# Applying Data Augmentation for Disambiguating Author Names*

**Luciano V.B. Espiridião[1,2], Laura L. Dias[2], Anderson A. Ferreira[2]**

[1]Instituto Federal de Minas Gerais (IFMG) – Belo Horizonte – MG – Brazil

[2]Departamento de Computação (DECOM)
Universidade Federal de Ouro Preto (UFOP) – Ouro Preto – MG – Brazil

{luciano.espiridiao,laura.lima1}@aluno.ufop.edu.br anderson.ferreira@ufop.edu.br

***Abstract.*** *Author name ambiguity is one of the most challenging issues that can compromise the information quality in a scholarly digital library. For years, researchers have been searched for solutions to solve such a problem. Despite the many methods already proposed, the question remains open. In this study, we address the issue of producing a more accurate disambiguation function by means of applying data augmentation in the set of data training. We also propose a SyGAR-based data augmentation approach and evaluate our proposal on three collections commonly used in works about author name disambiguation task. The experimental results showed scenarios where improvements are possible in the author name disambiguation task. The proposal of data augmentation outperforms other data augmentation approach, as well as improves some machine learning techniques that were not specifically designed for the author name disambiguation task.*

## 1. Introduction

Author name ambiguity in scientific publication records is a challenging problem that can decrease the information quality in Digital Libraries (DLs). This problem has been receiving the attention of many researchers for years and many solutions have been proposed to solve or reduce it [Ferreira et al. 2012b, Ferreira et al. 2020, Hussain and Asghar 2017, Sanyal et al. 2019]. However, no proposal solved the problem, leaving challenges open.

Author Name Disambiguation (AND) task aims to solve the author name ambiguity problem and can be seen as a specific Entity Resolution (ER) task. AND tries to distinguish authors with similar names or to identify publications belonging to the same author but registered under different names. Researchers can have similar (even the same) names or write the names in different ways, either abbreviating or omitting some term. In addition to these situations, spelling errors can still be observed, as [Oliveira 2005].

According to the type of approach, the proposed methods can be split into two groups: author grouping and author assignment methods [Ferreira et al. 2012b]. Both approaches can act in a supervised or unsupervised manner. Supervised methods need a set of labeled examples, i.e., with the identification of the author or the similarity between two records, to train a classification/regression function. Those unsupervised, on the other

*This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CNPq, FAPEMIG and supported by the UFOP.

hand, do not need labeled examples for training. They generally use techniques for assessing the similarity between instances in the disambiguation process, trying to group the most similar ones as if they belong to the same author, or, in an iterative way, try to infer a disambiguation function.

In general, supervised methods show better results compared with unsupervised methods [Ferreira et al. 2012b]. However, they require labeling a large set of examples, which is costly. In addition, there are problems related to the imbalance in the bibliographic collections due mainly to the lack of examples for several authors. To address this situation, we may use some data augmentation technique [Wei and Zou 2019] that is capable of generating new synthetic examples and add them to the training set.

We observed that most studies on AND did not focus on particular aspects of the collections, such as the fact that these collections are extremely unbalanced - many authors publish little and few authors publish a lot. For the successful application of supervised techniques, it is desirable to have a good number of labeled examples for each class. Since a few studies exploiting these challenges in depth, it seemed promising to investigate it. Recently, [Kim and Kim 2018] published a study that assesses the impact of unbalance between the number of positive and negative examples in training AND algorithms. The study evaluated how the ratios of negative to positive training data can affect the performance of machine learning algorithms in disambiguation author names.

Thus, the main contributions of this paper are: (1) a data augmentation approach for author assignment supervised methods capable of generating new synthetic data examples compatible with the author publication profile; and (2) an experimental evaluation of applying data augmentation in AND task.

The rest of this article is organized as follows. In Section 2, we describe the related works that motivated the proposal. In Section 3, we detail our proposal for data augmentation in AND task. Section 4 describes our experimental evaluation and discusses the results. Finally, in Section 5 we describe our conclusions and discusses possible directions for future work.

## 2. Related Works

After analyzing several surveys on disambiguation methods [Ferreira et al. 2012b, Ferreira et al. 2020, Hussain and Asghar 2017, Sanyal et al. 2019], we noticed the lack of work applying data augmentation on AND task and focus our search on studies that show the contribution of data augmentation in the AND task and on works about data augmentation in similar tasks. We briefly list some of such works most similar to ours.

In [Zhang et al. 2016], the authors propose to apply data augmentation on the text classification problem using a thesaurus as Thesaurus[1] and Wordnet[2] from Convolutional Neural Network (CNN) performed at character level. As the previous work, Wang and Yang [2015] use word embeddings[3] with the k-NN (k nearest neighbors) clustering algorithm and cosine similarity function to find similar words, in the embedding space, to a target word. In [Kobayashi 2018], the author proposes the Contextual Augmentation

[1]https://www.thesaurus.com, https://pt.wikipedia.org/wiki/Tesauro

[2]https://wordnet.princeton.edu/

[3]https://en.wikipedia.org/wiki/Word_embedding

(CA) method, which uses a bidirectional Recurrent Neural Network (RNN) (BiLSTM) for data augmentation, considering the context of the target words. In [Cheng et al. 2013], an active selection technique called Active Data Augmentation is proposed to add synthetic data to the training set.

