Testing the Robustness of Authorship Attribution Algorithms

An honors thesis for the Department of Computer Science

Shirish Pokharel

Tufts University, 2015

Table of Contents

Introduction

Juola (2007) defines authorship attribution as the science of inferring characteristics of the authors from the characteristics of documents written by that author. The simplest kind of authorship attribution problem is the one in which we are given a small, closed set of candidate authors and are asked to attribute an anonymous text to one of them. Usually, it is assumed that we have copious quantities of text by each candidate author and the anonymous text is reasonably long.

The field is also called ‘authorship profiling’, or ‘stylometry’. The field itself predates modern computers – attempts to use word frequency to fingerprint texts have been around since the late nineteenth century – but computers were used to ‘fingerprint’ documents most popularly in the 1960’s, to identify the author of certain unattributed Federalist Papers (to be discussed later). From the late 60’s to mid-nineties, there was certain optimism about the possibility of discovering latent properties of texts that are invisible to humans. In the early 90’s, proofs from document ‘fingerprint’ software were even admissible in certain courts in the US.

The optimism was waning though—lack of the serious computing power to work with hundreds to thousands of features led to stagnation in the field. Rudman (1998) criticized the authorship attribution studies saying: “*Non-traditional authorship attribution studies—those employing the computer, statistics, and stylistics, had enough time to pass through any “shake-down” phase and enter once marked by scientific, and steadily progressing studies. But after 30 years and 300 publications, they have not.*”

Since early 2000's, the field has seen a resurgence, thanks to the growing power of processors, and also to Juola's 2004 (Juola 2004, 2006) competition on authorship attribution; Juola also brought increasing interest to the field. The purpose of the competition organized by Juola, was to establish a collection of best practices in the field. Stamatatos' survey paper (Stamatatos 2009) showed the breadth and depth of approaches researchers had been using since the early 2000's.

To avoid confusion, we will use LDA-A (Latent dirichlet allocation— algorithm) to refer to the algorithm as developed by Blei et al. (2003) and LDA-M to mean the technique Linear Discriminant Analysis— Method (McLachlan, 2004).

Historical Overview

The issue of trying to identify the characteristics of an author from the language she uses has been around since the days of the Old Testament(Judges 12:5-6, Old testament), where the Gileadites identified a single salient feature of Ephraimite speech — the inability to pronounce a particular sound — that could be used to divide the speech of one group from another.
The arrival of modern statistics made more sophisticated analysis possible, and modern computers and corpora have made investigating such questions algorithmically possible. With the advent of corpus linguistics, authorship attribution has become popular and productive area of research.

The approaches taken to authorship attribution have been continuously evolving. Statistical analysis of published works by Holmes (1992) had suggested that average word length might be a distinguishing characteristic of writers. Later studies showed that average word length was neither stable within a single author, nor did it distinguish between authors. Since then, other statistics have been suggested and discarded, such as average sentence length, average word length, average number of syllables per word, distribution of parts of speech, and type/token ratios. Measures of ‘vocabulary richness’ such as such as Simpson’s D index, Yule’s “characteristic K” and Honore’s ‘R’ were considered as serious techniques in extracting authorial profiles till the mid-nineties, when increased corpus size and processing power proved that they too were not particularly usable in identifying authors (Miranda-Garcia et al, 2005). The underlying theory behind such approaches is that an authorial ‘fingerprint’ can be extracted as a summary statistic from a given text, and that different authors will vary, noticeably and consistently, along this statistic.

The Federalist papers are a set of newspaper essays published between 1787 and 1788 by an anonymous author named “Publius” in favor of the ratification of the newly proposed Constitution of the United States. It has since become known that “Publius” was a pseudonym for a group of three authors: John Jay, Alexander Hamilton and James Madison. It has since been generally accepted that of the 85 essays, Jay wrote five, Madison wrote 15 and Hamilton wrote 51, with 3 essays written jointly by Madison and hamilton. The other 12 essays, the famous “disputed essays” have been attributed to both Madison and Hamilton.

Modern analysis has almost unanimously assigned authorship of disputed essays to Madison on basis of traditional historical methods. This was made purely on the basis of statistically inferred probabilities and Bayesian analysis. The case was one of the bigger successes of stylometric analysis. The work, undertaken by Mosteller and Wallace in the nineties (Argamon et al., 2005) suggested that a small number of most frequent words in a language (‘function words’) could usefully serve as indicators of authorial style.

A noted episode of failure of authorship attribution is the case of ‘A Funeral Elegy’ in the 1990s. Don Foster found (Foster, 2001), using stylometric analysis, that the poem’s author ‘W.S.’ was in fact William Shakespeare. Traditional Shakespearean scholars reacted with scorn and disbelief on the basis of traditional objections such as writing style and content. They argued that it was unthinkable that “the supreme master of the language,” at the height of his career, could have written a work of “unrelieved banality of thought and expression, lacking a single memorable phrase in its 578 lines”. Later applications of stylometry by other scholars uncovered other evidence contesting Shakespearean attribution and suggesting errors in Foster’s analysis of the *Elegy*. Their analysis was found satisfactory by all parties and the work was attributed to John Ford. By 2002, Foster himself had come to accept the attribution.

The field has seen more successes in recent years however. Patrick Juola’s 2013 success (Kolowich "The Professor Who Declared, It's JK Rowling") in unmasking the nom-de-plume of the author behind the book *The Cuckoo’s Calling* is particularly notable. Using the text of the book, he used four features of the text: word-length distribution, character 4-grams, recurring word pairings, and the use of ‘function words’ (or stop words, a list of commonly-used words) (Wilbur et al., 1992) and compared them against books by other popular authors to distinguish J.K Rowling as the secret author of the book. This brought him an instant attention from popular media. J.K Rowling was soon revealed as the identity by the publishing house soon after, and the book has been published under her name since.

Interestingly the same success story is also the tale of limitations inherent in the field. First, it is a bane of the field that many attribution systems are heavily dependent on the training corpus, considerably limiting the range of their potential usage. In this case, Juola did not use Rowling’s more popular *Harry Potter* series as training data but *The Casual Vacancy,* Rowling’s other post-Potter novel. Additionally, the results were not unequivocal: the case was stronger for

Rowling than for three other contemporary authors. The primary task had not been author identification, but confirmation (of data acquired through more conventional means), and stylometric techniques had been successful at that.

Coulthard (2013), in writing for Journal of Law & Policy suggests that unlike forensic phoneticians, forensic linguists are never going to have reliable population statistics to enable them to talk about “the rarity of particular linguistic features”. He argues though that recent works have opened the way to derive reliable and usable data about individual linguistic usage that can be applied in cases of authorship attribution. With modern tools, linguists can make statements about frequency and likelihood of occurrence and provide rigorous probability statistics. He suggests doing so would increase the admissibility of linguistic evidence in the legal courts of justice.

Juola (2007) explains the usefulness of authorship attribution in politics, journalism and law. He argues that automatic and objective inference of authorship is now possible, thanks to improved statistical techniques and availability of computer-accessible corpora. He sees this as an important additional tool for investigative journalism and scientific analysis of documents and close readings of documents, which have traditionally given good results. Papers on authorship attribution routinely appear at conference and papers ranging from linguistics and literature through machine learning and computation, to law and forensics. Despite — or perhaps because of — this interest, the field itself is somewhat in disarray with little overall sense of best practices and techniques

Justification for Analysis

Several issues inherent in authorship attribution algorithms make this paper relevant. As has been mentioned earlier, many such algorithms are highly sensitive to the corpus used. This significantly limits the breadth of potential usage of such algorithms. Moreover, even within the same content domain (e.g. news articles), various formats of expression seem to respond differently to different approaches.

Finally, most of the existing work has been done with very large sample documents to train with, and a limited number of suspect authors. While more recent research has explored (more on that in literature review) into authorship attribution with limited training data and large number of authors, there is certainly room for more work to be done.

This project has a reasonably constrained domain. We wish to test authorship attribution algorithms on medium-length (300-1200 words) newspaper articles. We limit our training set to 6000-12000 words per author, so as to maintain a reasonably good balance between precision and recall. The number of candidate authors is in hundreds.

For this project, we have considered only those newspaper articles written by a single author. We consider newspaper articles as examples of ‘adversarial’ pieces of writing. Each newspaper article is likely to have gone through several levels of editors, who could have removed features that could be used to identify the original author and added features different from the rest of the article. Every newspaper also has its own style, to which all articles published in it must conform. The imposition of style is a barrier that we believe likely removes a lot of identifiable stylistic features. Several authorship attribution strategies use characteristics such as misspellings, positioning and style of punctuation, and use of certain spellings over others. Such methods are unlikely to be useful in identifying newspaper articles.

