Author Identification: Using Text Mining, Feature Engineering & Network Embedding

Anku Adhikari (aadhikr2), Sharan Subramaniyan (ssbrmny2)
University of Illinois, Urbana Champaign

Abstract
Authorship analysis is a challenging area that has been developed through centuries and with research done widely scattered across multiple disciples of mainly computational linguistics, text mining, data mining, stylometry and machine learning. Conventional techniques from the past relied heavily on stylometry and text-based content analysis of document text for authorship analysis. More recent developments use network embedding training focus heavily on document attributes to build a network and predict the author. We propose a system that incorporated the strengths of both text mining and network embedding methods and utilizes both the document text and document attributes fully when available. In this paper, we describe the system overview and implementation in detail and discuss how by supporting a more multi-faceted information based network embedding, it can be possible to get improved results. Finally we discuss our results and suggest some future improvements in terms of results, speed and performance for our system to handle larger corpus

Keywords
Author identification, authorship attribution, author profiling, data mining, text mining, feature engineering, machine learning, network modeling

1. INTRODUCTION

1.1 Definition
Authorship analysis problems can be classified and defined as follows [1][2]:

- **Authorship identification/ authorship attribution**: Finding the most likely author(s) for an article or document or the likelihood of an author writing a piece.
- **Author profiling or characterization**: Finding author profile or characteristics like gender, socio-cultural background and language familiarity, etc.
- **Similarity detection**: Analyzing the similarity between pieces of work to determine if possibly produced by single author without necessarily determining the author. Used often in plagiarism detection.

The terms “author identification”, “authorship attribution” and author profiling” is often used interchangeably in some cases and have a heavy overlap as author identification usually relies on author profiling and attribution as the first step. Author identification problem can be extended to check for anonymity testing where we check whether the author is fairly anonymous and hard to predict or is predictable and hence makes the author subject to bias sometimes (eg in scenarios of double-blind reviews).

In this paper we primarily explore author identification, but due to overlap in the areas we explore the same in brief below.

1.2 Background and History
Development in this area of author identification/attribution have roots in stylometry and its study. Stylometry is the study of linguistic style often with the objective of author profiling, author attribution and authorship identification/verification.
Stylometry was used as early as the 4th century to determine document authenticity [3]. A notable very recent work from 2015 is where a play “Double Falsehood” was identified as being William Shakespeare’s work [4].

As stylometry paved the path towards author profiling/ attribution and identification, there is one historically famous case for author identification from the 1780s being the authorship of the disputed federalist papers, that has often been used as an example problem for many earlier works in author identification works [5][6] and used often by researchers to further improve and evaluate their methods.

Current author identification techniques go beyond analyzing stylometric features and focus on modeling other document information using network modeling or other advanced data/text mining and machine learning techniques and for digital document use various information like graphics, emoticons, colors, layouts and so on and handles multiple information sources of heterogeneous nature[7][8].

2. STRUCTURE OF THIS PAPER

In this paper, in Section 3 we cover the related works in this area and classify them based on applications they target and also based on the methodologies they use. In Section 4 we define our project objectives. In Section 5 we describe our methodology, data source used and system overview as well as implementation of the system. In Section 6 we describe the results we obtained from our system. In Section 7 we discuss the results and compare it to results from other similar method. In Section 8 we give the conclusion of our project and summarize our findings and contribution. In Section 9 we discuss future work possibilities for extending the project that we will further explore and recommend some techniques that could be explored to improve results further. In Section 10 contains our references.

3. RELATED WORKS

3.1 Classifying Related Works Based On Application Domains

Authorship analysis has a many potential applications in areas of literature, digital content forensics, program code author, crime prevention, law enforcement, etc. Some major potential applications are in areas of authorship dispute resolution, plagiarism detection, ghost writer detection, ancient literature author identification, law enforcement based ransom note/article/email author identification, etc. and has been of interest to many scholars, scientist, lawyers, law enforcement officers, etc. dealing with such problems in real-life.

Since the area is very broad and there are varied applications, there have been many research works across many disciplines and different related works have focused their system designs and application goals to such specific problems.

