Defining Forensic Authorship Attribution for Limited Samples from Social Media

Robert Frye, David C. Wilson
University of North Carolina at Charlotte
9201 University City Boulevard
Charlotte, NC 28223

Abstract
Forensic author attribution (FAA) has its roots in stylometry research and has most recently grown to include intelligent systems approaches. But to be useful in the legal context, there are additional considerations for the design and development of such systems. In this paper, we provide a survey of the existing research in author attribution, and we propose research characteristics required to move the science towards acceptance in forensic applications, as required in court. The paper goes on to present our initial research on optimizing a common n-gram approach in the context of the PAN12 author attribution dataset and a large Twitter dataset. We conclude with a discussion of our findings and future work toward open questions in FAA.

Introduction
Accurately identifying individuals involved in Internet-based criminal activities based on stylistic cues in their communications is an important technical challenge for law enforcement. Approaches employ techniques for Authorship Attribution (AA) (Stamatatos 2009), and more generally Stylometry (Afroz et al. 2014). Application of Stylometry and AA techniques to criminal cases is known as Forensic Authorship Attribution (FAA) (Rocha et al. 2017), and carries additional requirements for use in this context. A key ruling by the United States Supreme Court has limited the use of evidence gained by authorship attribution techniques to the investigative role only (Chaski 1997). In the 1993 ruling of Daubert v. Merrell Dow Pharmaceuticals, Inc., the standard for expert testimony was clarified to state that it must be based on techniques shared by well-known and established sciences, like physics, biology, and chemistry. The Court-defined characteristics include:

• Calculation of known or potential error rate
• Methods of empirical testing
• Standardized procedures for applying the techniques involved in the science
• Peer review and publication
• General acceptance by the scientific community

Meeting the Daubert criteria is a significant challenge for AA research. Thus AA approaches have typically been used as investigative tools, but have not generally been accepted as a forensic science as defined by courts (Chaski 1997).

In order to help address the issue of AA acceptance in the forensic context, our research is investigating improved techniques for the task of accurately identifying authors of short texts from the Internet. In particular, our goal is to refine FAA techniques and practice to meet requirements for court use. This paper begins with an overview of AA research, with a focus on recent contributions to forensic short text analysis. It goes on to describe our proposed framework for an ensemble approach to FAA, which incorporates requirements toward satisfying the Daubert criteria for forensic scientific testimony. The paper then presents our initial work in n-gram-based AA methods, with a discussion of results and conclusions.

Related Work
The short nature of messages exchanged on modern social media makes identifying bad actors more difficult than with longer texts. Author attribution techniques have been available since at least 1962 (Ellegård 1962), but the earliest methods dealt with large training and test corpus sizes and limited candidates. In the 1990s, a minimum of 500 words was required for attribution with an acceptable degree of accuracy (Layton, Watters, and Dazeley 2010), and although much research has been completed to reduce this minimum sample size, identifying authors of short texts like Twitter messages remains difficult.

Rudman (1997) identified over 1,000 papers discussing potential style markers and features. Here, we focus on those which have been specifically evaluated for the task of short text classification. Stamatatos (2009) characterized AA methods as either machine-learning-based or similarity-based. In machine learning AA, classifiers are constructed from known writings of candidates and are then applied to documents from unknown authors. In similarity-based AA, a calculated metric describes the distance between an unknown document and known works of candidates, usually combined into one document for each author. Koppel, Schler, and Argamon (2011), for example, employed a similarity-based approach using 2,000 words of known text to identify the authors of 100-500 word blog snippets in their
corpus. They also identified four factors which affect AA precision: # candidate authors, size of unknown text, size of known text, and threshold score for an author to be declared unknown.

Support Vector Machines (SVM) are a commonly employed machine-learning approach. De Vel et al. (2001) and Diederich et al. (2003) first applied SVM in AA for email and German newspapers, respectively, and numerous other researchers have continued refinement. For example, Qian et al. (2014) included SVM and character n-grams in a tri-training ensemble AA approach. To address issues with large training data and high-dimensional feature sets, Wu (2012) developed a variant of SVM called Power Mean SVM (PMSVM).

AA research commonly employs n-grams as enabling features, for example in SVM classification (Afroz et al. 2014; Brocardo et al. 2013). Character n-grams are subsets of a text string consisting of n characters. Another common variant is to employ word-level n-grams. Kešelj et al. (2003) use this approach to find k-nearest neighbors in the Common N-Grams (CNG) approach (Kešelj et al. 2003), and to calculate similarity between text samples (Stamatatos 2007). Sapkota et al. (2015) examined subsets of character and word n-gram types in AA content-specific and cross-domain corpora, and found using character n-grams for classification was more accurate than using word n-grams. Brocardo et al. (2013) achieved promising results with n-gram based supervised learning for author verification. (Layton, Watters, and Dazeley 2010) employed an n-gram approach for analysis of Twitter data, with accuracy over 70% for 3 ≤ n ≤ 6 and author profiles of 120 tweets. Parts-of-speech (POS) tag n-grams may also be substituted for word n-grams and used for classification. (Subramanya, Petrov, and Pereira 2010; Rocha et al. 2017).