In [Wei and Zou 2019], the authors propose a method called Easy Data Augmentation (EDA), which uses four simple operations for data augmentation in collections applied to the text classification problem. These operations consist of: synonym replacement (SR) - $n$ random words in each sentence are replaced by their synonyms; random insertion (RI) - synonyms of target words (random) in a sentence are inserted in the same sentence; this is done $n$ times; random swap (RS) - two words in sentences are swapped randomly; this is repeated $n$ times; and random exclusion (RD) - random removal for each word in the sentence with a probability of $p$.

In our work, unlike those listed above, we intend to augment data by using a generative model, capable of capturing the authors' publishing profiles and use them to produce new synthetic citation records.

## 3. SADA – SyGAR-based Author Data Augmentation

In this section, we describe our proposal for data augmentation applied to the AND task. To enable the generation of synthetic but realistic citation records[4] for a given set of authors and their publications, i.e., to generate new records that are not the repetition or shuffling of attribute terms (as seen in some related studies), we need to infer the authors' publication profiles from a given collection. Thus, it is possible to create new synthetic examples following these profiles. Thus, we hope that these new synthetic records resemble, in terms of the distribution of terms of their attributes, with the original records.

SADA is based on SyGAR [Ferreira et al. 2012a], a tool designed for generating collections of synthetic citation records, aiming to aid in the evaluation of AND methods. It aims at filling the scarcity of labeled examples and avoiding the high cost of manual labeling. For the generation of synthetic data, SyGAR infers profiles[5] of each author's publication from a collection used to "learn" the profiles of each researcher. Therefore, it generates new synthetic data that follow the inferred profiles, in addition to simulating publication records by new (unknown) authors. Originally, SyGAR was proposed to generate an entire synthetic collection and was not used in the context of data augmentation. In this study, we propose another approach to SyGAR, using it with some adaptations, for generating new synthetic records in the training set of the original collections. As result, we have a new collection that merges original examples and new synthetic examples.

It is possible to highlight the operation of SADA in two steps: In the **fisrt step** (Figure 1 (a)), SADA obtains the authors' publication profiles from the input collection exactly as it is done in [Ferreira et al. 2012a], looking for: the distribution of topic popularity [Blei et al. 2003] based on the terms from the work titles and from the publication venue titles; collaboration between authors, using the probability distribution of the names of the co-authors of the citation records; the distribution of frequency of use of the names (and surnames) of the main author in publications; and the terms of the venues.

[4]Citation record is defined by [Oliveira 2005] as a structure that contains the metadata related to a citation.

[5]Obtained from the distribution of topic popularity [Blei et al. 2003].

In the **second step** (Figure 1 (b)), SADA generates new examples for the training set, using the profiles inferred in the previous step. For instance, to generate a title for a new citation record: first, the number of terms in this title is randomly defined from the distribution of the number of terms in each title; then, the terms are randomly generated from the frequency distribution of all the terms used by the author (class). In addition to the knowledge acquired in the profile discovery process, it is possible to simulate some interesting situations in the creation of new synthetic records. For instance, the probability that an author will publish with a new co-author or the probability that an author will change their topic of interest. In this step, parameters are used to control the total number of new examples and generate new publication profiles.

SADA also has some parameters that are used for both the first and the second step. These parameters are described below:

- **AUG** - defines the number of records that will be augmented for each record in the original collection;
- **Strategy** - defines which strategy will be used for data augmentation: **MCC** (min class count) - for this strategy, the minimum number of examples that a class (author) should have must be informed, in order to generate **AUG** examples for each original record; **ALL** - if this strategy is used, **AUG** examples will be generated for each record in the collection.

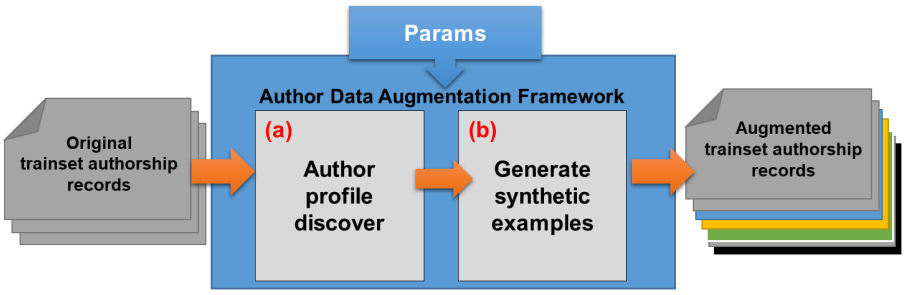


**Figure 1. SyGAR-based Author Data Augmentation (SADA).**

To clear on how the data augmentation approach works, consider the example in Table 1. In this table, the first two lines consist of original examples and the other ones have records augmented by SADA with the following configuration: $MCC\_5\_AUG\_5$. Noticed that the generated examples "captured" an author's publication profile and use this information to generate new training examples based on the setup.