The works of research that have been done with limited training data and large number of authors have often used ‘artificial’ text samples. For example, for their 2004 paper, Argamon et al. (2004) used 25-minute-long ‘stream of consciousness’ essays from 600 college students across 5 countries. They also regularly use often ‘personalized’ writings — such as blog entries, movie reviews, personal essays, and so on — to train their samples. While in essence a perfect algorithm would work equally well regardless of how the pieces of writings are written, different feature extraction methods do make assumptions about the text to rightly adjust parameters for machine learning algorithms.

I propose that such assumptions are different for text written with a personal voice, and those with a more ‘generalized’ voice. While I do not test this feature independently, and it is up to future research to test the claim, this would likely be an important factor during the analysis. If existing algorithms that perform relatively well fail on newspaper articles, this could be one of the explanations for such a failure.

In addition, I also do a cross-cultural analysis by controlling for the language (English), but taking newspaper articles from different countries. This is important because authors like Argamon and Stamatatos have claimed the relative advantages of using features like function words (stop words). With this analysis, I test (as part of larger test), the veracity of such claims within English but across different cultural context. Were this to vary considerably across cultures, this feature would not be stable across the English language itself and thus would be a less useful parameter in authorship attribution studies. On the other hand, if it were stable within the same cultural context for all cultures, but different between different cultures, this could be an important finding. It would mean such relatively computationally cheap strategy could be used to establish the cultural background of the author of a given piece, and would help in authorship attribution in intercultural context. This could be generalized to any algorithm that depends on some kind of feature extraction from a given text.

As we discuss data and methods, we need be cautious in the interpretation of the results of the experiments, which may be obvious but needs to be stated nonetheless: this is not an exhaustive study and comparison of cross-cultural and cross-contextual modalities authorship attribution. Since I test three different algorithms in articles from nine different sources from eight countries, with a total of ~3000 authors, there are a many variables that have not been controlled for. Thus, this study should be taken as just a stepping stone, an exploratory research that will mainly raise questions that, when answered, will offer solutions to issues in the field, rather than exhaustively answering any particular question.

The difficulty of authorship attribution to datasets extending to thousands of authors is particularly difficult—and thus has a potential for great rewards. While Argamon et al. (2006) have claimed to reach an accuracy of 46% with more than 1,000 authors, the data has been not made public, and the results have not been replicated. Zukerman et al. (2011) have used Latent Dirichlet Allocation (LDA-A) with Gibbs sampling to reach an average accuracy of 10% with 2,000 authors, and that they consider being 'state-of-the-art'. It is possible that we may have reached a plateau in gaining authorial information from limited text. It is also quite likely, though, that a combination of existing methods could gain improvements. In this context, using text with authors from wide variety of geographical and social backgrounds globally could lead to a better understanding of the workings of existing algorithms.

Literature Review

This portion will do a general overview of recent literature in authorship attribution. The field as a general would be too broad to do an exhaustive overview of, so it further zooms into the kind of literature relevant to the parameters set by the project: limited training data, large number of authors, and adversarial-type data.

The reviewed papers are particularly significant and relevant for this paper for inclusion of the following parameters following reasons: the choice of machine learning/modelling algorithms used, the number of, and types of features extracted for the analysis, the nature of the dataset involved, and the number of authors/posts per authors.

Stamatatos (2009) does a comprehensive survey of modern authorship attribution methods. He identifies four distinct authorship analysis tasks co-existing with authorship attribution: Authorship Verification (deciding if a text was written by an author or not), plagiarism detection (finding text similarity), authorship profiling or characterization (extracting information of the author of a given text), and detection of stylistic inconsistencies (as may happen in collaborative writing).

He presents a comprehensive review of the approaches to quantify the writing style. These begin with lexical features, such as token based (word length, sentence length, etc.), vocabulary richness, word frequencies, word n-grams, and errors. Then there are character-level features such as character types, fixed-length character n-grams, variable-length character n-grams, and compression methods (where a compression algorithm with the training document as the dictionary is used to compress a target document, and the size of the resulting file is taken as the ‘distance measure’). He then describes syntactic features of text, such as Part-of-speech, chunks, sentence and phrase structure, rewrite rules frequencies, and syntactic errors. The semantic features of text include synonyms, semantic dependencies, and functional words (or stop words). Finally, there are the application specific properties of text, that include information about the structure of the text that can be used to identify it, content-specific information that needs specialized dictionaries, and language-specific properties of text.

Stamatatos also discusses the various methods of feature selection for authorship attribution that have been discussed in the literature. Extracting features of a text can considerably increase the dimensionality of the feature set. In such cases, feature selection algorithms are applied to reduce the dimensionality of the representation: this helps avoid over fitting on the training data. However, such feature selection can be misguiding: the best features may strongly correlate with one of the authors due to content-specific rather than stylistic choices. In such a case, the features identified by feature selection algorithms may be too corpus-dependent with little general use. Stamatatos argues that the most important criterion for selecting features in authorship attribution tasks is their frequency. He cites Koppel, Akica, and Dagan (2007) who propose an additional important criterion for feature selection in authorship attribution: the instability of features. For example, words like ‘and’ and ‘the’ are very stable since there are no alternatives for them. On the other hand, words like ‘benefit’ or ’over’ are unstable because they have a variety of alternatives for them. Therefore, they argue, that instable features are more likely to indicate stylistic features of the author. He also mentions the concept of ‘entropy gain’ or ‘information gain’ from using a feature (using PCA, or Principle Component Analysis) as an alternative tool.

Various attribution methods used in academic and commercial settings are discussed in the Stamatatos paper. An exhaustive discussion of the methods described is beyond the scope of this paper. However, he divides the methods into two groups: profile-based methods, where individual texts help define the ‘profile’ of an author against which every new text is compared, and instance-based methods, where each training text is individually represented as a separate instance of authorial style. He also deals with hybrid approaches attempting to combine characteristics of profile-based and instance-based methods.

Rangel et al. (2013) identify 18 approaches to authorship attribution amongst the submissions received for CPAN2013 authorship attribution task. Of the 18, 10 methods used stylistic features, described earlier. Of those, 5 methods further used POS tags and 3 used HTML properties of the given documents. 7 participants used readability features, of which one used them exclusively. Emoticons were used by 2 participants, and discarded from one.

Further, different content features, such as Latent Semantic Analysis, bag of words, TF-IDF, dictionary-based words, topic-based words, entropy-based, words, and so on, were used by 11 participants. Named entities (1), sentiment words (1), emotion words (3), and slang, contractions and words with character flooding (4) were also used variously by participants. One participant used the text to be identified as query for a search engine. Another participant introduced a high variety of corpus statistics to build unsupervised models, and four participants used n-gram models. Finally, one participant introduced advanced linguistic features such as collocations and another participant used second order representation based on relationships between documents and profiles.

All approaches used supervised machine learning methods. A vast majority used decision trees. Three approaches used Support Vector Machines, two approaches used logistic regression, and others used Naive Bayes, Maximum Entropy, Stochastic Gradient Descent and random forest.

Rangel et al. find difficulty in establishing a correlation between the used features in different approaches and obtained results, mainly due to shared features between the approaches. However, they note that the second order representations based on relationships between documents and profiles was the winner of the task, and use of collocations the winner for English task (versus Spanish). They also point out that amongst all the approaches using n-gram features, only one performed better than the median. They observe that use of sentiment words, emotion words, and use of slang words does not seem to improve accuracy in the general case, but state the difficulty in establishing a solid correlation. It is also to be noted that the runtime of the submissions for the task ranged from 10.26 minutes to 11.78 days.

The diversity of the different approaches to the authorship attribution problem, and the lack of a clear winning algorithmic approach amongst them even for within a given domain show the potential amount of work to be done in the field. The running times of the algorithms make it clear again the unlikeliness of the robustness of a single algorithm (or even any given ensemble of algorithms) across multiple domains and fields. In any scenario, the paper is a treasure-trove for different algorithmic approaches to authorship attribution.

Koppel et al. (2010) consider authorship attribution as found in the ‘wild’: where the set of known candidates is extremely large (possibly thousands), might not even include the actual author and the known texts and the anonymous texts might be of limited length. They claim that even in such difficult cases, similarity-based methods along with multiple randomized feature sets can be used to achieve high precision. They claim to show the precise relationship between attribution precision and four parameters: the size of candidate set, the quantity of known-text by the candidates, the length of anonymous text, and a certain robustness core (the confidence by which the algorithm could tell if the author was one of the candidates) associated with an attribution.

They use a set of 10,000 blogs harvested from the internet, and for each blog, choose 2000 words of known text and a snippet, consisting of the last 500 words of the blog, such that the two texts are disjoint. Then they determine which — if any — of the authors of the known texts is the author of a given text.

Koppel et al. use character 4-grams as a naive method, and compare it against an ensemble method of 4-grams and randomized feature sets. They manually manipulate two variables: the size of iteration, and the size of randomized sets, to get optimal results. Using 40% (=100 000) available features per iteration, they achieve 87.9% precision with 28.2% recall.

