JCBI
Journal of Computing & Biomedical Informatics SICAIET
ISSN: 2710 - 1606 2024
Research Article
Collection: Artificial Intelligence and Emerging Technologies

Author Identification Using Machine Learning

Sikandar Ahmad Khan[1], Muhammad Asad[1], Haroon Asif[1], Amjad Ali[2*], and Muhammad Ahsan Jamil[3]

[1]Department of Computer Science, National College of Business Administration & Economics Multan Campus, Multan, 60000, Pakistan.
[2]Department of Information Technology, Bahauddin Zakariya University, Multan, 60000, Pakistan.
[3]Institute of Computing, Muhammad Nawaz Sharif University of Agriculture, Multan, 60000, Pakistan.
[*]Corresponding Author: Amjad Ali. Email: Amjadsaeedi@bzu.edu.pk

Academic Editor: Salman Qadri **Published:** February 01, 2024

Abstract: Identifying the writers of a piece of writing, whether anonymous or not, is a procedure that focuses solely on the writing style and not on the content itself. Most of the time, writing and speaking style may also be seen as techniques of underlying sentence construction, which can be evaluated using aspects such as vocabulary, length of sentences, and sequence of words, richness, and word frequency usage. The primary goal of this article is to examine and apply a variety of categorization approaches to research articles that analyze author identity and the content of those texts that are in dispute. Researchers' earlier work is also discussed and elaborated upon. After that, we were able to exhibit better findings from our experiments. With feature spaces, the SVM technique is particularly well-suited to the aforementioned problem. Across all experiments, it was found that the SVM was effective at determining the authorship of research publications. SVM was used to classify two sets of data in that study. Sections A and B of the experiment are referred to as Experiment A and B, respectively. 500 research papers from conferences and the majority of them from Google Scholar are included in Experiment A. After applying data mining techniques to the gathered dataset, we have a final set of 400 research papers. A, B, and C are the three subsets of the 400 research papers for high performance that were further subdivided. Our model was able to train quickly and evaluate good performance on these limited research criteria in these datasets as a result of an increase in both the number of authors and the number of publications that were included in the dataset.

Keywords: Author identification; SVM; NLP; Machine Learning; Feature Extraction.

1. Introduction

Authorship recognition has received a lot of interest lately from academics [1]. It has been done for centuries to determine who wrote papers with scribbled writing [2]. The enormous amount of written material has become accessible digitally and is kept in a number of unorganized forms [3, 4]. Author recognition is a crucial part of text mining. It can be difficult to derive useful data from unorganized or partially organized sources [5]. Text mining is frequently employed to examine a sizable volume of unorganized information and derive insightful conclusions [6]. Textual analysis is the process of taking unorganized or partially organized data and extracting useful data from it. This method creates analytical text frameworks that gather or categorize particular details according to the training information by using machine learning and methods for natural language processing [7, 8]. The author's recognition technique is shown in Figure 1.

The process of determining the likely writer of unidentified papers from a pool of prospective writers is referred to as author recognition. A written work categorization challenge typically takes into account author recognition. Pre-processing an inventory is the first step. Next, characteristics are extracted, and selected, and the written information is transformed into a feature matrix. Machine learning uses feature development as a crucial stage for predicting the model's performance.

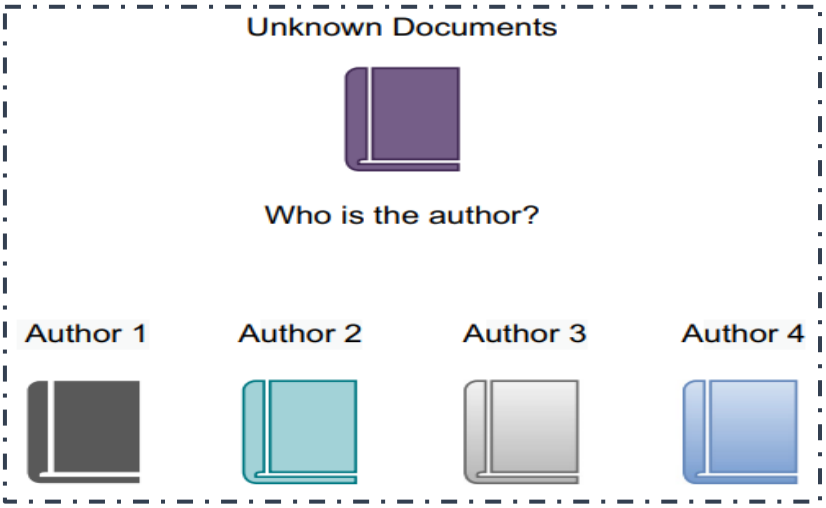


Figure 1. Authors Identification Process

Technologies for author recognition are currently being established in many fields, including evaluation detection methods [10] and hacker law [9]. Encryption recognition, sign identification, and identifying breaches all include AI. The trickiest and most significant part of author identification is identifying the key characteristics that best describe the author's written technique. Precise author identification could be possible if the primary characteristics can be extracted. Many investigators have studied this area and put forth a number of solutions [11, 12]. For author identification, the most crucial traits, including lexicon [13], morphological [14], content-specific [15], and stylometric traits [16], are utilized. Additionally, there are numerous methods for extracting word anchoring features that are employed with NLP text categorization and data mining tasks [17–19]. With this method, the pertinent traits are extracted from the written information. Additionally, it offers an acronym matrices databases, which is mainly utilized to improve the efficiency of ML systems' classification functions [20].

A comprehensive number of author authentication techniques have been suggested over the past ten years [21, 22, 23, 24]. Additionally, a pertinent shared job was incorporated in a number of earlier PAN versions [25, 26, 27, 28, 29]. Multiple factors affect how well author authentication techniques work. Cross-topic including cross-genre author authentication were both taken into consideration in PAN 2015, and the findings showed only moderately precise outcomes, particularly with Dutch articles and evaluations [26]. Writing messages, or non-professional literature released web by enthusiast researchers, adhering to various communities (or fanfiction influenced by specific highly regarded works), was utilized in the most recent two versions of PAN [28, 29]. Regarding this job, a substantial training set with over 350,000 validation cases was assembled, allowing the use of potent deep-learning techniques [30].

2. Literature Review

The study of author identification is a fairly recent subject of study. The initial logical solution to the issue came up in the research of Mendenhall, which investigated the attribution of works claimed to Bacon, Marlowe, and Shakespeare in the latter part of the nineteenth century, in 1887. Because of novel possibilities for criminal justice, literary study, current culture, and business, the issue of author identification has been increasing in prominence (Kestemont et al., 2019) [31]. Two groups of procedures are primarily used in the recognition technique: the initial group comprises techniques derived from the analysis of statistics, which include the study of principal components (Jamak et al. 2012) [32].

Employ characteristics that are probably unrelated to the text's subject in order to accomplish excellent author identification efficiency. A lot of design indicators have been employed for this purpose, ranging from beginnings centered around basic characteristics that include phrase length and terms depth

(Yule, 1944), to contemporary and pertinent research founded on words with functions (Boukhaled and Ganascia, 2015) [33], punctuation symbols (Martin-del-CampoRodriguez et al., 2019) [34], Part-of-Speech labels (Pokou, Fournier-Viger and Moghrabi, 2016) [35], parsing trees and character-driven characteristics (Sapkota et al., 2015) [36]. Depending on the retrieved traits, someone can divide extant Arabic-language compositions into two distinct groups (Al-Ayyoub, Alwajeeh, and Hmeidi, 2017) [37].

The lexicon technique, which determines the characteristic matrix for every sentence via the word appearances in it, is included in the initial group. The next group, which depends on the greater complexity of style indicators, computes specific attributes by attempting to identify more deeply and pertinent features of the language. Lastly, those with an interest are directed to (El Bakly et al, 2020) [38] for a greater overview of the various publications NLPinAI 2022 - Specialised Conference on NLP in Intelligent Systems 490 and concerns on the Arabic author authentication challenge.

Research on the English tongue makes up a significant portion of the scholarship [4, 5, 6, 7, 8]. Japanese [9], Mongolian [10], Persian [11], Albanian [12], Indian [13, 14], Brazilian [15], Russian [16, 17], German [18], and Arabian [19] are only a few of the cultures that have research conducted in them. Various sorts of information were employed for author authentication jobs when current research was analyzed. There exists research on articles from newspapers [4, 15, 18, 19], poetry [13], novellas [11, 12, 16], emails [20], music lyrics [21], computer code [22], tweeting, posts on blogs, and newsgroups [8, 9, 23], as well as other types of writing. Various sources of information have occasionally been analyzed or integrated [17, 25].

