


# Stylometric Authorship Attribution of Collaborative Documents

Edwin Dauber , Rebekah Overdorf, and Rachel Greenstadt  
Mailing Address: 3141 Chestnut St, Philadelphia, PA 19104, USA  
Phone: 215-895-2669

Drexel University, Philadelphia, PA, 19104, USA  
egd34@drexel.edu

**Abstract.** Stylometry is the study of writing style based on linguistic features and is typically applied to authorship attribution problems. In this work, we apply stylometry to a novel dataset of multi-authored documents collected from Wikia using both relaxed classification with a support vector machine (SVM) and multi-label classification techniques. We define five possible scenarios and show that one, the case where labeled and unlabeled collaborative documents by the same authors are available, yields high accuracy on our dataset while the other, more restrictive cases yield lower accuracies. Based on the results of these experiments and knowledge of the multi-label classifiers used, we propose a hypothesis to explain this overall poor performance. Additionally, we perform authorship attribution of pre-segmented text from the Wikia dataset, and show that while this performs better than multi-label learning it requires large amounts of data to be successful.

**Keywords:** stylometry, authorship attribution, machine learning, multi-label learning

This is a regular submission.

## 1 Introduction

Multi-label machine learning models are designed to assign multiple labels to an unlabeled sample in a classification task. These methods are well studied and have been used to great success in different real world learning problems in many distinct areas of research, such as image recognition and text categorization. In this work, we study the multi-label problem in the context of authorship attribution.

Authorship attribution methods have been used successfully to uncover the author of documents in many different domains and areas. These methods can be used to compromise privacy and uncover the author of any anonymous text on the web. There is an important caveat, however, to the use of current state-of-the-art authorship attribution techniques. While they are very effective with documents written by a single person, they are not designed to handle collaboratively written documents. With the rise of Internet collaborative writing platforms such as Wikipedia<sup>1</sup> and GoogleDrive<sup>2</sup>, the development of new techniques to handle multi-authored text is necessary.

Collaboration has also been considered as a stylometric defense [3]. By either having another author rewrite text to obfuscate it or writing collaboratively, standard stylometric methods fail to identify the correct author. We present an analysis of new stylometric methods specifically designed for multi-label classification that address this type of obfuscation.

Our contributions are as follows. We define five variations of the multi-label stylometry problem based on the availability of training data, test both traditional single-label stylometric techniques and multi-label classification techniques as methods to solve our variations on authentic collaborative documents collected from the Internet, and identify successes and limitations of these techniques. Specifically, we identify one of these variations, which we call *consistent collaboration*, for which these techniques are promising, at least for small closed-world scenarios, and we demonstrate that these techniques are insufficient as-is to solve the other four variations for even small closed-world scenarios. We also present a hypothesis to explain the performance on these different variations. We then show that account attribution using pre-segmented texts is possible given sufficient data and present an analysis of the level of separation in collaboration on these real-world documents, as a way of predicting the viability of supervised segmentation as an alternative to multi-label stylometry.

We formally define the multi-author stylometry problem in Section 2. We examine previous work related to multi-authored documents and Wikipedia in Section 3. We discuss our dataset in Section 4 and our methodology in Section 5. We demonstrate the results of our experiments in Section 6, discuss our results in Section 7, and discuss future work in Section 8.

## 2 Problem Statement

We consider two problems in which the authors of a collaborative document are in question. In the first problem, the only documents of known authorship are non-collaborative, single-authored documents. In the second problem, multi-authored documents of known authorship are available.

### 2.1 Non-Collaborative Training Documents

We define two variations in which the available training documents are non-collaborative.

---

<sup>1</sup> <http://en.wikipedia.org>

<sup>2</sup> <https://drive.google.com>

*Complete suspect set:* Non-collaborative documents of known authorship are available for each suspect. More formally: given a set of  $n$  authors  $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ , and a set of documents  $\mathcal{D}_i$  for each  $A_i$  which we know to be written by only that author; we want to identify the  $k$  authors of a document of unknown authorship  $d$ .

*Partial suspect set:* Non-collaborative documents of known authorship are available for some of the suspects. More formally: given a set of  $n$  authors  $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ , and a set of documents  $\mathcal{D}_i$  for each  $A_i$  which we know to be written by only that author, and a document of unknown authorship  $d$  written by  $k$  authors, of which  $c$  authors are in our suspect set, we want to identify those  $c$  authors.

## 2.2 Collaborative Training Documents

In the case where suspect authors have collaborative writings, we consider a subproblem in which all documents have the same number of authors. This problem has three variations.

*Consistent collaboration:* The suspect set consists of pairings or groups of authors who are suspected of collaboratively writing the document in question together. Formally: given a set of  $n$  author groups  $\mathcal{G} = \{G_1, G_2, \dots, G_n\}$ , where  $G_i = \{A_1, A_2, \dots, A_m\}$  and each  $G_i$  has a set of documents  $\mathcal{D}_i$  which we know to be written collaboratively by  $\{A_1, A_2, \dots, A_m\}$ , identify the true group of authors  $G_t \in \mathcal{G}$  of a document of unknown authorship  $d$ . This provides us with a best-case scenario in which we know all of the possible combinations of authors of  $d$  and have sufficient training data.

*Mixed collaboration:* Collaborative documents written by some of the suspect groups are unavailable, but other collaborative works by suspect authors are available. Formally: given a set of  $n$  author groups  $\mathcal{G} = \{G_1, G_2, \dots, G_n\}$  where  $G_i = \{A_1, A_2, \dots, A_m\}$  and each  $G_i$  has a set of documents  $\mathcal{D}_i$  which we know to be written collaboratively by  $\{A_1, A_2, \dots, A_m\}$ , identify the true group of authors  $G_t$  of a document of unknown authorship  $d$ , such that  $G_t$  may or may not be an element of  $\mathcal{G}$ . This provides us with an average-case scenario for which we know some of the possible combinations of authors of  $d$  and have sufficient training data for some of them while having limited training data for others.

*Inconsistent collaboration:* Collaborative documents written by the suspect groups are unavailable, but other collaborative works by suspect authors are available. Formally: given a set of  $n$  author groups  $\mathcal{G} = \{G_1, G_2, \dots, G_n\}$  where  $G_i = \{A_1, A_2, \dots, A_m\}$  and each  $G_i$  has a set of documents  $\mathcal{D}_i$  which we know to be written collaboratively by  $\{A_1, A_2, \dots, A_m\}$ , identify the true group of authors  $G_t \notin \mathcal{G}$  of a document of unknown authorship  $d$ . This provides us with the worst-case scenario that the authors of  $d$  have not collaborated in the past or such data is unavailable.