## 4. Experimental evaluation

In this section, we describe the collections, evaluation metrics and experimental setup, and also discuss the experimental results to evaluate our proposal.

### 4.1. Collections

For the experimental evaluation, we use some known collections described in the literature [Müller et al. 2017, Hussain and Asghar 2017]. Our collections were obtained from

**Table 1. Examples of data augmentation applied to the Kisti collection before and after data augmentation for the same class (author).**

| Aug. method | Example | | | |
|---|---|---|---|---|
| | Author name | Coauthor names | Title | Venue |
| ORIG | amar gupta | peter van der putten | why the information explosion can be bad for data mining, and how data fusion provides a way out | sdm |
| ORIG | amar gupta | satwik seshasai;ashwani kumar | an integrated and collaborative framework for business design: a knowledge engineering approach | dke |
| SADA | amar gupta | peter van der putten | information provides for out the a fusion business how an provides approach how framework be mining way design | dke |
| SADA | amar gupta | ashwani kumar | collaborative framework the information data data framework can a the fusion the approach the engineering collaborative a the | sdm |
| SADA | a gupta | satwik seshasai;peter van der putten | design how for data for a mining mining data data framework collaborative | sdm |

the DBLP and BDBComp digital libraries, named KISTI, DBLP and BDBComp. Sometimes, such collections are split into disjoint Ambiguous Groups[6]. Ambiguous groups improve the disambiguation method efficiency by reducing unnecessary comparisons.

**BDBComp**. This collection has records taken from the 10 largest ambiguous groups in the BDBComp digital library and has 361 citation records and 205 distinct authors [Cota et al. 2007]. In this collection, the number of publications per author is around 1.76, with little variation between ambiguous groups. 50% of the records consist of authors with only one publication and 75% with up to two publications. The distribution of publications by authors demonstrates that this collection has, on average, few examples for most authors, which makes it, although small, a challenging scenario for both disambiguation methods and the evaluation of data augmentation techniques.

**DBLP**. This collection was proposed in [Han et al. 2004], and processed and corrected in [Cota et al. 2007], with 220 distinct authors and $4,270$ citation records, distributed in 11 ambiguous groups. This collection has more examples and more distinct authors, and the ratio between records and authors is also higher (19.41 on average), with a small variation between ambiguous groups. This collection is more balanced compared with the other ones. This is an interesting scenario to evaluate data augmentation methods as, supposedly, there are enough publications for all authors. Therefore, it is important to investigate how data augmentation reflect in the disambiguation of this collection.

**KISTI**. Organized by [Kang et al. 2011], the largest collection used by us, has $41,671$ citation records split into 882 ambiguous groups and 6921 distinct authors. Its average number of publications per author is 6.02, but there is a wide variation in the size and distribution of authors on the ambiguous groups. This collection has lots of records, but it looks more like BDBComp than DBLP. Just like BDBComp, it is also a very unbalanced collection. However, contenting a few authors with many publications.

[6]Groups of authors with similar names. For instance, groups formed by all records whose authors' names begin with the same initial letter of the first name jointly with the last surname

### 4.2. Metrics

As our evaluation metrics, we use the $k$, $pairwise$-F1 and $b^3$-F1 metrics. For this study, we use the definition of the metrics defined in [Kim 2019][7].

**K**. The $K$ metric [Lapidot 2002] is defined in terms of Average Cluster Purity ($ACP$), and Average Author Purity ($AAP$). The $K$ metric is the geometric mean between $ACP$ and $AAP$. It is defined by the equations:

$$ACP = \frac{1}{N}\sum_{i=1}^{|P|}\sum_{j=1}^{|T|}\frac{n_{ij}^2}{n_i}; \quad AAP = \frac{1}{N}\sum_{j=1}^{|T|}\sum_{i=1}^{|P|}\frac{n_{ij}^2}{n_j}; \quad k = \sqrt{ACP \times AAP}$$

where $N$ is the total number of citation records in the ambiguous group; $T$ is the number of clusters, manually built on the ambiguous group; $P$ is the number of automatic clusters generated by an AND method for this ambiguous group; $n_{ij}$ is the total number of records in the automatic cluster $i$ that belong to the same manually built cluster $j$; $n_i$ is the total number of records in the automatically generated cluster $i$; and $n_j$ is the total number of records manually generated in the cluster $j$.

**Pairwise-F1**. This metric is the $f$ measure applied on the pairwise precision and recall metrics.

$$pP = \frac{|pairs(P) \cap pairs(T)|}{|pairs(P)|}; \quad pR = \frac{|pairs(P) \cap pairs(T)|}{|pairs(T)|}; \quad pF1 = 2 \times \frac{pP \times pR}{pP + pR}$$

where $pP$ is the precision based on pairs; $pR$ is the pairwise recall; $pF1$ is the harmonic mean between $pP$ and $pR$; $pairs(T)$ are pairs of records manually groups as belonging to the same author; and $pairs(P)$ are pairs of records automatically grouped by a disambiguation method as belonging to the same author.