3. Proposed Methodology

The four components of our suggested AI approach are:

1. The first step is to preprocess the reviews.

2. In order to extract features from the preprocessed reviews, we use Feature Representation, NLP. Training and prediction are two stages that are involved in a review's classification process.

3. An NLP feature extractor is used to preprocess the training data, which consists of several reviews, before generating features from the data.

4. After that, a machine learning classifier is trained using these attributes. Predicting fresh reviews' author (label) is then made possible using this classifier (inputs).

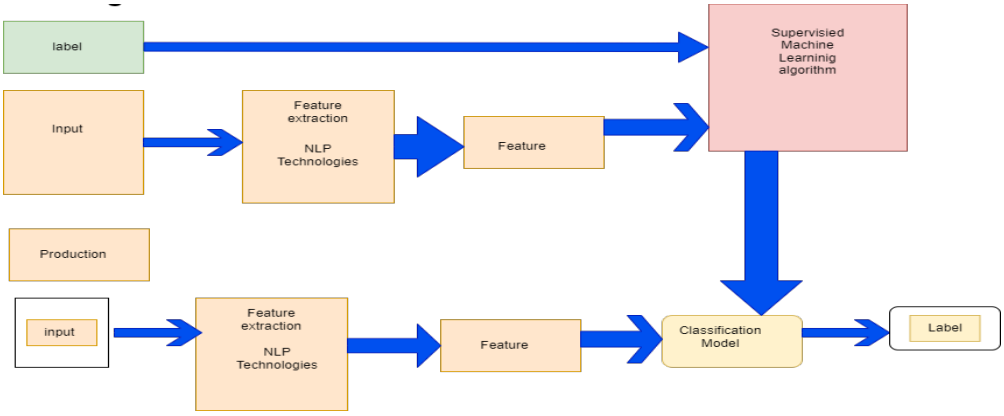


Figure 2. Overview of a\Authorship identification

3.1 Dataset Collection

A corpus is required in order to carry out authorship identification. According to the most recent explanation, a corpus can be anything from a collection of messages to source code, and according to this hypothesis, the corpus comprises examination papers. In any case, the term "research paper" covers an extremely broad scope; our corpus is comprised of exploratory articles in software engineering that are related to the topic of data recovery. There are two distinct information sets that are dependent on the corpus. The first of these includes merely research articles from data recovery algorithms such as "ECIR" and "SIGIR", "CIKM." The other dataset is made up of a some research papers written by creators working in the data recovery field. The reason for having two isolated informational collections was so that we could observe how the preparation intentions functioned when given varied amounts of information for preparation and testing. This was the rationale behind having two separate informational collections. In

the primary informative index, we only have a small number of authors who have produced a respectable number of research publications per author. In the second informational collection, we have a tremendous number of contributors, and each contributor has contributed a predetermined amount of study papers. When it came time to organize the first dataset, the one that had a predetermined number of creators, the exploratory papers were filed away in separate organizers. Each envelope was given the name of the person who had created the candidate, and because we were able to perform a search based on the organizer's name, there was no need for any sort of programmed creator name extraction. In spite of this, the research articles that were included in the later collection of information were not structured in this way. As a consequence of this, we wanted a method for this dataset that would allow us to isolate the inventor's name from the examination paper.

3.2 Preprocessing

It was necessary to extract text from an exam paper that might be used in the corpus after the initial step of building it. It was necessary to do programmed text extraction in order to remove stylometric highlights from all of the examination papers in the corpus. Two different approaches were considered for preprocessing PDFs. The first option was to use a Java package to directly extract-text by PDFs, and second option was to use OCR, also known as optical character recognition [34]. Apache's PDFBox1 and a design-conscious content extractor for the logical research h papers, well known as LAPDF [44], that were two participants evaluated for the task. An important drawback of PDFBox was the disturbance caused when extracting text from tables, pseudocode, and figures in research articles. Other than causing some noise in the output of the content extractor, its performance was excellent. When it came to the recurrence of turbulence in the yield text, the LAPDF library didn't suer similarly to PDFBox as it had before. Due to a lack of consideration for authors, article features, figures, and tables as well as pseudocode in LAPDF-library, this rationale is based. LAPDF, on the other hand, had a drawback in that it could only deal with two-segment logical articles, not single-segment articles. Next, the ABBYY2 optical character recognition was tested. When it comes to information, PDFBox has the same level of agitation as slow and sluggish. In addition, OCR was not used in the execution since it was deemed too expensive. There were two viable options that both had substantial drawbacks. A noisy dataset would result from using PDFBox, whereas single-section articles such as examination would be disregarded if LAPDF were selected instead. We were able to reduce the amount of noise in our dataset by using the ECIR-supported article selection strategy.

3.3 Author Search

Additionally, we expected to be able to extract the creator names of the exam paper from the PDF for the second dataset with a large number of creators and limited preparation and testing information. The underlying assumption was that the metadata in a PDF might be used to identify the authors of a particular exam paper. Although the vast majority of these PDFs featured auto-generated metadata generated by the compiler, the producers occasionally refreshed the metadata with the correct information. As a result, the idea that the metadata may be used to determine who wrote an exam paper was dismissed outright.

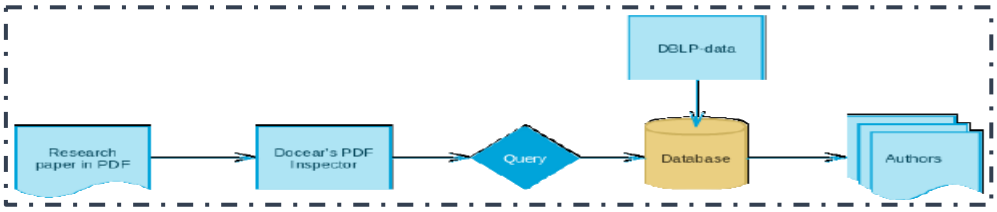


Figure 3. Process Flow

3.4 Feature Extraction

Multiple sets of features were obtained from the final extraction process, and they have been saved in (ARFF) format so that WEKA can work with them [36]. The 193 features in this file are a mix of syntactic and lexical features. Appendix A presents the overall features in the syntactic features file and lexical features type. To reduce processing time and prevent the dimensionality curse, no n-gram features are used. Because in [53] they defined a list of words and also stated that there is no predefined list of function words that each author may use, no sufficient distinct list for functional words was created in terms of syntactic properties.

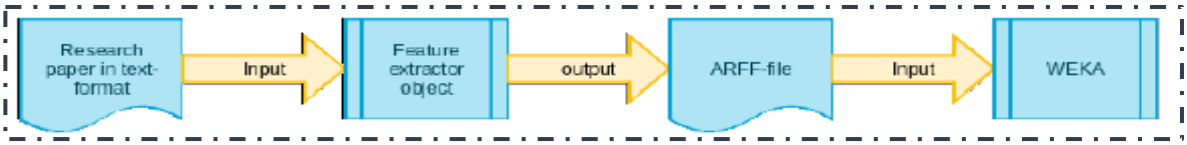


Figure 4. Feature Extraction

3.5 Processing Tool

The experimental results were obtained through the use of the data mining programme WEKA. A number of machine-learning algorithms are available in WEKA to perform the mining operation. Java initiatives can also benefit from the use of these mining-related algorithms. It provides capabilities for predictive modelling, data analysis, and visualisation [61]. For the mining process in WEKA, we used datasets in the ARFF format. In addition to methods, WEKA provides several options for dividing the data into training and testing sets. [62].

4. Results

4.1 Datasets

Data is required for every machine-learning task. For this purpose, algorithms are applied. Machine learning is used in the same way to identify the author of a piece of work. This repository had a significant amount of data that, in its raw form, would be difficult for any classifier to manage. Research articles from many sources make up the data. "Experiment A" and "Experiment B" are subdivided further in the experimental component. Table 1 provides a full breakdown of each experiment's data. In order to acquire accurate results for very large data sets, it is necessary to divide the data into two separate experiments, which is a sensible decision.

4.2 Experiment A

The experiment was a success! Dataset A, Dataset B, and Dataset C represent the three datasets that make up a segment. There are 200 publications in dataset A. Only ten authors who contributed as first authors were chosen, hence an author can only be identified as the first author in classification. The first author does the bulk of the description, style, and functional word selection, thus it stands to reason that he or she would be the best choice for this section. Dataset B contains the names of the top three authors of chosen research articles. Unless the author is one of the first three writers to contribute, the model will not be able to learn about the author's identity. Only authors with 20 or more articles are included in this dataset.