## 2.3 Pre-segmented Text

We consider one more problem in this paper, in which we have text which has already been segmented by anonymized author. Specifically, we use the revision history to segment the wiki articles by user account at the sentence level. In this case, we want to attribute the author's account. More formally: given a set of  $n$  authors  $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ , each of whom has a set of documents  $\mathcal{D}_i$  which we know to be written by only that author; we want to identify the author of an account  $a$  containing a set of  $k$  document segments  $a = \{s_1, s_2, \dots, s_k\}$ .

## 3 Background and Related Work

### 3.1 Multi-Label Learning

There have been a number of proposed techniques for multi-label learning that we consider in this work. All of these methods have been tested on various multi-label problems, but to the best of our knowledge, none of them have been proposed for solving the collaborative authorship attribution problem.

Multi-Label k-Nearest Neighbors (MLkNN) [21] is a lazy learning approach derived from the popular k-nearest neighbors classifier that utilizes MAP estimation based on the count of each label in the nearest neighbor set. Because of the likelihood estimation, this method performs well at ranking authors by likelihood of being one of the collaborators. It is also cheap, computationally, which is especially beneficial in authorship attribution when linking identities on a large scale, for example, Wikipedia.

While MLkNN is an adaptation of an algorithm to assign multiple labels to a sample, the following methods transform the problem to achieve multi-label classification. That is, they transform a multi-label classification problem into a single-label classification problem.

The most straightforward of these methods is binary relevance (BR) [19]. Binary relevance trains a binary yes or no classifier for each label. While this method is straightforward, it serves as a baseline since many methods easily outperform it. Label powerset (LP) [19] for example, instead creates a single-label classifier with each possible combination of the labels as one of the new labels, which captures label dependencies. We take advantage of this method especially, because authorship attribution of collaborative writings does not only include authors appending their writing together, but also editing or co-writing each other's work.

Another problem transform method is Hierarchy Of Multi-label classifierS (HOMER) [18]. HOMER is a multi-label learning method that recursively breaks down the label set into smaller label sets creating a balanced tree structure where all but the leaves each represent a single multi-label classification problem. Another method is RANdom k-labELsets (RAkEL) [20]. The RAkEL algorithm randomly selects a  $k$ -sized subset of labels  $m$  times and trains an LP classifier on each. Each iteration yields a binary solution for each label in the  $k$ -sized subset of labels. The average decision for each label is calculated and the labels with averages above a certain threshold are considered positive. We attempt to use these methods, but they offer no noticeable accuracy improvement over the basic methods.

Madjarov et al. wrote an experiments paper with various multi-label learning algorithms and datasets [12]. These datasets included 6 datasets for text classification. While some datasets proved difficult, others were less so. One dataset involving classifying airplane problems from aviation safety reports yielded exact match accuracy of 81.6% and example-based accuracy, which measures the percentage of correctly predicted labels, of 91.4%. From this, we can see that, depending on the specific problem, multi-label learning can be very applicable to text.

Prior work in multi-label authorship attribution is limited to de-anonymizing academic submissions. Payer et al. proposed a framework called *deAnon* to break the anonymity of academic submissions [15]. Along with common features used in stylometry (e.g. bag-of-words, letter frequencies), they included information about which papers were cited. They use an ensemble of linear SVMs, a common classifier used in authorship attribution; MLkNN, a multi-label classifier; and ranking by average cosine similarity. From 1,405 possible authors, the ensemble classifier obtained a 39.7% accuracy that one of the authors was the first guess and 65.6% accuracy than an author is within the first ten guesses.

Our work differs from this for a few reasons. First, we leverage the clear ground truth of Wikia's revision history to set up controlled experiments. We also compare other proposed multi-label techniques described previously against ranking techniques. We extend our evaluation to include a sample of multi-label metrics. These differences lead us to obtain better results and demonstrate by comparison the results we would obtain

not only against ranking techniques but also against results on single-authored documents in our domain of interest.

It is not always the case that, when given a multi-authored document, we want to know the set of contributing authors. In some cases, we want to know which authors wrote which pieces. In this case, methods that break apart the document in question can be very useful. This has been achieved through a sliding window approach [8] and sentence level classification [2,11]. However, both of these techniques were developed for large texts, as opposed to the short texts typically found on the internet. So, while they may be applicable for collaboratively written books, they are poorly suited as-is for use on wiki-scale text.

### 3.2 Single-Author Stylometry

In the case in which we know all documents in our dataset have only a single-author, we formally define the problem of authorship attribution as follows: given a set of  $n$  authors  $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ , for each of whom we have a set of documents  $\mathcal{D}_i$  which we know to be written by that author, we want to identify the author of a document of unknown authorship  $d$ . This problem has been studied extensively [1,4,7,14] and we borrow feature sets and methods from prior work. Juola wrote an extensive review of authorship attribution literature[10]. Because of the high accuracies reported by many of these works, we would consider that multi-authored stylometry might be an application for which multi-label learning could be applied.

The Writeprints feature set [1] is a set of popular features used in authorship attribution. It includes lexical, syntactic, content, structural, and idiosyncratic features. We use a subset of these proposed features.

Linear support vector machines (SVM) are often used in stylometry for classification and produce a high precision and high recall [7] for this problem. Later studies, including [1], similarly found that linear SVMs were a good classifier for stylometry. For our single-label technique, we also use a linear SVM.

In the specific domain of Wikipedia, authorship identification is studied as a way to combat sockpuppets [16] and vandalism [9]. Sockpuppet detection, however, has been studied through the text and metadata on talk pages and not on the text of articles or text written collaboratively. While vandalism detection does study the style of specific edits in the article text, the goal is not to determine authorship, collaborative or otherwise.

## 4 Data

Our dataset was collected from the Star Wars Wiki, Wookieepedia, a Wikia site focused on Star Wars related topics<sup>3</sup>. The dataset was collected by crawling the wiki through the Special:AllPages page, going through the revision history of each article. Our dataset includes 359,685 revisions by over 20,000 authors distributed over 29,046 articles. However, many of those authors had fewer than 5000 words from first revisions, allowing us no more than 75 authors for single-authored only experiments, and fewer for experiments training on single-authored documents and testing on multi-authored documents. While this suspect set is too small to make claims of scalability, it does allow us to showcase the overall difficulty of the problem and overall ineffectiveness of the existing techniques.