**b$^3$F1**. This metric works as follows: Precision ($b^3P$) evaluates if, given an instance of a citation record belonging to an author, it is predicted (automatically) as a citation record by the same author (manually defined), whereas recall ($b^3R$)) measures the extent to which all instances of citation records from the same authors are actually predicted. The value $b^3F1$ is the harmonic mean between those two values.

$$b^3P = \frac{1}{N}\sum_{t \in T}\frac{|P(t) \cap T(t)|}{|P(t)|}; \quad b^3R = \frac{1}{N}\sum_{t \in T}\frac{|P(t) \cap T(t)|}{|T(t)|}; \quad b^3F1 = 2 \times \frac{b^3P \times b^3R}{b^3P + b^3R}$$

where $t$ is an instance in $T$; $T(t)$ is a manually generated cluster that contains the instance t; $P(t)$ is the automatically generated cluster that contains the $t$ instance; and $N$ is the total number of instances in $T$.

### 4.3. Experimental setup

To evaluate the impact of data augmentation in the disambiguation task, we perform experiments on the original training sets (ORIG) and on training sets augmented by adding records provided by EDA and SADA. The EDA was also used as a baseline since, despite being a relatively simple technique, it provided good results in the sentence classification task [Wei and Zou 2019].

EDA uses the $\alpha$ value to determine the percentage of changes, on each operation (SR, RI, RS and RD), performed in a given sentence; the $p$ parameter determines the probability of a removal occurring in the RD operation; and the $n_{aug}$ parameter determines

[7] https://github.com/lucianovilasboas/nd_metrics

the number of new sentences to generate. We keep the values of $\alpha$ and $p$ equal to the best values experimented in [Wei and Zou 2019] and with $p = \alpha$ (in RD). For the $n_{aug}$ parameter, the values range from 1 to 10. EDA generates the work and publication venue titles for newly synthetic citation records. The author and coauthor names were repeated from the original records.

To evaluate the disambiguation results on both scenarios (with and without data augmentation), we use the disambiguation method proposed by [Santana et al. 2017] (supervised version), as well as three other Machine Learning techniques (Naive Bayes, Logistic Regression and Random Forest) also used in AND problems [Hussain and Asghar 2017].

The method proposed in [Santana et al. 2017] (called INC) operates as an incremental disambiguation method based on the most similar cluster. It can be used as both supervised or unsupervised method. To disambiguate, this method weights each attribute[8] from a citation record in order to give more importance to the attributes with greater discriminative power. Since a new input is disambiguated, it can be incorporated into the training set. However, it can be reclassified if new inputs provide stronger evidence of its authorship. Weights, in the supervised version, are learned by performing the cross-validation process using the training set records.

For the use of the other methods (based on the Naive Bayes, Logistic Regression and Random Forest techniques), the citation records were organized in a single record where each term from the author or coauthor names, and terms from the titles and the venues were combined in an $N$-dimensional vector (being $N$ is the vocabulary size of the unique terms from the set of records in each collection). We perform a dimentional reduction using the Singular Value Decomposition (SVD)[9]. For the author and coauthor names, we perform Bag of Words (BoW) and, for the other attributes, we perform Term Frequency - inverse Document Frequency (TF-iDF) to generate original vectors.

In order to evaluate the impact of data augmentation on the effectiveness of AND methods, first, we perform experiments on the original training data (ORIG), i.e., without applying data augmentation, and, next, we perform disambiguation methods learned with training set increased by applying the EDA and SADA data augmentation techniques. In both approaches, EDA and SADA, we used the same parameter values. In this section, we discuss and analyze the results using the parameter values that provide the best results. To evaluate the results, we use the ten-fold cross validation strategy and we compare the performance of the methods using Student's t-distribution with 95% confidence.

For each parameter and its respective value, we use the same nomenclature for both data techniques: <**MCC_C**> (min class count) considers only classes/authors with at least $C$ examples per author; <**ALL**> generates new records for each record from the original training set; <**AUG_N**> generates $N$ records for specified author.

### 4.4. Results and discussion

In this section, we describe, analyze and discuss the experimental results.

**Comparing SADA with EDA.** Figures 2, 3 and 4 show the results in the BDBComp,

[8] Author and coauthor name terms and title and publication venue terms.

