Managing Quality, Identity and Adversaries in Public Discourse with Machine Learning

A Thesis
Submitted to the Faculty
of
Drexel University
by
Michael Brennan
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
November 2012
Dedications

To my advisor, Dr. Rachel Greenstadt. If she had not come to Drexel University, I am certain that this document would not exist. Rachel helped me understand how I could apply my skill and interest in Computer Science towards greater good in the world. All students deserve an advisor that is as supportive and invested in them as Rachel has been to me.

To my family and friends, whose support keeps me going. In particular to Mom and Dad, who taught me how to be hard working, compassionate, and most importantly, be myself. To Megan Webb, who understands me, supports me, and believes in me. And to all of my friends who have lent an ear as I navigated my way through graduate school and continue to navigate my way through life.

To my undergraduate advisor, Dr. Christopher Rasmussen, who fostered my interest in research and experimentation. To my former advisor, Dr. Frank Lee, who believed that this day would come as long as I found the right path, and gave me the space and support to find it.

To my committee members, Dr. Anthony Joseph, Dr. William Regli, Dr. Dario Salvucci, and Dr. Doug Tygar. I am honored to have you join me as I close this chapter and begin the next one.
Acknowledgments

My supporting authors for much of this work, including Rachel Greenstadt, Sadia Afroz, Stacey Wrazien, Diamond Bishop.

I was also a supporting author to Sadia Afroz for the work covered in Section 3.4.8 and to Diamond Bishop for the work covered in Section 3.5.

I am also grateful to those who continue to carry this work forward including Sadia Afroz, Aylin Caliskan, Andrew McDonald Bekah Overdorf, Ariel Stolerman, and the other members of the Privacy, Security and Automation Lab at Drexel University.

This work was supported by Intel through the ISTC for Secure Computing, DARPA through grant N10AP20014, and the National Science Foundation.
# Table of Contents

**List of Tables** ........................................... vii  
**List of Figures** .......................................... viii  
**Abstract** ................................................... x  
1. **Introduction** ............................................. 1  
   1.1 Related Work ............................................... 3  
   1.2 Contributions .............................................. 4  
2. **Managing Quality and Topic Identification with Machine Learning** .......... 5  
   2.1 Management and Scaling Issues in Crowdsourced Quality and Topic Identification . . 6  
   2.2 Automation of Crowdsourced Ratings on Slashdot ................................. 8  
      2.2.1 Related Work: Slashdot, Recommender Systems, Online Communities .......... 9  
      2.2.2 The Slashdot Community ................................ 9  
      2.2.3 Approach, Data, Features .................................. 11  
      2.2.4 Evaluation .................................................. 15  
      2.2.5 Understanding the Results .................................... 18  
   2.3 Automation of Crowdsourced Topic Identification on Twitter ...................... 19  
      2.3.1 Related Work: Twitter, Topic Modeling ..................................... 21  
      2.3.2 Approach, Data, Features ...................................... 22  
      2.3.3 Evaluation .................................................... 25  
      2.3.4 Understanding the Results ...................................... 30  
   2.4 Future Work ................................................... 32  
   2.5 Conclusions: Can Machine Learning Improve Quality and Topic Identification Online? 33  
3. **Managing Identity and Anonymity with Stylometry** ................................. 34  
   3.1 Scaling and Anonymity Issues in Managing Identity with Stylometry ............... 36  
   3.2 Current Trends in Authorship Recognition ........................................... 37
3.3 Adversarial Stylometry ........................................ 55
3.3.1 Early Work in Adversarial Stylometry ...................... 56
3.3.2 The Role of Stylometry in Privacy and Anonymity .......... 61
3.3.3 Anonymous Speech and Public Discourse ................... 64
3.3.4 Detecting Deception in Stylometry ......................... 65
3.4 Circumventing Authorship Recognition ........................ 66
3.4.1 Methodology .................................................. 66
3.4.2 The Brennan-Greenstadt Corpora ............................ 68
3.4.3 Circumvention Approaches .................................. 69
3.4.4 Methods and Feature Sets ................................... 70
3.4.5 Evaluating Manual Circumvention Approaches ............... 73
3.4.6 Evaluating Circumvention By Machine Translation .......... 78
3.4.7 Measuring Writing Style Modification ....................... 82
3.4.8 Detecting Deception in Writing Style ....................... 84
3.4.9 Discussion and Future Work ................................. 88
3.5 Stylometry in Social Media ................................. 90
3.5.1 Related Work: Short-Form Message Analysis and Authorship Recognition ........... 92
3.5.2 Methodology and Data ....................................... 93
3.5.3 Evaluation .................................................... 94
3.6 Conclusions: How Does Stylometry Affect Online Identity and Anonymity Management? 99
4. Building Data Sets and Obtaining Baselines for Adversarial Research ........ 101
4.1 Challenges in Obtaining Adversarial Text Data ................ 102
4.1.1 Related Work: The Mechanical Turk Approach .............. 102
4.2 Creating a Stylometry Corpus with AMT: The Extended-Brennan-Greenstadt Corpus 103
4.2.1 Quality Control in the Extended-Brennan-Greenstadt Corpus .......... 104
4.2.2 Evaluating the AMT-based Corpus .......................... 105
4.3 Creating a Crowdsourced Ratings Baseline with AMT ................ 105

TABLE OF CONTENTS
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Overall Accuracy Chart.</td>
<td>15</td>
</tr>
<tr>
<td>2.2</td>
<td>Bad/Neutral/Good Confusion Matrix.</td>
<td>16</td>
</tr>
<tr>
<td>3.1</td>
<td>Baseline and Extended Feature Sets From Writeprints^1.</td>
<td>46</td>
</tr>
<tr>
<td>3.2</td>
<td>Identification Task (% Accuracy) From Writeprints^1.</td>
<td>49</td>
</tr>
<tr>
<td>3.3</td>
<td>Similarity Detection Task (% F-Measure) From Writeprints^1.</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>Results for models 1 and 2 from Clark and Hannon^2.</td>
<td>55</td>
</tr>
<tr>
<td>3.5</td>
<td>The methods and feature sets examined in this study.</td>
<td>71</td>
</tr>
<tr>
<td>3.6</td>
<td>Writeprints-Static Feature Set. adopted from the Writeprints approach^1.</td>
<td>74</td>
</tr>
<tr>
<td>3.7</td>
<td>The table shows performance of different feature sets in detecting regular and adversarial writing samples. The Writeprints feature set with SVM classifier provides the best performance in detecting deception.</td>
<td>86</td>
</tr>
<tr>
<td>3.8</td>
<td>This table shows the features that discriminate deceptive documents from regular documents. The top discriminating features according to Information Gain Ratio are mostly function words.</td>
<td>87</td>
</tr>
<tr>
<td>3.9</td>
<td>Imitated document prediction result: A classifier trained on imitation and obfuscation data set can detect imitated articles from the International Hemingway Imitation contest and the Faux Faulkner contest.</td>
<td>87</td>
</tr>
<tr>
<td>3.10</td>
<td>Feature Set Part I</td>
<td>95</td>
</tr>
<tr>
<td>3.11</td>
<td>Feature Set Part II</td>
<td>96</td>
</tr>
</tbody>
</table>
List of Figures

2.1 * = Estimate. Data and graph created by ITU World Telecommunication / ICT Indicators database. .................................................. 7

2.2 Precision, Recall, and F-Measure for the top 10 trends of each day in our data set using Mallet. .......................................................... 26

2.3 Precision, Recall and F-Measure for the top 10 trends of each day in our data set using the WCNB classifier. .......................................... 27

2.4 Results for MNB vs. WCNB, averaged across all three days. .................. 28

2.5 Comparing the addition of public non-trending tweets. Averaged across all three days. .... 28

2.6 Results for using more than just word frequencies in tweets. Averaged across all three days... 29

2.7 Comparing the original and cleaned tweets averaged across all three days. ............ 30

3.1 An example from Clark and Hannon demonstrating how values are calculated to compare two known sentences to an unknown sentence. In this case the unknown sentence is correctly attributed to author X because of the highly salient feature of choosing the word “verdant”2. .................................................. 53

3.2 A gradual decline in accuracy in distinguishing between two authors is observed as the most salient features are removed before building the classifier. But when text is compared to an obfuscation passage by the same author the accuracy drops much faster4. .... 65

3.3 As the number of modified features increases, the difference in accuracy in distinguishing between two authors increases at a slower pace. This shows that modifying more features can be an effective way to hide true authorship and avoid being detected4. ............ 65

3.4 The Writeprints approach versus Writeprints-Static on the Brennan-Greenstadt corpus. 73

3.5 Baseline accuracy. ................................................................. 75

3.6 Detection of obfuscation attacks. ............................................... 75

3.7 Accuracy in imitation detection. ................................................ 77

3.8 Success of imitation attacks. ................................................... 77

3.9 Google translation results. ........................................................ 79

3.10 Bing translation results. .......................................................... 79

3.11 Changes made in obfuscation and imitation passages in the Extended-Brennan-Greenstadt corpus. The amount to the right of the y-axis represents increases of a feature, and to the left are decreases. .................................................. 82
3.12 Accuracy when varying the bundle size. ............................................ 97
3.13 Accuracy when varying the number of users. ................................. 97
3.14 Blog Features vs. Tweet Features. .................................................. 98

4.1 Brennan-Greenstadt and Extended-Brennan-Greenstadt with the Basic-9 Neural Network approach. ................................................. 106
4.2 Brennan-Greenstadt and Extended-Brennan-Greenstadt with the Synonym-Based approach. ............................................................ 107
4.3 Brennan-Greenstadt and Extended-Brennan-Greenstadt with the Writeprints Static SVM approach. ...................................................... 108
4.4 Assessing the accuracy of quality assessments by all workers and only those who passed quality control. ............................................ 110
4.5 Assessing the accuracy of quality assessments by AMT workers. ........ 111
4.6 Assessing the accuracy of AMT worker submissions based on quality rating. ................................................................. 112
4.7 Assessing the accuracy of AMT worker submissions based on social spam rating. .............................................................. 113
4.8 Assessing the precision and recall of AMT workers in identifying social spam when varying the decision threshold. .......................... 115
Abstract
Managing Quality, Identity and Adversaries in Public Discourse with Machine Learning
Michael Brennan
Rachel Greenstadt, Ph.D.

Automation can mitigate issues when scaling and managing quality and identity in public discourse on the web. Discourse needs to be curated and filtered. Anonymous speech has to be supported while handling adversaries. Reliance on human curators or analysts does not scale and content can be missed. These scaling and management issues include the limits of crowdsourced comment rating systems, flaws in crowdsourced topic identification systems, and the identification or anonymization of large numbers of authors. Scaling rating and topic identification systems results in missing relevant quality discourse. Authorship recognition without automation is time consuming, costly, and does not scale. Using machine learning to automate authorship recognition gives rise to serious privacy and anonymity concerns, and deception in writing style can be difficult to detect. This work replicates comment ratings and topic identification of crowdsourcing systems that currently rely on human participation to be effective. This work also demonstrates the ability of machine learning to identify authors quickly and accurately, as well as methods to circumvent this identification. Finally, this work presents novel data sets as a foundation for future research in comment classification and adversarial authorship recognition.
Chapter 1: Introduction

The web has opened up unparalleled opportunities for participation in public discourse. This has occurred rapidly and participation has skyrocketed in recent years, with significant societal impact. This expanding scope of public discourse has created a need for systems that curate, filter, organize and secure content. Managing this expanding scope, however, gives rise to a variety of scaling and management issues due to limitations in human capacity. These issues can be mitigated through the use of machine learning.

Challenges in managing discourse include the limitations of crowdsourcing as a means of identifying quality and relevant discourse in commenting systems, flaws in crowdsourced topic identification systems that cause discussion content to be overlooked, and the identification or anonymization of large numbers of authors through contextual or linguistic means. There are also scaling issues in researching this field due to the lack of large robust data sets with accurate baselines.

Identifying high quality and relevant content amongst large amounts of discourse is often relegated to human-based crowdsourced rating systems or no means of discourse management at all. For those that utilize crowdsourcing methods to manage discourse, machine learning can assist in identifying content that is overlooked. This can occur because of the low visibility of quality discourse that is generated too long after the inception of the discourse, a lack of adherence to guidelines and ethics agreed upon by an online community, and missing identifying markers such as keywords that serve to segment and categorize the discourse at hand. These issues are particularly common in shorter forms of discourse such as comments on news websites like Slashdot and the 140-character-maximum messages on Twitter.

The use of machine learning can mitigate issues related to authorship recognition by replacing time-consuming and costly human expert analysis with extremely fast and cheap machine-learning-based classifiers. This raises new challenges related to privacy and anonymity which can also be mitigated by intentional modification of writing style and improved by using machine learning to
preserve obfuscated writing styles over time.

Studying the role of machine learning in managing adversarial discourse brings about new challenges in evaluation. By its nature, adversarial writing is intended to betray an accurate understanding of a situation. For example, a false review on a restaurant rating website is intended to be mistaken for a legitimate review. This component of adversarial discourse, however, makes it challenging to build and evaluate classifiers. The primary obstacle lies in acquiring baseline measurements of quality in large data sets that can be used to train classifiers and determine their accuracy. Small data sets can be created with human experts, but greatly expanding this process is not tenable due to time and economic constraints.

This work evaluates the effectiveness of machine learning techniques in identifying quality discourse, demonstrating that a machine learning classifier can identify quality content with up to 85% accuracy when compared to the ratings issued by human moderators in a crowdsourced rating system. It also evaluates the effectiveness of such classifiers in identifying relevant content without using human-labeled data as a feature for the classification process. This work demonstrates that we can identify the correct topic of discussion in a social media data set over 80% of the time when compared to the work of human participants.