We chose to use this dataset because it is the largest English language Wikia and has enough data to run controllable experiments with authentic collaborative documents. Additionally, it has the property that text is naturally organized into topics so we can control for the topic vocabulary, ensuring that we are classifying authorship and not just topic or subject of the writing. This dataset also contains articles of a range of sizes, from under 100 words to a few over 3,000 words. Most importantly, this dataset has clear ground truth in

---

<sup>3</sup> [http://starwars.wikia.com/wiki/Main\\_Page](http://starwars.wikia.com/wiki/Main_Page)

the form of revision histories. However, some of the potential problems from Wikipedia persist in Wikia, including the possibility of sockpuppets and the various writing guidelines, including rules of style.

For the *mixed collaboration* and *consistent collaboration* cases, we note that the number of potential suspects is actually much larger. This is because most collaborative groupings are rare, occurring only once or twice in the entire dataset, and therefore in order to have sufficient training data for any given author, many other authors need to be introduced into the dataset. As such, we do not have firm control over the number of suspect authors, but will make note of the number of suspects when presenting results. We also have limited ability to control for the number of training samples per author and so will also present the total number of training samples in the dataset. It is important to note that while the total number of suspects may be large, the number of actually significant suspects is closer to the 75 authors for which we had single-authored training data, and in some cases may be even less. This is because most authors only contribute to a few documents at the truncated level which we observe. These documents are used to boost the amount of training text and range of collaborative training groups available for the other authors. Due to lack of data and collaborative groupings for these rare authors, the chances of any given sample being attributed to them is unlikely, unless in combination with their collaborators.

We attribute the overall small number of principal authors to the wiki environment. In general, wikis have a few very active members and include many people who make occasional edits and corrections. Therefore, it is not surprising that most authors have very little data available.

#### **4.1 Training and Testing Data**

For experiments with single-authored documents, we collected data only from first revisions of articles to guarantee that documents have only a single-author. We gathered 5,000 words of text for each author, chunked into 500 word documents, appending articles as necessary. If text from an article would extend beyond 500 words, we truncated the article and discarded the remaining text so that cross-validation would not train and test on parts of the same article. We used chunks of 500 words because this is a value which has been found to work well in the past to balance number of documents and presence of style [5].

For multi-authored data, we chunked in the same manner as above, with the caveat that we controlled for the split of authorship in the multi-authored documents. We truncated the revision history as soon as we had sufficient authors for the experiment. We set thresholds for authorship based on the number of authors, and if the threshold was not met we took only the initial version as part of our single-authored dataset.

We also performed some experiments attributing pre-segmented samples of text. For this dataset, we determined authorship on the sentence level by locating the first revision in which the sentence appeared. We then took consecutive sentences by the same author as a sample, and restricted the dataset to samples between 100 and 400 words.

#### **4.2 Collaborative Examples**

In this subsection, we demonstrate the collaborative process on two short documents to increase understanding of the possible forms collaboration can take in this setting. We use colors to denote text originating in different revisions or revision sets. In the interest of conserving space, for each set of consecutive revisions by the same author we take only the last such revision.

In the Alpha Charge page edits, we can see that sometimes collaboration takes the form of editing and expanding. Notice that the first author wrote most of the text, but the second author changed the first word and expanded the end of the first sentence. Segmentation methods would be forced to lose information on the first sentence, because it is the work of two authors but can only be assigned to one.

In the Bark Mite page edits, we can observe a very different kind of collaboration. Here, notice that the first author wrote two sentences. The second author added some front matter, which would be placed in a table on the wiki to better define the subject of the page. The third author then adds a single long sentence to the end of the article, which makes up over half of the words in the article. This kind of collaboration is more receptive to segmentation, and a suitably powerful segmentation algorithm with sufficient data would lose little to no information.

### Example Revisions

#### Alpha Charge

This is an article stub with little special content.<sup>a</sup>

This is the first revision set by the first author.

The alpha charge was a discreet type of explosive, often used by Trever Flume. Alpha charges came in full, half, and quarter charge varieties, offering different blast strengths.

This is the final revision by a second author.

~~The~~An alpha charge was a discreet type of explosive, often used by Trever Flume due to the explosives lack of noise and smoke. Alpha charges came in full, half, and quarter charge varieties, offering different blast strengths.

#### Bark Mite

This is an article stub with a table as well as text.<sup>b</sup>

This is the first revision set by the first author.

Bark mites were arthropods on Rori and Endor. They ate bark, and made large hives in trees and caves.

This is the second revision set by a second author.

~~Arthropod~~ ~~Trees~~ ~~Bark~~ Bark mites were arthropods on Rori and Endor. They ate bark, and made large hives in trees and caves.

This is the final revision by a third author.

~~Arthropod~~ ~~Trees~~ ~~Bark~~ Bark mites were arthropods on Rori and Endor. They ate bark, and made large hives in trees and caves. ~~Bark mites appeared in the video game Star Wars Galaxies, a massively multiplayer online-role playing game developed by Sony and published by LucasArts, prior to its closure on December 15, 2011.~~

<sup>a</sup> [http://starwars.wikia.com/wiki/Alpha\\_charge](http://starwars.wikia.com/wiki/Alpha_charge)

<sup>b</sup> [http://starwars.wikia.com/wiki/Bark\\_mite](http://starwars.wikia.com/wiki/Bark_mite)

## 5 Methodology

For all evaluations for the multi-authored text, we use the Writeprints Limited feature set, extracted through JStylo [13]. We experimented with many different multi-label classifiers, and will only be presenting the best results. In addition, for all experiments with multi-authored testing documents we use a best-case scenario evaluation of a linear SVM which takes the top  $m$  predicted authors for a testing document written by  $m$  actual authors out of the set of  $n$  suspects. For real application, this would prove optimistic, since techniques would be needed to compensate for not knowing the exact number of authors.

For the evaluations of the pre-segmented data, we use a partial normalized version of the Writeprints feature set, also extracted through JStylo. We also re-extract features for the first revision dataset using this set to directly compare to the pre-segmented samples. Table 1 shows the number and type of features used for both feature sets.