[9] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

**Table 2. The $K$, $B^3$, and $Pairwise - F1$ results on the BDBComp, DBLP and Kisti collections (ORIG $vs.$ SADA). Best results (and statistic ties) are highlighted in bold.**

| | | ORIG | | | SADA | | |
|---|---|---|---|---|---|---|---|
| | | k | pF1 | $b^3F1$ | k | pF1 | $b^3F1$ |
| BDBComp | NB | 0.497 (0.045) | 0.052 (0.035) | 0.409 (0.051) | **0.582** (0.039) | **0.095** (0.036) | **0.517** (0.051) |
| | LR | 0.863 (0.054) | 0.421 (0.140) | 0.857 (0.057) | **0.931** (0.033) | **0.607** (0.133) | **0.930** (0.032) |
| | RF | 0.849 (0.051) | 0.389 (0.145) | 0.843 (0.052) | **0.918** (0.054) | **0.556** (0.174) | **0.917** (0.054) |
| | INC | 0.941 (0.028) | **0.742** (0.167) | 0.939 (0.028) | **0.956** (0.035) | **0.820** (0.099) | **0.956** (0.035) |
| DBLP | NB | **0.349** (0.061) | **0.238** (0.101) | **0.346** (0.060) | **0.374** (0.078) | **0.266** (0.115) | **0.372** (0.077) |
| | LR | **0.724** (0.100) | **0.690** (0.167) | **0.721** (0.099) | **0.725** (0.101) | **0.687** (0.173) | **0.721** (0.102) |
| | RF | **0.568** (0.102) | **0.509** (0.164) | **0.565** (0.100) | **0.609** (0.098) | **0.552** (0.171) | **0.603** (0.096) |
| | INC | **0.876** (0.039) | **0.863** (0.074) | **0.875** (0.039) | **0.867** (0.039) | **0.852** (0.082) | **0.865** (0.039) |
| Kisti | NB | 0.735 (0.034) | 0.572 (0.066) | 0.718 (0.038) | **0.817** (0.030) | **0.707** (0.051) | **0.812** (0.032) |
| | LR | 0.904 (0.023) | 0.872 (0.042) | 0.901 (0.024) | **0.945** (0.016) | **0.946** (0.026) | **0.944** (0.016) |
| | RF | 0.910 (0.026) | 0.909 (0.039) | 0.909 (0.027) | **0.938** (0.020) | **0.947** (0.031) | **0.938** (0.020) |
| | INC | **0.950** (0.034) | **0.934** (0.068) | **0.950** (0.034) | **0.949** (0.035) | **0.932** (0.075) | **0.949** (0.036) |

DBLP and KISTI collections under our evaluation metrics. Under $ACP$, $pP$ and $b^3P$, both EDA and SADA improves the average results. But, under $AAP$, $pR$ and $b^3R$, the improvements do not always happen.

On the results applying EDA, as it practically duplicates existing examples changing some terms, a few new information is added by new synthetic examples which leads to several ties considering the baseline (ORIG) or even a worsening in some situations. But, in the case of SADA, this situation does not occur since when considering the authors' profiles in the data increase process, our approach manages to add more information.

**Results in BDBComp.** In BDBComp, SADA improves the methods' results in almost all scenarios (see Table 2 and Figure 2) under the $K$ and $b^3$ metrics. On the other hand, applying EDA, we have several statistical ties under all metrics. SADA provides significant gains under the three metrics. For instance, under the $pF1$, the improvements are around $11\%$, $43\%$, $44\%$ and $83\%$, for INC, LR, RF and NB, respectivelly, compared with the use of the original training sets. The BDBComp collection has several authors publishing few works. Thus, such authors are underrepresented in the original training set. Applying data augmentation, we add new examples in the training set, improving the author representativeness, providing more evidence for learning the disambiguation function.

**Results in DBLP.** In the DBLP collection, data augmentation did not improve the results. But, on average, there exist improvements under the precision metrics ($ACP$, $b^3P$ and $pP$), according to the plots presented in Figure 3. The DBLP collection has several examples for each author. Thus, the original training sets already produce good learning functions, in contrast to the examples provided by the BDBComp collection.

It was noticed that in training sets with more examples for each class, that adding more examples did not yield better results, since models tend to generalize properly when there are enough examples. As this collection has many authors and many publications per author, the strategies for generating new examples need to consider a higher initial value $C$ in $MCC\_C$, as most authors in this collection already have at least five or more publications. Therefore, strategies such as $MCC\_[1...4]\_AUG\_N$ had no effect on the disambiguation results in relation to the original collections. However, using $MCC\_5\_AUG\_5$ provided getting closer to the averages with higher values of $ACP$, $pP$ and $b^3P$.