They discover that although results degrade rapidly with decreasing snippet size, even for as few as 100 words, they get precision of 71% at 10% recall. Performance seems to degrade significantly between 500 and 400 words. As for known-text length, they find that while increased size of known texts improves results, the difference in performance between 2000 and 1500 words of known text is marginal. They note that passable results can be achieved even for snippets as short as 100 words. They find that the four parameters: snippet size, known-text size, number of candidates and score, account for most of the variability in coverage and precision, so that for any given attribution one can assign a fairly accurate estimate of the likelihood that the attribution is correct.

The authors do note that, while as expected, accuracy increases as the number of candidate authors diminishes, this happens only up to a point. Their paper paves the way for accurate authorship attribution even within limited training data, and is thus relevant to this research.

Moreover, the fact that they use n-grams for their analysis makes it more interesting vis-a-vis existing literature. Therefore, their paper is one of the papers this experiment has hoped to replicate experimentally.

Taking Koppel et al.'s work (as mentioned earlier) further, Narayanan et al. (2012) experimentally demonstrate the effectiveness of their authorship attribution techniques with as many as 100 000 candidate authors. They argue that their result has serious implications for anonymity and free speech — an anonymous blogger or whistleblower may be unmasked unless they take steps to obfuscate their writing style.

For their experiment, the authors assemble a dataset of 2.4 million posts taken from 100, 000 blogs. They describe the various features that make it difficult to scale classification beyond a few hundred authors. They rule out otherwise-excellent methods such as Linear Discriminant Analysis (LDA-M) (similar to SVD mentioned earlier) because they are unable to take advantage of normalization techniques due to the sparseness of data. They find that Naive Bayes works 'surprisingly well' for the classification. They extract a total of 1188 lexical, syntactic and character features of the text, and find that Naive Bayes on normalized feature set gives them the best performance. That they use a limited training data makes their setup idea for replication in this case.

They argue that despite having demonstrated the viability of their techniques in cross-context setting, a more thorough investigation of different classifiers is necessary, along with an analysis of the impact of training and test set size. They state that understanding and modelling *how* writing style is affected by context has the potential to bring major gains in accuracy.

The results of their experiment are impressive. Out of 100,000 authors, in about 20% of trials, the author of the sample is ranked first among all 100, 000. Moreover, in the median case, the likely authors of a sample can be narrowed down to a set of 100-200, a reduction by a factor of 500-1000. While, as was discussed during the defense of this paper, such results will yet not stand in the court of law, it is sufficient for an intelligence agency (malicious, or otherwise) to significantly narrow the number of potential suspects to be kept under surveillance. Narayanan et al. discuss the implications of such a practice in great detail.

Because of the very impressive nature of their results, this research has tried to emulate their results to compare against other existing methods too.

Stamatatos (2013) begins with the observation that while many authorship attribution studies have demonstrated the effectiveness of character n-gram features for representing text, most of them examined the simple case where the training and test corpora are similar in terms of genre, topic, and distribution of texts. He tests the robustness of n-gram based authorship attribution methods under cross-genre and cross-topic conditions. He also provides a set of guidelines to tune an authorship attribution model according to the properties of training and test corpora.

The author makes an important — and bold— claim with regards to the usefulness of n-grams in representing documents. He argues that when the available texts are not on the same thematic area, and shows that contrary to intuition, character n-grams are more robust features than frequent words when the thematic area or the genre of the text is not controlled.

This paper is used as a justification to experiment with n-gram based approach to authorship attribution described in Koppel et al. N-gram based methods are usually seen as naive, basic, and often suspects in being effective in simple dataset but less useful otherwise. Stamatatos (2013) discusses the cyclical attitude towards n-grams for authorship attribution in detail.

Seroussi et al. (2011) propose employing Latent Dirichlet Allocation in authorship attribution. In it, they claim 'state-of-the-art' performance for both 'few and many candidate authors'. Since LDA-A does not inherently work with an explicit 'author profile', they suggest two implementations: LDA-s (single) where all available texts for an author are treated as a single document, and the word-topic model is created on that, and LDA-m (multiple) where every text is treated separately and a trial document is assigned the author whose documents it minimizes the total distance with. They do note that this requires a large sample from each author to model correctly.

Seroussi et al. (2013) use topic models as suggested in their earlier paper to perform authorship attribution. They use their techniques to successfully predict sentiments and categorization of the test documents, and claim to match the state-of-the-art performance. However, their results are not further reproduced. This research has also tried to replicate Seroussi et al.'s results, as explained in both their papers.

Abbasi et al. (2008) develop the 'Writeprints' technique for identification and similarity detection anonymous identities. It is a Karhunen-Loeve transforms-based techniques (Maccone, 2009) that uses online datasets spanning different domains: email, instant messaging, feedback comments, and program code. They claim their results outperform benchmark techniques like SVM, Ensemble SVM, and PCA, on the identification and similarity detection tasks with accuracy as high as 94% when differentiating between 100 authors. They use four datasets: email messages from the Enron corpus, buyer/seller feedback comments extracted from Ebay, programming code snippets extracted from java technology forum, and chat logs taken from cyberwatch. We will observe that while it can be used as a reference in feature extraction, the kind of data Writeprints uses (rather personalized) is quite different from the one we will be dealing with, and thus the direct method is not relevant to the current paper.

Afroz et al. (2014) explore the performance of authorship attribution algorithms on challenging datasets. They consider two scenarios: one which involves text written by an unknown cybercriminal and a set of potential suspects and second which involves users with multiple accounts on a forum, where the analysis engine outputs possible doppelganger accounts operate by the same user. They claim a 77-84% accuracy in the first scenario, and 94% recall with 90% precision on blogs.

To do so, they used a linear Support Vector Machine (SVM) with sequential Minimal Optimization(SMO). They perform 10-fold cross-validation, with at least 4500 words per author as training data, and at least 500 words per author as test data. The feature set they use involves a combination of syntactic and character-level features.

They find that while their accuracy increased with more words per author, on average, it did not improve when more than 4500 words were used per author while training. They also observe that the features with highest information gain ratios were different for the different languages they used. For example, while the use of foreign words had a high information gain in English, it was not so for Russian. Similarly, the use of uppercase letters was significant only for German. This suggests that any feature-extraction based authorship attribution algorithm is likely to be sensitive to the nature of the language in question, and raises the possibility of different content domains within the same language having noticeably different features relevant.

Afroz et al. use the Blogger dataset used by Narayanan et al. that resulted in 200 blogs suitable for the experiment. They used 100 blogs as training dataset, and 100 as a test dataset.

This paper uses Afroz et al.'s results to set the minimum training data for each author.

Afroz (2013) tackles the issue of deception in authorship attribution, and works without making the assumption underlying the significant majority of research in the field: that authors do not intentionally change their writing style. She shows that such methods fail to identify authorship when the assumption does not hold. She uses the Brennan-Greenstadt adversarial dataset, and a similar dataset collected particularly for her research. The articles in the datasets are written by authors who were specifically told to hide all stylistic markers (among, and within different articles) that might be used to identify them. She shows that linguistic cues that can detect deception in the existing adversarial dataset. They can also detect indication of masking in the documents attempting to imitate a different author's writing style. She shows how long-term deceptions such as the blog posts from 'A Gay Girl in Damascus' are different from such short term deceptions. She finds such deceptions to be robust to her classifier but more vulnerable to traditional techniques, suggesting that long-term manual deception of authorship attribution algorithms is not trivial, and given enough sample writings from an author, it is not extremely difficult to identify author even for an adversarial dataset.

She considers various possible means of stylistic deception: obfuscation (writing a document such that one's personal writing is not recognized), imitation (imitating someone else's style), and machine translation.

Afroz uses various machine learning algorithms to identify stylistic deception, and focuses on the SMO SVM as it outperforms other classifiers.

She uses an adaptation of the the Writeprints feature set as proposed by Zheng et al. (including lexical, syntactic, and content specific features). She discovers that the pattern of use of function words in obfuscated documents shows a clear distinction from their usage in original documents.

In identifying stylistic obfuscation, Afroz finds that accuracy of identification of an author does not improve after training on more than 4500-5000 words per author.

Afroz's paper's relevance comes from its attempt to tackle adversarial data, and we try to emulate her methods in our experiments.

Luyckx et al. use a corpus with 145 different authors to show the effects of many authors on feature selection and learning, and show robustness of memory-based learning approach in authorship attribution and verification with many authors and limited training data when compared to eager learners.

They use a 200 000-word corpus that consists of 145 student essays of about 1400 words each, about a documentary on Artificial Life, hereby keeping markers of genre, register, topic and age relatively constant. They see an observed decrease in accuracy with an increase in the number of authors. They report an accuracy of 50% and suggest that studies that are reporting over 95% accuracy on a 2-author study are overestimating their performance and the importance of the features selected.