**Table 1** Feature Sets

Feature type	Count (single-authored and multi-authored)	Count (pre-segmented)
Basic Counts	1 (characters)	2 (characters, words)
Average Characters per Word	1	1
Character Percentage	3 (digits, total, uppercase)	3 (digits, lowercase, uppercase)
Letter Frequency	26	26
Letter Bigram Frequency	$\leq 50$	$\leq 676$
Letter Trigram Frequency	$\leq 50$	$\leq 1000$
Digit Frequency	10	10
Digit Bigram Frequency	$\leq 100$	$\leq 100$
Digit Trigram Frequency	$\leq 1000$	$\leq 1000$
Word Length Frequency	variable	variable
Special Character, Punctuation Frequency	variable	variable
Function Word Frequency	$\leq 50$	$\leq 512$
Part of Speech Tag Frequency	$\leq 50$	$\leq 1000$
Part of Speech Bigram Frequency	$\leq 50$	$\leq 1000$
Part of Speech Trigram Frequency	$\leq 50$	$\leq 1000$
Word Frequency	$\leq 50$	$\leq 1000$
Word Bigram Frequency	$\leq 50$	$\leq 1000$
Word Trigram Frequency	$\leq 50$	$\leq 1000$
Misspelling Frequency	$\leq 50$	$\leq 1000$
Special Word Counts	0	3 (unique, large, used twice)

This table demonstrates the types and amounts of various features used in the two feature sets we use in this paper. Bigrams refer to sequential pairs, while trigrams are sequential triples.

## 5.1 Experimental Design

We begin by establishing the effectiveness of stylometry techniques in the Wikia domain on documents by single-authors. We do this by performing 5-fold cross-validation on our single-authored dataset. The purpose of this experiment is to establish a baseline of the performance of our techniques in this domain for solving the traditional authorship attribution problem.

For each variation we defined, we test both the single-label linear SVM and a wide range of multi-label classifiers. We evaluate *complete suspect set* and *partial suspect set* using a train-test technique. We had 60 authors for each experiment, with 9 single-authored training files each. For both of these experiments, the best multi-label classifier was a label powerset classifier with a linear SVM as the base classifier and a threshold of 0.5, so for all result analysis of these experiments we will examine this classifier as well as the standard linear SVM.

We evaluate *consistent collaboration* through 5-fold cross-validation. We evaluate *inconsistent collaboration* and *mixed collaboration* through a train-test technique, which we also use for *complete suspect*



*set* and *partial suspect set*. For the training data for the *inconsistent collaboration* and *mixed collaboration* cases, we use the *copy transformation* in which each document is counted once for each label (author) to whom it belongs [17]. For *consistent collaboration* and *mixed collaboration*, the same label powerset classifier with linear SVM base and threshold of 0.5 was the best multi-label classifier. However, for *inconsistent collaboration* a binary relevance classifier with naive bayes (NB) base classifier was the best multi-label classifier.

For *mixed collaboration*, we have on average 3.7 training collaborative groups per author, each with on average 3.4 training documents. On average, 5% of the test documents have author groups distinct from those in the training.

Additionally, we have experiments on pre-segmented data. Here, we use a linear SVM as our classifier, and perform cross-validation experiments. We adapt a technique proposed by Overdorf and Greenstadt for tweets and reddit comments and by Dauber et al. for pre-segmented source code to perform account attribution, as well as performing simple attribution of individual samples [6,14]. For account attribution, we average the SVM output probabilities for the samples belonging to the account in order to attribute the samples as a group. We experiment with account sizes of 2, 5, and 10 samples. We perform experiments with 10 training samples per author, ranging from 10 to 50 authors, for each. We also experiment with the effect of adding more training samples, and perform experiments using an account size of 10 with both 20 and 30 training samples.

## 5.2 Evaluation Metrics

In the multi-label classification case, simple accuracy as a metric does not give sufficient information to understand the performance of the classifier. Traditional accuracy corresponds to an "exact match" of guessed and correct authors. Indeed, this metric has been proposed and tested in the case of academic papers under the name *guess-all*. In multi-label machine learning literature, *guess-all* is referred to as *subset accuracy* [19]. A broader metric, *guess-one*, measures the frequency with which we correctly predict any author of the document in question. However, *guess one* does not exactly match to any multi-label learning metric, so while we consider subset accuracy, we do not use *guess one*.

Subset accuracy is considered ineffective at portraying the actual success of the classifier due to ignoring the complexities of how a classification can be partially correct in multi-label learning [15]. Therefore, we also consider *example-based accuracy* (EBA), which describes the average correctness of the label assignments per example. It is calculated by taking the average of the number of correctly predicted labels divided by the total number of actual and predicted labels per example. This shows how many authors we have correctly predicted on average per example. In real-world applications, both subset accuracy and EBA have value in determining the believability of the predictions of our classifiers.

Finally, in order to compare directly between our linear SVM and multi-label techniques, we calculate a version of EBA for our linear SVM which considers the top  $m$  ranked authors as predicted labels. As a result, for the SVM each two-authored document will have an accuracy contribution of 0,  $\frac{1}{3}$ , or 1. In the more general case, the accuracy contribution for partially correct attributions ranges from  $\frac{1}{2m-1}$  when only one of our selected labels is correct to  $\frac{m-1}{m+1}$  when we only select one incorrect label. For a multi-label classifier with  $n$  labels, the accuracy contribution of each document for which we were partially correct can range from  $\frac{1}{n}$  when we choose all incorrect labels and one of the correct labels to  $\frac{m}{m+1}$  when we select all of the correct labels as well as an additional label.

## 6 Results

### 6.1 Single-Authored Baseline

In order to set a baseline and to form a context for multi-author stylometry, single-authored documents in the same domain must be analyzed. With traditional methods used in other single-authored stylometry problems, we analyze *first edits* of a Wikia page, guaranteeing a single author wrote all of the text. With a SVM classifier and Writeprints Limited feature set, described in Section 5, 5-fold cross validation achieved an accuracy of 51.3% with 10 authors and 14.2% with 75 authors. Note that accuracy here is number of correct classifications over the total number of classifications, so it is most similar to subset accuracy in that a correctly classified instance is completely, and not partially, correct.

We notice that even in these purely single-author results, our accuracies are lower than those reported in other literature [1,4]. We believe that this is in part due to the rules of style adhered to by Wikia editors. To some extent, Wikia authors attempt to mutually imitate each other in order to have an encyclopedic tone and unified style.