Table 1. Description

Experiment A			
Dataset	Author contribution	Research papers	Total No of Research Papers
Dataset A	First Author	20	100
Dataset B	Among 1st four authors	30	200
Dataset C	Any author among the contributors	30	200

There are a total of 400 papers included in this collection, 20 for each of the authors. When compared to dataset A, which had just 10 articles for each author, this data set has more than doubled the number of published authors. To train the model with a total of 400 papers, dataset C uses an author with 30 publications included in the search query, although datasets B and A differ significantly in that the extent of contribution. Dataset B has only three authors who made a contribution to a research publication; however, dataset C has contributions from all of them. Research papers are encouraged and considered in training datasets for all authors who contributed something. Contribution parameters were also established by the strength of the writer's contributions to the paper. It is worth noting that the data in this dataset is identical to that in dataset B. This dataset includes 200 research papers, 30 of which have been published in peer-reviewed journals.

4.2.1 Experiment A results

The data for Experiment A was gathered from a variety of online sources and academic journals in Google Scholar. All 500 downloaded papers have been normalized into ARFF file format so that they can be run in WEKA, and 400 of these articles have been used in this process. The values mentioned in the A, B, and C datasets were used in this experiment. These datasets include anywhere from five to twenty contributors. From 05 publications to 20 publications for any author who is a first contributor, or one of the first three contributors, or any of the contributors. Model performance can only be evaluated on a variety of test and train values, which is why datasets are categorically divided among different authors according to their level of participation.

4.2.2 SVM results for Dataset A

Each author has ten papers in this dataset in which they are the initial author. This dataset includes 200 research papers. Just those authors whose names appear in these research articles' first paragraphs are required to be identified by the support vector machine's task. Using a different test and training set for the same dataset mitigates uncertainty about the dataset's complexity. Only five papers by each author were considered in the first training and implementation iterations, and the same was preserved in the train and test datasets. The second technique is to use the full author set, which in this dataset includes 10 papers written by each author as a primary author. We used an even and odd mixed series for our model's test and training sets, respectively, in order to accurately describe our model. Adding more authors to the classification cycle gets difficult for support vector machines as the size of the training and testing classes increases and the model's comprehension of the combined data becomes more complicated.

Table 2. SVM performance metrics for Dataset A with 05 papers by each author

Performance metrics for Dataset A for 5 papers by every author			
Number of Authors	Accuracy %	F-measure	Kappa Statics
02 Authors with 05 papers of every	97	0.97	0.97
04Authors with 05 papers of every	95.00	0.86	0.82
05Authors with 05 papers of every	85.02	0.83	0.74
07Authors with 05 papers of every	58.34	0.54	0.53
10Authors with 05 papers of every	47.92	0.48	0.33

To explain how SVM performs with different values of test and train dataset, Table 2 shows that accuracy decreases as the number of authors' publications increases. At the same time, it's the most accurate with the least amount of author effort. When the two authors with the most articles in the corpus are selected in a train set, the output is at its highest. There is a 99 percent accuracy rate, which is among the best model performance. As the number of authors who each contributed to five publications increased, the accuracy of 95.00 percent was shown, which decreased to 85.02 percent for 05 authors, 58.34 percent for 07 authors with 05 individual publications, and 10 authors with 05 first-author contributions. According to these findings, the accuracy of the classifier falls as the number of authors involved increases (the addition of more data).

As the number of authors with the same number of publications increases, the F-measure for SVM decreases. F-measure is maximized while the number of writers is kept to a minimum while the contribution from each author is equal. When two authors are selected in a train set, the output is at its highest. In this case, the F-measure is 0.99, making it one of the highest model measures available. With example, the F-measure drops from 0.82 to 0.83 for the addition of the fourth author, and then drops to 0.57 for the addition of the seventh, and finally drops to a very low 0.45 for the addition of the tenth author. These findings show that as the number of authors contributing a total of 05 papers increases, so does the clas-

sifier's F-measure. Because the Kappa statics gauges' class imbalance concerns as well as the logical relevance amongst classifiers for inter-rated reliability, these values are critical for model performance.

Table 3 shows the values of Kappa statics for dataset A when SVM is applied to it with various test and train dataset values, showing that the Kappa statics for SVM drop as the number of published authors grows. It has the highest Kappa statistics with the fewest authors and an equal contribution to the papers. When two authors with five publications each are selected in a train set, the output is at its highest. It has one of the highest model measures of Kappa statics at 0.99. Increasing the number of writers from two to four shows Kappa statics of 0.80, which reduces to 0.75 for the fifth and seventh authors, 0.51 for the eighth and ninth, and 38 for the tenth and twelfth authors, respectively. Increasing the number of authors with a 05-paper contribution each increases the Kappa statics of the classifier's value.

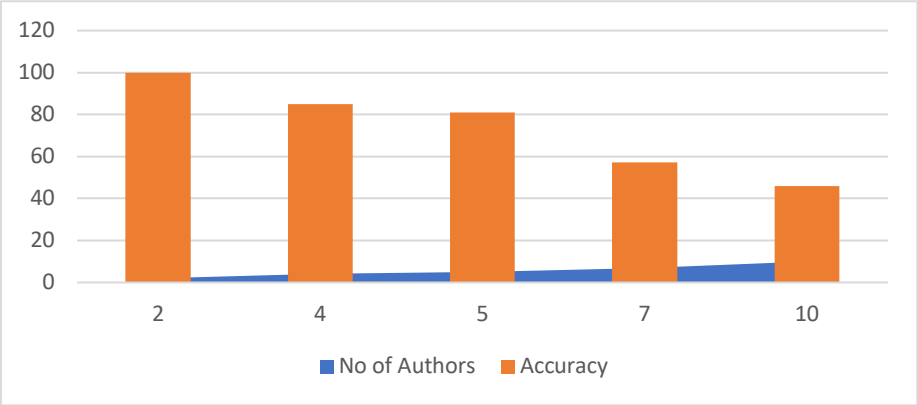


Figure 5. SVM accuracy for Dataset A (05 paper/author)

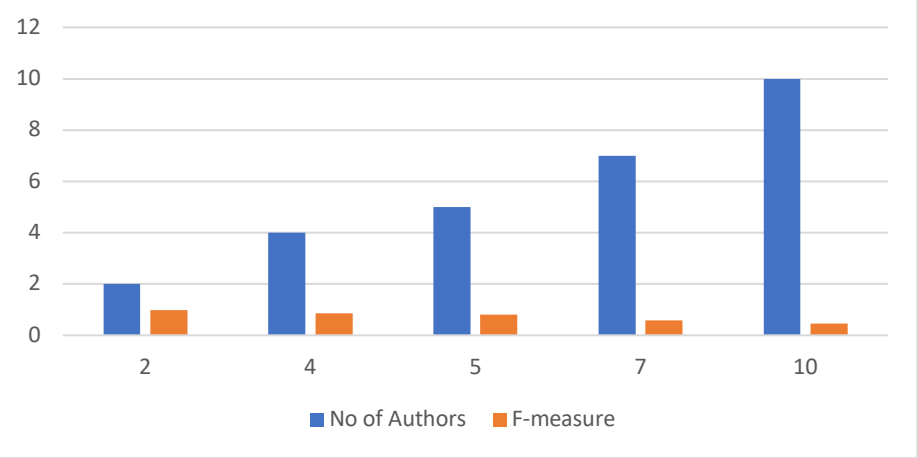


Figure 6. SVM F-measure for Dataset A (05 paper/author)

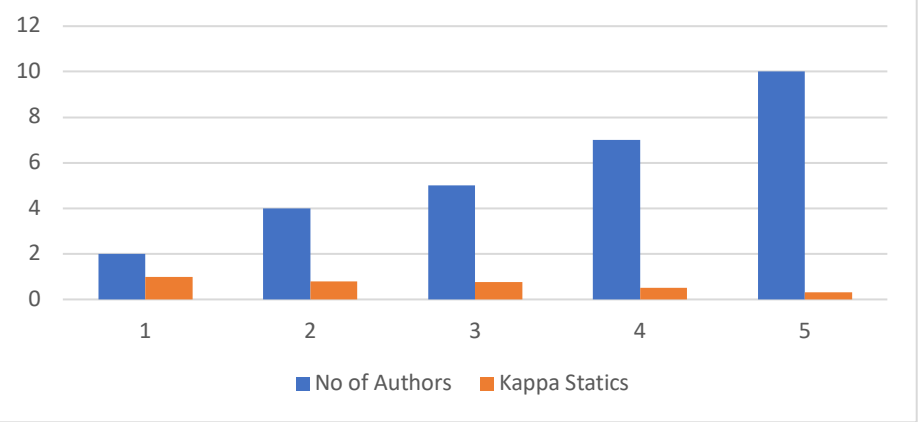


Figure 7. SVM kappa stats for Dataset A (05 paper/author)

In figure 5, the accuracy of dataset A is depicted as a percentage. The accuracy of SVM decreases as the number of authors in the train and test sets increases but the amount of contributions remains constant. When the model's F-measure ranged from 0.99 to 0.45 in figure 7, the same thing happened. This showed that as the data values increased, our method's error detection rate decreased but its precision decreased. In figure 8, the kappa stats are depicted. As the number of training samples increases, the machine's relative advantage over the classifier drops, resulting in a lower overall level of performance.