Data and data collection

A total of about 2.7 million articles were collected, with the combined size of 7.5 Gigabytes. The sources of those articles were 8 different newspaper websites from 7 different countries in Asia (and Turkey). Along with the text, the date of collection, the authors of the articles, the original date of publishing, and the links to the articles were collected.

The scraping algorithm ran from Nov 2014 to January 2015. The collection process was as follows.

The chosen newspaper sites were mapped, and a custom scraper for each website was created to visit all article pages and to extract the relevant parts of the webpage.
The scraping algorithm was then run every 10 minutes for every newspaper, with several days' worth of articles being scraped in every 'batch'. Scraping 'locks' were used so that an unfinished previous batch wouldn't let a new batch job to run, to avoid double scraping intersected days.

The data was stored in MongoDB database, hosted online on Compose.io (formerly MongoHQ). The scraper was written in python scripting language, using the BeautifulSoup library.

The Scraped websites were as follows:

Newspaper	Country Published	No. of Articles
Cambodia Daily	Cambodia	24 674
Dawn	Pakistan	892 966
Daily Express Archive	India	170 308

Daily Express	India	548 466
Jakarta Post	Indonesia	231 564
Daily Mirror	Sri Lanka	38 517
Philstar	Philippines	173 405
The Kathmandu Post	Nepal	149 274
Today’s Zaman	Turkey	303 507

Table 1: Data sources and their sizes. For more details, refer to Appendix 1.

All the sites have articles up to January 2015. However, the starting dates of their publishing differ considerably. As a group, the dataset has newspaper articles starting from 1991.

Dawn and *Daily Express* have several times more articles than others because of several reasons. First, they are significantly larger in terms of population than most other countries. Second, they have an active newspaper/news website scene. Third, since both are multicultural, English is the lingua franca in both, and spoken by a large number of people. Because of that, they have a large number of English articles. Cambodia Daily has relatively smaller number of articles because the archive was weekly-based, even though it is itself a daily. We hypothesize that the relatively smaller population and small population of English speakers may have contributed to the small number of articles, even though the range of its articles (1998-2015) is quite large.

For this project, we were interested in only those articles that were written by a single author, with explicit byline. Thus while the general database was quite large, the extracted set of data was quite limited. All the articles with empty bylines were ignored, and so were those with names of news agencies, and in-house reports. A database of ‘to be ignored’ authors/groups was made, and it was used to filter out the undesired documents.

Authors with 30 or more articles were the ones considered for the analysis. The argument is this: with a worst-case word-count of ~400 words per article, we would have >12 000 words per author, enough for such analysis according to most of the available literature on the issue. We use 25 articles (>10, 000 words) per author to train our algorithms, and the rest to test them. We limit the number of articles used for training for 2 reasons: first, we want to accurately replicate algorithms tested under realistic constraints, and because by limiting the number of training articles, we are limiting the precision-recall skew and the detailed analysis it would need later.

The number of authors filling such criteria scaled according to the number of articles, with a few exceptions. Zaman has 220 authors satisfying our requirements, while the Kathmandu Post, which has half as many articles, has almost twice that many authors. Both Dawn and Indian Express have authors in the 800’s. The exact number of authors fulfilling all the criteria is not mentioned here because there still exist some noise in the author-name data that needs filtering out. The number of average author per newspaper averaged around 300.

Due to the inconsistency in web design, extracting author names was more tedious than was expected. There was inconsistent and irregular use of author tags in some pages, while some would have article text overflowing into author field but only in few articles. This process was more time consuming that it warranted.

After the eligible authors were discovered, their articles were extracted and sanitized. Sanitization involved removing of multiple spaces, paragraphs, and so on, so we were left with only sentence structures and punctuations. While some papers (Argamon et al., 2007) have suggested the use of spacing and paragraphs as useful features in identifying authors, this particular project would not fit the feature set being used in this project. The sanitization was done by tokenizing documents using delimiting characters, and stringing them back together using the space character.

Other sanitization work was not done in the general dataset, because the Narayanan (2012) paper would need all possible features of the text. Thus, the data was used as-is for the Narayanan experiment. For the other algorithms, the text was turned to lower case and only alphabets and digits were considered. It is possible such sanitization may have led to decreased performance for the two algorithms in question. However, it had to benefits: first, to accurately replicate the circumstances described in the papers in question, and second, to reduce the number of features in question. Without reducing the analysis space by basically ‘collapsing’ the possibilities based on the character case and punctuation, we saved significant running time and memory through computation. For example, for lower-case based 3-grams there are 27^3 = ~25000 possibilities, with 4000 commonly used ones. If mixed cases are allowed, the possibility goes up to 52^3 = ~125000. An inclusion of 10 digits and 10 punctuations takes that up to 72^3 = ~370, 000. Therefore, the practicality of conducting experiments was traded in for possible loss in accuracy.

Different sources had differently peculiar text, and that had to be accounted for in the analysis. For example, some sources had the date and location of the text’s publishing inside the text itself while others didn’t. Others had secondary content (such as ‘content boxes’ popular in magazines) included with the main text, which implied a larger use of numbers compared to normal text. Some had captions of the photos included in the text inseparable from the text.

The peculiarities created very observable 3-gram patterns in the texts: the frequency of 3-gram for month names (jan, feb, mar, etc) were all present in the top 100 trigrams for those newspapers that put the date published on the body of the articles. This was not true for any of the newspapers that did not have that. Hence, a mere analysis of top-100 3-grams was sufficient to distinguish the ‘group’ of newspaper. We understand that the analysis is unnecessary for this identification: a mere ‘check’ of presence of the identifying date in the text would suffice, without having to calculate the relatively expensive n-gram analysis. Regardless, the n-gram analysis is important because it removes the necessity of having the domain-specific knowledge for any particular newspaper, and gives us a more generalized tool. It will be seen as a common thread running throughout this paper: while some methods may seem unnecessarily roundabout way of doing certain analytics, doing them enables us to 1) have a generalizable algorithm without delving into domain-specific knowledge and 2) understand the workings of our algorithm and reason through them, as opposed to having a ‘black-box oracle’ that works but does not further our knowledge of the text in any way,

After the text and author names were sanitized, they were stored in text files for temporary processing (they are stored permanently in MongoDB database), for the sake of simplicity. After consulting the referred literature, it was decided that each author profile would be represented by a concatenation of all the text written by the concerned author — versus each document acting independently, clustered around an author profile. This goes back to our discussion of LDA-m and LDA-s, and the Seroussi et al. discovery that LDA-s was more effective. Individual lines contained the entire corpus for each author in test and training data files. Another text file had an author name in each line, each corresponding to the same line in test and training files.

Algorithm Implementations and Experimental Results

Whenever existing implementations of the algorithms described were found, they were used instead of being implemented from scratch.

For the LDA-A paper, an existing implementation in C++ called GibbsLDA++ was used (Phan et al., 2007). The alpha and beta parameters, and the number of iterations was reproduced from reference paper. The output word-topic probably distributions were clustered using Hellinger distance as implemented by the Python library NLTK (Natural Language ToolKit) (nltk.org, accessed January 17 2015). It is important to point out here that the analysis did not use the existing NLTK training corpus so no ‘cross-contamination’ from corpus to test documents was likely. Results were measured and collected through custom Python code.

The Narayanan et al. (2012) algorithm was implemented entirely by (this) author, with the paper serving as the reference for the implementation. Parts-of-speech tagging, function words, and tokenization were based on the Stanford CoreNLP corpus and training set for the Python NLTK package. Since parts of speech and tokenization are universal features in English, and no literature found by the author has considered the non-universality of function words, it is assumed that such processing (using an unrelated corpus) did not have negative impact. The naive bayes classification used after the extraction of features was again based on Python’s NLTK’s ‘classify’ module.

The improved version of the n-gram-based version of as described by Koppel (2010) paper was implemented entirely by the author, in Python. To randomly select a required number of articles, an article set was first shuffled using the Fisher-Yates shuffling algorithm (Fisher et al., 1948) and first ‘n’ desired number of articles was used. The inbuilt NLTK cosine similarity function was used. The threshold and others parameters were exactly as described in the reference paper.

For every algorithm, the following was done: 25 articles were used for training per author, and 5-15 articles were used for testing, totaling a total of 40 articles used per article. For authors who had a large number of articles (some had over 500), the training and test articles were selected randomly to avoid the possible ‘temporal-based’ and ‘topic-based’ contamination of the corpora which could associate an author too strongly with certain topics or words associated with certain time periods. The authors were compared only against other authors within their newspaper, and the corpora used were limited to the newspapers of the concerned authors.

Because the used data was limited, and the selection of training/test data was already random, cross-validation was not performed. Moreover, it was assumed that comparing across models using cross-validation for each would be overestimating the performance for each model (with limited data) and would not be extra helpful in the analysis. Limited processing power further made such validation too time-consuming for the set limit: all algorithms were given a maximum of 6 hours to complete, and it was observed in primary experiments that cross-validations often ran for longer than that.