**Results in KISTI.** KISTI is the largest collection evaluated, but publish distribution as

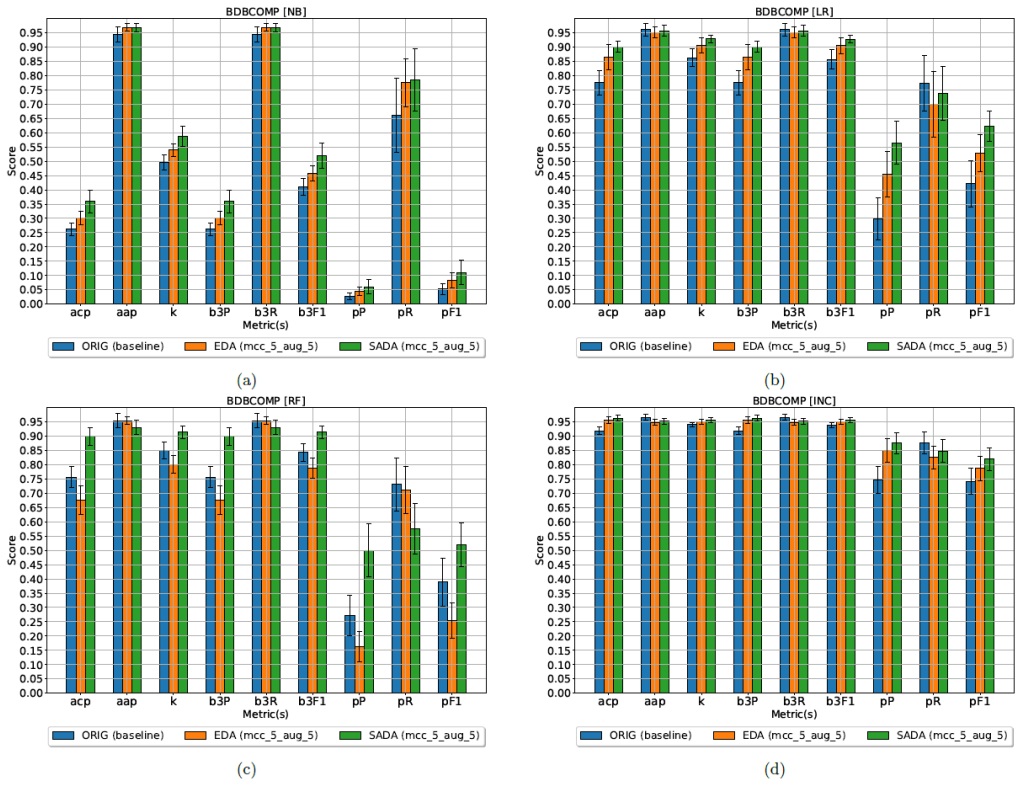


**Figure 2. Bar plots jointly the confidence interval (95% confidence) for NB, LR, RF and INC for AND Methods applyed on the BDBComp collection.**

BDBComp, i.e., KISTI has several authors publish a few works. We notice that, in ambiguous groups with many author publishing few works, performing data augmentation provides better results. Otherwise, data augmentation does not improves the results.

Usually, data augmentation in KISTI collection improves results under the precision and recall metrics (see Figures 4 (a), (b) and (c)) for all evaluated methods but INC that decreases the recall metrics (see Figure 4 (d)).

**Analyzing the purity of the disambiguated clusters.** We notice that, augmenting the training set allows the method to produce clusters more purity, i.e., increase the $ACP$, $pP$ and $b^3P$ values, mainly in the BDBComp and KISTI collections (see Table 3). We also notice that increases on the $K$, $pF1$ and $b^3P1$ metrics are directly related with those increases. Ideally, we would like both the precision and recall values close to $1.0$. But, if we need to chose one, we prefer to have high precision values (or purity), indicating publications of the same author are not grouped with publications of other authors. Usually, separating publications grouped together in the same cluster is more complicated than jointing publications of the same author that are separated by the disambiguation process.

## 5. Conclusions and future works

In this study, we propose SADA, a data augmentation technique for the supervised author name disambiguation task. SADA generates new synthetic records based on the data from the training set of the original collections. The generation of new records follows the authors' publication profiles. Usually, our experimental evaluation showed improvements on performing author name disambiguation methods when we augment the training set.