Koppel, Schler, and Argamon (2009) describe an ensemble TF/IDF approach for AA using the following 3 variants: content tfidf – restricted to content words; content idf – binary idf restricted to content words; and style – tfidf on stylistic features (function words and strings of non-alphabetic, non-numeric characters). For each variant, cosine similarity was used to compare snippets, features, or sets of features. Compression-based methods for AA (e.g., (Rocha et al. 2017)) are based on the idea that texts from the same author are more efficiently compressed than texts from different authors. Rocha et al. (2017) implemented the PPM-5 variant of Prediction by Partial Matching (PPM) compression and found better performance than PMSVM for fewer than 500 training tweets.

**Forensic Authorship Attribution**

FAA differs in important ways from AA. Rocha et al. (2017) notes, “In contrast to other authorship attribution tasks found in the broader world of signal processing like active authentication, forensic authorship attribution does not assume a claimed identity before analysis to verify a match.” In FAA, we assume the unknown author of a text may be one or more authors from a known set, or may be a person for whom no known texts have been attributed. In other words, we may not have a sample of the actual author’s texts with which to compare our text of unknown origin, presenting unique challenges beyond those present in traditional AA.

For example, much of early AA research had large, well-known, readily available datasets as observed in the works of (Merriam 1980) (Shakespeare) and (Burrows 1989) (Jane Austen). For short texts in modern communication, such a substantial per-author corpus is not as readily available. To alleviate this problem, (Mikros and Perifanos 2013) merged all tweets by an author in each training and testing set into unified sets with positive results; however, their smallest author sample had 3,692 words, exceeding the 500 word minimum for AA established in the 1990s (Layton, Watters, and Dazeley 2010).

By 1997 approximately 1,000 style markers for AA had been proposed (Rudman 1997), and in the 20 years since, numerous combinations of different style markers have been rejected as best able to determine unknown authors. In spite of this, the Qsum technique of (Morton and Michaelson 1990) was used in several court cases to cast doubt about true authors of confessional statements, including the pardon of Nicky Kelly in April 1992 by Ireland. Numerous researchers later evaluated Qsum and deemed the technique unreliable (Holmes 1998), so while these first attempts at FAA produced results favorable to some parties, they failed to meet the Daubert standard through lack of general acceptance in the scientific community. Whether as researchers we hope to have our work used in court or as an investigative tool, in order for FAA to advance, inclusion of the Daubert requirements could move FAA towards general acceptance by legal systems and the scientific community. Rocha et al. (2017) defined other specialized considerations for FAA, as follows. First, testing sets are difficult to control, and may be limited to one small sample. Second, training methods must tolerate noise, as data is often uncontrolled and noisy. Third, a rigorous process must be used, for accuracy, efficiency, and legal considerations. Fourth, error rates must be discovered and recorded for each algorithm, to satisfy legal requirements. Fifth, authors may attempt to mask their true identity.

**Ensemble AA Approach for FAA**

Rocha et al. (2017) described a generalized framework for forensic AA research with a Twitter corpus. Here we propose an extended perspective on this framework to include ensemble FAA and to help satisfy the Daubert requirements. Ensemble-based approaches leverage combinations of individual approaches to provide better overall performance. For example, (Burke 2002) described hybrid approaches for recommender systems, and we propose to adopt the same ensemble concepts towards creating better AA techniques. The potential for ensemble approaches was noted in (Rocha et al. 2017), suggesting “...combining the output of the PMSVM, Random Forest, SCAP and PPM models to ideally arrive at a more accurate result, compared to examining each model in isolation.” Our proposed approach follows, with extensions from (Rocha et al. 2017) noted in italics:

1. From social media, extract messages of suspected authors to serve as training data.
2. Pre-process messages as required by the selected AA algorithm(s).
3. Pre-process the message(s) of unknown origin in the same manner as was used to pre-process the known messages.
4. Process the dataset with the TF-IDF-based ensemble approach of (Koppel, Schler, and Argamon 2009).
5. Extract features to use for classification (word, character, and part-of-speech (POS) n-grams, as well as other lexical and syntactic features may be used).
6. Train multiple classifiers using extracted features (SVM, random forests, compression-based).
7. Use trained classifiers to identify the author of the message of unknown origin.
8. When a majority of classifiers agree on one result, report the agreed upon author as the “correct” author. When no agreement is reached between the classifiers, report the author identified by the most accurate classifier (as supported by the most recent research for each approach) as the author.
9. When the ensemble TF-IDF and ensemble classifiers agree, accept the result as correct. When they do not agree, accept the result of the classifier ensemble.
10. Calculate precision, recall, accuracy, and (actual or estimated) error rates.
11. Evaluate and report results.