As part of the second iteration, ten research papers are due from each author listed as the first author in Dataset A. Table 3 summarizes the total performance of each author's 10 articles, as well as the three performance metrics that were used in the analysis.

Table 3. SVM shows metrics for Dataset A

Performance metrics for Dataset A for 10 papers by each author			
Number of Authors	Accuracy %	F-measure	Kappa Statics
05 Authors with 10 papers on both	82.02	0.87	0.79
10Authors with 10 papers on both	69.26	0.69	0.68
15Authors with 10 papers on both	66.95	0.64	0.57
17Authors with 10 papers on both	58.82	0.57	0.53
20Authors with 10 papers on both	57.69	0.54	0.51

Research articles in this section will have a variety in authorship from one to ten authors, depending on the number of contributors we have chosen for each publication. 05, 10, 15, 17, and 20 writers were randomly picked for their research papers. It is 82.02 percent accurate, 69.23% accurate, 66.91 percent accurate, 58.81% accurate, and 57.69 percent accurate, in that order. As the number of writers grows, so does the amount of data that can be used to train the algorithm. In addition, the fact that additional data for taking makes it more efficient in the testing phase has also been shown.

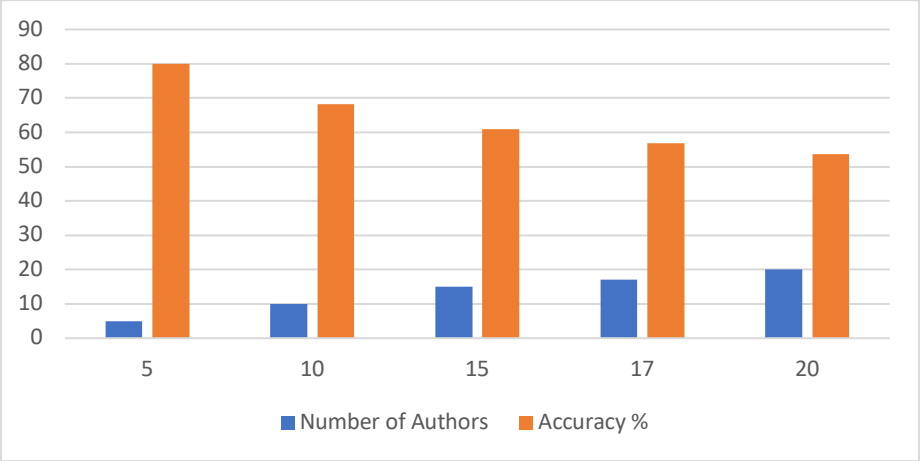


Figure 8. Dataset for Accuracy part A (10 paper by each author)

It has an F-measure value of 0.87 for five authors, ten authors, ten papers by each author, fifteen authors, six papers by each author, seven papers by each author, and the F-measure value of 0.59 for ten writers, ten papers by each author. Maximum F-measure value for 05 authors with 10 papers by each author is available here. There are 5 authors in 10 publications by each author, and the kappa statistic value is 0.79, which is the highest value in this table. 0.65 For ten writers; 0.57 for 15 authors; 0.53 for 17 authors; and 0.51 per author for 10 authors with 10 papers.

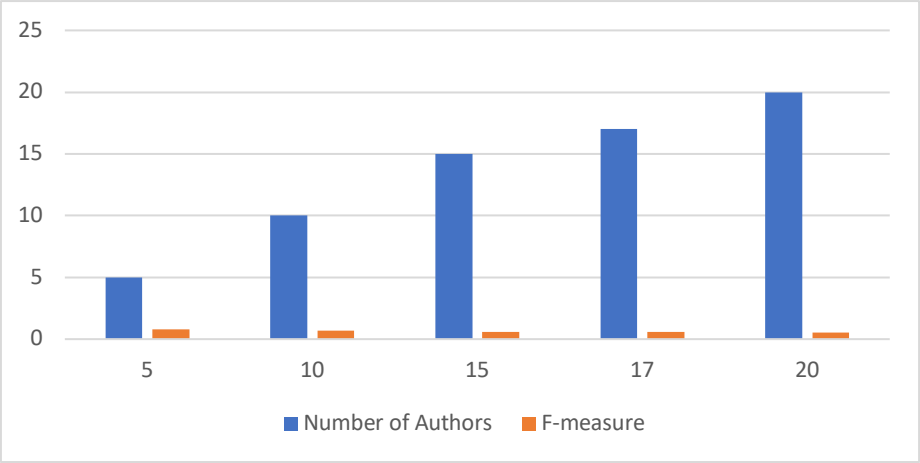


Figure 9. Dataset for F-measure part A (10 articles by each author)

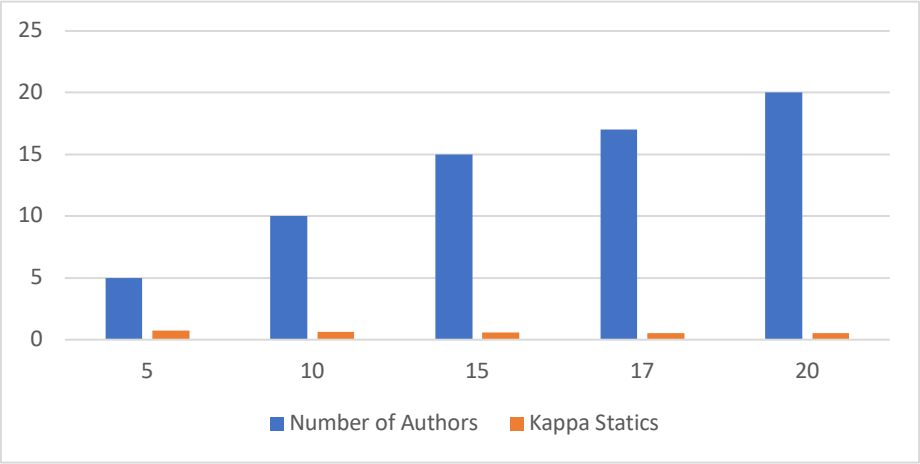


Figure 10. Dataset for Kappa statics part A (10 papers by each author)

4.3 Experiment B

Instead of using any of the previously downloaded papers, "Experiment B" relied on research papers that students had either submitted themselves or were provided by their professors or supervisors. This section contains papers from various conferences related to information retrieval. There are separate training and testing phases in these conferences that are tailored to the needs of machine learning. For the training phase, machine learning algorithms can learn information and patterns from provided data values if they are trained in well-arranged settings. Using the testing set, the model's performance accuracy may be determined by comparing it to the training set's results. Even the tiniest bits of data may now be analyzed by machines, thanks to the latest breakthroughs in machine learning. Detailed breakdowns of the complete dataset are shown in Table 4. Testing samples are obtained from some of the most important conference sources, whereas most of the variables relying on our selection procedure, either percentage spit or cross-validation, are involved in training samples.

Table 4. Train and test set in experiment B

Experiment B	
Train set	Test Set
CIKM 2012, CIKM 2013, SIGIR 2005, SIGIR 2011, CIKM 2005, SIGIR 2012, SIGIR 2013, ECIR 2012, ECIR 2013 and RIAO 2007, CIKM 2011.	SIGIR 2014, CIKM 2014, ECIR 2015 and ECIR 2014.

Data from these experiments will be used to create the experiment B dataset. These are the subject of numerous academic articles. In most situations, machine learning algorithms are unable to complete the training part in the required time frame due to the fact that most research papers have 5 or more contributors, making it impossible for the algorithm to differentiate the associated article for its true single source. Some researchers have their name listed in only one research paper in this data set. In authorship identification, only those authors who have published more than two papers in selected journals are considered Iterations of data normalization are required, and some data values that were not suitable for training and used more resources than usual are removed from the initial phase. In the second iteration, the algorithm was unable to read any pdf articles, and as a result, these papers were removed from the dataset. A normalized dataset (table 5) results from the process of normalization being repeated.