Each model was considered to be tested on each dataset, totaling 9*3 = 27 potential experiments. However, after running primary experiments, certain experiments were considered unhelpful, and thus only 19 total experiments were run. This issue will be discussed in detail in the following pages.

Observations on the LDA-A paper

Early experiments for the LDA-A gave disappointing results. ‘The Kathmandu Post’ dataset was used to test the algorithm with 310 qualifying authors. The Hellinger Distance between the word-topic distribution for an author profile and test documents was the distance used: the lower the distance, the greater the likelihood of authorial match. On average, an author’s test documents came as the 152'nd likeliest candidate for his/her profile, suggesting the algorithm was not significantly better than chance than in identifying the authors. However, on multiple trials, the likelihood that the actual author came up in the top 6 identified authors was 20 % — significantly more than the expected 1.1 % (adding the probability that the author could be any of the six authors) that would occur due to chance. Despite this, the result did not meet the standards set by the reference paper, and the expectations set while choosing the algorithm to be tested for. In addition, the training time exceeded the allocated six hours on Sunfire 61/62 Halligan servers. Therefore, the algorithm was not selected to be run on the larger data set.

The cause of the poor results is likely to be the limited size of the training set. As the reference states, one of the cons of using this methodology is the need of large training data to create an accurate model. The number of training articles could have been increased (as was discussed with Prof. Blumer during experimental phase), but the number of candidate authors with requisite number of training articles would be quite small. Considering the scope of this research — newspaper articles with limited training data and a large number of potential authors — the algorithm was considered unsuitable for the current purpose.

Future projects in the field could potentially explore the relationship between training data and the results for LDA. Ways to optimize the generalized algorithm for increased performance in this particular domain might be possible, possibly by varying the parameters and the number of iterations. Due to the potential nature of the project such explorations were considered unfit for this paper.

Observations on Experiments from Narayanan Paper

1188 features from documents were extracted for the Narayanan paper experiment, and normalized, as described in the paper. The authors do an extended discussion on the implementation, and made replicating it quite reasonably easy. Naive Bayes was used on resulting feature sets to predict the authorship of the unattributed documents.

The accuracy of prediction varied observably over different datasets, suggesting that the methodology is sensitive to variables uncontrolled in the text in this paper. While a Linear Discriminant Analysis (LDA-M) was not performed to identify the features that contributed more significantly to the sensitivity, as Narayanan et al. do to identify the entropy gain for each feature, the reference paper’s findings provide a hint on the behavior of our results. The authors discovered that frequency of the character “` “, the number of characters, and frequency of period and comma were the features that showed the largest entropy gain in their experiments. The datasets that saw poor performance from this also happened to be the ones that were poorly formatted (had dates and other non-essential information in text fields). The features that were the most telling of authors’ style were made less relevant by carelessness in formatting the articles. Sanitizing the data better — a rather time-consuming and manual process to manually check and verify the inconsistencies in text — might lead to improved and more consistent performance for this algorithm.

It was further observed that the results for best performance still did not match the performance claimed by Narayanan et al for comparable training data size. Narayanan et al. claim an accuracy of 20% for training size of 25 articles — the accuracy averaged less than half of that throughout all the datasets, as the table following will show.

There are several possible causes to this; the most likely one is that the nature of data masks the stylistic markers of the individual authors as extracted by Narayanan et al.’s algorithm. It is also possible that the paper’s conditions were not accurately replicated despite a pained attempt to follow the detailed explanation, that even the ‘sanitized’ data had markers that ‘polluted’ features, or that the paper itself presents the selected best-performing results. In any case, the failure to achieve comparable results even with the best possible presentation of data suggests that there is a need for greater investigation into authorship attribution for newspaper articles as described in this paper.

A possible explanation of the variable performance of the algorithm with different data sources is that the different sources have different ‘nature’ of authors: some newspaper might be more inclined to allow their contributors/reporters to contribute on a wide variety of topics, while others might be inclined to limit them to their field of specialty. A wide topic distribution for any given author is suggested as the possible causative factor in decreasing the accuracy of this method.

Work done by Sapkota et al. (2014) strongly hints at the possibility. They discover that thematically-trained algorithms perform attribution considerably better than cross-thematic training. Adding training data in general improved performance, regardless of the topics involved. They also find that training on diverse topics is actually better than training them in a related themes, a direct contradiction to the claim made earlier. We offer the following explanation to the contradiction: while, with a relatively large training set, training in diverse topics might improve general accuracy as they claim, doing so with limited and thematically noisy data, and testing with similarly noisy test data does not. Put differently, the random selection of limited training and test data for authors with large number of articles created performance bias against datasets containing such authors, as authors who wrote more were likelier to write on more diverse topics, and show stylistic differences across time. Moreover, Sapkota et al. use data from email, essay, blog, chat, and phone interview in their analysis: we claim the difference in the domains makes for an unfair comparison.

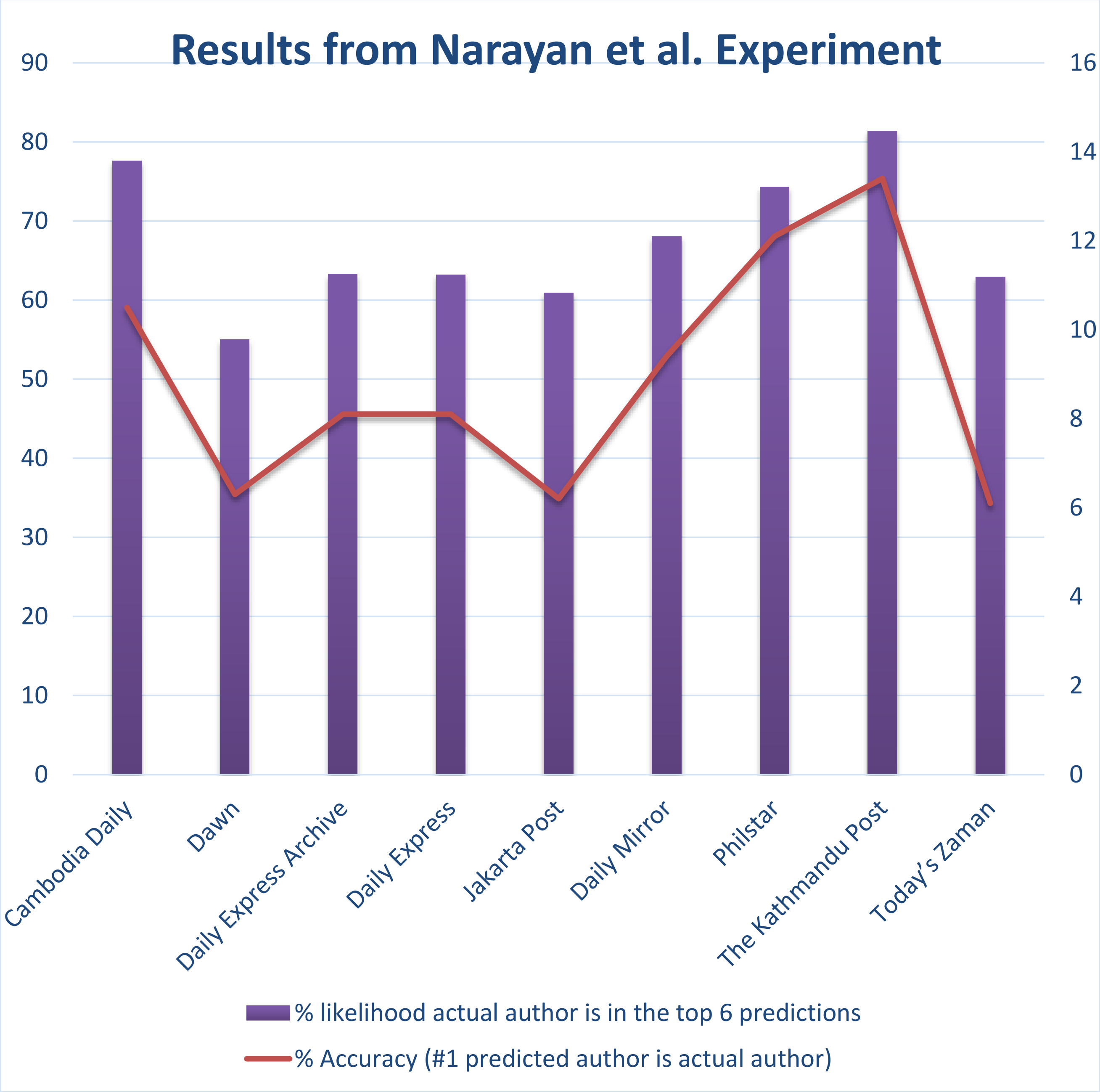
The above leads us to suggest a way to test the topic-variety sensitivity of different authorship attribution algorithms. Topic identification — or text categorization — is a reasonably well-solved problem in Computer Science (Joachims, 1998). By looking for a negative correlation between an algorithm's performance across varied datasets, and the category diversity of different datasets, we would be able to identify the corpus-dependent sensitivities of different authorship attribution algorithms. This could even be used as a criterion to choose between different approaches to authorship attribution. While Sapkota et al. (2014) discuss thematic and topical differences related to attribution tasks, they do not explore this avenue. Works in literature exploring this possibility in detail were not discovered by the author, suggesting the potential for two things: 1) the need/usefulness of a varied sets of texts from different contexts but similar domains in authorship attribution, and 2) exploring different means to evaluate authorship attribution algorithms.