This work also demonstrates the effectiveness of machine learning when dealing with scaling issues that arise in managing identity through writing style. It shows the effectiveness of multiple classification methods in identifying the identity of anonymous authors based on writing style—an approach called stylometry. Up to 40 authors can be identified with over 80% accuracy. This includes techniques that have previously only been applied to small numbers of authors. These results can be troubling from a privacy perspective, and as a result this work also demonstrates how humans can circumvent such identification by manually modifying writing style. These manual modifications reduce the effectiveness of stylometry techniques to that of random chance. Machine learning can also be used to aid in this circumvention process by helping authors understand the myriad features that reflect an individual’s writing style.

This work also introduces two new data sets for studying adversarial public discourse. The
first is a 45-author corpus for adversarial stylometry and the second is a data set of comments from Slashdot.org containing many intentionally deceptive or disruptive comments, ratings by the Slashdot community, and verification by a range of outside participants. Both were created by leveraging large numbers of human participants with the guidance of a few human experts. A framework for creating adversarial data sets of this type is also detailed.

1.1 Related Work

Work surrounding the management of quality and identity in discourse is widespread, but research in utilizing machine learning to fully or partially automate such management is not.

Quality management of discourse is most often accomplished through crowdsourced rating systems. Crowdsourcing is defined as “the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call”\(^7\). Some of most popular websites on the Internet, such as YouTube, Facebook, and Reddit, utilize some form of crowdsourcing in order to identify quality discourse content. There is an abundance of research in examining the challenges that these communities and systems face, such as the potential for users to leave communities due to information overload\(^8\) and the issue that the passage of time plays in disempowering quality discourse\(^9\). A missing piece in this research, however, is an examination of how machine learning may be able to replicate and even improve components of these systems.

A related area of research and common form of content and quality management using automation is in the topic of recommender systems, which take input from a range of participants to provide tailored content, generally to a specific user\(^10-12\). While this is a field that utilizes machine learning in the context of quality, recommender systems are focused on tailoring content to an individual rather than improving the quality and utility of overall discourse for a community.

Identity and anonymity in discourse is also a widely researched field but one that, for the most part, has only taken into consideration location-based privacy concerns. The role of anonymity in public discourse has been widely recognized as important\(^13\), and research on technical methods for maintaining anonymity is widespread\(^14,15\). There has also been a wide range of research on
identifying individuals based on the linguistic style present in their writing\textsuperscript{1\textendash}16. But with a few exceptions, the intersection of these two fields and the implications for privacy and anonymity had gone largely unexplored. Early work in the field we have identified as adversarial stylometry includes work demonstrating a minimum of 6,500 words being required to confidently determine the identity of an author\textsuperscript{17}. There is also work demonstrating that minor modifications of feature vectors of writing style profiles can circumvent authorship recognition\textsuperscript{4}. But until now, no researchers have examined the potential for human adversaries to circumvent authorship recognition, nor have they applied deception to actual text. Existing research has also not yet utilized statistical machine learning to understand the changes made to writing in successful circumvention attempts.

Building data sets to study the potential for machine learning in public discourse and protecting anonymity is a challenge due to the need for adversarial data with a known ground truth. For example, in adversarial writing style analysis we need to be able to study and compare passages that are written legitimately with those written deceptively. In order to do this we must know the true identity of the deceptive writer. Such data is difficult to come by, especially on a larger scale, so we created a process for generating such data efficiently utilizing Amazon’s Mechanical Turk service and employing a series of specifically designed quality control measures.

1.2 Contributions

This work applies statistical machine learning to the challenges of moderating quality and protecting identity in public discourse. We employ machine learning in problems currently solved only by crowdsourcing for large-scale quality analysis and topic identification in online discourse. Machine learning is also employed to prove the devastating effects that stylometry can present for anonymity, and then demonstrates the potential for individuals to circumvent such anonymity concerns by intentionally modifying writing style. We also provide new data sets for research in these topics, as well as a framework for building large adversarial data sets for research purposes using Amazon Mechanical Turk.
Chapter 2: Managing Quality and Topic Identification with Machine Learning

We are witnessing a transition from a world in which gatekeepers and editors filter content before it is published, to a world full of vast amounts of user-generated content in which information filtering is done after publication. Online communities have developed a variety of community-based filtering and rating mechanisms to help maintain quality and manageability. James Surowiecki notes that several conditions must be present (including diversity and independence) for a crowd to be wise. It is an open question whether these crowdsourced filtering mechanisms represent “the wisdom of crowds” or “the censoring mob.”

In crowdsourced filtering systems, users rank content provided by other users on the validity or usefulness within their particular context. The goal is that “good” content will rise to prominence and “bad” content will fade into obscurity. These mechanisms are heterogenous in their design, ranging from the simple (up or down vote) to the byzantine. The principles underlying these mechanisms are not well understood and they have several known weaknesses. First, the content must have a large enough initial audience to produce a rating. Even on popular sites, the critical mass necessary to produce ratings dissipates quickly, so that comments on articles not added quickly are destined for obscurity. Second, these mechanisms are often gameable. Work by Annalee Newitz showed how Digg could be manipulated to get arbitrary content on the front page. Third, despite the use of these mechanisms, information overload persists. Last, these mechanisms often enshrine existing voices and overlook content from new or marginalized voices.

Similar and often overlapping to the issue of content quality is the issue of content relevance. In crowdsourced topic identification systems, it is the users who identify relevant topics for content. The most common form of identifying topics in this manner is through keywords, which may or may not be elements of the content itself.

Relying on the crowdsourcing-based system for both quality and relevance filtering comes with
issues, particularly when scaling them to large quantities of discourse. This chapter first identifies some of those issues as they relate to weaknesses in human participation and the potential for machine learning. It then demonstrates how machine learning can be used to replicate the filtering and topic identification abilities of the crowd, which sets the groundwork for automation resolving scaling issues in the future.

2.1 Management and Scaling Issues in Crowdsourced Quality and Topic Identification

The number of people with Internet access has skyrocketed in the past decade, as can be seen in Figure 2.1. This has brought with it an increasing number of participants in content creation and discourse on the web. In 2005 the most popular tech news and discussion site was Slashdot, with about 4.3 million unique visitors per month\(^1\). Interest increased and Slashdot was then topped by Digg, then Digg was topped by Reddit as the most popular site in this category. Reddit was seeing over 35 million unique users per month in January of 2012\(^2\).

With the infusion of participation comes an increase in content. This presents a challenge: individuals come to these sites in large part because they find value in the discussion element. An increase in the discussion, however, results in a decrease in the utility of reading through all elements of it due to information overload. Readers do not find much value in perusing through thousands of comments when only 20 or 30 provide information relevant and useful to them. As a result, many of these sites have implemented some form of crowdsourcing to deal with these scaling issues of discourse on the web.

These systems are necessary because without them, information overload reduces utility as users are overwhelmed by noise. But scaling these crowdsourcing systems presents challenges. How we filter is not value-neutral, and shapes the nature of online spaces and the opportunities they afford online speakers. The nature of crowdsourcing quality and content determines which voices will reach a large audience and which will be silenced by obscurity. The result of a flawed system is that it prevents users from viewing or learning about content they would have wanted to read. In that way,

\(^1\)http://www.barbarafrench.net/2005/02/07/special-report-the-state-of-analyst-weblogs-part-2/
\(^2\)http://blog.reddit.com/2012/01/2-billion-beyond.html
these systems can become a form of censorship as it might mean the disempowerment of voices that objectively should be heard in a discussion.

The challenges that arise do not necessarily insinuate the malevolent motive that the term “censoring mob” brings. For example, a community based on voting which then reorders discourse in accordance with votes quickly makes it very difficult for new discourse to enter into the conversation and rise to prominence. This is because as the existing unsorted content is viewed and voted on, subsequent viewers are presented with an echo chamber where they see comments that were previously voted up, neglecting those with fewer votes which include the most recent additions to the discourse.

These challenges can be summarized as follows:

1. **Crowdsourcing quality discourse identification unintentionally quiets certain voices.**

   These voices may be those of people who arrive too late to a discussion, opposing viewpoints,
2. **Crowdsourced approaches to topic identification results in overlooked discourse.**

Relying on crowdsourcing for topic identification is prone to mistakes that can cause discourse to go overlooked.

The potential for these weaknesses of crowdsourcing to be addressed by automation first requires an understanding of the abilities of machine learning to replicate the ratings of an online community. We demonstrate that this replication is possible with a high degree of accuracy. We also demonstrate the ability of machine learning to replicate crowdsourced topic identification with a similar degree of accuracy, and to much greater effect than traditional topic segmentation. This work presents a baseline for augmenting crowdsourcing systems with machine learning in order to increase the utility of discussions and reduce oversight of relevant discourse.

### 2.2 Automation of Crowdsourced Ratings on Slashdot

We studied the automation of crowdsourced ratings through the Slashdot (slashdot.org) community, identifying a combination of features that allow us to extract comments that the community will rate as good with high (82%) precision\(^3\). Furthermore, we can segment comments into good, neutral, and bad categories with 76% accuracy. We found that author reputation and contextual features were the most salient; however, we also discovered many salient linguistic features, which when used alone can extract good comments with 57% to 63% accuracy, depending on the inclusion of humorous posts. These features provide insight into what types of posts garner high versus low ratings.

Our approach uses statistical machine learning techniques to predict outcomes automatically and objectively gain insight into the workings of these filtering mechanisms. We show that though understanding the content of a piece of writing is difficult, determining whether an online community will find it interesting is more tractable. This work demonstrates that it is possible to design mechanisms by which individuals, communities, and intelligent software agents can collaborate to

---

\(^3\)Throughout this section, “precision” refers to the precision averaged across all classes. Additionally, recall in all cases was within one percent of the precision.
explain and improve social, bottom-up filtering, and expand the range of the possible in terms of the values these systems can reflect and the communities they can serve.

2.2.1 Related Work: Slashdot, Recommender Systems, Online Communities

There are many well-studied issues surrounding online communities that rely on peer rating systems to determine the relevance of the content. Information overload can easily occur in online communities when it becomes difficult for users to filter the relevant and interesting content. This can encourage users to leave the community. In addition, considerable time can pass before the fair and poor comments are identified in communities like Slashdot. This research illustrates why it is important to have a valid and accurate rating system.

A well-studied superset of our topic is recommender systems, which use collaborative filtering, learning, or a combination of both to rate and recommend products or content. This research has noted the difficulty in identifying the combinations of measures to use in a comparative evaluation. There has been considerable interest and work by Paul Resnick on designing recommender systems that are not vulnerable to manipulation. Another large body of related work deals with assigning reputation to individuals rather than their work.

Other related work has taken an ethnographic approach to studying the phenomenon of online spaces, some of it looking directly at the Slashdot community. In “Social Network Sites: Definition, History, Scholarship,” danah boyd and Nicole Ellison tackle the question of defining social networks and online communities.

The work differs from related research in that it deals with the task of learning how a community selects quality discourse and applies it to automatically improve and facilitate discussions. This research lays the groundwork for next steps in augmenting identification of quality public discourse with machine learning based on an understanding of how the community works.

2.2.2 The Slashdot Community

Slashdot (slashdot.org) is a technology news site and online community. Readers of the site submit articles which are reviewed by a team of editors, who select the best ones to post as the news items...
for that day. The community then discusses the articles through a comment system. Each news post has its own comment series. Due to the large number of comments each article receives, Slashdot has implemented a crowdsourcing system for users to rank the comments on how relevant they are to the article and to other users. The Slashdot community is a useful starting point because their community rating system is particularly rich in metadata. The system uses a scale from -1 to 5, with 5 signifying the comments most worth reading. Comments that receive a very low score are typically hidden, while comments with a higher score are highlighted. This is beneficial because it prevents a user from sorting through an abundance of useless data in order to access relevant commentary. In addition to the numerical rating posts can also be given a rating description such as “Insightful” or “Informative” if they are good and “Offtopic” or “Flamebait” if they are bad, among many others.

There are many other nuances to the Slashdot system such as moderator points and metamoderation that are not discussed in this chapter but complete details about how the Slashdot rating system works can be found in their FAQ

Slashdot was chosen because the richness of the rating scheme helps make the site a valuable testbed for an agent intended to augment a collaborative filtering system. Our goal, however, is not to augment Slashdot’s system alone. We believe that this research can open the door to augmenting systems with more obscure or less robust rating systems.

A sample news story and highly rated (5) comment based on it follow:

Subject: Rabbit Ears To Stage a Comeback Thanks To DTV
Post: Jeffrey Breen writes Like Monty Python’s Killer Rabbit, cheap indoor antennas seem harmless to satellite and cable providers. But with the digital TV transition in the US, rabbit ears can suddenly provide digital-perfect pictures, many more channels, and even on-screen program guides. Already feeling pressure as suddenly budget-conscious consumers shed premium channels, providers must now get creative to protect their low-end as well.

Date: Saturday, February 14, @04:55PM
Tags: business competition usa entertainment tv story

---

4 http://slashdot.org/faq/com-mod.shtml#cm520
5 http://news.slashdot.org/article.pl?sid=09/02/14/2025245
6 http://news.slashdot.org/comments.pl?sid=1128309&cid=26858873

Chapter 2: Managing Quality and Topic Identification with Machine Learning
Excerpt 1: A Sample Post.

Not rabbit ears (Score:5, Informative)

by Show+Me+Altoids (1183399) on Saturday February 14, @04:57PM (#26858873)

Rabbit ears don’t pick up UHF signals; they are for VHF which is going away. It’s the “loop” part of current antennas which will receive UHF.

* 78 hidden comments

Excerpt 2: A Sample Comment.

Participants in Slashdot discussions may be anonymous or registered users. Anonymous posters suffer a built-in penalty of having all their comments start with a score of zero, whereas all registered users start with a score of one. Registered users also have access to a host of additional features on the site such as the ability to become “friends” with other users, set up a personal profile, and obtain the privilege of rating comments to help shape discussions.