We compile major works aimed at real-world problems dealing with authorship analysis and classify them into different domains and list of application it is intended to address in Table 1.

3.2 Classification Of Related Works Based On Methods And Features

A general framework which encompasses most authorship attribution studies before network mining, involved coming up with a set of stylometric features to represent the document and then exploring various combinations of learning algorithms to predict authors for unseen documents. Author identification / attribution task has typically been viewed as a text categorization or classification problem, which starts with data preparation and cleaning, and
then feature engineering (extraction and normalization), representing a document as a feature vector, splitting the data into training and testing and then training classification models to predict authors of unknown papers from within a list of available in the authors from training set.

Table 1: Domain areas and applications of author identification and related research works

<table>
<thead>
<tr>
<th>Application Domain</th>
<th>Problem</th>
<th>Notable References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature Analysis and Disputes</td>
<td>Author Identification/ Authorship attribution</td>
<td>[9][10][11]</td>
</tr>
<tr>
<td></td>
<td>Authorship Disputes</td>
<td>[6] [12] [14] [22]</td>
</tr>
<tr>
<td></td>
<td>Author Verification</td>
<td>[13][15][16][17][18][19][20][21][30]</td>
</tr>
<tr>
<td></td>
<td>Author Profiling</td>
<td>[21][23][24][225][26]</td>
</tr>
<tr>
<td></td>
<td>Detecting Stylistic Inconsistencies</td>
<td>[27] [28]</td>
</tr>
<tr>
<td>Inauthenticity Detection</td>
<td>Plagiarism Detection</td>
<td>[28][29][31][32][33]</td>
</tr>
<tr>
<td></td>
<td>Ghost Writer Detection</td>
<td>[28][30]</td>
</tr>
<tr>
<td>Privacy Evaluation</td>
<td>Bias analysis (e.g.: Double blind conference review setting)</td>
<td>[34][35]</td>
</tr>
<tr>
<td>Digital Content Forensics</td>
<td>Emails Forensics</td>
<td>[36][37][38][39]</td>
</tr>
<tr>
<td></td>
<td>Blog content analysis</td>
<td>[37][42]</td>
</tr>
<tr>
<td></td>
<td>Online message (social media) authorship</td>
<td>[37][40][41]</td>
</tr>
<tr>
<td>Program Code Authorship</td>
<td>Program code authorship forensics</td>
<td>[33][43][44][45][48]</td>
</tr>
<tr>
<td></td>
<td>Code authorship disputes</td>
<td>[44][48]</td>
</tr>
<tr>
<td></td>
<td>Malware author identification</td>
<td>[46][47]</td>
</tr>
<tr>
<td></td>
<td>Proprietary software and legacy code disputes</td>
<td>[33][43][44][45][46][47][48]</td>
</tr>
<tr>
<td>Crime Prevention and Law Enforcement</td>
<td>Ransom Note Forensics</td>
<td>[49][50]</td>
</tr>
<tr>
<td></td>
<td>Cyberspace security</td>
<td>[37][38][40][42][51]</td>
</tr>
</tbody>
</table>
Recent methods using network modeling utilize mainly document attribute information such as the venue where the document was published, date of publication, type of publication, etc. and pair it with some document text based keywords that are extracted. But they do not focus heavily on extracting more patterns from the text that could provide more author characteristics. Then further network modelling, machine learning methods are applied to train, test and finally attribute the authorship.

Based on these different information extraction focus, we can also broadly categorize the into two categories:
1. Document text based approach
2. Document attribute based approach

We found a previous survey that did a very detailed comparison of methods till 2012 [1]. However we did not see any recent surveys attempting to summarize and compare the more recent works. So to give a better overview of major recent developments, we have summarized the recent methods proposed after 2012 till date in Table 2.

Most of the methods prior to 2016 before network embedding for author attribution was introduced were mainly based on linguistic structures and stylometry which focus heavily focus on feature engineering and extraction from the text body of the document and mine and analyze only that data, without analyzing other document attributes like the date it was written, where it was published, type of publication, etc. Among the features extracted, there are mainly 4 kinds of features 1) Lexical features 2) Syntactic features 3) Structural features and 4) Context specific features.