The results obtained for this set of experiments are tabled below. Results are rounded up to the nearest tenth.

Data Source	% Accuracy (#1 predicted author is actual author)	% likelihood actual author is in the top 6 predictions
Cambodia Daily	10.5	77.6
Dawn	6.3	55.0
Daily Express Archive	8.1	63.3
Daily Express	8.1	63.2
Jakarta Post	6.2	60.9
Daily Mirror	9.4	68.0

Philstar	12.1	74.3
The Kathmandu Post	13.4	81.4
Today’s Zaman	6.1	62.9

Table 2: Results from the Narayanan et al. paper experiments



Graph 1: Results from Narayan et al. experiment

Observations on Koppel et al's improved n-grams algorithm

Using randomized variation of 4-grams feature sets — as explained in the Koppel et al. — and 40 % of the available features (=100, 000) as used in the original paper, we find that while the performance achieved is not as high as described, the results are more consistent across different data sets than the other methods tested. The parameters *k1,* and *k2* were used based on the original paper: k1 = 100 and k2 = 40%. An addition had to be made to the original algorithm to fit the measurement metrics used in this paper: whereas the paper asks for choosing the most-occurring top match, we had to calculate and keep track of the six most-occurring top matches for each iteration.

While precision and recall are used by Koppel et al., to evaluate the results of their experiments, we continue with our existing evaluation methods to keep consistency with other algorithms. The algorithm is run over each data set for two times each (despite the fact that the algorithm already involves iterating many times on a randomly chosen feature set), and results are averaged out. This set of experiments took by far the longest time, with each experiment taking longer than 5 hours in the Tufts homework server. This can likely be attributed to the large feature space being used, and the relative slowness of Python in handling such tasks. The fact that we were tracking the top-six matches instead of the topmost match added at least some extra overhead in the calculations, which added up over all the iterations.

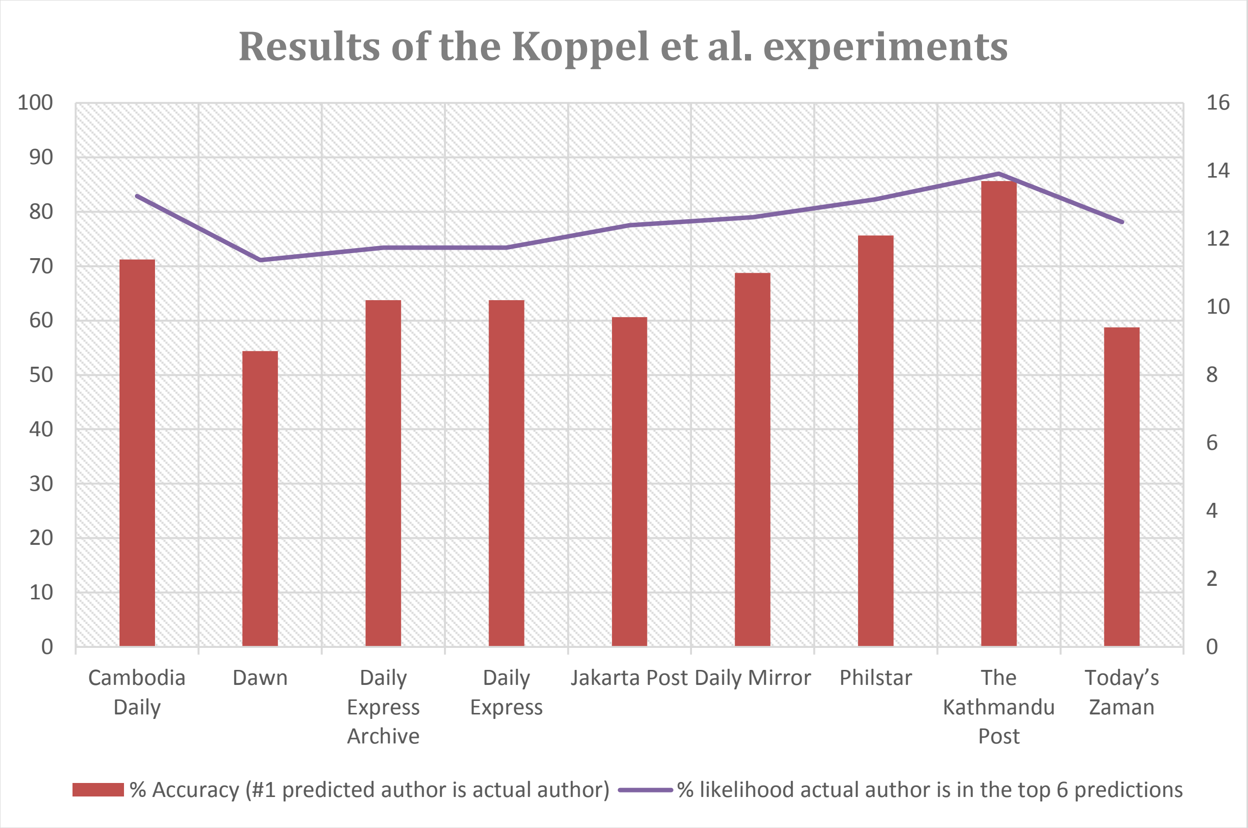
As can be observed from Table 3, the results from different sources are more similar than in the previous experiment, suggesting the algorithm is more robust to corpus-dependent results. The accuracy has improved for all the sources — and most significantly for those that had the poorest accuracy in the previous experiment. In addition, the variance in accuracy across the various data sources has decreased.

An interesting pattern in both tables 2 and 3, but observed more strongly here, is that the % accuracy and % likelihood the author is in top 6 choices are not strictly correlated. The gains achieved in Table 3 compared to Table 2 are not equally distributed in the two measures either. This suggests that while the ‘distinctness’ of more authors is discovered, the ‘random’ selection of articles (in both the experiments) and features (in this particular experiment) results in the accuracy not scaling equally in identification of the author as #1 suspect since articles to span temporally and topically, and thus stylistically. We present this as supporting the claim we claimed earlier. With that as the basis, we predict that increasing the number of training articles per article will result a larger gain in the likelihood that the author is in top n % of authors chosen than the likelihood that the #1 chosen author is the actual author. While multiple runs of each experiment could remove the effects of potential sampling bias, such runtime was too time-consuming to be run for the purposes of this project.

Here we make an observation regarding the number of candidate authors. The number of candidate authors varied between 220 and 700 between various publications, and averaged around 300. The number was larger for *Dawn* and the two *Express* sources, while others hovered generally around the average. Perhaps because they are in the same order of magnitude, we see no significant influence of the number of candidate authors in accuracy, though we do observe that the three above sources generally see less accurate predictions than others. The variable was not considered significant enough to warrant an expository discussion. Considering the difficulty in identifying individual authors and group of authors, the exact number of authors for each publication is not presented. In case of interest in the data, further information will be made available.