Table 5. Experiment B dataset

Final Dataset for Experiment B	
Unique authors identified	407
Train set	1684
Test set	708
Maximum number of contributions in dataset per author	35
Minimum number of contributions in the dataset per author	04

In the final selection phase, there are 1683 training examples and 707 testing instances in the dataset. Because of the large number of authors who had only contributed a single paper to the corpus, it had grown enormously. The data values for authorship identification were made easier to train with this standardization. A previous range for writers who have made at least two contributions to the publication in question, up to a maximum of 30. For the model to function properly, this strategy offered more information patterns in comparison to several contributions than a single instance regarding any data value could provide. Using this method, you can figure out how a certain author chooses his or her words. As a result, the model's performance suffers as a result of slower response times and greater utilization of resources.

4.3.1 SVM Results for Dataset B

The first three authors' texts are easier to identify than the first author's alone. Each author's dataset contains 20 papers in which he/she ranks among the first three authors in this selection phase. This dataset contains a total of 400 research papers. Using only the names of the top three authors of each of these studies, support vector machines must determine which tests belong to which authors. Using distinct datasets for testing and training helps clear up any confusion caused by the dataset's complexity. The second method makes use of the entire dataset's initial three authors, each of whom has 20 publication contributions to their name. It is stated in this part that there must be a minimum of 20 authors to reflect the authorship identification model, but we employed an even and odd mixed series for testing and training purposes. For support vector machine classification, adding additional authors with the same number of publications in the classification cycle gets harder since the train and test class sizes increase and model comprehension of combined data becomes more complicated. Each author in this dataset must have contributed at least 20 research publications, and only the top three authors and a random number of authors from these lists were selected. Table 6 shows the performance metrics for Dataset B, as seen in the figure.

Table 6. SVM Performance for Dataset B by the first three authors

Number (N) of Authors (A)	**Accuracy %**	**F-measure**	**Kappa Statics**
(10) A with 20 Articles of each	60.37	0.63	0.55
(15) A with 20 Articles of each	59.10	0.57	0.47

(20) A with 20 Articles of each	55.87	0.51	0.45

In this area, we've included anywhere from one to twenty different authors. Those authors who have contributed at least 20 articles to the field of research have been picked, and their names appear in the top three authors of selected research publications. A total of 10, 15, and 20 authors were selected, each with a specified quantity of contributions. In order to test our model's ability to correctly identify authors in a variety of text files from various sources, we need to select a large number of writers who have been published in academic journals.

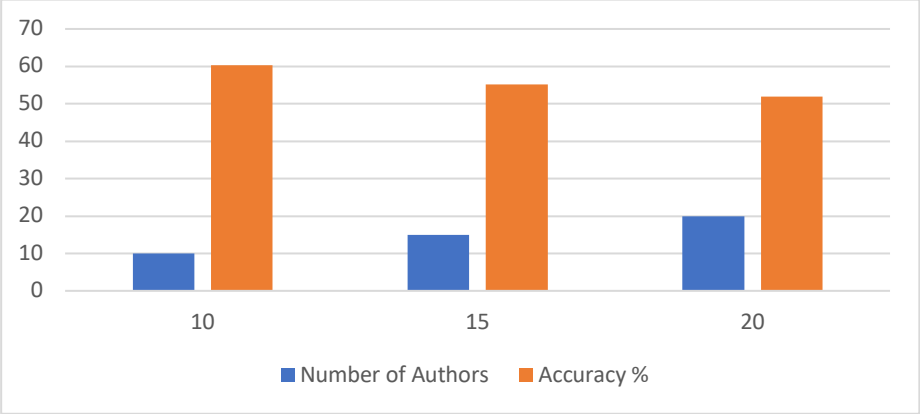


Figure 11. Accuracy for Dataset B (First three authors)

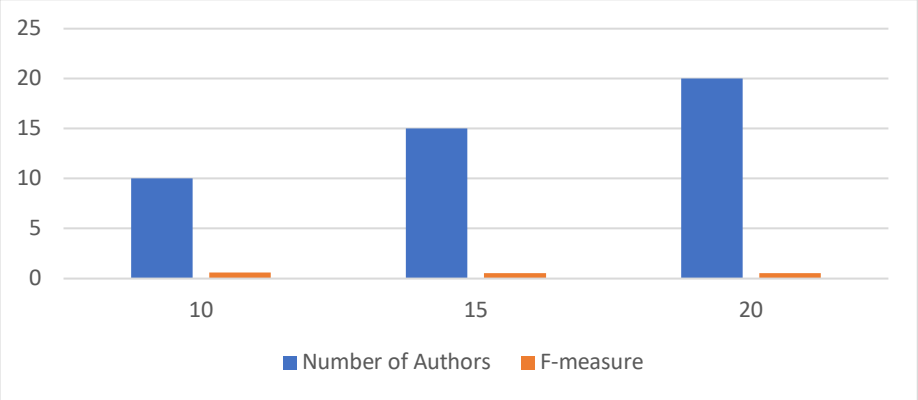


Figure 12. Dataset F-measure for part B

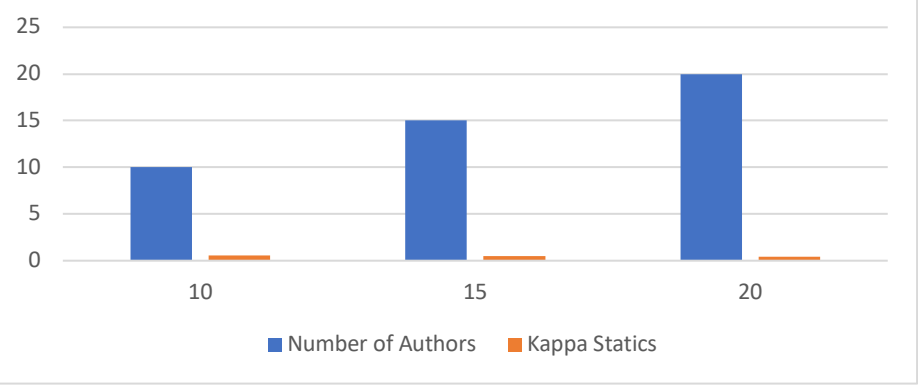


Figure 13. Dataset for Kappa statics part B

4.3.2 SVM result for experiment B

Experiment B, which is characterized as training and testing class for a classification method to better understand our model, utilizes data from a variety of conferencing sources. It is still difficult to run WEKA on a dataset that contains multiple PDF files originating from a variety of online sources. This is due to the

fact that the enormous data files remained in the training phase for the majority of the time, and in many instances, an accurate estimation of when the training phase would conclude could not even be made. When compared to experiment A, the dataset under consideration here contains a significantly greater number of examples for both training and testing. According to the original data values presented in a number of study articles, there are a total of 408 unique authors, and each of them is responsible for a minimum of two publications under their name. The number of articles that an author can have in this dataset is limited to no more than 30. It is determined how effective the classifier's training was by exposing it to 707 instances taken from the model's total of 1683 training classes. The sparse and dispersed data features of this dataset were the root cause of the poor performance of the dataset as well as the lengthy validation period. Table 7 contains a representation of a large dataset that contains text drawn from a variety of subject areas. The heap size for the SVM configuration was increased from 40.0 to 2000.0 so that it could accommodate more processing RAM for the purpose of storing the attribute values for the validation phase of the study. This was done as a direct result of the size of the dataset.

Table 7. SVM for Experiment B

Number of Authors	Accuracy %	F-measure	Kappa Stats
407Authors with 02 to 30 papers of each	11.09	0.17	0.08

Because this dataset comprises a lot of huge data files, these results are significantly lower than those from experiment A. It took a long time for the SVM algorithm to process this dataset. It's not acceptable to have a low level of accuracy because of its lack of relevance to the final product. SVM results show that a text may be categorized into its original relevance about its creator, regardless of where it came from. In this dataset, the SVM method has a relatively high rate of misclassification. Using this dataset, the model was able to accurately identify 70 out of 707 test instances.

4.4 SVM Results for Dataset C

The data in this collection has been filtered to make it easier to identify the creator, who could be any of the dataset's contributors. This dataset includes all authors who made a contribution to a research publication. This dataset has a random number of writers. As compared to the first author or the first three writers, the model identification tasks get easier. Each author's dataset contains 20 papers in which he/she ranks among the first three authors in this selection phase. This collection includes 400 research papers. Using a support vector machine, the aim is to identify tests by the name of the author for only those authors who are the authors in these research publications. Using distinct datasets for testing and training helps clear up any confusion caused by the dataset's complexity. The entire group of 20 authors, each of whom contributed 20 papers to this dataset, was employed in this approach. We employed an even and odd mixed series for the test and training set implementation of our authorship identification model in this part, which has a minimum number of authors of 20. Support vector machine (SVM) algorithms can be improved by using additional data, such as author numbers, to distinguish between original authors based on their profile as domain expertise and whether or not they contributed to the work. All authors listed in the author's list, as well as a random number of others, contributed to this dataset, with each author having 20 contributions throughout the 20 papers. Dataset C's performance metrics are shown in Table 8.