### 6.2 Non-Collaborative Training Documents

For *complete suspect set* and *partial suspect set*, we ran experiments using 60 authors with 9 single-authored first edit training documents per author. We used the same 60 authors for both problems, with different test instances, and experimented ranging from 2-authored documents to 4-authored documents. For *complete suspect set*, all test instances only had authors from within the suspect set, and for *partial suspect set* all test instances had at least one author in the suspect set and at least one author outside the suspect set. That means that for the 2-authored documents EBA and subset accuracy are identical for *partial suspect set*.

Figure 1 shows the results of our experiments for these problems. We do not show the subset accuracy results for *complete suspect set*. This is because we only have non-zero subset accuracy for 2-authored documents for this case. The linear SVM taking the top two authors had subset accuracy of 4.3% and the binary relevance classifier with naive bayes as the base had subset accuracy of 1.5%. Along with the low EBA results, which cap at 23.2% for label powerset with a linear SVM base and 21.7% for a linear SVM taking the top two authors and get worse as the number of authors increase, this shows that predicting the authors of a collaboratively written document from singularly written documents is not practical.

The fact that the EBA results for *partial suspect set* are similar to the results for *complete suspect set* suggests that in the general case these problems aren't very different. The notable difference comes from subset accuracy, due to the fact that for *partial suspect set* some samples reduce to identifying if a suspect author is one of the authors of the document. We show that while this still is a hard problem, it is easier than identifying the set of authors of a collaboratively written document from singularly written documents. The other notable trend in the results is that as we add more authors to the testing document, accuracy decreases. This suggests that single authored training documents are less effective the further the testing document gets from being single authored.

### 6.3 Consistent Collaboration

In Figure 2, we examine the results of the *consistent collaboration* experiments. We note that the number of suspects is not held constant here, and neither is the number of overlapping suspect groups. However, we can make some observations by comparing to the results from purely single authored results. We can note we have far better accuracy on consistent collaboration pairs than purely single authored documents with comparable numbers of suspects, and that the magnitude of the difference increases as we have more

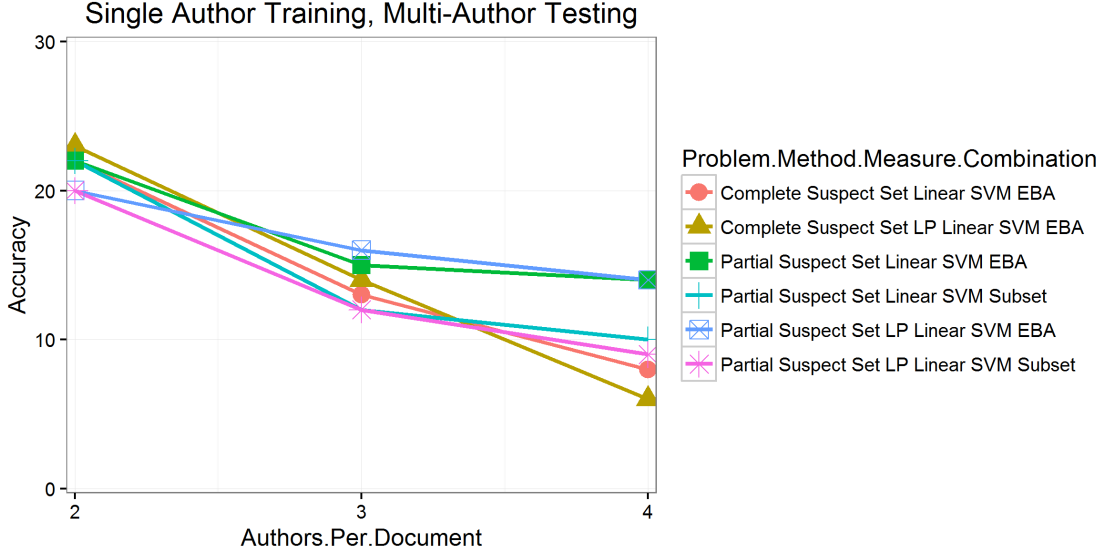


Fig. 1: This graph shows the results of training on single authored documents and testing on multi-authored documents, with the number of collaborators per test document on the x-axis and accuracy on the y-axis. There are a constant 60 suspect authors with 9 training documents each. Linear SVM attributions were performed by taking the top  $x$  authors for a document by  $x$  authors.

authors collaborating on the document. The two primary factors which could account for this are the number of collaborators and the amount of overlap between collaborative groups, which decreases in our dataset as we increase the number of collaborators. While these results are not conclusive due to lack of data, they suggest that *consistent collaboration* is one subproblem of multi-authored authorship attribution which current tools can deal with.

One likely explanation for these observations is that collaborators' styles blend into the overall style of the document. As a result, collaboration groups would have a more distinct style than individuals, and as the groups grow they become more distinctive. Another is that as collaboration groups grow, the percentage contribution by any one member decreases, reducing the influence of overlapping members and of more difficult to attribute members. While it would take more evaluation on more datasets to confirm these hypotheses, they would explain these observations, and if true would mean that this particular subproblem is generally easy among authorship attribution tasks, which presents a significant privacy risk to any people who have frequent collaborators in both the public and anonymous spaces.

#### 6.4 Mixed and Inconsistent Collaboration

Figure 3 shows the results of both the *mixed collaboration* and *inconsistent collaboration* cases. We note that the number of suspects and amount of training data are not held constant here. However, we can still make some important observations. The primary observation is that, regardless of the changes in the number of suspects or the number of collaborators per document, EBA for *mixed collaboration* is higher or approximately equal to EBA for *inconsistent collaboration*, which is greater than or approximately equal to subset accuracy for *mixed collaboration*. Subset accuracy for *inconsistent collaboration* is not shown because it is