Data Source	% Accuracy (#1 predicted author is actual author)	% likelihood actual author is in the top 6 predictions
Cambodia Daily	11.4	82.9
Dawn	8.7	71.1
Daily Express Archive	10.2	73.4
Daily Express	10.2	73.4
Jakarta Post	9.7	77.5
Daily Mirror	11.0	79.0
Philstar	12.1	82.3
The Kathmandu Post	13.7	87.0
Today’s Zaman	9.4	78.1

Table 3: Results from Koppel et al. experiments



Graph 2: Results from Koppel et al. experiments

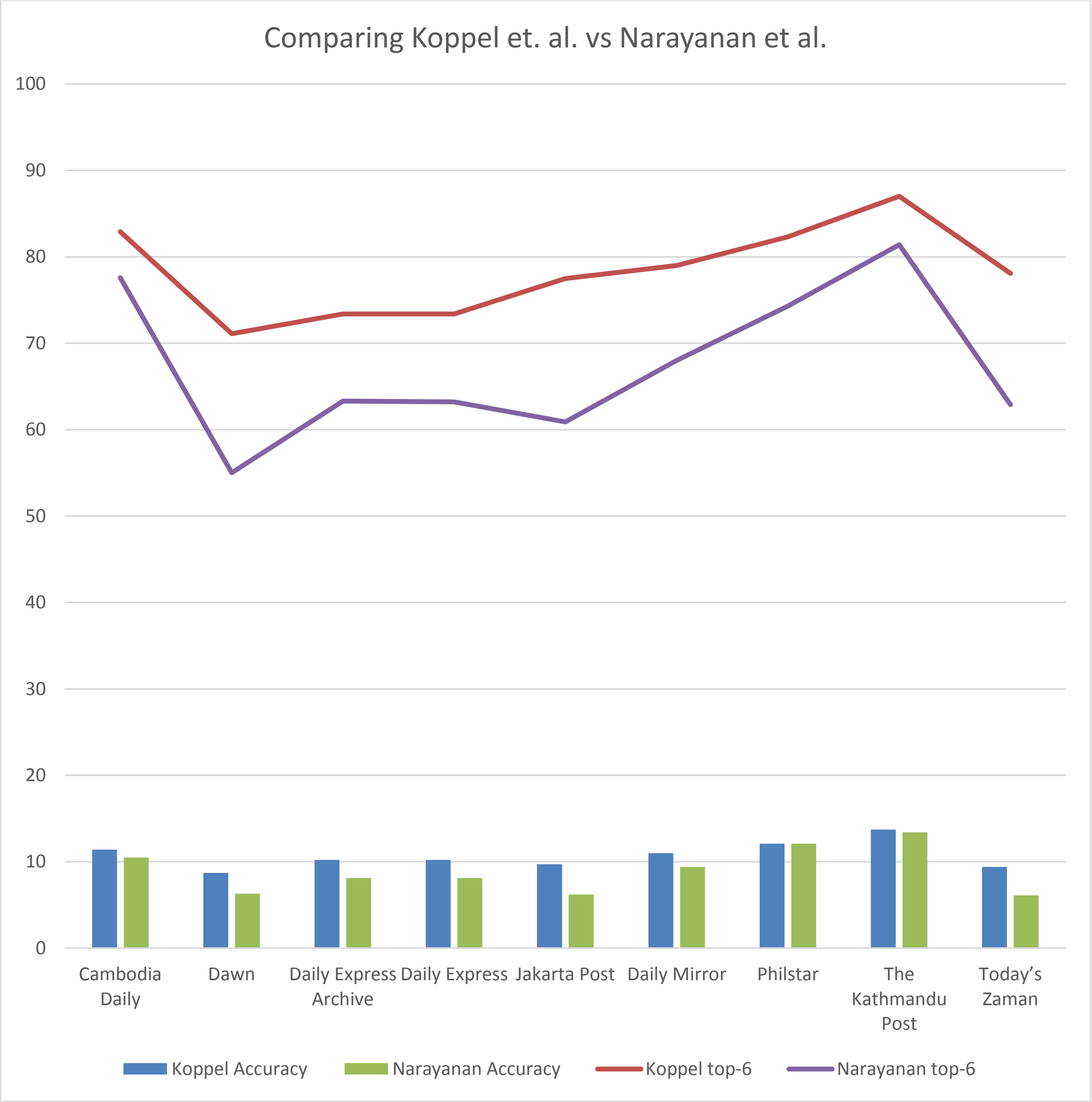


Chart 3: Difference in performances of Koppel et al. and Narayanan et al.

Justification for the lack of an ensemble method

In its proposal phase, the project had offered to combine elements of different algorithms to see if an ensemble of different tested tools would improve performance over any individual method. In retrospect, this was not done, because of the unsuitability of the tested algorithms as described below.

The nature of LDA-A makes it unsuitable to be combined with the other two algorithms. While the others use distance in multidimensional Cartesian space as the distance measure for clustering, LDA-A uses Hellinger distance, which is used to quantify distances between two probability distributions. Additionally, the relatively unimpressive results obtained as a result of experiments made the point of using it to bolster any other algorithm moot.

Koppel et al.'s n-gram based method is powerful against conventional n-gram approaches precisely because of its 'random-bag' approach. On the other hand Narayanan et al's method uses character-level frequencies the conventional way, normalizing within the same feature amongst all candidates, and across the various extracted features. With that, introducing a probabilistic measure to his method would not allow us to preserve features within the same author.

Thus, the choice of algorithms to be tested, made after the initial proposal was submitted, made it difficult to construct an ensemble method and test it.

However, after having looked at the results, it seems the entire point behind the need of an ensemble method has been cleared. We wished to offer improvements to existing algorithms: comparing the Narayanan et al., vs Koppel et al., we have identified and confirmed (the likely) causes of the weaknesses of both algorithms as implemented and tested in this paper. Thus, we feel justified in having used a variety of algorithmic approaches to analyze different data sources for identifying the sensitivities and robustness of the approaches.

On the chosen metrics, and the lack of rigorous statistical analysis

The conventional measures for machine learning algorithms are precision and recall, in addition to accuracy. Precision-recall are often graphed against each other, and multiple such curves are generated by varying controlled variable to identify the ideal value for the variable such that the precision and recall give-take is maximized.

That approach would have been unsuitable for this project because we are not varying an underlying variable in an algorithm, but changing between variables themselves. In addition, in this project we have decided to use variables identified as ideal by Afroz et al. (2013) and Koppel et al., among others.

The accuracy (or the number of #1 chosen authors who were the actual authors) was chosen to maintain consistency with the Narayanan et al. and the LDA-A paper who both use the measure. In addition it can be seen as an easy-to-understand number to easily compare algorithms, given that the number of training and testing sets are balanced properly for all authors (as was done for this paper).

The likelihood of an author’s appearing in the top 6 was used as a measure because it was present in data for the Koppel paper, and seemed like an understandable way to present data obtained from the LDA-A paper.

We offer no particularly rigorous statistical analysis of the results (or present such statistics) because of the nature of the data currently used. As we have seen, wild fluctuations in the algorithms' performance across various datasets already put the exact numbers into suspicion. Any statistical analysis is only as rigorous as the data underlying it: without a guarantee that the performance numbers are in the expected ballpark, we would merely be fudging numbers and getting differences in performances that could be unrelated to the algorithms' performance.

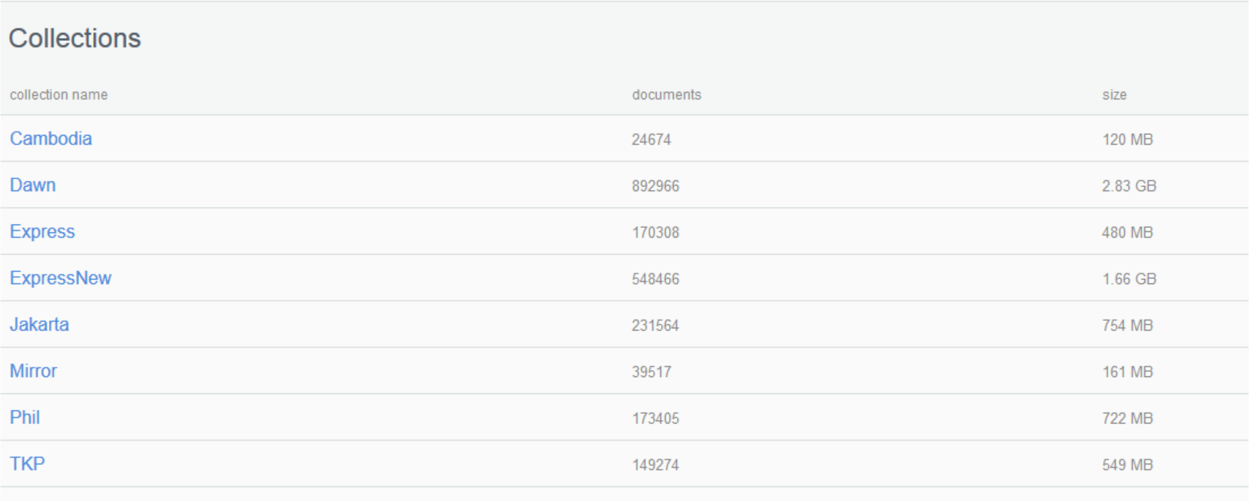
Data and its potential

The 2.7 million articles used in this project were scraped particularly for this particular purpose. That it originates from newspapers, across 8 different countries, is likely to make it useful for all sorts of textual analysis. It would allow duplication of results from this paper, and allow easy entry into similar research without having to worry about the data. This would likely be useful not only for computer science research, but likely also for history, political science, and linguistics research. While all the works collected are copyrighted, they would definitely be used for academic research, and could be maintained by someone on campus. Therefore, it is recommended that the data be made available for interested students or faculty.

Conclusion

This paper replicated the algorithms as described in three distinct approaches to authorship attribution, and tested them against newspaper articles from different regions. We identified the potential sensitivities and robust factors associated with those algorithms using their performances against different data sets, and other algorithms. This was an exploratory paper into such analysis, and has identified several potential areas where further research might be warranted.