Historically, SVM and PCA have been the most popular techniques applied mainly for feature extraction and processing in both older works as well as recent ones. MDA, LDA and decision tree have remained popular. Bayesian network, genetic algorithm, naive Bayesian, Markov model, term frequency and topic modeling based methods took off sometime in the early 2000s. Recently some research works have explored the usage of neural network based techniques, kNN, heterogeneous network embedding and OBA2.

### 3.3 Discussion of Popular Methods

#### 3.3.1 Stylometric Features Analysis

The area author analysis has strong ties with the field of stylometry based analysis of document to identify a certain style, language usage “fingerprint” of an author. This has even been called “human stylome” by Halteren et. al. who used machine learning methods early on to try to establish that authors have their own unique distinguishing styles [10].

Lexical, Syntactic, Character Based, Structural, Content specific (n grams) are some of the types of stylometric features used for the learning tasks. This is explored in many prior survey papers [56] [57].

#### 3.3.2 Machine learning techniques

Feature learning and stylometry based methods generally rely on machine learning based approaches to identify the author. Some typical features that methods rely on are as follows:

- Stereotypical dialectological variation (e.g.: American versus British English) [1]
- Syntactic constructions that indicate psychological or medical conditions [52]
- Multivariate features based methods. For example instead of one specific term, looking at the collective impact of average word length, vocabulary, etc. [1].

Some simple univariate features have been shown to be insufficient in distinguishing
authors by themselves and recently more complex methods have been proposed using multivariate features [1][53].

Some examples of Machine learning techniques popularly used so far are PCA, distance measures (cosine similarity, KLD distance), SVM, LDA, decision tree, neural network, frequent pattern, k means, EM clustering, naïve Bayes, Bayesian networks.

3.3.3 Text mining techniques

The document text can be represented as vector of word frequencies without focusing on contextual occurrences of these words. This is similar to the Bag of Words model. Similarity or bag-of-words based author identification methods all basically utilize general text mining methods to process the text and identify the author based on terms used, similarity in text sequences and styles, etc. Common words (called function words) like pronouns, articles, prepositions are removed. If context is to be considered, the adjacent words of a word is taken into account (n-gram of words) as a textual feature.

3.3.4 Network Embedding techniques

Organizing and processing large scale data networks have been recently attempted using feature engineering with network embedding. The information is modelled as a network with nodes and edges. Any prediction or identification of missing link information is done by data mining on rest of the network in either supervised or unsupervised manner depending on the dataset and application.

Network embedding can be broadly classified into:

1. Homogeneous network embedding
2. Heterogeneous network embedding

Network embedding techniques are newly emerging area in data mining and most recently [54]. They have been applied to author analysis and attribution problems [34] paired with feature engineering and have shown better results than conventional simply feature engineering or stylometry analysis and text mining based techniques.

In one of the most recent major papers in this area of author identification by Chen et. al. [34], double blind anonymized technical papers were processed using network embedding techniques that were aimed for heterogeneous bibliographic networks. In this setting, both the reviewer and author information are hidden from each other in order to eliminate bias. In this recent work by Chen et. al. [34], heterogeneous network embedding, resulted in better results that the popular feature engineering based techniques. This is one work that utilized non-linguistic data mainly, as it utilized other attributes of the document such as conference venue, date, etc. A major limitation in the work by Chen et. al [34] seems to be imposed by the dataset they used (AMiner citation network), which was severely limited in fields such as references.

In another older work in this area by Hill et. al [55], author identification was done by citation matching using methods from social network analysis, graph theory, and bibliometrics. They used the KDDCup 2003 Physics paper archive and using only intra-database citation alone was able to identify the author between 25% to 45%. This tells us there is a lot of value hidden in the citations, often due to the trend of technical papers to cite their own research group’s previous work. This also shows the significance of incorporating document attributes to gain more information about a document whenever available.