2.2.3 Approach, Data, Features

This approach uses statistical machine learning to gain insight into the mechanisms by which online communities filter and censor content from the bottom up. We began by mining the Slashdot community for features that would allow us to replicate the community rating system. We used a combination of information gain, intuition, and trial and error to identify feature sets that would yield high accuracy. We evaluated the features using several machine learning algorithms including neural networks, support vector machines (SVMs), and Bayesian approaches. The best results were found using SVMs and the results in this chapter reflect this. Ultimately we were able to study the salience of reputation-based, social, and linguistic features to gain insight into the behavior of community filtering as practiced by the Slashdot moderating community.

Features The features we used to classify Slashdot comments are divided into two groups: linguistic features and contextual and author reputation features. The linguistic set represents features related to the words, their meanings, and the structure of the text. Most of the linguistic features
were extracted from the comments using the Linguistic Inquiry and Word Count (LIWC) software\(^7\), a text analysis database designed by psychologists to study the various emotional, cognitive, and structural components in text\(^31\). The contextual and author reputation sets do not represent the content of the comments. Instead, they are based upon contextual information of the comment, such as when it was posted or how much discussion it generated, or information about the reputation of the author such as his or her recent comment ratings. The full list of linguistic, contextual and author reputation features and their descriptions is available on the web\(^8\).

**Contextual and Author Reputation Features.** These features were primarily based on observations made by sifting through the comment database. For example, we observed that comments that were made a very long time after the original post were much less likely to receive a high rating. This is likely due to the fact that as the day moves on, fewer people are reading the discussion section of old posts. Further features developed in an attempt to exploit the author reputation metrics that exist in the system, such as the average score of the 24 most recent comments by an individual author. Some of these features are specific to the Slashdot community and cannot be directly applied to the rating system for other online communities, but analogous features can be found in many of them. Author reputation features for YouTube, for example, might include how many subscribers they have, how many videos they have posted, and the average ratings of those videos. Some of the features used include:

- \textit{timeDifference} between original post and comment.
- \textit{subComments}: number of replies under the comment.
- \textit{posterIdNumber}: how long a poster has been on Slashdot.
- \textit{posterAcceptanceRatio}: percentage of articles that the user submitted that were accepted and posted as news.

While all of the features we have looked at are useful at gaining insight into how the filtering process works, some features are only in evidence significantly after the comment has been posted.

\(^7\)http://www.liwc.net/
\(^8\)http://psal.cs.drexel.edu/files/Slashdot_Features.pdf

**Chapter 2: Managing Quality and Topic Identification with Machine Learning**
(for example, subComments). These ex-post features are indicators of how the community is rating these comments, but they cannot be used in any mixed-initiative system that combines machine learning and collaborative filtering. It is also worth noting that some of the most salient features, such as timeDifference, may represent flaws in the collaborative filtering scheme. It is difficult to say if good comments tend to be timely or if late comments simply do not get the benefit of moderator eyeballs.

**Linguistic Features.** The motivation behind including linguistic features was our hypothesis that comments that receive higher ratings generally exhibit higher quality writing. We expanded on this with further linguistic analysis based on ideas such as the hypothesis that comments with overall positive sentiment would be more likely to receive a high score. Some of these hypotheses proved to be true, as is explained in the Discussion section. The linguistic features included thirty features from LIWC, seven unigrams, and an additional six features we derived and extracted from the comment text such as the number of words appearing in both the comment and original post.

The advantage to using linguistic features is the ability to easily port them across a variety of rating systems. As the end goal of this research is to create an agent that can augment a collaborative filtering system, effective linguistic features would be an important part of making such an agent as portable as possible. Some linguistic features include:

- **Comment Sentiment**: Ratio of positive to negative emotion words.
- **Swear Words**: Percentage of swear words in the comment.
- **First Person Pronouns**: Percentage of words that are first person pronouns (i, my, mine).
- **Post Word Count**: The number of words in the comment that also appear in the original news post.
- **Word Count**: The total number of words in the comment.

**Approach** All classification was performed using a Support Vector Machine classifier that used a Gaussian radial basis function. The continuous features were discretized into four bins before
classification. A single experiment consisted of generating the data set and feature space, randomly selecting an equal number of comments from each class, and running the SVM classifier through 10 fold cross-validation. We took samples from a data set of 528 comments, or 1173, depending on whether we divided the data set into two classes or three. This variation is due to keeping the class distribution equal as changing the score range for each class affected the maximum number of comments per class. Accuracy measurements were obtained by running each experiment five times, randomly generating a new data set and feature space for each iteration. Classification was performed using the WEKA toolkit\(^9\) and the LIBSVM library\(^10\).

We evaluated the ability of our feature set to predict the community rating of comments made on Slashdot news stories on the dates of Saturday, February 14th\(^11\) and Monday, February 16th 2009\(^12\). We chose these days based on the assumption that the community may behave differently on weekends and weekdays due to the number of people trapped behind a computer monitor for the 9-5 workday. Additionally, we restricted the comments we analyzed to just the first-tier replies, meaning the comments that were direct replies to the original post and not comments that were replies to other comments. We did this based on the assumption that comments made further along in each thread were less likely to be viewed by the whole community and thus less likely to accurately represent the general opinion of the community on what comments were good or bad.

The classification experiments sought to answer a number of questions. What features seem to represent the ways in which the community determines the quality of a comment? Is it possible to predict the original community rating of a comment based on a selection of both linguistic and author and community-specific features? Are linguistic features alone useful in determining the quality of a comment? Are “funny” comments more difficult to automatically classify than those which have been labeled “informative”?

\(^9\)http://www.cs.waikato.ac.nz/ml/weka/
\(^10\)http://www.csie.ntu.edu.tw/~cjlin/libsvm/
\(^11\)http://slashdot.org/index.pl?issue=20090214
\(^12\)http://slashdot.org/index.pl?issue=20090216
**Table 2.1:** Overall Accuracy Chart.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Score to Class Distribution</th>
<th>Feature Set</th>
<th>Overall Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>[-1,0,1,2] [3,4,5]</td>
<td>Linguistic + Contextual + Reputation</td>
<td>82%</td>
</tr>
<tr>
<td>2</td>
<td>[-1,0,1,2] [3,4,5]</td>
<td>Linguistic</td>
<td>63%</td>
</tr>
<tr>
<td>3</td>
<td>[-1,0] [1] [2,3,4,5]</td>
<td>Linguistic + Contextual + Reputation</td>
<td>76%</td>
</tr>
<tr>
<td>3</td>
<td>[-1,0] [1] [2,3,4,5]</td>
<td>Linguistic</td>
<td>42%</td>
</tr>
</tbody>
</table>

2.2.4 Evaluation

While the Slashdot rating system allows for comments to be rated from -1 to 5, we found that attempting to classify a comment as belonging to a specific score class is not very useful; there is too much noise involved and the benefits of classifying something as a 4 instead of a 5 are negligible toward achieving the overall goal of improving the quality of discourse. So we looked at two different methods of categorizing the comments: extracting the good comments and ignoring the rest, and dividing the comments into “good”, “neutral”, and “bad” categories.

**Extracting Top Comments**  The first task of our classifier was to extract the best comments without attempting to further classify everything else. We considered a comment to be of the highest quality if it had a rating equal to or higher than three. Using our extended feature set we were able to determine whether or not a comment was rated in this highest set by the community with 82% accuracy. This is an important result despite being relatively straightforward as a classification task because it demonstrates the ability of a machine learning system to perform the most important task for a collaborative filtering system meant to enhance the level of discourse about a topic: highlighting the elements of the discussion that are most relevant and worthwhile. We looked at modifying our definition of a top comment to be only those with a score of 4 or 5, but found negligible improvements.

**Predicting Bad, Neutral, and Good Comments**  Only extracting the good comments is not necessarily enough for an effective agent meant to augment collaborative filtering systems. We do not necessarily want to penalize comments that would be deemed by the community to be simply “average.” Because of this, it can be important to make a distinction between multiple levels of
**Table 2.2:** Bad/Neutral/Good Confusion Matrix.

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Bad</th>
<th>Neutral</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad</td>
<td>324</td>
<td>35</td>
<td>32</td>
</tr>
<tr>
<td>Neutral</td>
<td>23</td>
<td>301</td>
<td>67</td>
</tr>
<tr>
<td>Good</td>
<td>33</td>
<td>89</td>
<td>269</td>
</tr>
</tbody>
</table>

comment quality. We examined the ability of our classifier to specifically segment the comments between “bad”, “neutral”, and “good” posts. A “bad” post is one with a score of -1 or 0, a “neutral” post has a score of 1, and a “good” post has a score greater than or equal to two.

We were able to classify the comments with an overall accuracy of 76%, significantly higher than random-chance classification of 33%. An example confusion matrix for one of the random tests can be seen in Table 2.2. This confusion matrix shows that in addition to the relatively high accuracy, the misclassifications are skewed in a way that makes sense. For example, almost three times as many “good” comments were classified as “neutral” as were classified as “bad.” Since a comment classified as “bad” in the Slashdot community receives penalties such as being automatically hidden, it is much more desirable for a “good” comment to be classified as “neutral” instead of “bad.”

“Funny” vs. “Insightful” vs. “Informative” We believed that some posts will be difficult to automatically classify. Humor, for example, is notoriously difficult for automated systems, though there is some promising work on the subject. Funny Slashdot comments can be identified by the community in the same way “Insightful” or “Informative” comments are. We used this metadata to determine if “funny” comments were easier or harder to classify. We confirmed our original beliefs by finding that when only “funny” posts are included in the “good” class (since only good posts have the potential for being described as funny), the overall accuracy drops from 76% among 3 categories to 65%. Furthermore, the precision for the “good” class specifically drops from 70% to 52%.

Linguistic vs. Contextual and Reputation Features The most salient features in our set are contextual features such as subCommentCount and author-reputation features like poster-RecentScore. We found that when we looked at linguistic features alone they were not as effective as the contextual and reputation-based features but were still quite salient in determining the com-
community rating of a comment. This is especially true when comments that are classified as “funny” are left out. Humorous comments often have a very different linguistic makeup when compared to “informative” or “interesting” comments, leading to linguistic features being less effective when classifying them.

In the case of extracting “good” comments with a score greater than or equal to three, linguistic features alone yielded an accuracy of 55%. If we removed “funny” comments, however, that accuracy rose to an average of 63%. Segmenting the comments between “good”, “neutral”, and “bad” yielded an accuracy of 42%. Once again, if we removed the “funny” comments we saw an increase to 46%.

While these numbers are not as significant as our earlier results, they demonstrate that augmenting collaborative filtering with linguistic features that can be extracted across most community filtering systems is possible.

**Salient Features.** The most salient features all made use of author reputation information, rather than post content. These were features like the number of news posts submitted by the author, the number of friends the author had, the ratio of posts accepted for publication on the site, the length of time the author had been active on Slashdot, and the aggregate score for other comments posted by the author.

Following these features in salience were features related to the properties of the discussion itself, namely the number of sub-comments generated by the comment and the promptness of the comment relative to the article being posted.

The most successful set of linguistic features selected were the pronoun-based features, particularly first-person pronouns that indicated a well-received post. We identified other salient linguistic features but they were considerably less effective than pronoun usage. The next most salient features are the length of the comment (longer being better), the number of words the comment had in common with the post, the number of commas, and the lexical density.

**Misclassification.** There is still considerable work to be done to identify classes of posts that were difficult to classify. However, good short posts were often misclassified, especially when they were posted by an anonymous author. The data suggests, though more analysis is needed, that good
posts missed by the classifier often reflect comments of authors with little or negative reputation on the site. In general, anonymous posts were easier to classify than attributed posts as they were more likely to be rated bad.

2.2.5 Understanding the Results

While contextual and author reputation features did provide both good results and insight into the filtering mechanisms of the Slashdot community, there is something unsatisfying about using these features. One goal of a filtering system should be to elevate the good comments of new, rare, or often less useful commenters. While the bulk of good comments may be recognizable by author reputation alone, the value of community filtering is in recognizing when that is not the case. However, the use of these features show how the structure of a community site and the structured metadata it provides can make this hard classification problem tractable.

Features based on the text alone would allow more democratic filtering and also allow improved filtering of anonymous writing. However, using linguistic features poses a number of difficulties, based on the traditional hardness of natural language processing. Word sense disambiguation was a challenge, as was context. Despite these difficulties, we were able to identify several linguistic features that were salient, showing that determining if a piece of writing is likely to be viewed as “informative” or “insightful” to a community does not necessarily require understanding.

Learning to predict the rating behavior of an online community has identified features that are correlated with high (and low) ratings. Merely replicating the metric, however, does not separate correlation from causation. Are these features truly what the community is looking for in its rating or are these features just correlated with other hidden features that identify good posts? One way to shed light on this is to test if exposing these features to users helps them craft more interesting and better received posts. We plan to perform such user studies in our future work.

Analyzing how frequently certain features appear within good, neutral, and bad comments provides insight into the specific mechanisms that the Slashdot community uses when rating a comment. Some features provided more obvious feedback while others supplied surprising insights. For example, one might expect posts that contain more swear words to be ranked poorly and our data
supports this claim, as swear words appeared 57% more frequently in bad posts than in good posts. This indicates that the more swear words a person uses in their post, the more likely the Slashdot community will give it a lower rating.

Other features, however, produced unexpected results, such as second person pronouns and first person pronouns. The results showed that second person pronouns appeared 26% more frequently in bad posts while first person pronouns appeared 34% more frequently in good posts. This could indicate that the Slashdot community rates comments higher if the author of the comment takes ownership of the writing by using first person pronouns instead of second person pronouns.