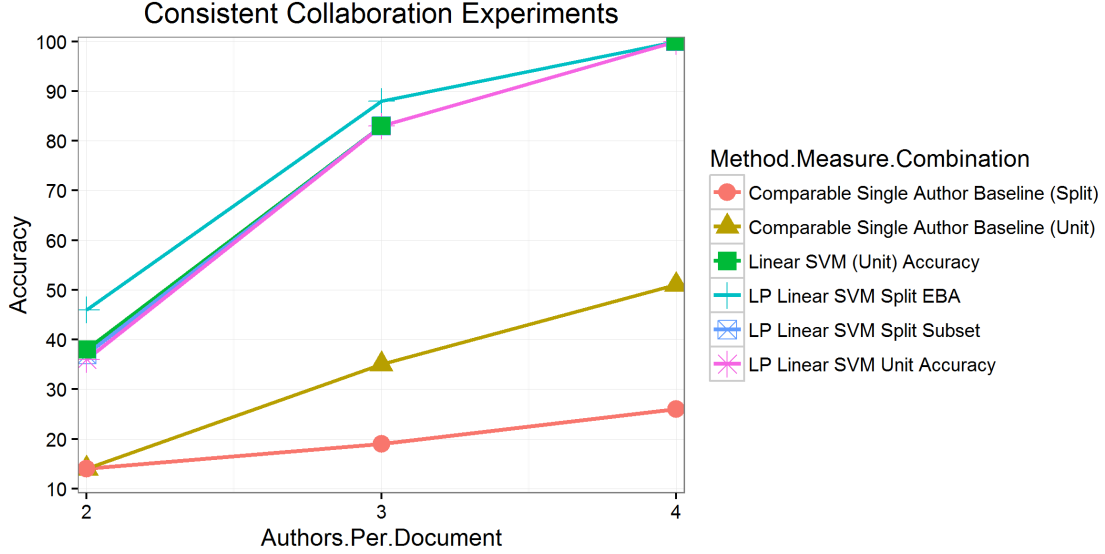


Fig. 2: This graph shows the results of the consistent collaboration experiments. The difference between Split and Unit is that Unit investigates the group of authors exclusively as a set, while Split investigates the group of authors as separate entities. For 2-authored documents, we had 400 documents from 116 authors in 134 pairs. For 3-authored documents, we had 58 documents from 49 authors in 22 triples. For 4-authored documents, we had 20 documents by 28 authors in 8 groups. Beyond that, we had too little data to continue. Additionally, we show the accuracies for the closest size suspect set to both Split and Unit cases from the single authored experiments for comparison purposes. For 2-authored documents, both of those are 75 suspects. For 3-authored documents, this is 50 suspects for Split and 20 suspects for Unit. For 4-authored documents, this is 30 suspects for Split and 10 suspects for Unit.

only non-zero for the linear SVM at 2-authors per document and 3-authors per document, and for each of those it is 1.5%.

This trend is not surprising, given two basic facts. First, EBA is a much easier metric than subset accuracy, as discussed in Section 5. Secondly, *inconsistent collaboration* is a strictly harder special case of *mixed collaboration*. More interesting is the fact that the best performing multi-label classifier for *inconsistent collaboration* was a binary relevance classifier based on naive bayes, while for all other experiments it was the label powerset classifier based on the linear SVM. Combined with the results from *consistent collaboration*, this suggests a reason why multi-label classification does not work well in the general case for authorship attribution.

Label powerset is a classifier which attempts to treat combinations of labels as single labels in order to make the multi-label learning problem into a single-label learning problem. In contrast, binary relevance transforms the multi-label problem into a binary classification problem for each label. The fact that normally label powerset works better, and that *consistent collaboration* seems to work well, suggests that for stylometric authorship attribution the combination of authors causes a shift in features distinctive to the combination, which can no longer be easily linked back to the styles of the original authors individually by traditional techniques. Therefore, when training data is lacking for combinations of authors, as occurs somewhat for

*mixed collaboration* and completely for *inconsistent collaboration*, we are either left with a less well-trained label powerset classifier or forced to fall back on an ineffective binary relevance classifier. This also shows why training on single-authored documents and testing on multi-authored documents works poorly, since that is a similar process to that of binary relevance, without the benefit of having training documents with input from other authors.

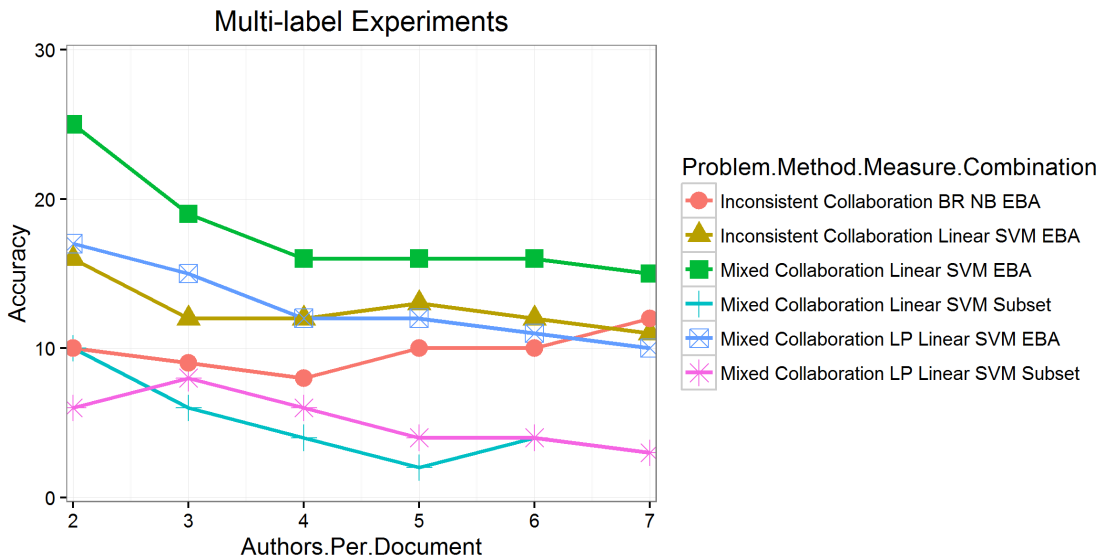


Fig. 3: This graph shows the results of the *mixed collaboration* and *inconsistent collaboration* experiments. For all experiments, there are many suspect authors serving as distractors with only a couple of training instances, due to the small number of occurrences for most collaborative groupings. For 2-authored documents, we had over 360 training instances and over 360 suspect authors. For 3-authored documents we had over 320 training instances and over 470 suspect authors. For 4-authored documents, we had about 360 training instances and over 630 suspect authors. For 5-authored documents, we had over 420 training instances and over 840 suspect authors. For 6-authored documents, we had over 470 training instances and over 1030 suspect authors. For 7-authored documents, we had over 500 training instances and over 1200 suspect authors. Due to lack of training data, most of these suspects have little impact.