Inspired by these previous research works, in our approach we will be using both types of information (features and attributes) whenever available for better results.
Table 2: The most important studies on authorship analysis  
(a continuation of Bouanani and Kassou[1])

F1: Lexical features; F2: Syntactic features; F3: Structural features; F4: Content-specific features; A1: Authorship identification; A2: Authorship characterization; A3: Similarity detection

<table>
<thead>
<tr>
<th>Year</th>
<th>Reference</th>
<th>Authors</th>
<th>Corpus</th>
<th>Domain</th>
<th>Features</th>
<th>Language</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>[63]</td>
<td>A1</td>
<td>Nonfiction literature, poetry</td>
<td>Literary</td>
<td>F1</td>
<td>Multi-language</td>
<td>SVM, kNN</td>
</tr>
<tr>
<td>2013</td>
<td>[66]</td>
<td>A1</td>
<td>Twitter(short messages)</td>
<td>Social Media</td>
<td>F1</td>
<td>English</td>
<td>SVM, linear kernel</td>
</tr>
<tr>
<td>2015</td>
<td>[65]</td>
<td>A1</td>
<td>Reuters Corpus Volume 1, Guardian daily newspaper</td>
<td>News</td>
<td>F1</td>
<td>English</td>
<td>n-gram classifiers</td>
</tr>
<tr>
<td>2016</td>
<td>[64]</td>
<td>A1</td>
<td>Twitter</td>
<td>Social Media</td>
<td>F1</td>
<td>English</td>
<td>Supervised learning</td>
</tr>
<tr>
<td>2016</td>
<td>[61]</td>
<td>A3</td>
<td>Political news articles</td>
<td>News</td>
<td>F1, F4</td>
<td>Tamil</td>
<td>Context based word embedding</td>
</tr>
<tr>
<td>2017</td>
<td>[58]</td>
<td>A1, A2</td>
<td>Literary genre dataset (Project Gutenberg, PAN 2014)</td>
<td>Literary</td>
<td>F2, F3</td>
<td>English</td>
<td>SVM with linear kernel</td>
</tr>
</tbody>
</table>
4. OBJECTIVE

Our objectives in the project are as follows:

Study related works broadly across all disciplines
1. Classify and understand methods used and their pros and cons
2. Design a system that combines best techniques and is able to utilize more of available data compared to other state-of-the-art methods today. We believe that utilize both the document text and document attributes will provide us more information to disambiguate the author.
3. Implement the system combining text mining and network embedding methods
4. Train and then test and evaluate on a big corpus

5. METHODOLOGY

5.1 Data Source

As a data source for training, testing and evaluating our system we use the ACL anthology network corpus, 2014 [69]. Some statistics related to the ACL AAN data source as follows:

- Number of papers=23766
- Number of authors=18862
- Number of venues=373
- Number of paper citations=124857
- Number of author collaborations=142,450

5.2 System Overview

After the initial document processing and, cleaning we break down the system into the following three phases:

PHASE I: Feature extracting/text mining: Using n-gram, stop words, stemming, etc

PHASE II: Network embedding and Training: Using Modified LINE (Tang et. al)[54]

PHASE III: System Test and Evaluation: Cosine Similarity Test over vector values for document with unknown author

Our system overview is shown in Figure 1.

5.3 Phase I: Feature Extraction And Text Mining

In the first phase we take the data subset that we extract from the corpus and did cleaning and preprocessing to remove artifacts on. Some initial steps before running feature extraction are data selection and parsing, followed by data cleaning and reformattting if needed.
Figure 1 shows the top level graphical representation of the author analysis/identification problem.

After this we perform feature extraction by frequent pattern mining. We use n-gram frequent pattern and consider unigram, bigram and trigram patterns and perform stop word removal and also allow stemming. For implementing this we use the nltk packages in python and specifically nltk collocation pattern to perform collocated n-gram extraction. For stemming PorterStemming has been used which is also a python nltk package. Besides the n-gram key phrases we extract and compile the links between other features such as author-paper pairs (only for train set), reference list and citation networks between authors, co-author network, etc.