**Initial Study: Common N-Gram FAA**

The (Stamatatos 2007) approach to authorship attribution uses a concatenation of the training texts for any given author to produce pairs of character n-grams and their normalized frequencies of occurrence within the training profile. The top $L$ pairs form the profile for a given author’s writing style. A profile for each test sample is then created and compared to the author profile for each candidate author using a dissimilarity function. A k-nearest neighbor (knn) approach is then applied to predict the author, with $k = 1$. Because this application of CNG uses character-based n-grams, the method is language-independent and also captures the author’s use of uppercase and lowercase letters, as well as lexical and syntactical features. As social media platforms are available in numerous languages around the world, CNG is highly appropriate for use in FAA of social media.

In spite of these advantages, CNG does require proper tuning of two system parameters. The first is the size of $n$, the number of characters in the n-gram, with the best practice defined as $3 \leq n \leq 5$. The Stamatatos CNG method also requires the tuning of the profile size $L$, with best results generally found where $1000 \leq L \leq 5000$. Stamatatos also identified four potential similarity measures for comparing author profiles, with $d_1$ and $d_2$ having roughly equivalent results, so we used the $d_2$ similarity measure in our evaluation.

**Improving CNG**

With these limitations in mind, we optimized the CNG algorithm implemented by Potthast et al. (2016) to generate a programmatic value for $L$ based on average $L$-value across training profiles, reasoning we can calculate an average $L$ and only consider the top 10%, 25%, or 50% of the profile. We applied our modified version of this CNG implementation to the PAN12 dataset, with $n = 5$ and $L$ optimized to the top 10%, 25%, and 50% of n-gram/frequency pairs. In each test case average accuracy improved, with the largest improvement found at the top 25% of $L$, where average accuracy improved from 81.88% to 87.80%, when compared to the original implementation with $n = 5$ and $L = 2000$.

After successfully replicating the previous findings with the PAN12 dataset to verify our implementation, we downloaded a Twitter corpus with 10,580,280 messages from 6,311 authors, consisting of data files with anonymized author IDs, messages with URLs and retweets removed, and a separate parsing of the messages into their part-of-speech (POS) tags. We compared the accuracy of the original CNG algorithm to our optimized CNG (OCNG) algorithm with this Twitter dataset. For comparison, we used the CNG algorithm as implemented by Potthast et al. (2016), with $n = 5$ and $L = 1000$ and compared this to our OCNG. For optimization, we calculated the average size of $L$ across all training samples and set $L$ equal to 25% of the average. We created 10-fold randomized datasets from the Twitter corpus, with training set sizes for each author of 30, 45, 60, 90, and 120 tweets per author. We selected random tweets from each author with 1, 2, 4, and 6 tweets in an unknown testing file for each of our candidate authors, and unknown tweets were selected from messages excluded from the training data. We did not include training samples for additional authors, and only built training profiles from our candidate authors. Our results are summarized in Figure 1.

![Figure 1: Using OCNG for authorship attribution of Twitter Data, for 5 and 10 candidate authors, with training set sizes of 30, 45, 60, 90, and 120 tweets.](image-url)
set. We performed multiple two factor analysis of variance (ANOVA) tests to compare the results of the CNG and OCNG approaches as the number of unknown tweets and number of training tweets changed. The results were interesting, but did not yield significantly different results between approaches. In fact, the original algorithm appears to perform slightly better in some cases. However, detailed examination of the significance p-values for two-factor ANOVA across changing different numbers of candidates showed a potential avenue for future research. The number of candidate authors shows lower impact on significance as training sample sizes increase, so with larger training profiles, we should be able to accurately identify correct authors from among larger candidate pools. We also notice with larger unknown sample sizes, training profile size is more significant, as seen with the 6 tweet unknown line bending under the statistical significance value of $p \leq 0.05$. This information may help to establish an important evidence threshold for presenting findings in court. For example, further research may establish that a sample of 6 tweets from unknown users and 90 tweets from a suspect may be enough to match or rule out a candidate with 95% confidence. Future research will investigate open questions in AA and FAA, such as identifying the most effective features and different combinations of underlying methods to improve accuracy.

**References**


