big Data 2019), Dec. 2019.

[24] A. Liaw and M. Wiener, “Classification and Regression by randomForest,” *R news 2.3*, vol. 2, pp. 18–22, 2002.

[25] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[26] J. S. Cramer, “The Origins of Logistic Regression,” *SSRN Journal*, 2003, doi: 10.2139/ssrn.360300.

[27] D. G. Kleinbaum, M. Klein, and E. Rihl Pryor, *Logistic Regression: A Self-Learning Text*, 3. ed. New York, NY: Springer, 2010.

[28] R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[29] D. Wetcher-Hendricks, *Analyzing Quantitative Data: An Introduction for Social Researchers*. Hoboken, N.J: Wiley, 2011.

[30] W. R. Klecka, *Discriminant Analysis*, Nachdr. Newbury Park, Calif.: Sage Publ, 2003.

[31] J. C. Griffis, J. B. Allendorfer, and J. P. Szaflarski, "Voxel-Based Gaussian Naïve Bayes Classification of Ischemic Stroke Lesions in Individual T1-Weighted MRI Scans," *Journal of Neuroscience Methods*, vol. 257, pp. 97–108, Jan. 2016, doi: 10.1016/j.jneumeth.2015.09.019.

[32] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, "Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection using Random Forest and Gaussian Naïve Bayes," *PLOS ONE*, vol. 9, no. 1, p. e86703, 24 2014, doi: 10.1371/journal.pone.0086703.

[33] J. Joyce, "Bayes' Theorem," Jun. 2003, Available: https://stanford.library.sydney.edu.au/entries/bayes-theorem/.

[34] H. Zhang, L. Jiang, and J. Su, "The Optimality of Naive Bayes," in *In Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference*, pp. 562–567.