## 6.5 Authorship Attribution of Pre-segmented Text Samples

Figure 4 shows the results of the experiments with pre-segmented text samples. Not shown in the graph is the result of a single experiment with accounts of 10 samples and 10 suspect authors with 90 training samples each, which had accuracy of 63.6%. Along with the results in the graph, we can conclude that, like shown in [6] with source code, both the number of training samples and the number of samples in the account to be attributed are important to increasing accuracy. Unlike the work with source code, which showed a relatively modest number of samples needed to reach high accuracy, in this work we show that we would need more samples than are present in our dataset to reach high accuracy. However, we do show that we can surpass the

base accuracy for standard stylometric chunks with at least 20 training samples and 10 account samples to attribute.

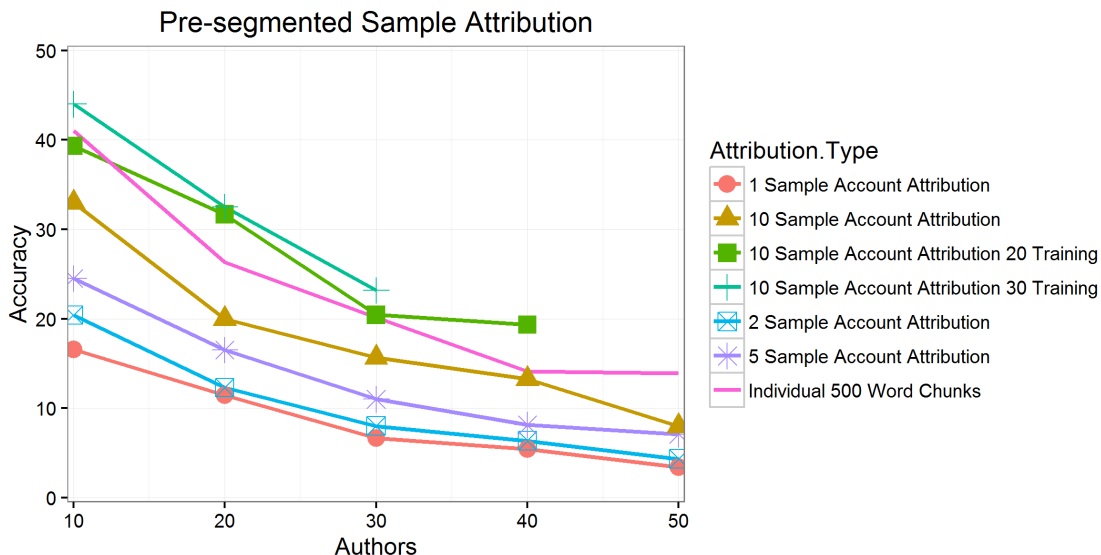


Fig. 4: This graph shows the results of experiments on pre-segmented text samples, with a comparison to traditional chunking performed on our single author first edit dataset. The samples were identified on a per-sentence basis, with a sample consisting of a set of consecutive sentences originating from the same author. Samples used for experimentation were between 100 and 400 words with normalized features, and for consistency we used the same feature extraction process for our comparison chunks. The line labeled Individual 500 Word Chunks is this comparison experiment, and uses 10-fold cross-validation with 10 chunks per author. The experiments labeled with 20 and 30 training samples were performed with 3-fold and 4-fold cross-validation respectively, and end early due to lack of authors with sufficient data. The remaining experiments were performed with 2-fold cross-validation and 10 training samples.

## 7 Discussion

For the *consistent collaboration* case, we notice that the subset accuracy and example-based accuracy of the multi-label techniques are similar. The fact that example-based accuracy is somewhat higher than subset accuracy here suggests that some, but not many, of the mis-classifications are between overlapping groups. We note that while this case is multi-authored, the authors occur in repeating groups, making it closer to a single-authored case overall.

However, in the other cases, example-based accuracy is clearly better than subset accuracy for all approaches. This indicates that once we lose the similarity to single-author stylometry, it becomes noticeably harder to make the exact correct predictions than to make partially correct predictions, just as in other applications of multi-label learning.

Applications which are single-author, or are multi-author but reducible to single-label problems, are best handled with an SVM. Applications which are purely multi-author are best handled with a multi-label classifier. However, for any multi-author problem, it is essential to have multi-authored training data. While we can obtain some correct predictions from multi-authored documents in which the combinations of authors in the training documents does not reflect the combinations of the documents we want to evaluate, if we do not know that the authors we are interested in only collaborate with certain people it is best to have as wide a range of collaborative groups as possible.

In our experiments, label powerset was the best multi-label classifier. Combined with our results between subproblems, we hypothesize that, stylometrically speaking, collaboration causes a new style to emerge distinct to the set of collaborators. This would mean that in order to achieve good results using these multi-label classifiers we would need sufficient training data for all combinations. In other words, this hypothesis would mean that of the five problems we have defined, only *consistent collaboration* can yield good results in typical use. However, our own experience shows that it can be difficult to gather sufficient data to enable this special case.

Based on our experiments, we believe a defense like the one proposed in [3] to be an effective tool against the current limitations of stylometry. Because their defense relies on crowdsourcing, rather than using contacts, they avoid both the *complete suspect set* and *consistent collaboration* cases. If the crowdsourcing participants rarely rewrite or their rewrites are difficult to identify for training data, then this defense forces the *partial suspect set* case. If the crowdsourcing participants rewrite often, and their rewrites can be identified, then this defense allows no better than the *mixed collaboration* case, and if used sparingly either forces linkability as explored in the original paper or the *inconsistent collaboration* case. Because each of those cases yield poor accuracy, it is unlikely that an analyst would be able to achieve a confident attribution against this defense. However, we stress that this is based on current limitations, and future breakthroughs may still show that this defense is insufficient.

## 8 Future Work

We are interested in combining single-authored and multi-authored documents in both the training and test sets. In doing this, we hope to determine if we can lessen the burden of acquiring training data while expanding the range of document authorship sets which can be correctly classified.

Our dataset is small, so we also would like to evaluate on a larger dataset, potentially gathered from Wikipedia. Our current results suggest that scalability might be a greater problem for stylometry on collaborative documents than for conventional stylometry. More importantly, we wish to determine if there is a point at which training an SVM on combinations of authors becomes computationally impractical, even if the training data was present.