Broadly we have mainly two kinds of features extracted that either describes the paper or the author. The paper features and author papers that we retrieved, processed and used as inputs are as follows:

A. Paper Features:
   - Document title (Parsing, retrieval from metadata)
   - Venue of publication (Parsing, retrieval from metadata)
   - Key Phrases for document (python nltk consecutive n-gram extraction, stop word removal, stemming)
   - References/Citations by paper (Parsing, retrieval from data source)

B. Author Features:
   - Author frequent phrase distribution (python nltk consecutive n-gram extraction, stopword, stemming).
   - Author Citation Network (Parsing, retrieval from metadata)
   - Author Collaboration Network (Parsing, retrieval from metadata)

We then separate the train and test sets after this, which both have features extracted and are ready for training or testing.

5.4 Phase II: Network Embedding And Training

For using the input data and to create network embeddings and training we use a heterogeneous network embedding method called LINE [54] that we modify for this application. We changed the original LINE embedding methods[54] to customize it for author identification applications and also added supports for more meta paths based training if meta paths are available. For this project we modified LINE working together with 2nd author of paper who is a current graduate student at UIUC (Meng Qu).
In the modified LINE, there are more features and more meta paths for 1\textsuperscript{st} and 2\textsuperscript{nd} order evaluations to support the paper and author features that we described in earlier sections.

Modified LINE inputs that our system takes are the features extracted from Phase I that includes the author and paper features as shown in Figure 2. Modified LINE will train this and provide an author embedding vector list and paper embedding vector list for each author and paper mentioned in the train set. We run Modified LINE with parameter set to take 100 million samples (sample=100), with size of embedding vectors set to 50 and 4 threads used for running it in parallel.

5.5 Phase III: Author Prediction For Test Set
In Phase III we take the input embedding vectors from Phase II, which consists of the a) author embedding vector list and b) paper embedding vector list and use this as our trained model. We then take an input test paper with unknown author and generate embeddings for the test papers. We then compare the embedding vectors for the trained author models and test paper embedding values. We use cosine similarity over the embedding vector values to detect the distance between a possible author and the test paper.
We then obtain the results for each test paper in this way and rank the top 10 most similar author from our trained model, based on closest cosine similarity distance seen.

We define accuracy in our system based on whether one or more actual author(s) appears on this top 10 ranked list after testing. If it falls in this list, it is considered a “hit” else a “miss”.

\[
\text{Test set Accuracy} = \frac{\text{No. of hits}}{\text{No. of paper}} \times 100\%
\]

6. RESULTS
We randomly sample a subset of data that includes 1216 authors and coauthors and get a subset generally of the size of 3648 papers for testing and evaluation. We sample it down to a smaller subset as we find that the current dataset is too big and requires too much computation which we did not have access to. We run our tests in this subsystem and obtain an average accuracy result of 66.66%.

7. DISCUSSION
As our dataset is a fairly new 2014 dataset, there was not any other works that used this for author identification that we could directly compare the results against.
We however briefly analyze and compare the results for some related methods that either use network embedding as well or just heavily utilize citation information, the combination of both which our system is able to support well.

Comparing our results with previous works such as KDD 2003, Hill et.al [55] which uses just citation information and obtains 40% average (ranging from 27 to 85% accuracy) we find that our system performs better. But they obtained these results on a data set of 9k and more, and a direct comparison is not possible.

Another closely related method is the Task-Guided and Path-Augmented Heterogeneous Network Embedding for Author Identification by Chen et. al [62]. It obtained an accuracy of 0.83 for MAP@10 and 0.92 for Recall@10. Though the results seem better, this is obtained only on a subset of 1k samples from a dataset having 1 million entries. So direct comparison is not possible as we find that with increase in samples, the possibility of false positives and false negatives increases dramatically in our system.

For direct comparison, further evaluation using same datasets and sample sizes maybe needed, but some of the previous datasets are outdated and may not serve as a good example compared to more recent datasets that we found to be better organized, more complete without much missing information and required less cleaning.

8. CONCLUSION

In this project we contribute with both organization and classification of the related works in the area, developed in a scattered manner spanning across multiple disciplines such as text mining, data mining, stylometry, machine learning, computational linguistics etc. Providing a broad overview of this area, classifying methods based on the applications targeted as well as based on the technical methods and features that they relied on we hope to provide a detailed one stop reference to all the research works till date in various scattered areas.