### 2.3 Automation of Crowdsourced Topic Identification on Twitter

Twitter is a microblogging and social networking service that allows users to add friends and send messages to the friends that follow them in the form of “tweets”–messages of up to 140 characters. Twitter also keeps track of trending topics on their service. These trends indicate topics that are being discussed heavily on Twitter and are tracked by keyword. For example, if “Egypt” is a trending topic, then looking at all posts under that trend will reveal any post with the keyword “Egypt” or “#Egypt.” Commonly used terms, such as “coffee”, are removed and the top 10 trending topics are displayed on the Twitter homepage.

Effectively organizing trends and the tweets that belong to them is significant because of a wide range of parties that benefit from doing so. Recent events such as the earthquake in Haiti and the Arab Spring–as well as new research show the increasingly important role of social media and Twitter specifically. There are also traditional beneficiaries of increased understanding of social media, such as advertisers and the individual users. The level of interest in Twitter trending topics specifically can be seen through whatthetrend.com, a site recently created by Twitter to help people understand what the trends are and why they are being discussed.

Given the importance of identifying trending tweets, a major problem that trend classification faces is a relatively crude method for identifying which tweets are part of a specific trend. Using only the trend keyword to identify a trend will miss relevant and potentially important pieces of discourse. Take, for example, the #bahrain trend which was popular early on in the Arab Spring...
due to the protests occurring on February 15th, 2011. Here is a tweet identified as being part of the #bahrain trend:

“#Bahrain 's Pearl Sq looks like Tahrir Square all over again. Police gathered in force but standing back for now.”

And here is another tweet that is clearly related to the same topic but would not be identified as being part of the #Bahrain discussion due to the lack of the proper keyword:

“just got back from manama and the scenes are overwhelming women are being part of the protest wow jst wow”

Manama is the capital of Bahrain and the protests of the day were centered there. In addition, it appears to be a first-hand report that notes the role women are taking in the protests. This is especially relevant given the serious women’s rights issues at stake in Bahraini society. This illustrates how identifying trends related to a tweet through keywords alone is insufficient. Important elements of discourse can be easily overlooked. The recent boom in social media being used both as an organizational tool and a journalism tool shows the necessity for more comprehensive means of organizing discourse in these mediums.

The research in this section explores this need by demonstrating identification of which trending topic a tweet is a part of without knowing if any trend keywords are present in the tweet. We are able to successfully perform this task with high accuracy (85%), which shows the potential for augmenting existing tagging systems based on collective human intelligence with machine learning to get a more comprehensive view of discourse related to a specific topic.

The rest of this section will first examine recent work related to topic modeling in social media. It will then outline the data set we created for this research and explain the classification method for identifying which trend a tweet belongs to using an efficient, modified Bayesian approach that corrects for systemic weaknesses in a traditional naïve Bayes classifier. The evaluation section explains a variety of ways that we tested the results, demonstrating the strength of the approach.

More information about Women’s Rights issues in Bahrain can be found at http://www.bahrainrights.org/en/women
in identifying trending tweets. Finally, we discuss additional strengths and weaknesses and outline future work.

2.3.1 Related Work: Twitter, Topic Modeling

There has been a great deal of interest in topic modeling and analysis, especially on microblogging platforms like Twitter. One of the most recent examples is the work by Ramage et al. that uses topic models to characterize information needs of Twitter users\textsuperscript{34}. This research is effective in better representing content on Twitter to the users who want it, but it still relies on the existing system of identifying trending tweets by keywords.

Cheong and Lee's paper “Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base” is an ethnographic study that looks at how Twitter trends behave and how different levels of hot topics progress through time, and attempts to understand the underlying characteristics of trends and trendsetters\textsuperscript{35}. The authors distinguish between different types of trends (long, medium, and short term) and the demographics of the users that help create the trend (gender, Twitter client used).

Leskovec et al. studied the flow of information and news through social media and mainstream media in their paper “Meme-tracking and the dynamics of the news cycle”. This paper followed changing quotes as they passed through the media and the authors able to coalesce topics based around the same quote even if the words were not identical matches\textsuperscript{36}. This research also sets the tone for the importance of tracking topics through the blogosphere and mainstream media, leading directly to the question of how these issues are discussed and tracked in social media like Twitter.

Chang et al.'s paper “Reading Tea Leaves: How Humans Interpret Topic Models” presents a new method for assigning documents to topics in large document collections\textsuperscript{37}. Their work provides insight in developing methods of identifying which content belongs to a particular topic or trend. Phelan et al. propose a novel technique for identifying news stories of interest based on Twitter feeds. Their motivation is that recommender systems take time to develop, requiring a “critical mass” of stories before accurate predictions can be made\textsuperscript{38}.

Bernstein et al.'s work on Eddi groups tweets based on subjects explicitly or implicitly mentioned
in a tweet. While this has real applications for trend classification, its reliance on search engines as a knowledge base means that it won’t have the same effectiveness on Twitter-specific trends. Instead it is geared specifically toward “tweets that break news, post URLs, or comment on events”. Our system uses Twitter alone as the knowledge base for trend classification, which means we obtain high accuracy for classifying Twitter-based trends such as “#idontappreciate” and “#nawimstraight” as seen in Section 3.

Effectively understanding and analyzing social media for trending topics benefits everyone. Recent events and new research have shown the increasingly important role of social media though events like the recent uprising in Egypt and research projects like Ushahidi\textsuperscript{14}, which crowdsource social media and short messages (SMS, email) in order to map crisis information for aid workers, first responders, and the news media.

2.3.2 Approach, Data, Features

To determine whether or not it is possible to identify which trend, if any, a tweet is a part of we created a data set based on current Twitter data, extracted a feature set from the tweets and users in the data, and implemented an improved version of a naïve Bayesian classifier to apply to it\textsuperscript{33}.

Our system relies on word frequency counts in both the individual tweets and the information provided in the profile of a tweet’s author as well as the time zone provided by the author’s profile. Trending topic keywords are not included in the feature set for any tweet. In addition, the weights of user profile word frequencies are reduced by 60%. This number was determined through experimentation that showed equal weighting of profile, and tweet word frequencies did not optimize results. This makes sense, as users may choose to write about a wider range of topics than their profile suggests.

The data set used in this research was created specifically for this project. The final set included tweets in English from the top 10 current trends on Twitter obtained on three different days (June 2nd through June 4th, 2010) for a total of 40,757 tweets belonging to a combined 30 trends. In addition, 2,947 tweets were collected on June 5th from the public timeline, from which all tweets

\textsuperscript{14}http://www.ushahidi.com

Chapter 2: Managing Quality and Topic Identification with Machine Learning
with any reference to the 30 trends were removed, leaving 2,471 public, not-trending tweets. The
text in each tweet was parsed into word and punctuation tokens, which were then used for token
frequencies.

Between these tweets there were 29,881 unique users who were in the Twitter system (some users
that posted tweets no longer existed because they were flagged as spam accounts or removed by the
user). Profile information was collected for each of these users and word frequencies were extracted
for all words in the user description. The time zone was also pulled from each user to be used as
the replacement for the geo-location information that was missing in almost every tweet.

A “clean” data set was also created from the original data. This set included only tweets with
greater than 15 words and punctuation tokens, and did not include more than one trend keyword.
This data set was reduced to 23,939 tweets.

For all data, keywords relating to the trending topics were removed so as not to influence the
classification task–leaving them in would heavily skew the precision towards correct classification as
those words would clearly indicate they belonged to a specific trend. The keywords to remove were
taken directly from the trending topic name itself, so in the case of “oil spill” both the word “oil”
and the word “spill” were removed from tweets. If the trending topic was simply “idontappreciate”
then that is the only token that was removed; the word “appreciate” alone was not removed.

The (T)WCNB Classifier We used a novel classification approach that is well suited for the task
at hand with the Transformed Weight-normalized Complement Naïve Bayes classifier (TWCNB).
The advantage to this approach is the speed of training a Bayesian classifier while correcting for
some of the major weaknesses that a naïve Bayesian approach can have when dealing with data sets
that may have incongruous numbers of instances per class. Our approach ended up leaving off the
“transformed” step for this application, as described below.

TWCNB is an improvement over the standard Multinomial naïve Bayes approach that looks at
the frequency of each word in a document and calculates the probability that document is part of each
class based on the occurrence of those words in that class\textsuperscript{40}. Newer machine learning approaches
such as Support Vector Machines (SVMs) are heavily favored over Bayesian approaches, but the
viability of the Bayesian approach resurfaced with research by Rennie et al. demonstrating the
effectiveness of the TWCNB classifier\(^{33}\).

TWCNB improves upon systemic problems with the Bayesian approach and problems in modeling
text. The problems it mitigates are the skewed data bias (more training data for one class biases class
boundary weights) and weight magnitude errors that come with text classification based on word
frequency ("San Francisco" is weighed twice as heavily as "Boston" because it has two words even
though each phrase conveys the same amount of information). The three text modeling corrections it
makes are transforming term frequency to make multiple occurrences of the same word in a document
not seem so unlikely, transforming document frequency to lessen the influence of commonly appearing
words, and normalizing word counts so long documents don’t negatively affect probabilities\(^{33}\).

The TWCNB approach was looked at again by Kibriya et al.\(^{41}\). They claim that not all of the
modifications may be necessary for optimal performance on some data sets. We also found this to
be true with our data set. Correcting for the term frequency actually reduced the precision of our
classifier. This step was intended to make multiple occurrences of the same word in a document
have less of a skewing effect. For this element of TWCNB’s transformation component, Rennie et
al. show that this bias starts to be a factor when words occur four times or more\(^{33}\). The same
word occurring more than four times in a 140-character tweet is much less likely than in a document
consisting of hundreds or thousands of words. Given these results, it’s not surprising that including
that element of the classifier may not be effective on our data set. The classification approach for
TWCNB can be summarized as:

\[
\text{class}(t_i) = \arg\max_c \{ \log(Pr(c)) - \sum_n f_{ni} \log(\frac{1 + \sum_{k=1}^{\vert C \vert} \frac{F_{nk}}{N_k}}{N}) + \sum_{k=1}^{\vert C \vert} \sum_{x=1}^{N} F_{xc} \}, k \neq c \land k \in C
\]

Where \(F_{nk}\) is the frequency of word \(n\) in class \(k\) and \(f_{ni}\) is the frequency of word \(n\) in tweet \(i\)
and \(N\) is the size of the vocabulary. And that can be compared to the traditional multinomial naïve
Bayes (MNB) approach:

\[
\text{class}(t_i) = \arg\max_c \{ \log(Pr(c)) + \sum_n f_{ni} \log(\frac{1 + \frac{F_{nc}}{N}}{N}) \} + \sum_{x=1}^{N} F_{xc}
\]
The primary difference between the approaches is that TWCNB looks at the features as they appear in all classes except the current class in question, whereas the MNB approach looks at the word frequencies for each class individually.

The word frequency transformations followed the following equation where \( f \) is the word, \( D \) is the number of tweets, and \( df \) is the number of tweets \( f \) occurs in.

\[
\text{transform}_{f} = \frac{\log(f+1) \times \log(\frac{D}{df})}{\sqrt{\left( \sum_{x=1}^{N} F_{xn} \right)^2}}
\]

The creators of this approach demonstrate it raises the effectiveness of the Bayesian approach nearly to that of SVMs on the data sets they tested it on. The final version of the classifier used in this approach is simply WCNB as it does not include word frequency transformations. We include this description and discussion of the transformation element as it may be of use in similar applications that deal with longer passages like those occurring in many forms of online discourse.

**Why Not A Support Vector Machine?** The biggest achievement of TWCNB is that its effectiveness is comparable to that of a support vector machine (SVM). We confirmed this in our own initial findings early in the study, demonstrating that the 1% to 4% decrease in precision when using TWCNB over SVM that Rennie et al. find extends to our data. With our very large data set we found that training an SVM can take up to one hundred times longer than the Bayesian approach. This is important in Twitter because the generation of trending topic feeds needs to be done in near real-time. Given the scale of a social media platform like Twitter we feel that small difference precision between SVM and TWCNB approaches is not worth the trade-off in training time.

### 2.3.3 Evaluation

We evaluated our data in a variety of ways to thoroughly analyze the strengths and weaknesses of our approach. We first established a baseline using standard unsupervised topic-modeling using the original tweet data without the trend keywords. We then analyzed our approach, breaking down the effectiveness on different days, which trends were easier to pick up, and the improvement of our WCNB approach over traditional WNB and complete TWCNB.
A Topic-Modeling Baseline  To derive a baseline for our analysis, we also analyzed how well basic, unsupervised topic modeling\textsuperscript{42} fared at the task of detecting and classifying tweets. We used the topic-modeling functionality in Mallet 2.0.5\textsuperscript{43}. Mallet estimates a distribution of topics using Gibbs Sampling. We ran Mallet with default parameters, creating three models of 10 topics each. The results can be seen in Figure 2.2.

The topics generated by the topic-modeling software did not always correlate with the trending topics. For example, in one of the models, topic 3, “amp unlimited http iphone tethering ipad makes announces customers big iphones coming users kills tiered end drops plan read” and topic 7 “http ly bit news wireless phone stop usage att caps leak eu pricing give url blog eat cap offering” corresponded to the trending topic “Data Plans”, whereas no topics in that model corresponded to the trend #zimbabwe. We assigned a topic to a trend based on the class that the plurality of tweets assigned that topic belonged to (thus, adding an element of supervision to the problem).

The topic model represents a document—in this case a tweet—as a combination of topics. We considered the topic assigned the highest proportion to be the class assignment. It is likely that a supervised topic model\textsuperscript{44} or a model using a combination of multiple topics might do better. However, these results are intended to provide a base level of performance. Each model was evaluated by precision, recall, and f-measure.
We also combined the models for each day using a voting procedure. For each tweet, we assigned it to a trend if two of the three models agreed upon where it should be assigned and left it unassigned otherwise, resulting in increased precision.

The results were characterized by high variance. They performed much better on certain trends such as “Rachel Corrie” and “Data Plans” than others like “#everlastingfriends” and ”#idontappreciate.” News-oriented topics tended to garner better results than social topics.