Table 1. Support Vector Machine Performance for Dataset C using any authors

Quantity of Authors	Accuracy %	F-measure	Kappa Statics
The (4) Authors with (20) paper of each	79.34	0.80	0.72
The (8) Authors with (20) paper of each	67.24	0.67	0.63
The (12) Authors with (20) papers of each	58.23	0.59	0.53
The (16) Authors with (20) papers of each	49.1	0.50	0.46
The (20) authors with (20) papers of each	48.3	0.49	0.45

Authors were chosen from a group of 20 in this section. The number of authors who have each contributed at least 20 research publications, together with the names of those writers, has been narrowed down to a select group. We chose four, eight, twelve, sixteen, and twenty writers from a manuscript having the aforementioned number of contributions. However, all of these are not identical, and the rationale for selecting numerous authors identify the recital effectiveness of our model when it is given diverse text-files of research articles by the different validation and sources is gained to properly classify authors. As more authors are included in the data values, the identification process becomes more difficult because the model is unable to tell which authors belong to the true author. There are at least 20 research papers written by each of our authors, who were chosen from a pool of hundreds of applicants for consideration. A random sample of four, eight, 12, and 20 authors were chosen to write research articles. Accuracy levels range from 79% to 67% to 68% to 58% to 58% to 48%... Because it has more data to train on, the accuracy decreases as the number of writers increases.

In addition, the fact that additional data for training makes it more efficient in the testing phase has also been shown. It has a 0.80 F-measure value for four authors with 20 papers by each author, 0.67 for eight, 0.59 for 12, and 0.50 F-measure values for 16 writers with 20 papers by each author. This paper has a maximum F-measure score of 4. Each author has 20 papers. With four authors and 202 papers, the kappa statics performance value is 0.72, which is the highest in this table. When there are eight authors, the score drops to 0, 63; when there are twelve authors, it drops to 0, 53; when there are sixteen authors, it drops to 0, 45; and when there are twenty authors, it drops to 0, 45. The total performance for a randomly selected dataset is better than that of the first author and the first three authors collectively.

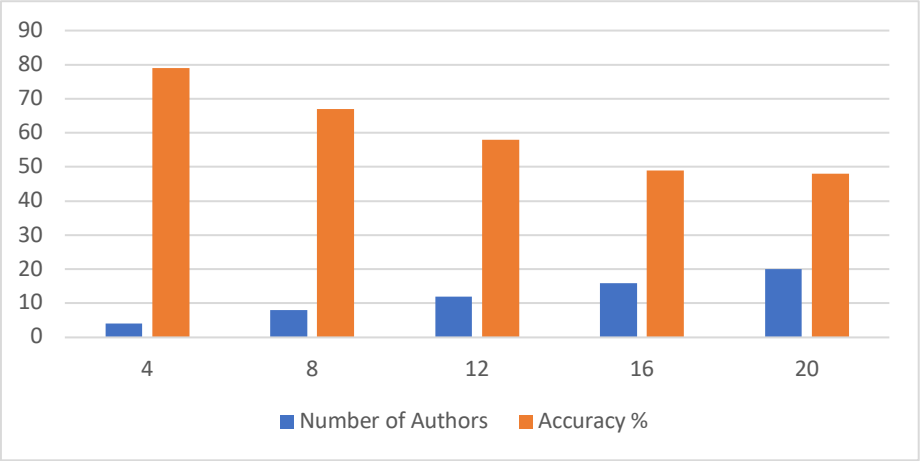


Figure 14. Dataset accuracy part C (Somebody of authors)

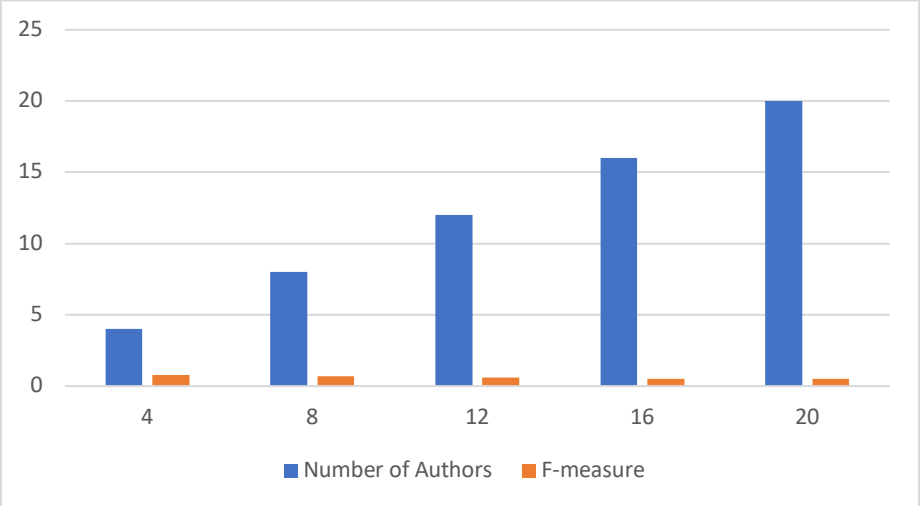


Figure 15. Dataset for F-measure part C (Anyone of authors)

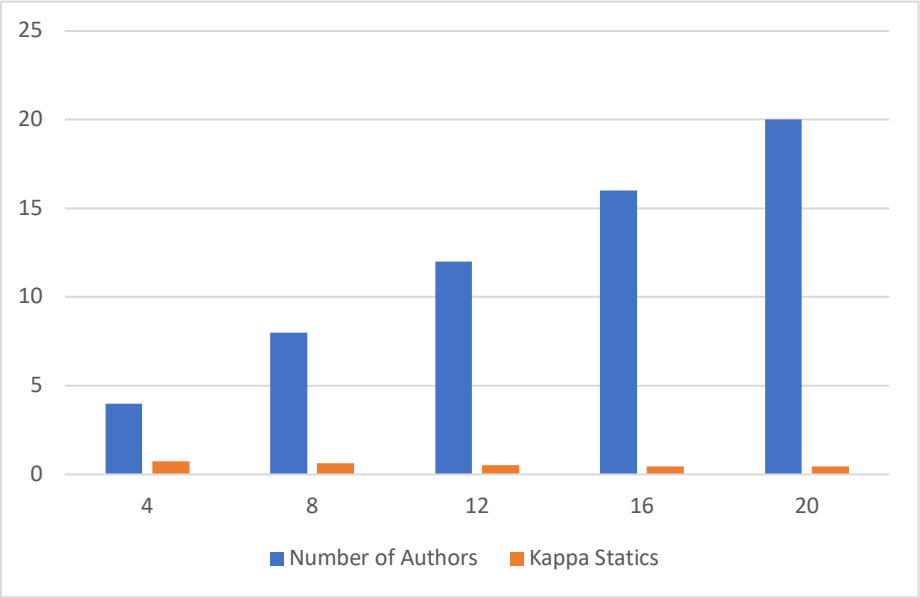


Figure 16. Dataset for F-measure part C (Anyone of authors)

