Introduction

Researchers have studied ways of identifying anonymous authors for over a century. Word-length distributions were used to determine who among Shakespeare, Marlowe, and Bacon the author of disputed plays was. Perhaps the most famous application of linguistic forensics came when Ted Kaczynski, dubbed the “Unabomber,” was identified by family members who read his 35,000 word manifesto. Modern applications in the cyber-world include identifying anonymous posters on criminal network message boards and tracing email-based threats or scams. These problems differ in difficulty from traditional identification problems—unlike comparing a play to the attributed works of Shakespeare, often in the cyber-world we do not have a labeled set of messages with which to compare our anonymous one. Instead, it is more likely that we must use another source, leading to the need for further studies and solutions for cross-domain authorship analysis.

Related Work

Research has increased focus on tackling shorter and shorter data sets, increasing the difficulty of the problem. Eder's 2013 study showed that word samples of 1000 words were unreliable. Authorship identification has progressed from believing problems with text lengths under 1000 words to be difficult to tackling sets of under 250 words[][][][]. Burrow's Delta procedure was used to distinguish authors of texts greater than 1500 words but was demonstrated useful for reducing the likely candidate list for texts as short as 100 words [].

Hirst and Feiguina used syntactic label bigrams to distinguish the Bronte sisters with accuracies of around 90% even for training sets as small as 200 word blocks. In a cyber setting, Kucukyilmaz et al. achieved 99.7% accuracy in predicting authors of chat messages from a set of 100 possible matches. Solorio et al. achieved 62% accuracy predicting web forum posts from 100 authors. Zheng et al. used SVMs to achieve accuracies of between 70 and 95% on forum posts, based on the number of suspect authors (5 to 20) and the number of messages per author (10 to 30). Bhargava et al. used between 200 and 300 tweets per author to classify between 10 to 20 authors with accuracies around 80 to 90% . Layton et al. found that their model for classifying tweet authorship improved non-significantly after 120 tweets per author. Brocardo et al. investigated email authorship verification for 25 to 100 message blocks per author and just 250 or 500 characters per block.

In addition to dealing with the shorter lengths of forum posts and tweets, studies on electronic texts have also had to deal with nosier data in the form of misspellings, improper grammar, missing punctuation, acronyms, emoticons, and Internet lingo. Many studies, like MacLeod and Grant which incorporated features such as 'lower case I', 'G clipping', 'accent stylization', and 'whole word number homophone substitution' and Layton et al. which preprocessed hashtags and @reply symbols in tweets, have focused on tackling this issue.

Despite the difficulty presented by this noise, these preceding problems are all considered to be relatively simple compared to real world uses. While these problems have progressed from distinguishing a handful of authors each with lengthy, well-formed texts to distinguishing up to 20 or so authors with shorter and noisy electronic texts, a true authorship problem, as maintained by Koppel et al., may have thousands of candidate authors that may or may not include the correct author and very limited texts for each candidate. Koppel et al. use space-free character 4-grams for features and iterate combinations of features to achieve a coverage of 42% and precision of 93%, given a 500 word sample to attribute to 1000 possible authors. Narayanan et al. were able to predict the author of an unknown blog post from 100,000 candidates with 20% accuracy.

In a break from previous studies, Narayanan et al. do not control their test documents for topic, thus addressing a cross-context setting. But as they note, their work, which uses blog posts to classify an unknown blog post, does not address a cross-domain setting. With one of the main problems in authorship analysis being the establishment of a verified corpus with which to compare an unknown document, it is easy to see how a cross-domain solution would be useful. When linking a suspect to anonymous forum posts, for example, there will not be any labeled posts with which investigators can use to create a corpus for comparison. This paper identifies this problem and explores the feasibility of a cross-domain solution.

Between 1978 and 1995, a man named Ted Kaczynski killed three people and injured 23 others with homemade bombs. Nicknamed the Unabomber, he was finally caught when his sister-in-law and brother read his 35,000 word manifesto, *Industrial Society and Its Future*, and recognized the writing style, particularly the use of the phrase “cool-headed logicians”.