Our Results  The effectiveness of classifying tweets as part of a trend was tested across different classifiers, feature sets, and data sets. The “standard” classification setup used as a basis for comparison against all of the other methods is the WCNB classifier applied to three days of data with 10 individual trends for each day (no public “no-trend” timeline included), all three major features (tweet text, user description text, and time zone), and the original data set that has not been scrubbed in any way.

As we can see in Figure 2.3, the classifier is quite effective in classifying the tweets. Depending on the day the precision ranges from 80% to 86%, the recall from 76% to 84% and the F-Measure from 78% to 80%.\(^{15}\) In these figures we see that the variance in accuracy between trends can be significant with some trends, such as “Gano Mexico” exhibiting great differences in precision and recall.

\(^{15}\)Figures illustrating these results can be found at https://psal.cs.drexel.edu

Figure 2.3: Precision, Recall and F-Measure for the top 10 trends of each day in our data set using the WCNB classifier.
Figure 2.4: Results for MNB vs. WCNB, averaged across all three days.

Figure 2.5: Comparing the addition of public non-trending tweets. Averaged across all three days.

Recall and others like “Janey” and “Zimbabwe” showing almost perfect recall or precision. The graphs primarily show that accuracy remains relatively uniform with only a few poorly classified trends.

The WCNB classifier showed noticeable improvements over a standard multinomial naïve Bayesian classifier. The improvements are displayed in Figure 2.4. Precision saw a measurable improvement of about three points and recall saw a more dramatic gain of about eight points.

All the experiments up to this point have only tasked the classifier with distinguishing between 10 trends, and all tweets are known to be a part of one of those trends. For the next task we obtained about 2,400 tweets that were not part of any trend in the original data set and introduced
them into the data set to see if the added noise would negatively affect classification. It did, but not significantly. The precision and recall dropped by three points and two points respectively, as can be seen in Figure 2.5. This is a promising result, as any real-time classifier will have to deal with large numbers of tweets unrelated to any one trend. Our results here are only a start, however, with the non-trending tweets numbering only twice that of any one trending tweet class. In reality it may be far more.

As expected, incorporating the user description and time zone features also had a positive effect on the accuracy of the classifier. Figure 2.6 shows the difference in accuracy when not using the user and time zone features. The difference is not huge but it is consistent across classes and data sets and demonstrates the effectiveness of including additional features beyond simple tweet token frequency.

Figure 2.7 shows the increased precision that results from cleaning the data set of junk tweets such as those with multiple trends that are often used for spam purposes. Interestingly, the precision increases noticeably while the recall actually decreases slightly.

Finally, we tested the ability of the classifier to perform at the level of an experienced human. We manually classified 100 random individual tweets from Wednesday’s data along with the public timeline and obtained results very similar to that of the classifier. We correctly classified 77% of
the tweets. In most cases the mistakes that were made were not the same as the mistakes made by the classifier. Further study should be done to verify these results as this is a fairly small test set. These results appear to indicate, however, that the effectiveness of the classifier may be at least that of human classification.

2.3.4 Understanding the Results

While these results are specific to Twitter, one could easily see how they can be extended to other domains. The highly effective means of classifying tweets without utilizing the corresponding keyword is an important discovery as it means that real improvement upon such systems is easily within reach. We believe research like this will help demonstrate the effectiveness of machine learning while helping to demystify the concept as a whole by showing clear, straightforward applications that have been overlooked.

The previous section laid out a number of positive findings, but one somewhat surprising result was that the use of a clean data set did not have a greater impact on precision, and in fact reduced recall. We expected the clean data set to improve overall accuracy as we observed that many of the misclassifications that occurred were a result of short tweets or spam tweets that contained many trending topic keywords in them in order to maximize readership of the tweet. But what is more important in this context: precision or recall? The answer might change depending on the topic at
hand, both in terms of content and volume of discussion.

Similarly, we were surprised at the relatively small decrease in accuracy when the public non-trending tweets were included in the data set. We thought those tweets would be difficult to classify but the precision for classifying non-trending tweets consistently stayed between 75% and 85%.

**Examining Misclassifications** This is only an anecdotal evaluation of misclassifications, but we thought it would be interesting to demonstrate some of the wrongly classified tweets based on our observations throughout the experiments. The name at the end of each example is the wrongly attributed class.

#EverLastingFriends argue.. mayb even fight dont talk for awhile but at the end of the day call n say &quot;wat u doin&quot; (#dontappreciate)

The words “don’t” and “dont” often appeared in tweets wrongly classified as “dontappreciate” since they are not on the trend keyword remove list.

Check out Eclipse soundtrack! (L) http://bit.ly/bgQWo

This is a classic example of a tweet with too little information. After “eclipse” and “soundtrack” are removed, there is very little information available for the tweet to be classified by.


Despite requesting only English tweets, other languages often came through. This one should have remained unclassified.

Why is Zimbabue trending? (Almagro)

Mentioning just the keyword and the word “trending” is a common occurrence that leads to misclassification. Future work may want to incorporate common Twitter terms such as “RT” and “trend” into the remove list.

this is f’n outrageous! - http://cli.gs/h4UJGz #hiphopwashighschool Guachaca Zimbabue #everlastingfriends Joran Data Plans Almagro (#everlastingfriends)
This is a classic example of spam. Many trending topics are included in the tweet along with a spam link.

### 2.4 Future Work

A number of questions have arisen from our results that pave the path for future work. When looking at replicating ratings on Slashdot, how well do the features identified in the data set do as trust metrics for other communities? If they are trained on Slashdot data versus data from that community, do other features work better? How do the filtering mechanisms (granularity, democracy, etc) relate to which features are best? What about community demographics?

We already know that short and funny posts are more difficult to classify than longer posts that offer information or insight. However, there is more work to be done in understanding why certain posts are misclassified and whether difficult-to-classify posts can be detected automatically.

Understanding the differences in community standards and procedures could help communities cross-pollinate their discourse. If the community-rating mechanisms for two communities could be approximated automatically, then these automated mechanisms could bring relevant content to the attention of new communities. We plan to explore how the features we’ve identified translate to other communities and weather other features or algorithms work better.

We anticipate hard limits to the accuracy of filtering mechanisms based on text-based features that do not actually understand natural language. However, the question of whether one can determine if a piece of discourse will be of interest to a community is not precisely the same as understanding it as has been demonstrated by this research.

Upon examining topic identification on Twitter, there are a few clear next steps such as performing a more in-depth analysis in obtaining features that can distinguish between the misclassifications. Additional user features such as friend lists, response tweets, retweets (re-posting someone else’s tweet to your own account), and more could be mined to improve classification results. For example, if a user’s tweet is retweeted by users who have participated heavily in a certain trend, that may give insight into the potential trend of the original tweet.

Further refinement of the approach outlined in this research could add a lot to parsing and
understanding discourse of important topics in social media. Tools to better handle and understand
the massive amount of social communication on the Internet are essential as networks like Twitter
and Facebook grow. Machine learning has a big role to play in the development of those tools.

2.5 Conclusions: Can Machine Learning Improve Quality and Topic Identification Online?

The research in this section demonstrates that machine learning approaches can be very successful
at replicating the ratings and topic identification garnered through crowdsourcing systems. The
coming research challenge is in incorporating these approaches into a combined system that uses
crowdsourcing and machine learning in tandem to provide accurate ratings with a minimum of
overlooked content.
Chapter 3: Managing Identity and Anonymity with Stylometry

In online discourse, the challenges of managing anonymity and the presence of adversaries are prevalent. The need for managing identities can also be important, such as distinguishing between unique authors regardless of their real-life identity. The specific challenges of linguistic authorship recognition are notable and often overlooked, as are the abilities of automation to aid in answering these challenges.

When looking at a publication of unknown authorship, stylometry is used to answer the question “who wrote this document?” Stylometry is a form of authorship recognition that specifically deals with the linguistic features found in text. By comparing the features found in an unknown document with those of a series of potential authors, it is possible to determine the true author with a very high probability.

The field of stylometry has been used to great effect by historians and literary detectives to identify the authors of the Federalist Papers, Civil War letters, and Shakespeare’s plays. In many historical matters, authorship has been unintentionally lost to time, and stylometry can be used to determine true authorship. But it also can have important applications to modern writing, especially in the forensic sense. For example, stylometric techniques are currently used as evidence in courts of law in Britain, the U.S., and Australia.

“In some criminal, civil, and security matters, language can be evidence...When you are faced with a suspicious document, whether you need to know who wrote it, or if it is a real threat or a real suicide note, or if it is too close for comfort to some other document, you need reliable, validated methods.” – Institute For Linguistic Evidence

But the implications and uses of detecting the writing style of an author go beyond criminal courts. Users of the Internet with the desire to publish anonymously may also desire to hide their writing style in order to circumvent stylometry techniques. The goal of these authors is to keep
their identity private. In many parts of the world it is the anonymity of the internet that is the key ingredient to exposing crime, government corruption, human rights abuses, and simple discourse about issues that may be considered controversial by local or state authorities in some countries and communities, but are considered a basic human right in other parts of the world. Just as it can be argued that the creation of effective stylometry techniques is necessary for security and protection of individuals, it can also be argued that privacy and secure anonymous communication are essential for the same reason. Secretary of State Hillary Clinton brought a great deal of attention to this issue in January 2010 when she gave a major public speech about the importance of freedom on the Internet:

“And we must also grapple with the issue of anonymous speech. Those who use the Internet to recruit terrorists or distribute stolen intellectual property cannot divorce their online actions from their real world identities. But these challenges must not become an excuse for governments to systematically violate the rights and privacy of those who use the Internet for peaceful political purposes.” – Hillary Clinton

With the growth of stylometry as an accurate means of determining authorship, the question of how it affects privacy and anonymity in the information age is becoming very important. Privacy and anonymity are held in high regard by many activists, journalists, and law enforcement officers. The introduction of adversarial stylometry and the use of circumvention techniques on stylometry are possible means of ensuring the privacy and anonymity of an individual.

While stylometric methods existed before computers and artificial intelligence techniques, the field is currently dominated by AI techniques such as neural networks and support vector machines. Stylometry is an excellent example of a field that has benefited greatly from the fruits of AI research and continues to be a good demonstration of how to practically apply new research in AI.

This chapter will first review the field of stylometry by detailing how the stylometry problem is formulated, approaches to solving it, some of the history behind it, and the current state of the art. It will then introduce the new field of adversarial stylometry and examine the privacy and anonymity implications of this area of research. It will then demonstrate the ability of adversaries

---

**Chapter 3: Managing Identity and Anonymity with Stylometry**
to circumvent authorship recognition as well as the ability to detect such adversaries. Finally, it will investigate the ability for stylometry to affect the type of short-form discourse discussed in the previous chapter.

3.1 Scaling and Anonymity Issues in Managing Identity with Stylometry

The use of stylometry in managing identity brings with it a series of scaling issues and gives rise to privacy and anonymity concerns. Some of these issues are resolved by machine learning, while others are brought on through utilizing machine learning.

Performing authorship recognition through manual means by human experts, while an effective means of identifying authors, is not scalable to the large amounts of discourse that are present on the Internet. Effective stylometry techniques based on machine learning approaches can distinguish authorship with a high degree of accuracy in an order of seconds. This also brings to light challenges in large-scale stylometry, such as building effective means of authorship recognition across multiple domains.

A potentially more significant issue that comes up when scaling effective machine-learning-based stylometry methods is the ability to identify authors of speech that was intended to remain anonymous. Current approaches to privacy and anonymity do not take into consideration the potential for writing style analysis to determine identity. And while it is possible for individuals to mask their writing style through manual circumvention techniques, another challenge that arises is the difficulty in maintaining that alternative writing style without the help of machine learning.

A summary of these challenges:

1. **Human-based authorship recognition is not scalable without extremely high cost.**

   This work does not analyze the accuracy of machine-based stylometry when compared to human experts, but does demonstrate high accuracy at a fraction of the cost of human-based approaches.

2. **Scaling user-based discourse analysis across domains is challenging.** Automated linguistic approaches to reputation analysis and identifying patterns in discourse can aid systems
in promoting strong content and identifying adversaries.

3. **Automated, scaled authorship recognition is a privacy threat to anonymous speech.**

   Widespread identification of authors through writing style is immune to commonplace location-based approaches to anonymity.

4. **Maintaining a consistent alternate writing style is difficult.** Research shows that maintaining an alternate writing style over a long period of time without assistance is very challenging, but that establishing a consistent alternative style can be aided by machine learning.

   We see that the use of machine learning can help greatly with the first issue of scaling authorship recognition to larger data sets and higher numbers of potential authors. We demonstrate the ability of machine learning to scale to up to 45 unique authors using writing style alone. Recent work has demonstrated an ability for authorship recognition to scale up to thousands of unique authors with high accuracy. Scaling authorship recognition across domains is also a challenge when scaling identity management with stylometry. This work presents analysis on performing this task in short-form social media. Scaling authorship recognition, however, also brings serious privacy concerns and could have troubling implications for anonymous speech. We show how individuals are able to circumvent stylometry, despite the scalability of machine-learning-based methods for identifying authors. Ongoing work based on the adversarial stylometry framework presented here also goes on to enable maintenance of consistent alternative writing styles with machine learning.

### 3.2 Current Trends in Authorship Recognition

Stylometry is a piece of the greater problem of authorship attribution. Patrick Juola, an expert in computer linguistics and authorship recognition, says there are three basic problems of authorship attribution: determining who, of a known set of potential authors, wrote a document in question; determining who wrote a document with no certain prior information on potential authors; and determining qualities about a document or author based on a sample of text, such as determining how many authors were involved or what the nationality of the author is. Juola says the term
“stylometry” is generally used for the last of the three, but it is more common for the term to be applied to authorship recognition when using linguistic features to answer any or all of the questions he has presented. A linguistic feature is one that deals with the language of the text, including both the grammar (i.e. syntax and structure of words) and semantics (the meaning). Therefore, linguistic features do not include contextual clues such as the time a document was written or physical clues such as handwriting style.