While wiki articles are one type of collaborative document, they are not the only one. We would like to extend our evaluation to other collaborative domains, including more traditional text and source code. While source code has an easy collection from GitHub<sup>4</sup>, it is difficult to find a data set for traditional text with clear ground truth outside of the wiki setting. This is especially important since our single-authored baseline results are so poor, so we hope to find a collaborative data source which is easier to attribute in the base case.

We are also interested in investigating other multi-label learning techniques and different parameterizations. It is likely that optimizing the parameters and finding the optimal algorithm will greatly improve the results in the multi-label case. It is also possible that doing so will improve the single-label results and be

---

<sup>4</sup> <https://github.com/>

able to better compensate for non-representative training data, such as only having single-author training documents or only having collaborative groups which do not occur in the data of interest.

Additionally, we are interested in investigating the effects of changing our feature set, both by admitting more n-gram features and by pruning based on heuristics such as information gain. We would also like to experiment with different chunk sizes and amounts of training text, to determine if it is necessary to include more information to find authors' styles in the multi-author case.

Because we have identified a potential cause for the difficulty of multi-label stylometric attribution, we would like to further investigate to see if we can find a method which works around the issues we have identified. Alternately, we would like to find a way to perform supervised segmentation on documents of this small scale.

## 9 Conclusion

Collaborative documents require a different way of thinking about stylometry. With single-authored documents, the privacy concern comes from analysts collecting a corpus of the author's text and comparing a sensitive document to them. With collaboratively written documents, current techniques require the analyst to collect a corpus of documents written by collaborative groups.

We show that with sufficient training data, the *consistent collaboration* case is the only case in which multi-label stylometry is viable using currently available techniques. We also show that even in other cases, the multi-label learning algorithm which attempts to perform the same transformation, label powerset, performs the best as long as there is some data for the combination of authors. Because of this, we hypothesize that the feature values of collaborations are distinguishable by collaborative group, rather than by member of the group. From a theoretical standpoint, that would mean that label powerset is the correct type of multi-label classifier for the problem. However, in practice it is rare that sufficient data exists for training on all possible collaborative groups of interest. We conclude that this is the greatest difficulty for the application of conventional multi-label learning to stylometry.

Prior work has suggested that collaboration may provide an effective defense against stylometry. While we are not ready to conclude that stylometric attribution of small collaborative documents without training data written by the same collaborative group is impossible, it is clearly a much harder problem than conventional stylometry and requires the development of new techniques. Therefore, those seeking temporary anonymity may find it safer to have people with whom they have never written another publicly available document collaborate with them.

We also investigate the viability of performing segmentation in these situations. We show from the structure of the collaboration that while in some cases authors work on distinct sections of the document, in others authors work not only in the same section but on the same sentences. Therefore, while segmentation may work well in some cases, there are others for which it is difficult to fully capture the collaborative nature of the document with segmentation techniques. We also present results from attempts to attribute pre-segmented text. We demonstrate that, while it is harder to attribute individual segments than it is to perform traditional document attribution, once there are sufficient training and evaluation samples it is possible to attribute an account of such samples. Between these, we believe that supervised segmentation methods, especially with overlapping segments, may allow for a reasonable attribution in some cases, with the caveat that some information may be lost and that they need to be tailored to smaller text sizes than current unsupervised methods require.



## References

1. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)* 26(2), 7 (2008)
2. Akiva, N., Koppel, M.: A generic unsupervised method for decomposing multi-author documents. *Journal of the American Society for Information Science and Technology* 64(11), 2256–2264 (2013)
3. Almishari, M., Oguz, E., Tsudik, G.: Fighting authorship linkability with crowdsourcing. In: *Proceedings of the second edition of the ACM conference on Online social networks*. pp. 69–82. ACM (2014)
4. Brennan, M., Afroz, S., Greenstadt, R.: Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)* 15(3), 12 (2012)
5. Corney, M.W., Anderson, A.M., Mohay, G.M., de Vel, O.: Identifying the authors of suspect email. *Computers and Security* (2001)
6. Dauber, E., Caliskan, A., Harang, R., Greenstadt, R.: Git blame who?: Stylistic authorship attribution of small, incomplete source code fragments. *arXiv preprint arXiv:1701.05681* (2017)
7. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. *Applied intelligence* 19(1-2), 109–123 (2003)
8. Fifield, D., Follan, T., Lunde, E.: Unsupervised authorship attribution. *arXiv preprint arXiv:1503.07613* (2015)
9. Harpalani, M., Hart, M., Singh, S., Johnson, R., Choi, Y.: Language of vandalism: Improving wikipedia vandalism detection via stylometric analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. pp. 83–88. Association for Computational Linguistics (2011)
10. Juola, P., et al.: Authorship attribution. *Foundations and Trends® in Information Retrieval* 1(3), 233–334 (2008)
11. Koppel, M., Akiva, N., Dershowitz, I., Dershowitz, N.: Unsupervised decomposition of a document into authorial components. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. pp. 1356–1364. Association for Computational Linguistics (2011)
12. Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S.: An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition* 45(9), 3084–3104 (2012)
13. McDonald, A.W., Afroz, S., Caliskan, A., Stolerman, A., Greenstadt, R.: Use fewer instances of the letter "i": Toward writing style anonymization. In: *Privacy Enhancing Technologies*. pp. 299–318. Springer (2012)
14. Overdorf, R., Greenstadt, R.: Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *PoPETs* 2016(3), 155–171 (2016)
15. Payer, M., Huang, L., Gong, N.Z., Borgolte, K., Frank, M.: What you submit is who you are: A multi-modal approach for deanonymizing scientific publications. *IEEE Transactions on Information Forensics and Security* (2015)
16. Solorio, T., Hasan, R., Mizan, M.: Sockpuppet detection in wikipedia: A corpus of real-world deceptive writing for linking identities. *arXiv preprint arXiv:1310.6772* (2013)
17. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. *Int Journal of Data Warehousing and Mining* 3(3), 13 (2007)
18. Tsoumakas, G., Katakis, I., Vlahavas, I.: Effective and efficient multilabel classification in domains with large number of labels. In: *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*. pp. 30–44 (2008)
19. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: *Data mining and knowledge discovery handbook*, pp. 667–685. Springer (2010)
20. Tsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multilabel classification. *Knowledge and Data Engineering, IEEE Transactions on* 23(7), 1079–1089 (2011)
21. Zhang, M.L., Zhou, Z.H.: MI-knn: A lazy learning approach to multi-label learning. *Pattern recognition* 40(7), 2038–2048 (2007)