After this detailed survey and classification of this area, we analyze the kinds of data sources used and information extracted in previous methods and found they focused mostly on just document text or document attribute. Most methods in text mining and stylometry can also go hand in hand with generating features for heterogeneous networks and training it. So we designed and implemented a system that utilized both document text and document attribute when available, and combined text mining based features extracted with network embedding and training. Our system contains more network meta paths and allows a more multi-faceted analysis for author identification. The results from our system are comparable to some previous methods with similar evaluation data set sizes, but does not outperform other methods that run more computationally complex algorithms using heterogeneous network mining methods.

We hope to further analyze each meta path and allow varied weightage in the future to further improve results and also do clustering of data into small similar subsets with potential authors (authors from same area) before running the algorithm, to make it a more computational light and scalable system and reduce the possibility of false positives and false negatives that can occur, since the search space can reduces considerably by clustering. We discuss this future improvements to our system that we are currently building in the future work section as well. In summary, we have shown that combining text mining and heterogeneous network embedding methods and utilizing all aspect of the information about a document can build a more flexible and powerful system that
evaluates various features and identifies an author without being limited to only certain features.

9. FUTURE DIRECTION

We believe topic mining and clustering into similar and smaller subsets of authors to test a paper with an unknown author against could considerable improve the efficiency of our system and as well as reduce the computation power and improve the speed. We have already started exploring topic mining using SegPhrase[70] on the big corpus for this, but SegPhrase is computationally demanding if data size if big. However, after the first initial run of SegPhrase and segmentation of the train corpus, topics could be incrementally updated as newer papers are added further. SegPhrase could also improve the frequent patterns generated. Due to size of corpus and how the embedding and training handles it, powerful systems may sometimes be required. Classifying the data and testing it with a cluster can help improve speed.

Another possible modification to our system is changing LINE-based embedding to further take varied weighted values for various features instead of weighing all features equally. This would require further evaluation of which meta-paths contain the most reliable information by separately testing each meta-path and based on accuracy of authors it obtains, assigning weights to the metapath. We believe this is essential in improving performance and utilizing as much metapath information when available while also not just picking and limiting to certain meta paths like most previous works do, as there is no guarantee we will always have information to generate those meta-paths.

10. REFERENCES


[29] Stamatatos, E. Intrinsic plagiarism detection using character n-gram profiles. threshold, 2(1,500), 2009


forensics: Can we track code to its authors?
Computers & Security, 12(6), 585-595, 1993


binary code authorship attribution. Digital Investigation, 11, S94-S103, 2014

In Proceedings of the 4th ACM workshop on Recurring malcode (pp. 73-78). ACM, 2006

International Conference on (pp. 243-248). IEEE, 2007


In International Conference on Intelligence and Security Informatics (pp. 59-73). Springer Berlin

[52] C. Brown, M. A. Covington, J. Semple, and J. Brown, “Reduced idea density in speech as an
indicator of schizophrenia and ketamine intoxication,” in International Congress on

[53] D. Hoover, “Another perspective on vocabulary richness,” Computers and the

(WWW’15), Florence, Italy, 2015

[55] Shawndra Hill and Foster Provost. The myth of the double-blind review?: author
identification using only citations. SIGKDD Explor. Newsl. 5, 2, 179-184, 2003

[56] Cristani, M., Bazzani, L., Vinciarelli, A., Murin, V.: Conversationally-inspired
Stylometric Features for Authorship Attribution in Instant Messaging. ACM Multimedia, 2012

Profiling Techniques”, International Journal of Applied Engineering Research, Volume 11,

[58] de Roc Boronat, C., & Wanner, L. On the Relevance of Syntactic and Discourse Features

AND PREPROCESSING ON AUTHOR IDENTIFICATION, Anadolu University Journal
of Science and Technology A-Applied Sciences and Engineering 2017,Vol 18-1, pg 218-224,
2017.

Learning. In Machine Learning and Applications (ICMLA), 2016 15th IEEE International

Finding Semantic Relationship between Words in Tamil Language. Indian Journal of Science
and Technology, 9(45), 2016.

[62] Chen, T., & Sun, Y. Task-Guided and Path-Augmented Heterogeneous Network Embedding