4.5 Performance metrics

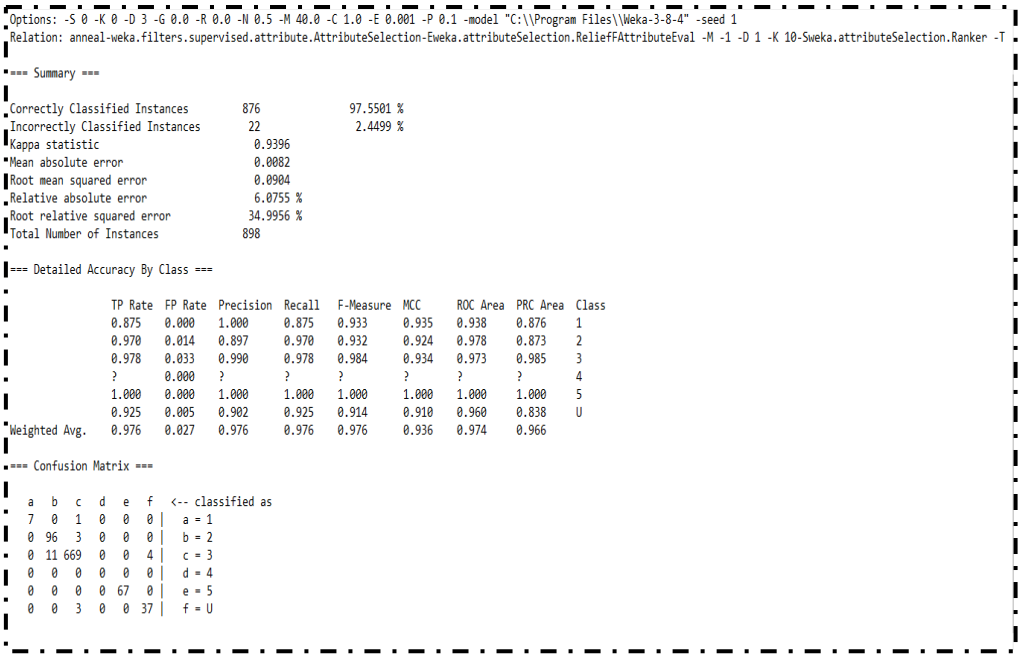
There are a variety of measures for measuring the performance of machine learning systems. Metrics can be used for a variety of purposes, depending on the nature of the output. As a rule of thumb, classification and misclassification rates of the applied model are taken into account while selecting a classifier. Table 9 displays the metrics that were chosen, along with a brief discussion of each. As a combination of several single evaluation values, these three metrics --indicate how important these selected measures are.

Table 9. Performance Metrics

Performance Metrics	
Metrics	**Description**
F-measure	Measure model accuracy Precision and recall combined view The harmonic mean of precision and recall
Accuracy	Correctly classified instances measure Represent the percentage of true results It is a ratio of true values to total value
Kappa Statics	Multiclass and imbalnce class representer Measure inter-rated reliability of classifier Statistical performance comparison measure

The model's overall accuracy is also a measure of the model's evaluation limit. Weka machine learning output is evaluated using a variety of measures depicted in Figure 13. Running a built-in dataset from the WEKA repository yields the performance figures and outcomes.

These are the key metrics for evaluating job performance that were discussed in detail in Chapter 3. Only three performance indicators, namely accuracy, f-measure, and Kappa stats, are employed in this study.



```
Options: -S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model "C:\\Program Files\\Weka-3-8-4" -seed 1
Relation: anneal-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10-Sweka.attributeSelection.Ranker -T

=== Summary ===

Correctly Classified Instances         876               97.5501 %
Incorrectly Classified Instances        22                2.4499 %
Kappa statistic                          0.9396
Mean absolute error                      0.0082
Root mean squared error                  0.0904
Relative absolute error                  6.0755 %
Root relative squared error             34.9956 %
Total Number of Instances              898

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.875    0.000    1.000      0.875    0.933      0.935    0.938     0.876     1
               0.970    0.014    0.897      0.970    0.932      0.924    0.978     0.873     2
               0.978    0.033    0.990      0.978    0.984      0.934    0.973     0.985     3
               ?        0.000    ?          ?        ?          ?        ?         ?         4
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     5
               0.925    0.005    0.902      0.925    0.914      0.910    0.960     0.838     U
Weighted Avg.  0.976    0.027    0.976      0.976    0.976      0.936    0.974     0.966

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   7   0   1   0   0   0 |   a = 1
   0  96   3   0   0   0 |   b = 2
   0  11 669   0   0   4 |   c = 3
   0   0   0   0   0   0 |   d = 4
   0   0   0   0  67   0 |   e = 5
   0   0   3   0   0  37 |   f = U
```

Figure 1. Performance Key Matrices in Weka

5. Conclusions

Identifying the writers of a piece of writing, whether anonymous or not, is a procedure that focuses solely on the writing style and not on the content itself. The primary goal of this article is to examine and apply a variety of categorization approaches to research articles that analyze author identity and the content of those texts that are in dispute. Researchers' earlier work is also discussed and elaborated upon. After that, we were able to exhibit better findings from our experiments. Authorship detection using machine learning is discussed in this article, and we demonstrate how to use a support vector machine in practice. With feature spaces of high dimensions and sparse data for individual cases, the SVM technique is particularly well-suited to the aforementioned problem. Across all experiments, it was found that the SVM was effective at determining the authorship of research publications. SVM was used to classify two sets of data in that study. Sections A and B of the experiment are referred to as Experiment A and B, respectively. 500 research papers from conferences and the majority of them from Google Scholar are included in Experiment A.

All articles with authors who only contributed to a single research paper were likewise deleted in the following iteration. We now have a final set of 400 research papers after all the data has been cleaned up. A, B, and C are the three subsets of the 400 research papers for high performance that were further subdivided. In dataset A, the only author of a research paper that can be recognized is the first author, whereas in dataset B, the first three authors of a research paper are the only ones that can be identified. Our model was able to train quickly and evaluate good performance on these limited research criteria in these datasets as a result of an increase in both the number of authors and the number of publications that were included in the dataset. Experiment B formalizes and assesses a second dataset as a complete dataset with a low yield due to the dataset's enormous number of files and various data domains. The poor yield is due to the fact that the dataset contains numerous data domains. Because we can utilize any words and letters as attributes in SVM, we are exempt from the need that select particular features. Aside from that, it makes no difference how the features are weighted or preprocessed because the outcomes of using numerous ways are the same.

References

1. Abuhamad, M. et al. Code authorship identification using convolutional neural networks. Futur. Gener. Comput. Syst. 95, 104–115 (2019).
2. Matalon, Y., Magdaci, O., Almozlino, A. & Yamin, D. Using sentiment analysis to predict opinion inversion in tweets of political communication. Sci. Rep. 11, 1–9 (2021).
3. Smith, T. B., Vacca, R., Mantegazza, L. & Capua, I. Natural language processing and network analysis provide novel insights on policy and scientific discourse around sustainable development goals. Sci. Rep. 11, 1–10 (2021).
4. Tamboli, M. S. & Prasad, R. S. Authorship analysis and identification techniques: A review. Int. J. Comput. Appl. 77, 11 (2013).
5. Zhang, K. et al. Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. IEEE Trans. Knowl. Data Eng. (2021).
6. Durazzi, F., Müller, M., Salathé, M. & Remondini, D. Clusters of science and health related twitter users become more isolated during the covid-19 pandemic. Sci. Rep. 11, 1–11 (2021).
7. Stappen, L., Baird, A., Schumann, L. & Bjorn, S. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. IEEE Trans. Afect. Comput. (2021).
8. Benzebouchi, N. E. et al. Authors' writing styles are based authorship identification system using the text representation vector. In 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), 371–376 (IEEE, 2019).
9. Iqbal, F., Binsalleeh, H., Fung, B. C. & Debbabi, M. A unified data mining solution for authorship analysis in anonymous textual communications. Inf. Sci. 231, 98–112 (2013).
10. Ziani, A., Azizi, N. & Guiyassa, Y. T. Combining random subspace algorithm and support vector machines classifier for Arabic opinions analysis. In Advanced Computational Methods for Knowledge Engineering, 175–184 (Springer, 2015).
11. Steinfeld, B. et al. The role of lean process improvement in the implementation of evidence-based practices in behavioral health care. J. Behav. Health Serv. Res. 42, 504–518 (2015).
12. Rabab'Ah, A., Al-Ayyoub, M., Jararweh, Y. & Aldwairi, M. Authorship attribution of Arabic tweets. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), 1–6 (IEEE, 2016).
13. Stamatatos, E. On the robustness of authorship attribution based on character n-gram features. J. Law Policy 21, 421–439 (2013).
14. Zheng, R., Li, J., Chen, H. & Huang, Z. A framework for authorship identification of online messages: Writing-style features and classification techniques. J. Am. Soc. Inform. Sci. Technol. 57, 378–393 (2006).
15. Mohsen, A. M., El-Makky, N. M. & Ghanem, N. Author identification using deep learning. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 898–903 (IEEE, 2016).
16. Sarwar, R., Li, Q., Rakthanmanon, T. & Nutanong, S. A scalable framework for cross-lingual authorship identification. Inf. Sci. 465, 323–339 (2018).
17. Stappen, L., Baird, A., Cambria, E. & Schuller, B. W. Sentiment analysis and topic recognition in video transcriptions. IEEE Intell. Syst. 36, 88–95 (2021).
18. Benzebouchi, N. E., Azizi, N., Aldwairi, M. & Farah, N. Multi-classifier system for authorship verification task using word embeddings. In 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), 1–6 (IEEE, 2018).
19. Stein, R. A., Jaques, P. A. & Valiati, J. F. An analysis of hierarchical text classification using word embeddings. Inf. Sci. 471, 216–232 (2019).
20. Mikolov, T., Yih, W.-t. & Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, 746–751 (2013)
21. E. Stamatatos, Authorship verification: A review of recent advances, Research in Computing Science 123 (2016) 9–25.
22. N. Potha, E. Stamatatos, improving author verification based on topic modeling, Journal of the Association for Information Science and Technology 70 (2019) 1074–1088. doi:https://doi.org/10.1002/asi.24183.