Stylometry exists because of the general belief that each individual has a fairly unique style to their writing, much like individuals have a unique fingerprint. Juola proposes that the reason for this is because every individual has to learn language on their own and therefore every person learns it slightly differently which introduces small differences in ways they communicate. The ways in which writing styles differ can be complex and subtle as is demonstrated later in this document when current effective methods of stylometry are examined in detail.

The story of determining the authorship of the famous anonymous Federalist Papers is a classic example of stylometry. Eighty-five papers were published anonymously in the late 18th century to persuade the people of New York to ratify the American Constitution. The authorship of 12 of these papers was heavily contested. To discover who wrote the unknown papers, researchers have analyzed the writing style of the known authors and compared it to that of the papers of unknown authorship. It was established early after they were published that they were written by either James Madison or Alexander Hamilton. The features used to determine writing styles have been quite varied. Original attempts looked at the length of words, whereas later attempts looked at pairs of words, vocabulary usage, sentence structure, function words, and so on. Almost all studies show the author was Madison. The Federalist papers have become a very popular corpus for testing new methods of stylometry due to their historical significance, availability, clear organization, strict domain, and an overwhelming amount of existing related research.

Stylometry has played a serious role in forensics over the years, especially in the 1990s with the introduction of the cusum, or “cumulative sum”, technique. This method measures the stability of a specific feature throughout multiple documents to establish authorship and was adopted in
many court cases, especially in England. But the accuracy of cusum came under fire, culminating in massive controversy about the use of the technique\textsuperscript{47}. It has since been modified with better accuracy that has been supported by researchers, but Juola cites this as an example of a black eye on the field of stylometry in general and claims that questionable accuracy is the biggest challenge to the field as a whole\textsuperscript{45}. Juola claims that there are three parts to the question of accuracy: technical accuracy of individual methods and how they are evaluated, the quality of the corpus used to obtain results, and researcher expertise. In addressing the last point he specifically deals with the lack of training that many researchers have in understanding how to conduct a proper empirical study.

There are some interesting hurdles in the field of stylometry that, while far from unique, certainly are not common. One such aspect is the fact that the writing style of individuals tends to change over time, and compensating for that is a difficult task. Over time, the entire body of test data for a specific author slowly becomes dated and less reliable.

There is no consensus on what method of stylometry is the most effective or even which features are the most important. The only true consensus is that the method chosen should reflect the text that is going to be classified. For example, counting the number of semicolons might be appropriate for classifying the author of code but not for classifying essays written by children in elementary school.

**How Stylometry Works**

Stylometry is both a study in linguistics and a clearly defined machine learning problem. And at the core of a machine learning problem is identifying and combining the features that represent the classes that are to be learned for the classification task. This is an especially interesting and important topic when formulating stylometry as a machine learning problem due to the virtually limitless number of features that are available.

**Linguistics and Natural Language Processing** Stylometry is partially a natural language processing (NLP) problem. As we will see in detail later, a lot of useful information about the author can be extracted from linguistic features obtained by processing text. A healthy understanding of
processing language is essential to constructing a meaningful stylometry technique. It is useful to view the basic parts of language knowledge outlined by Jurafsky and Martin:

- **Phonetics and Phonology** - The pronunciation and sounds of words. This is not relevant to stylometry.
- **Morphology** - Understanding how words break into components and the meaning of those components (for example, understanding that adding an “s” to some words makes them plural).
- **Syntax** - The meaning of structural relationships between words and how they are used in the construction of language.
- **Lexical and Compositional Semantics** - Lexical semantics deal with the meaning of words and compositional semantics deal with the meanings of words when they are used together (for example, the use of the word “end” when referring to the “end of the street” versus the “end of the 18th century”).
- **Pragmatics** - The relationship between the meanings of words and the intended action that will result from using them, such as understanding that using “please” before making a request is intended to increase the likelihood of the request being honored.
- **Discourse** - The ability to understand the meaning of words outside of a single scope, such as knowing who “he” or “she” is when reading a paragraph about an individual long after they have been explicitly identified.

While not all elements of language processing are directly relevant to stylometry, most of them are. Authors identify themselves through the unique way that they navigate each of these parts of language. Some ideas may be obvious, such as structural differences in sentences (overuse of specific punctuation, for example) or word choices that the author makes (displaying a different understanding of lexical semantics). Others may not be so obvious. An author may have a characteristic trait relating to pragmatics; maybe he or she has a uniquely blunt way of phrasing requests or asking questions that can suggest it is likely he or she authored an unknown document.
Feature Selection in Machine Learning  Central to any stylometry task is the problem of feature selection. There is a virtually unlimited number of possible linguistic features to extract from a given piece of text. Many types of features are finite, such as the distribution of letters in the English alphabet, but others have an incredibly large number of possibilities. Take, for example, the idea of the n-gram feature where each segment of n letters constitutes a possible item in the feature space. There are then $26^n$ possible n-gram features. Given the potentially massive feature space, selecting the features is one of the most crucial components of a stylometric system.

Blum and Langley suggest the best way to approach the problem of extracting the most valuable features is to frame it in terms of a heuristic search with each state being a subset of the total set of available features. A four-part approach is suggested in this method of feature selection. First, an approach to starting the search must be selected, such as forward selection (starting with no features and incrementally adding those deemed relevant), backward elimination (starting with all features and removing those that are irrelevant), or some variation on these approaches. Second, a method for conducting the search (greedy, best-first, etc.) must be chosen. Then the means of evaluating each state of the search must be selected, such as how well each feature splits examples from a training set. The last step is to decide the point at which the search ends, such as when new features have no impact on improving accuracy.

When faced with the “potentially bewildering” number of possible feature sets for a stylometry problem, the need for a robust approach to feature selection becomes clear. But it is also important to recognize that this is not a problem where each researcher begins at square one. While the number of possible features is endless, decades of research have consistently shown that certain types of linguistic features are far more informative than others in determining authorship. In this section, we will review some of the most important types of linguistic features for authorship recognition problems.

Function Words  A function word is a word with little to no meaning on its own and instead is used to define relationships between other words known as content words, such as nouns and verbs like “bed” and “sleep”. Examples of function words are prepositions, conjunctions, and articles such as...
as “of”, “and”, and “the”. They have become a focal point of stylometry research for both their demonstrated effectiveness\textsuperscript{1:45} and historical reasons, as they were used for the original Federalist papers analysis described earlier\textsuperscript{50}. Function words are effective because they are a topic-independent means of understanding how an author expresses ideas and relationships.

The use of function words spawned from the idea that synonym pairs could be used to distinguish between two authors, for example if author $A$ often used the word “thin” and author $B$ often used the word “skinny”. While a slightly modified version of this approach later proved to be fairly successful, the original Federalist papers analysis did not have enough synonym pairs to make a useful feature set. Function words, however, are pervasive throughout written text\textsuperscript{51} and are the most common words to appear across languages and collections of text\textsuperscript{45}.

**Vocabulary Features** Contextually, vocabulary can provide many obvious clues as to the identity of an author. Clear examples of this would be the use of a word in a document that is purported to have been written before the word entered common usage, or an indication of the nationality of an author based on the spelling of a word ("color" vs."colour")\textsuperscript{45}. But from a purely linguistic, statistical perspective, there are many vocabulary features that can indicate authorship without additional contextual clues like publication date or author nationality. Using vocabulary statistics as features can be used to attribute specific vocabulary usage qualities to authors. One such example is the “vocabulary richness” of a document, an estimate of the size of authors’ vocabularies. Another is the synonym-based approach used by Clark and Hannon\textsuperscript{2} that will be explained later in this document.

**Syntactic Features** Features based on syntax can indicate authorship by finding similarities in how an author chooses to grammatically construct the ideas he or she is trying to present. Syntactic features are not about meaning of ideas, but how they are represented. A very basic syntactic feature is punctuation. One author may tend to rarely use exclamation points, or construct longer sentences and thus use more commas. Another example is the use of part-of-speech (POS) tagging to decide what part of speech each word is (such as verb or a noun).
Syntactic features concerning the representation of words and ideas can also be combined with the meaning behind them. Word n-grams are a good example of this. An n-gram is a series of n words considered as a feature. For example, “play” is a unigram, “a play” is a bigram and “seeing a play” is a trigram. In looking at bigrams, clearly “a play” and “to play” have entirely different meanings. Juola points out that looking at an n-gram in this way can draw upon both the vocabulary and syntactic value in the feature.

Another form of n-grams are character n-grams. As Juola explains, this can overcome some of the caveats presented by standard word-based n-grams such as the inability of a word-based approach to recognize the difference between “play”, “player”, and “plays” without deeper analysis into the structural components of the words, known as morphological analysis. A four-gram approach, however, would recognize the common step “play” and likely do a better job of obtaining a feature representative of the author of such a document. The versatility of this approach can be seen in its successful use in authorship recognition for source code.

**Other Features** There are far too many salient features to give attention to all of them in this paper, but there are a few others worth mentioning. Structural features include the average length of sentences, the number of paragraphs, and the general organization of the writing. Other idiosyncratic features, which are used with great effectiveness in the Writeprints method discussed later in this paper, include word misspellings. They might also include revealing mistakes in grammar. More examples of features in all of these categories can be seen later in Table 3.1.

**Methodologies for Attributing Authorship**

While there are many different approaches that can be taken for authorship recognition, machine learning has become the dominant approach to handling the large feature space and high number of possible classifications that are consistent with stylometry. Because of this, this document will continue to focus on automated authorship recognition primarily as a machine learning problem but will mention other methods as well.
Supervised  The most common approach to modern stylometry is some form of supervised analysis of the feature space, mostly by using machine learning techniques. The authorship recognition task is generally to discover the author of an unknown document given some amount of information about the writing styles of potential suspects. Basic statistics were the simplest form of supervised analysis to be used in this context. One could look at relatively simple features such as the average number of words per sentence, length of paragraphs, and complexity of words, and compare the unknown document to the evidence in the training set. A benchmark method was the one developed by J. F. Burrows. It looked at 150 of the most frequent words in a series of poems and assigned scores to new documents based on the frequency of these words compared to the norm. But these methods are not enough to produce highly accurate automated approaches, especially with larger numbers of potential authors. Despite this, Burrows’s work is considered to be quite good and is still used as a basis for comparison against new methods of stylometry that are being developed.

Given the generally massive number of features that can be extracted from text (the Writeprints technique explained later uses tens of thousands), machine learning methods have become the most effective means of supervised analysis. A machine learning method that has found particular success in stylometry is the support vector machine (SVM). Juola cites a variety of work that demonstrates the effectiveness of SVMs in authorship recognition tasks, including somewhat recent research that shows they perform better than all other methods of machine learning that have been applied. While SVMs are becoming more heavily favored, general machine-learning techniques like neural networks, Bayesian classifiers, linear discriminant analysis and decision trees have all found their place as effective approaches to authorship recognition, given an adequate set of features.

Unsupervised  Unsupervised stylometry problems are characterized by segmenting a number of documents of unknown authorship into classes with the goal of the classes representing each author in the sample. A supervised approach is not possible in this scenario as there is no prior information about possible authors. Unsupervised analysis can be a difficult problem in stylometry due to the potentially large number of features. Another difficulty with unsupervised analysis is that the features that are extracted are not going to be independent of one another in most cases, and steps
must be taken to identify sets of independent features. The effects of both of these problems are mitigated by many researchers through principal component analysis (PCA), an approach that finds the axes that capture the greatest variance of the feature vectors and allows comparisons between authors by measuring the distance between the independent components after they have been projected into a lower dimensional space. Unsupervised analysis then treats each of these eigenvectors as a single feature. Other common unsupervised methods in stylometry are multidimensional scaling and clustering.

Creating a reasonably accurate unsupervised stylometric method is a very difficult task and because of this it is also sometimes framed as an easier “similarity detection” problem\(^1\). In this problem there is a known number of authors and exactly two documents attributed to each one. The goal is to find the pairs of documents by the same author using unsupervised analysis. The basic idea is the same—there is no information known about any author—but the additional information about the number of authors and number of documents per author makes the task easier.

**Case Study: Writeprints**

Abassi and Chen identify four major weaknesses in existing stylometry research\(^1\). First, there is a major gap between the amount of energy focused on supervised versus unsupervised stylometry, specifically in terms of the similarity detection problem. Unsupervised stylometry has seen significantly less research. Second, feature sets used by existing methods lack the breadth or depth to deal with more than 20 unique authors with sufficient accuracy. Third, only a very small number of approaches utilize individual feature sets for each possible author, despite their demonstrated effectiveness in machine learning and pattern recognition. Finally, insufficient research has been performed to examine how effective feature sets and various methodologies perform across different domains and which ones could be useful across multiple domains (including some domains that have not been studied at all, such as instant messaging).

**Features and Techniques** A major component of the Writeprints technique is the use of individual-author-level feature sets. Instead of using one feature set for all authors, different feature sets are
created for each individual author. For example, there may be feature sets for misspelled words and the feature “definate” is in the feature set for author A but not for author B, as author B has not been observed to make that spelling mistake.

A full list of features used by Writeprints can be found in Table 3.1. This table illustrates the baseline feature set, meant to represent current trends in stylometry, and the extended feature set, developed for Writeprints and used in conjunction with the concept of individual-author-level features.