Appendix 1: Data collected from various sources



Collections

collection name	documents	size
Cambodia	24674	120 MB
Dawn	892966	2.83 GB
Express	170308	480 MB
ExpressNew	548466	1.66 GB
Jakarta	231564	754 MB
Mirror	39517	161 MB
Phil	173405	722 MB
TKP	149274	549 MB

Bibliography and Reference

Abbasi, A., & Chen, H. (2008). Writeprints. *ACM Transactions on Information Systems*, *26*(2), 1–29. doi:10.1145/1344411.1344413

Afroz, S. (2013). *Deception in Authorship Attribution*. Drexel University, Philadelphia

Afroz, S., Caliskan-islam, A., Stolerman, A., Greenstadt, R., & McCoy, D. (2014). Doppelgänger Finder : Taking Stylometry To The Underground. *IEEE Symposium on Security and Privacy*. doi:10.1109/SP.2014.21

Argamon, S. (2007). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, *23*(2), 131–147. doi:10.1093/llc/fqn003

Argamon, S., & Koppel, M. (2009). a Systemic Functional Approach To Automated Authorship Analysis. *Journal of Law and Policy*, 299–315.

Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. *ACH/ALLC*, 1–3. Retrieved from http://tomcat-stable.hcmc.uvic.ca:8080/ach/site/xhtml.xq?id=162

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Trans. Inf. Syst. Secur.*, *15*(3), 12:1–12:22. doi:10.1145/2382448.2382450

Coulthard, M. (2013). On Admissible Linguistic Evidence. *Journal of Law & Policy*, *21*(2), 441–466. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=lpb&AN=88410034&site=ehost-live

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, 109–123. Retrieved from http://link.springer.com/article/10.1023/A:1023824908771

Fisher, Ronald A.; Yates, Frank (1948) [1938]. Statistical tables for biological, agricultural and medical research (3rd ed.). London: Oliver & Boyd. pp. 26–27.

Foster, Donald (2001). Author Unknown: Tales of a Literary Detective. Holt Paperbacks. Retrieved from Amazon.com

Goodman, J. (2002). Extended Comment on Language Trees and Zipping, 6. Retrieved from http://arxiv.org/abs/cond-mat/0202383

Grabchak, M., Zhang, Z., & Zhang, D. T. (2013). Authorship Attribution Using Entropy. *Journal of Quantitative Linguistics*, *20*(4), 301–313. doi:10.1080/09296174.2013.830551

HaCohen-Kerner, Y., & Margaliot, O. (2014). Authorship Attribution of Responsa Using Clustering. *Cybernetics and Systems*, *45*(6), 530–545. doi:10.1080/01969722.2014.945311

Hirst, G., & Feiguina, O. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, *22*(4), 405–417. doi:10.1093/llc/fqm023

Holmes, D. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. *… Royal Statistical Society. Series A (Statistics in Society)*, *155*(1), 91–120. doi:10.2307/2982671

Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, *37*, 151–178. doi:10.1023/A:1022673822140

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, *19*(4), 453–475. doi:10.1093/llc/19.4.453

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. *Artificial Intelligence: Methodology, Systems, …*. Retrieved from http://link.springer.com/chapter/10.1007/11861461_10

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features* (pp. 137-142). Springer Berlin Heidelberg.

Juola, P. (2007). Authorship Attribution. *Foundations and Trends® in Information Retrieval*, *1*(3), 233–334. doi:10.1561/1500000005

Kolowich, S. "The Professor Who Declared, It's JK Rowling." *The Chronicle of Higher Education*. 29 July 2012. Web. 13 Mar. 2015. <http://chronicle.com/article/The-Professor-Who-Declared/140595/>.

M. Koppel, N. Akiva and I. Dagan (2007). Feature Instability as a Criterion for Selecting Potential Style Markers, JASIST: Journal of the American Society for Information Science and Technology, Volume 57, Issue 11, Pages:1519-15250

Koppel, M., Schler, J., & Argamon, S. (2008). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, *60*, 9–26. doi:10.1002/asi

Koppel, M., Schler, J., & Argamon, S. (2010). Authorship attribution in the wild. *Language Resources and Evaluation*, *45*(1), 83–94. doi:10.1007/s10579-009-9111-2

Koppel, M., Schler, J., & Argamon, S. (2013). Authorship Attribution: What's Easy and What's Hard? *Available at SSRN 2274891*, *39*(2006), 317–331. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2274891

Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship Attribution with Thousands of Candidate Authors, 659–660. doi:10.1145/1148170.1148304

Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, *65*(1), 178–187. doi:10.1002/asi.22954

Li, J., Chen, H., & Huang, Z. (2006). A Framework for Authorship Identification of Online Messages : Writing-Style Features and Classification Techniques, *57*(3), 378–393. doi:10.1002/asi

Luyckx, K., & Daelemans, W. (2005). Shallow Text Analysis and Machine Learning for Authorship Attribution. *Technology*, 149–160. Retrieved from http://eprints.pascal-network.org/archive/00004945/

Luyckx, K., & Daelemans, W. (2008a). Authorship attribution and verification with many authors and limited data. *… of the 22nd International Conference on …*, (August), 513–520. Retrieved from http://dl.acm.org/citation.cfm?id=1599146

Luyckx, K., & Daelemans, W. (2008b). Authorship attribution and verification with many authors and limited data. *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*, *1*(i), 513–520. doi:10.3115/1599081.1599146

Maccone, Claudio, *A simple introduction to the KLT (Karhunen—Loève Transform)*, Deep Space Flight and Communications, Springer Praxis Books 2009, pp 151-171

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience.

Meina, M., Brodzińska, K., Celmer, B., Czoków, M., Patera, M., Pazecki, J., & Wilk, M. (2013). Ensemble-based classification for author profiling using various features Notebook for PAN at CLEF 2013. In *PAN - Uncovering Plagiarism, Authorship, and Social Software Misuse a benchmarking activity on uncovering plagiarism, authorship and social software misuse*. Retrieved from http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-papers-final/pan13-author-profiling/meina13-notebook.pdf

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. *Proceedings of EMNLP*, *4*(4), 404–411. doi:10.3115/1219044.1219064

Miranda-García, a., & Calle-Martín, J. (2005). Yule's characteristic K revisited. *Language Resources and Evaluation*, *39*(2005), 287–294. doi:10.1007/s10579-005-8622-8

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012). On the feasibility of internet-scale author identification. *Proceedings - IEEE Symposium on Security and Privacy*, 300–314. doi:10.1109/SP.2012.46

Phan, X. & Nguyen, C. (2007-2008). GibbsLDA++: A C/C++ Implementation of Latent Dirichlet Allocation. Retrieved from http://gibbslda.sourceforge.net/ on January 17, 2015

Qian, T., Liu, B., Chen, L., & Peng, Z. (2014). Tri-Training for Authorship Attribution with Limited Training Data. *Acl*, 345–351.

Rangel, F., Rosso, P., Koppel, M., & Stamatatos, E. (2013). Overview of the Author Profiling Task at PAN 2013. Retrieved from http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf

Sanderson, C., & Guenter, S. (2006). On authorship attribution via Markov chains and sequence kernels. *… . 18th International Conference on*, 18–21. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1699558

Sapkota, U., Solorio, T., Montes-Y-Gómez, M., Bethard, S., & Rosso, P. (2014). Cross-Topic Authorship Attribution: Will Out-Of-Topic Data Help? Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1228–1237, Dublin, Ireland, August 23-29 2014.

Seker, S. E., Al-Naami, K., & Khan, L. (2013). Author attribution on streaming data. *Proceedings of the 2013 IEEE 14th International Conference on Information Reuse and Integration, IEEE IRI 2013*, 497–503. doi:10.1109/IRI.2013.6642511

Seroussi, Y., Zukerman, I., & Bohnert, F. (2011). Authorship attribution with latent Dirichlet allocation. *Proceedings of the Fifteenth …*, 1–10. Retrieved from http://dl.acm.org/citation.cfm?id=2018957

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship Attribution with Topic Models, (May 2013). doi:10.1162/COLI

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*(3), 853–860. doi:10.1016/j.eswa.2013.08.015

Smith, P. W. H., & Aldridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method*. *Journal of Quantitative Linguistics*, *18*(January 2015), 63–88. doi:10.1080/09296174.2011.533591

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information …*, 1–28. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.21001/full

Stamatatos, E. (2013). on the Robustness of Authorship Attribution Based on Character N-Gram Features. *Journal of Law & Policy*, 421–439.

Tambouratzis, G., & Markantonatou, S. (2004). Discriminating the Registers and Styles in the Modern Greek Language-Part 1: Diglossia in Stylistic Analysis. *Literary and Linguistic …*, *19*(2), 197–220. Retrieved from http://llc.oxfordjournals.org/content/19/2/197.short

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, *18*(1), 45–55. doi:10.1177/016555159201800106