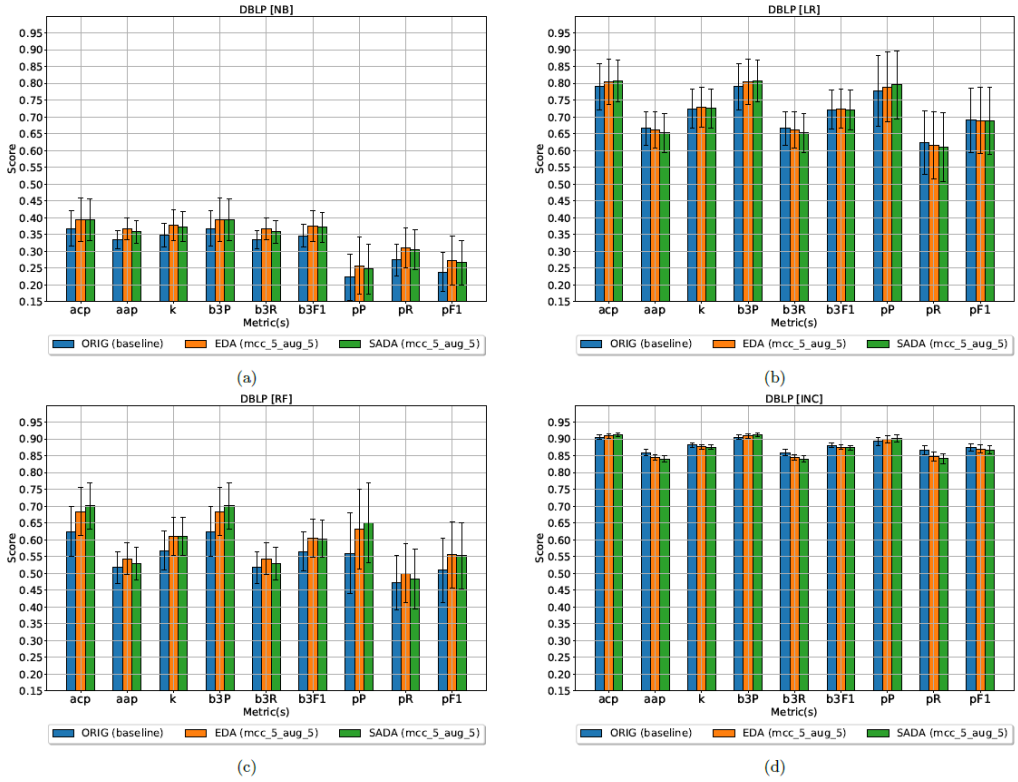


**Figure 3. Bar plots along with confidence interval (95% confidence) performing the NB, LR, RF and INC methods in the DBLP collection.**

**Table 3. The NB, LR, RF and INC results in BDBComp, DBLP and KISTI under the ACP, $b^3P$, and $pP$ metrics. Best results are highlighted in bold.**

| | | ORIG | | | SADA | | |
|---|---|---|---|---|---|---|---|
| | | ACP | pP | $b^3P$ | ACP | pP | $b^3P$ |
| BDBComp | NB | 0.262 (0.039) | 0.028 (0.019) | 0.262 (0.039) | **0.359** (0.070) | **0.061** (0.043) | **0.359** (0.070) |
| | LR | 0.775 (0.073) | 0.298 (0.127) | 0.775 (0.073) | **0.902** (0.035) | **0.565** (0.131) | **0.902** (0.035) |
| | RF | 0.757 (0.064) | 0.272 (0.121) | 0.757 (0.064) | **0.900** (0.052) | **0.500** (0.160) | **0.900** (0.052) |
| | INC | 0.918 (0.029) | 0.746 (0.140) | 0.918 (0.029) | **0.963** (0.042) | **0.876** (0.129) | **0.963** (0.042) |
| DBLP | NB | **0.368** (0.092) | **0.223** (0.119) | **0.368** (0.092) | **0.393** (0.106) | **0.247** (0.129) | **0.393** (0.106) |
| | LR | **0.790** (0.121) | **0.778** (0.181) | **0.790** (0.121) | **0.807** (0.108) | **0.796** (0.176) | **0.807** (0.108) |
| | RF | **0.625** (0.129) | **0.560** (0.206) | **0.625** (0.129) | **0.701** (0.118) | **0.651** (0.204) | **0.701** (0.118) |
| | INC | **0.902** (0.035) | **0.886** (0.057) | **0.902** (0.035) | **0.906** (0.031) | **0.892** (0.062) | **0.906** (0.031) |
| Kisti | NB | 0.595 (0.047) | 0.416 (0.067) | 0.595 (0.047) | **0.729** (0.042) | **0.570** (0.065) | **0.729** (0.042) |
| | LR | 0.836 (0.036) | 0.783 (0.064) | 0.836 (0.036) | **0.918** (0.022) | **0.914** (0.043) | **0.918** (0.022) |
| | RF | 0.871 (0.032) | 0.855 (0.051) | 0.871 (0.032) | **0.923** (0.024) | **0.923** (0.045) | **0.923** (0.024) |
| | INC | 0.956 (0.030) | **0.945** (0.094) | **0.956** (0.065) | **0.961** (0.041) | **0.951** (0.089) | **0.961** (0.060) |

A systematic evaluation on some collections using various configurations allowed us highlight some characteristics that could lead to the use of data augmentation in this domain. If the ratio between the number of examples and authors is small, usually data augmentation improves the results. Overall, data augmentation improves the purity of the disambiguated clusters produced as results by an AND method. In some case, it also reduces the fragmentation, i.e., publications of a same author split into several clusters.

Another important observation is that even machine learning methods that were not proposed to deal exclusively with the name ambiguity problem can benefit from the data augmentation technique developed in this study. The results suggest that, after data augmentation in the original collections, these methods can become more competitive.