23. S. Ding, B. Fung, F. Iqbal, W. Cheung, Learning stylometric representations for authorship analysis, IEEE Transactions on Cybernetics 49 (2019) 107–121.
24. B. Boenninghoff, R. M. Nickel, S. Zeiler, D. Kolossa, Similarity learning for authorship verification in social media, in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2457–2461. doi:10.1109/ICASSP.2019.8683405.
25. T. Gollub, M. Potthast, A. Beyer, M. Busse, F. M. R. Pardo, P. Rosso, E. Stamatatos, B. Stein, Recent trends in digital text forensics and its evaluation - plagiarism detection, author identification, and author profiling, in P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings, volume 8138 of Lecture Notes in Computer Science, Springer, 2013, pp. 282–302.
26. M. Potthast, T. Gollub, F. M. R. Pardo, P. Rosso, E. Stamatatos, B. Stein, Improving the reproducibility of pan's shared tasks: - plagiarism detection, author identification, and author profiling, in E. Kanoulas, M. Lupu, P. D. Clough, M. Sanderson, M. M. Hall, A. Hanbury, E. G. Toms (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings, volume 8685 of Lecture Notes in Computer Science, Springer, 2014, pp. 268–299.
27. E. Stamatatos, M. Potthast, F. M. R. Pardo, P. Rosso, B. Stein, Overview of the PAN/CLEF 2015 evaluation lab, in: J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. J. F. Jones, E. SanJuan, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings, volume 9283 of Lecture Notes in Computer Science, Springer, 2015, pp. 518–538.
28. J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. M. R. Pardo, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, volume 12260 of Lecture Notes in Computer Science, Springer, 2020, pp. 372–383.
29. J. Bevendorff, B. Chulvi, G. L. D. la Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, volume 12880 of Lecture Notes in Computer Science, Springer, 2021, pp. 419–431.
30. S. Bischoff, N. Deckers, M. Schliebs, B. Thies, M. Hagen, E. Stamatatos, B. Stein, M. Potthast, The Importance of Suppressing Domain Style in Authorship Analysis, CoRR abs/2005.14714 (2020). URL: https://arxiv.org/abs/2005.14714.
31. Kestemont, M. et al. (2019) 'Overview of the Crossdomain Authorship Attribution Task at PAN 2019.', in CLEF (Working Notes).
32. Jamak, A., Savatić, A. and Can, M. (2012) 'Principal component analysis for authorship attribution', Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy. Udruga za promicanje poslovne informatike, 3(2), pp. 49–56.
33. Boukhaled, M. A. and Ganascia, J.-G. (2015) 'Using function words for authorship attribution: Bag-ofwords vs. sequential rules', in Natural Language Processing and Cognitive Science. De Gruyter, pp. 115–122.
34. Martin-del-Campo-Rodriguez, C. et al. (2019) 'Authorship Attribution through Punctuation n-grams and Averaged Combination of SVM'
35. Pokou, Y. J. M., Fournier-Viger, P. and Moghrabi, C. (2016) 'Authorship Attribution using Variable Length Part-of-Speech Patterns.', in ICAART (2), pp. 354– 361.

36. Sapkota, U. et al. (2015) 'Not all character n-grams are created equal: A study in authorship attribution', in Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 93– 102.
37. Al-Ayyoub, M., Alwajeeh, A. and Hmeidi, I. (2017) 'An extensive study of authorship authentication of arabic articles', International Journal of Web Information Systems. Emerald Publishing Limited
38. El Bakly, A. H., Darwish, N. R. and Hefny, H. A. (2020) 'A Survey on Authorship Attribution Issues of Arabic Text', International Journal of Artificial Intelligent Systems and Machine Learning, 2, pp. 86–92.
39. A. M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 898-903, doi: 10.1109/ICMLA.2016.0161.
40. Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In Proceedings of the 27th International Conference on Computational Linguistics, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
41. Barlas, G., Stamatatos, E. (2020). Cross-Domain Authorship Attribution Using Pre-trained Language Models. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 583. Springer, Cham. https://doi.org/10.1007/978-3-030-49161-1_22
42. Ramezani, Reza. "A language-independent authorship attribution approach for author identification of text documents." Expert Systems with Applications 180 (2021): 115139.
43. Olga Fourkioti, Symeon Symeonidis, Avi Arampatzis, Language models and fusion for authorship attribution, Information Processing & Management, Volume 56, Issue 6, 2019, 102061, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2019.102061.
44. S. Okuno, H. Asai and H. Yamana, "A challenge of authorship identification for ten-thousand-scale microblog users," 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 52-54, doi: 10.1109/BigData.2014.7004491.
45. Z. Damiran and K. Altangerel, "Author Identification-An Experiment Based on Mongolian Literature Using Decision Trees." 2014 7th International Conference on Ubi-Media Computing and Workshops. IEEE, 2014. pp. 186-189.
46. Ramezani, Reza, Navid Sheydaei, and Mohsen Kahani. "Evaluating the effects of textual features on authorship attribution accuracy." ICCKE 2013. IEEE, 2013.
47. H. Paci, E. Kajo, E. Trandafili, I. Tafa and D. Salillari, "Author Identification in Albanian Language," 2011 14th International Conference on Network-Based Information Systems, pp. 425-430.
48. Pandian, A., V. V. Ramalingam, and R. V. Preet. "Authorship identification for Tamil classical poem (Mukkoodar Pallu) using C4. 5 algorithm." Indian Journal of Science and Technology 9.46 (2016).
49. Kale Sunil Digamberrao, Rajesh S. Prasad, Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi, Procedia Computer Science, Volume 132, 2018, Pages 1086-1101, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.024.
50. Oliveira W Jr, Justino E, Oliveira LS. Comparing compression models for authorship attribution. Forensic Sci Int. 2013 May 10;228(1-3):100- 4. doi: 10.1016/j.forsciint.2013.02.025. Epub 2013 Mar 24. PMID: 23597746.
51. Romanov, Aleksandr & Kurtukova, Anna & Shelupanov, Alexander & Fedotova, Anastasia & Goncharov, Valery. (2020). Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. Future Internet. 13. 3. 10.3390/fi13010003.
52. Fedotova, A.; Romanov, A.; Kurtukova, A.; Shelupanov, A. Authorship Attribution of Social Media and Literary Russian-Language Texts Using Machine Learning Methods and Feature Selection. Future Internet 2022, 14, 4. https://doi.org/10.3390/fi14010004
53. Sage, M., Cruciata, P., Abdo, R., Cheung, J.C., & Zhao, Y.F. (2020). Investigating the Influence of Selected Linguistic Features on Authorship Attribution using German News Articles. SwissText/KONVENS.

54. Otoom, Ahmed & Abdallah, Emad & Jaafer, Shifaa & Hamdallh, Aseel & Amer, Dana. (2014). Towards author identification of Arabic text articles. 2014 5th International Conference on Information and Communication Systems, ICICS 2014. 1-4. 10.1109/IACS.2014.6841971.
55. O. de Vel, A. Anderson, M. Corney, and G. Mohay. 2001. Mining email content for author identification forensics. SIGMOD Rec. 30, 4 (December 2001), 55–64. https://doi.org/10.1145/604264.604272
56. B. Kırmacı and H. Oğul, "Evaluating text features for lyrics-based songwriter prediction," 2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES), 2015, pp. 405-409, doi: 10.1109/INES.2015.7329743.
57. Upul Bandara, Gamini Wijayarathna, Source code author identification with unsupervised feature learning, Pattern Recognition Letters, Volume 34, Issue 3, 2013, Pages 330-334, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2012.10.027.
58. Alonso-Fernandez, Fernando & Belvisi, Nicole & Hernandez-Diaz, Kevin & Muhammad, Naveed & Bigun, Josef. (2021). Writer Identification Using Microblogging Texts for Social Media Forensics. IEEE Transactions on Biometrics, Behavior, and Identity Science. PP. 1-1. 10.1109/TBIOM.2021.3078073.
59. M. S. Atar, E. Esen and M. A. Arabaci, "Supervised author recognition with aggregated word embeddings," 2018 26th Signal Processing and Communications Applications Conference (SIU), 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404464.