**Table 3.1:** Baseline and Extended Feature Sets From Writeprints.

<table>
<thead>
<tr>
<th>Group</th>
<th>Category</th>
<th>Baseline</th>
<th>Extended</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>Word-Level</td>
<td>5</td>
<td>5</td>
<td>total word, % char. per word</td>
</tr>
<tr>
<td></td>
<td>Character-Level</td>
<td>5</td>
<td>5</td>
<td>total char., % char. per message</td>
</tr>
<tr>
<td>Letters</td>
<td></td>
<td>26</td>
<td>26</td>
<td>count of letters (e.g., a, b, c)</td>
</tr>
<tr>
<td>Character Bigrams</td>
<td></td>
<td>&lt;6/76</td>
<td></td>
<td>letter bigrams (e.g., aa, ab, ac)</td>
</tr>
<tr>
<td>Character Trigrams</td>
<td></td>
<td>&lt;17,576</td>
<td></td>
<td>letter trigrams (e.g., aaa,aab, aac)</td>
</tr>
<tr>
<td>Digits</td>
<td></td>
<td>10</td>
<td></td>
<td>digits (e.g., 1, 2, 3)</td>
</tr>
<tr>
<td>Digit Bigrams</td>
<td></td>
<td>&lt;100</td>
<td></td>
<td>2 digit number frequencies (e.g., 10, 11)</td>
</tr>
<tr>
<td>Digit Trigrams</td>
<td></td>
<td>&lt;1,000</td>
<td></td>
<td>frequency of 3 digit numbers (e.g., 100)</td>
</tr>
<tr>
<td>Word Length Dist.</td>
<td></td>
<td>20</td>
<td>20</td>
<td>frequency of 1-20 letter words</td>
</tr>
<tr>
<td>Vocab. Richness</td>
<td></td>
<td>8</td>
<td>8</td>
<td>richness (e.g., hapax legomena, Yule’s K)</td>
</tr>
<tr>
<td>Special Characters</td>
<td></td>
<td>21</td>
<td>21</td>
<td>occurrence of special char. (e.g., @#$%)</td>
</tr>
<tr>
<td>Syntactic</td>
<td>Function Words</td>
<td>200</td>
<td>300</td>
<td>frequency of function words (e.g., of, for)</td>
</tr>
<tr>
<td>Punctuation</td>
<td></td>
<td>8</td>
<td>8</td>
<td>occurrence of punctuation (e.g., !,?, )</td>
</tr>
<tr>
<td>POS Tags</td>
<td></td>
<td>&lt;2,300</td>
<td></td>
<td>frequency of POS Tags (e.g., NP, JJ)</td>
</tr>
<tr>
<td>POS Tag Bigrams</td>
<td></td>
<td>varies</td>
<td></td>
<td>POS tag bigrams (e.g., NP VB)</td>
</tr>
<tr>
<td>POS Tag Trigrams</td>
<td></td>
<td>varies</td>
<td></td>
<td>POS tag trigrams (e.g., NP VB JJ)</td>
</tr>
<tr>
<td>Structural</td>
<td>Message-Level</td>
<td>6</td>
<td>6</td>
<td>e.g., has greeting, has url, quoted content</td>
</tr>
<tr>
<td></td>
<td>Paragraph-Level</td>
<td>8</td>
<td>8</td>
<td>e.g., no. of paragraphs, paragraph lengths</td>
</tr>
<tr>
<td></td>
<td>Technical Structure</td>
<td>50</td>
<td>50</td>
<td>e.g., file extensions, fonts use of images</td>
</tr>
<tr>
<td>Content</td>
<td>Words</td>
<td>20</td>
<td>varies</td>
<td>bag-of-words (e.g., “senior,” “editor”)</td>
</tr>
<tr>
<td></td>
<td>Word Bigrams</td>
<td>varies</td>
<td></td>
<td>word bigrams (e.g., “senior editor”)</td>
</tr>
<tr>
<td></td>
<td>Word Trigrams</td>
<td>varies</td>
<td></td>
<td>word trigrams (e.g., “editor in chief”)</td>
</tr>
<tr>
<td>Idiosyncratic</td>
<td>Misspelled Words</td>
<td>&lt;5,513</td>
<td></td>
<td>misspellings (e.g., “believe,” “thougth”)</td>
</tr>
</tbody>
</table>

The Writeprints Algorithm  The algorithm has two steps. The creation step interprets the style of an author in detail by analyzing the writing over the feature set. This can be thought of as understanding the positive correlations between a piece of text and an author. The pattern disruption step identifies “red flags”–features that have zero relevance to an author–and uses them as a negative correlation between that author and a piece of text.

First a usage pattern—the Writeprint—must be created for an author based on their individual-
author-level feature set. A feature vector is created for all non-zero features and is projected into an $n$-dimensional space though a Karhunen-Loeve (K-L) transform, with $n$ being decided on with the Kaiser-Guttman stopping rule. K-L transform is a method of principal component analysis, except unlike PCA, it is not restricted to using one set of features for all classes (in this case, authors). This $n$-dimensional representation is the usage pattern for each author for a specific piece of text.

The main influence of the individual-author-level feature sets is in pattern disruption. The algorithm takes advantage of the fact that the presence of features an author does not use can be just as effective in an identification task as those the author does use. Two patterns must be created when two author identities are compared to one another because the feature sets are different. A pattern must be created for author $A$ using $A$’s text and $B$’s feature set and for author $B$ using $B$’s text and $A$’s feature set. Then the similarity between $A$ and $B$ is decided by looking at the Euclidean distance between these two patterns.

To return to a previous example, when looking at the similarity between author $A$ and $B$ in a segment of text, the presence of a misspelling of the word “definite” should increase the distance between the two. The amount by which the distance should be increased depends on the disruption value, $dp$. This value is not constant for all disruptor features—it varies depending on how significant the feature is.

Some features, specifically vocabulary features such as the use of the word “Colorado” versus the use of the word “folks,” may be more or less indicative of content than style and their disruption value varies with that based on a modified information gain formula: $dp = IG(c,p)K(syn_{total} + 1)(syn_{used} + 1)$. IG is the information gain of the feature regardless of its significance, $K$ is a constant that is used to control the overall magnitude of disruptor features, and $syn_{used}$ and $syn_{total}$ represent the number of synonyms that exist for that word and the number that have actually been used by the author. In this case there are more synonyms for “folks” than “Colorado,” so a disruptor feature based on “folks” leads to a higher disruptor value.

**Evaluation and Results** Abbasi and Chen evaluated the Writeprints technique on four different data sets, each posing their own challenges. The first is the Enron email corpus, a publicly available
data set of emails with an average of about 28,000 words per author. The second is a collection of eBay comments over a three year period totaling about 23,000 words per author on average. Third are snippets of code from the Sun Java Technology Forum with about 44,000 words per author. And finally, a chat corpus from CyberWatch with only about 1,500 words per author. One hundred authors were selected from each data set at random for the evaluation. There are many interesting differences between the data sets. For example, the first two are asynchronous communication, and the last is synchronous communication. The latter has not been studied to the degree of the former. There are also clear differences in the amount of information available about each author. Overall the most striking thing about the data sets is just how different they are. This is important to note given the fourth point the authors made about weaknesses in the field: there isn’t enough work looking at how stylometric techniques perform across domains. Some features may be more useful in certain domains and not others, and some overall stylometry techniques may be more applicable in some cases and not others.

**Identification Task** Each author’s text was split into 10 parts. Half of them were assigned to the author identity as part of the training set and half were considered to be unknown. The accuracy was determined by performing 10 fold cross-validation and was compared to three other techniques: an ensemble classifier that also uses extended individual-author-level feature sets, an SVM classifier using the extended group-feature set and an SVM classifier that uses the basic group-feature set. Tests were performed with 25, 50, and 100 unique authors.

In the asynchronous communication data sets (email and eBay comments), Writeprints outperformed the other techniques, sometimes quite significantly as the number of potential authors rose. For 100 authors, Writeprints was able to achieve 83% accuracy for the email set and 91% for eBay comments, as opposed to 40 to 77% and 83 to 91% for the other methods. Writeprints was on par with the other approaches in the code segment data set for 25 and 50 authors, and was significantly better at 100 authors, achieving 53.5% accuracy—12 points above the closest competitor. The synchronous chat data set proved to be difficult across the board with 50% accuracy for Writeprints at 25 authors going down to 32% at 100 authors, but this was still significantly better than all the
other methods. The full list of results can be seen in Table 3.2.

**Table 3.2: Identification Task (% Accuracy) From Writeprints**

<table>
<thead>
<tr>
<th>Test Bed</th>
<th>Techniques/Features</th>
<th>No. Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Enron Email</td>
<td>Writeprint</td>
<td>92.0</td>
</tr>
<tr>
<td></td>
<td>Ensemble</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>SVM/EF</td>
<td>87.2</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>64.8</td>
</tr>
<tr>
<td>eBay Comments</td>
<td>Writeprint</td>
<td>96.0</td>
</tr>
<tr>
<td></td>
<td>Ensemble</td>
<td>96.0</td>
</tr>
<tr>
<td></td>
<td>SVM/EF</td>
<td>95.6</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>90.6</td>
</tr>
<tr>
<td>Java Forum</td>
<td>Writeprint</td>
<td>88.8</td>
</tr>
<tr>
<td></td>
<td>Ensemble</td>
<td>92.4</td>
</tr>
<tr>
<td></td>
<td>SVM/EF</td>
<td>94.0</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>84.8</td>
</tr>
<tr>
<td>CyberWatch Chat</td>
<td>Writeprint</td>
<td>50.4</td>
</tr>
<tr>
<td></td>
<td>Ensemble</td>
<td>46.0</td>
</tr>
<tr>
<td></td>
<td>SVM/EF</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>37.6</td>
</tr>
</tbody>
</table>

The most important conclusion to draw from these results is the clear effectiveness of the extended individual-author-level feature sets. While Writeprints performed the best in almost all cases, second place was consistently the ensemble SVM classifier using the same feature sets. In addition, this is some of the most comprehensive stylometry research with large numbers of authors. The effectiveness of the technique in the more traditional setup of asynchronous communication in a standard identification task is highly encouraging for the prospects of authorship recognition on the web and in a large-scale setting. The Writeprints technique suffered the smallest drop in accuracy as the number of authors increased, suggesting it is highly scalable. The authors attribute this success in particular to the use of disruptor features.

**Similarity Detection Task**  Once again each author’s text was split into 10 parts and two new identities were created with five text sections each. This time, however, both identities are considered to be anonymous. So for each task, the similarity detection problem is dealing with 24 unknown identities and attempting to pair them together, whereas the identification problem is dealing with 12 known identities and attempting to find which one is the best fit for each unknown identity. Like the identification task, Writeprints was compared against a series of unsupervised stylometry techniques with both the baseline and extended feature sets and both group and individual-author-level features.
These methods included PCA with both baseline and extended feature sets and a Karhunen-Loeve transform with extended individual-author-level features. The remaining experimental conditions were identical to the identification task.

Unlike the identification task, Writeprints outperformed all methods in every single experiment. Usually the difference is quite significant, as can be seen in Table 3.3. Abbasi and Chen draw similar conclusions as they did earlier, such as the ability of Writeprints to scale more easily to large data sets and more subjects. Like the identification task, the best performance was in the asynchronous communication data sets (Enron emails and eBay comments). Writeprints clearly has made significant progress in demonstrating the effectiveness of unsupervised classification of a similarity detection task with a robust set of features and strong classification technique. This is probably one of the most significant results in the paper given the relatively small amount of attention this task has received by the research community.

**Table 3.3:** Similarity Detection Task (% F-Measure) From Writeprints

<table>
<thead>
<tr>
<th>Test Bed</th>
<th>Techniques/Features</th>
<th>No. Authors</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>Enron Emails</td>
<td>Writeprint</td>
<td>93.62</td>
<td>94.29</td>
<td>85.56</td>
</tr>
<tr>
<td></td>
<td>K-L</td>
<td>75.29</td>
<td>68.23</td>
<td>65.44</td>
</tr>
<tr>
<td></td>
<td>PCA/EF</td>
<td>70.32</td>
<td>56.33</td>
<td>50.82</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>64.32</td>
<td>48.49</td>
<td>34.33</td>
</tr>
<tr>
<td>eBay Comments</td>
<td>Writeprint</td>
<td>100.00</td>
<td>97.96</td>
<td>94.59</td>
</tr>
<tr>
<td></td>
<td>K-L</td>
<td>92.25</td>
<td>84.10</td>
<td>80.93</td>
</tr>
<tr>
<td></td>
<td>PCA/EF</td>
<td>81.19</td>
<td>77.32</td>
<td>72.25</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>75.65</td>
<td>70.02</td>
<td>60.19</td>
</tr>
<tr>
<td>Java Forum</td>
<td>Writeprint</td>
<td>90.13</td>
<td>85.02</td>
<td>76.87</td>
</tr>
<tr>
<td></td>
<td>K-L</td>
<td>77.76</td>
<td>67.63</td>
<td>60.27</td>
</tr>
<tr>
<td></td>
<td>PCA/EF</td>
<td>76.21</td>
<td>66.65</td>
<td>56.10</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>72.90</td>
<td>60.59</td>
<td>42.45</td>
</tr>
<tr>
<td>CyberWatch Chat</td>
<td>Writeprint</td>
<td>68.43</td>
<td>62.88</td>
<td>49.91</td>
</tr>
<tr>
<td></td>
<td>K-L</td>
<td>50.72</td>
<td>42.39</td>
<td>30.77</td>
</tr>
<tr>
<td></td>
<td>PCA/EF</td>
<td>40.00</td>
<td>33.30</td>
<td>19.80</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>39.43</td>
<td>28.62</td>
<td>20.10</td>
</tr>
</tbody>
</table>

One problem with these results is the difficulty in comparing the results between the two tasks. The identification task was measured in terms of accuracy (number of correctly classified identities divided by the total number of identities) whereas the similarity detection task was measured in terms of F-measure: \((2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})\). Recall is number of elements of a class that are correctly identified, and precision is the number of true positives of all elements...
classified as a specific class.