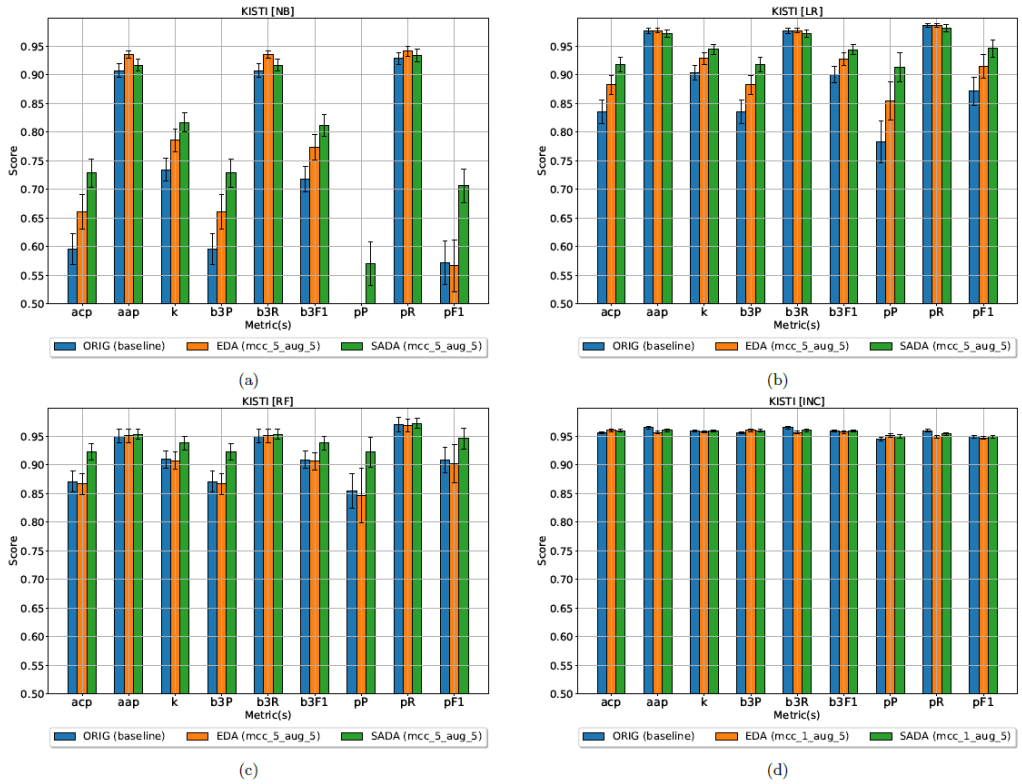


**Figure 4. Bar plots along with confidence interval (95% confidence) performing the NB, LR, RF and INC methods.**

In this work, we focus on evaluating data augmentation in the author supervised assignment methods. Thus, we intend to evaluate its impact on other author names disambiguation approaches, as well as other disambiguation methods and scenarios.

## References


Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cheng, Y., Chen, Z., Wang, J., Agrawal, A., and Choudhary, A. (2013). Bootstrapping Active Name Disambiguation with Crowdsourcing. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1213–1216, New York, NY, USA. Association for Computing Machinery.

Cota, R. G., Gonçalves, M. A., and Laender, A. H. F. (2007). A Heuristic-based Hierarchical Clustering Method for Author Name Disambiguation in Digital Libraries. In *Proceedings of the XXII Brazilian Symposium on Databases*, pages 20–34, João Pessoa, Paraiba, Brazil.

Ferreira, A. A., Gonçalves, M. A., Almeida, J. M., Laender, A. H., and Veloso, A. (2012a). A tool for generating synthetic authorship records for evaluating author name disambiguation methods. *Information Sciences*, 206:42–62.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2012b). A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Record*, 41(2):15–26.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2020). Automatic disambiguation of author names in bibliographic repositories. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 12(1):1–146.

Han, H., Giles, C. L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th JCDL*, pages 296–305, Tucson, USA.

Hussain, I. and Asghar, S. (2017). A survey of author name disambiguation techniques: 2010-2016. *The Knowledge Engineering Review*, 32:1–24.

Kang, I.-S., Kim, P., Lee, S., Jung, H., and You, B.-J. (2011). Construction of a large-scale test set for author disambiguation. *IP&M*, 47:452–465.

Kim, J. (2019). A fast and integrative algorithm for clustering performance evaluation in author name disambiguation. *Scientometrics*, 120(2):661–681.

Kim, J. and Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics*, 117:511–526.

Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. *CoRR*, 2:452–457.

Lapidot, I. (2002). Self-Organizing-Maps with BIC for Speaker Clustering. Technical report, IDIAP Research Institute, Martigny, Switzerland.

Müller, M. C., Reitz, F., and Roy, N. (2017). Data sets for author name disambiguation: an empirical analysis and a new resource. *Scientometrics*, 111:1467–1500.

Oliveira, J. W. A. (2005). Uma estratégia para remoção de ambiguidades na identificação de autoria de objetos bibliográficos. Master's thesis, Uiversidade Federal de Minas Gerais. Departamento de Ciência da Computação, Belo Horizonte, Brazil.

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., and Ferreira, A. A. (2017). Incremental author name disambiguation by exploiting domain-specific heuristics. *Journal of the Association for Information Science and Technology*, 68(4):931–945.

Sanyal, D. K., Bhowmick, P. K., and Das, P. P. (2019). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, 47(2):227–254.

Wang, W. Y. and Yang, D. (2015). That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. In *Proceedings of the EMNLP*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Wei, J. and Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the EMNLP-IJCNLP*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Zhang, X., Zhao, J., and LeCun, Y. (2016). Character-level Convolutional Networks for Text Classification. In *Proceedings of the NIPS*, pages 649–657, Cambridge, MA.