**The Impact of Writeprints** Writeprints is a significant paper in the field of stylometry because it effectively assesses the weaknesses in existing research and proposes a very accurate methodology for demonstrating that they can be overcome. The accuracy across multiple domains for such high numbers of potential authors is essentially unparalleled. The use individual-author-level feature sets will likely be adopted by many more researchers in this field as the high performance that results from them cannot be ignored. The use of pattern disruptors is a novel technique that also shows a lot of promise. And the usefulness of having a very large range of features to analyze and choose from cannot be understated after looking at the results from this technique.

**Case Study: The Synonym-Based Approach**

While the future of highly effective domain-independent stylometry must clearly take the route of using large numbers of features, there is still an important place for novel approaches that rely on one set of features, as they can later be incorporated into umbrella techniques like Writeprints. Clark and Hannon developed a great example of this by utilizing synonym-based features to identify authors. Their research is reviewed in this section.

The foundational hypothesis of this method is that the repeated choice of a specific word given a selection of potential synonyms is tightly related to writing style and can be used to identify an unknown author. The approach is to measure the selection of each word and weight that measurement by accounting for how common that word is and how many choices the author had. Clark and Hannon give the example of choosing the words “red” and “computer.” With “red”, the author has a wide range of choices to express the concept he or she is attempting to communicate. But there are not as many synonyms for “computer”, so it is likely not going to be as indicative of writing style. Furthermore, the choice to use the word “red” is not as indicative of writing style as the use of “scarlet”, as “red” is much more common. The process of weighting and the end result is explained in greater detail in Model 1.

Clark and Hannon developed three models throughout the evolution of their technique and
evaluated each of them individually. The first model is more rudimentary than the others but serves
to explain the basic concepts quite well. The second and third models expand upon the findings
from evaluating Model 1 in different ways.

**Model 1** The feature vector for Model 1, $f_1$, for a word $w$ contains two elements: the **number of synonyms** $s$ for the word, based on Princeton’s WordNet database\(^1\), and the “**shared text frequency**” $n$ for $w$ where $n = \min(n_1, n_2, ..., n_k)$ for $k$ authors and $n_i$ is the frequency of $w$ for
author $i$.

To compare an unknown author $u$ to a known author $k$ where $k \in T$ and $T$ is the set of all potential authors, a match score is calculated between $u$ and every $k$. The match algorithm is summarized as follows:

Initialize the match value $m$ to 0. For each pair of unique words $w_u$ and $w_k$ used by
author $u$ and author $k$ respectively, and where $w_u = w_k$, calculate the feature vector $f$
as described above and update $m = m + f[n] + f[s]$. The unknown identity $I$ is then
classified as: $I = \text{arg max}, k \in T \ match(u, k)$.

An example of the final process for deciding who authored a piece of text can be seen in Figure
3.1. This example demonstrates the difference between the weighted synonym-based approach and
a simple “bag-of-words” approach\(^2\).

**Model 2** Two major weaknesses presented themselves in the initial evaluations of Model 1.
First, the biggest indicators of authorship were a small set of words such as common pronouns and
helping verbs like “it” and “having”. The authors believed that word choices in these scenarios are
not adequate indicators of authorship. Second, words were being given equal consideration even
when one should be more indicative of authorship than another, such as “red” and “scarlet”, as
indicated earlier.

The first problem was resolved by implementing a stop-word list and taking into consideration
the global frequency of words. The stop word list consisted of 319 of the most common words in the
English language, different versions of those words obtained through morphological processing, and

\(^1\)http://wordnet.princeton.edu/

\(^2\)
Figure 3.1: An example from Clark and Hannon demonstrating how values are calculated to compare two known sentences to an unknown sentence. In this case the unknown sentence is correctly attributed to author X because of the highly salient feature of choosing the word “verdant”.

$S = \text{#uses} \times \text{#synonyms}$

$S = 1 \times 8 + 1 \times 11 + 1 \times 1$

$S = 20$

$S = 1 \times 26$

$S = 26$
a list of the most common 90% of first and last names from the 1990 US census.

The second problem was resolved by weighting each word based on the frequency in which it appears in the overall text. The feature vector for Model 2, $f_2$, for a word $w$ contains the two parts of $f_1$ plus the global frequency $g$ of $w$ and the sum $u$ of the global frequencies of all synonyms of $w$. The match value update equation in this model is therefore: $m = m + f_1[n] \ast f_2[u]/f_2[g]$.

These modifications strengthen Model 2 because it takes into account not just the synonym choices an author makes but how substantial those choices are compared to how often alternate words are used throughout the corpus. If everyone is using the word “scarlet”, it will be less significant than if almost no one else is using it.

**Model 3** The focus of model 3 is to combine morphological analysis into the process so that two forms of the same word will not be treated differently, such as some noun and a plural form of that noun. This is accomplished by stemming, which is the process of converting all words to their root form (for example, turning “plays” into “play”). Everything else is the same as Model 2.

**Evaluation and Results** Clark and Hannon used 1.3 million words from four authors who were all born within one year of each other (1802-1803) to evaluate their models. Each author ended up with six to seven thousand unique words. The text was preprocessed to remove potentially identifying information that is not based on linguistic style such as the table of contents. Results were calculated for sets of two, three, and four authors using four-fold cross-validation (where each fold is an equally sized piece of the author’s text).

The results for models 1 and 2 are summarized in Table 3.4. Model 3 is not included because it ended up having no significant impact on the accuracy of the system. Model 2 is clearly a huge improvement over Model 1. It has high accuracy and a gentle degradation as the number of authors is increased. In addition, Model 2 also provided the largest gap between the top choice and the second choice for assigning authorship to the unknown document. This was especially important when looking at the differences between Models 2 and 3, as they provided the same end accuracies, but the gap was larger for Model 2. This can be interpreted as Model 2 providing the highest confidence in the results.
Table 3.4: Results for models 1 and 2 from Clark and Hannon².

<table>
<thead>
<tr>
<th>Authors = 2</th>
<th>Authors = 3</th>
<th>Authors = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Model 1</td>
<td>0.625</td>
<td>0.625</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.979</td>
<td>0.979</td>
</tr>
</tbody>
</table>

The results from this technique are very promising. They demonstrate a very accurate method of stylometry based on one type of feature. It is research like this that can make the case for the inclusion of a new feature type into existing robust sets like the one presented in Writeprints. There are some shortcomings from this research, such as the somewhat limited number of authors, which is not uncommon for traditional stylometric research but is quickly becoming outdated.

**Other Notable Approaches**

Just as there are innumerable features that can be used to attempt authorship recognition, there are seemingly innumerable approaches that have been developed for the task. Though an SVM is now favored for most approaches, it is worth noting the introduction of neural networks into the field by Tweedie, Singh, and Holmes⁵⁴. Papers like theirs helped strengthen the growth of machine learning as the primary approach for stylometry. There have been other novel approaches such as the work by Oakes using ant colony optimization to look at the Federalist Papers⁵⁵. While the work is not overwhelmingly accurate, it is another example of exploratory research that supports the field in general. Zheng, Huang, and Chen also looked at applying authorship recognition to online text (a precursor to some of the work in Writeprints), and also evaluated a number of machine learning techniques including neural networks and SVMs, concluding that SVMs are the superior classification approach⁵⁶. Juola and Baayen present interesting work demonstrating the effectiveness of determining authorship using a measurement of “linguistic distance” through cross-entropy⁵⁷.

### 3.3 Adversarial Stylometry

Current approaches to stylometry are increasingly focused on larger-scale approaches than their predecessors. The ability to attribute authorship of an unknown document amongst a very large set of potential authors and applying that technique across multiple domains is desirable, as it is now
becoming possible. With the advent of these more effective techniques, concerns that have been raised in the past about protecting the privacy of individuals are being given a more serious look. This section examines these concerns in greater detail and introduces adversarial stylometry.

We define “adversarial stylometry” as the notion of applying deception to writing style in order to affect the outcome of stylometric analysis. There is great potential for this in terms of protecting privacy and concern that could significantly affect the effectiveness of newer, more powerful methods of stylometry. This section reviews some of the early work in the field and then explains our work on circumventing authorship recognition, which is the first formal work on the subject.

3.3.1 Early Work in Adversarial Stylometry

The question of whether or not it is possible to be deceptive in one’s writing style is one that has not received much attention until recently. Even in Juola’s recent book on authorship attribution, one of the largest and most comprehensive to date, the notion of deceptive writing style is only briefly touched upon—though he asserts that it is a fruitful and important area for future research.45

Early work in the field we now call adversarial stylometry includes research by Dr. Josyula Rao and Dr. Pankaj Rohatgi studying the impact of stylometry on pseudonymity and determining that 6,500 words of writing is enough to leak the identity of an author.17 Others have looked at the potential of automatically obfuscating a document to preserve anonymity and determined that in case of the Federalist Papers it took just 14 changes per 1,000 words to shift authorship from Madison to Hamilton.4 This work, however, does not modify the actual text in the documents. They instead modified the numerical feature vectors after they have been extracted from the original text.

Previous research in the field has also looked at authorship recognition and pastiche by comparing the work of Gilbert Adair to that of Lewis Caroll, whom he was trying to imitate by writing follow up stories to Alice in Wonderland.58 However, our work is the first to apply adversarial stylometry to actual documents written by humans to defeat stylometry and test the results against multiple methods and features sets. Patrick Juola validated the effectiveness of adversarial writing on stylometry by evaluating the Brennan-Greenstadt corpus with JGAAP2 and also demonstrated

2JGAAP is available at http://jgaap.com

Chapter 3: Managing Identity and Anonymity with Stylometry
some methods’ resistance to such writing samples\textsuperscript{59}.

**Obfuscating Document Stylometry to Preserve Author Anonymity**  
Kacmarcik and Gamon studied the potential for intentionally obfuscating writing style in the Federalist Papers data set by examining if they could influence the attribution of the unknown documents to Hamilton instead of Madison. They were concerned with three basic questions when facing the task of modifying a document to escape detection: How easily can the parts that must be modified be identified and changed? How much resistance do stylometry techniques offer? How much work has to be done to successfully obfuscate a piece of text?\textsuperscript{4}

Kacmarcik suggests that the first step to preserving anonymity is to make changes to the document that fix stylistic idiosyncrasies but will not alter the meaning of the document. Such changes include fixing spelling errors, formatting, and punctuation. These features can be used with great effect by modern stylometry techniques, as demonstrated by the Writeprints method.

Next, however, are the more complex stylistic markers, such as word choice and sentence structure. This study only examines word usage in an effort to demonstrate their hypothesis: that it is possible to obfuscate writing style with a small number of changes to the feature set warrants a closer look and merits further research. An example is given that illustrates the point very well: in the Federalist Papers, Hamilton uses the word “while” 36 times and the word “whilst” just once, whereas Madison never uses “while” and uses “whilst” 9 times. This is a clear indicator of style that will have to be corrected if authorship of an unknown Federalist Paper is going to be assigned based on vocabulary-based features\textsuperscript{4}.

When deciding how to modify a document, there are two approaches that can be taken. The first is to attempt to obfuscate the style by hiding within the set of potential authors, known as an obfuscation attack. In this way the modification would mean conforming to the standards put forth by the other authors in the set. The second is to attempt to mimic the style of a particular author, also known as an imitation attack\textsuperscript{60}.

This study was limited to modifying only word-frequency features, but deciding what words to modify is an essential part of the problem. The goal is to modify as few words as possible, so the
most salient word features must be chosen. To determine which word features were the best, a Decision Tree Root (DTR) method of ranking was used. To extract the top $n$ features, a decision tree is created $n$ times (though the whole tree doesn’t need to be created, only the root needs to be determined) and after each iteration the root feature is removed and added to the ranked feature list.

Simply taking the top $n$ features as determined by DTR ranking did not prove to be effective enough on all documents—about half of Madison’s documents were still attributed to him with high accuracy. In an attempt to broaden the effectiveness, the top DTR features were given a frequency threshold so only the most common word features were used, but they were still ranked by DTR. This proved to be much more effective than the original DTR ranked list.

Five different classifiers were constructed, all using a linear-kernel SVM. Each classifier dealt with a different set of features based on a different piece of research. The features ranged from 70 common function words based on the work of Mosteller & Wallace to selecting just three basic function words. For testing purposes, the feature vectors were modified directly instead of modifying the document itself.

The primary measurement for success in the evaluation of the obfuscation is reduction in the confidence measure. Measuring confidence is important here because in all cases every document was classified as having been authored by Hamilton instead of Madison, so a better measurement was needed to decide just how well the obfuscation attack performed. The confidence measure was reduced between 75% and 99%, with an overall average reduction of 86.86%.

The secondary measurement for success is the number of changes required per 1,000 words to cause a document to be attributed to Hamilton instead of Madison. The average number of changes required was 14.2 per 1000 words. It is hard to say whether or not this is an acceptable number of changes for an automatic obfuscation attack as it is easy to determine what words should be removed, but properly inserting new words into a document could be a difficult problem.

Looking back to the original three questions that were posed, this research seems to provide some answer to all of them. It is clear that it is quite easy to modify the feature vectors in order
to confuse an otherwise acceptable classifier, but it does not address the difficulty of automatically modifying the actual text of the document. Following that, it seems that the stylometry techniques studied here offer little resistance to these obfuscation attacks but only a small sample of limited techniques were studied. And finally, only a small amount of text needs to be changed to alter authorship attribution of these documents. Overall this study is not a conclusive definitive answer to the questions raised by examining the idea of adversarial stylometry, but it was the first step in discovering if there is any merit to the claim that malice and deception are an important area of research in stylometry.

**Can Pseudonymity Really Guarantee Privacy?** Rao and Rohatgi define pseudonymity as the ability of a user to conduct a series of interactions on the web that can all be linked back to a single pseudonym, or false identity, that cannot be linked to the true identity of the user\(^7\). They name a number of categories and examples of privacy enhancing technologies that aid this goal, such as anonymizing agents that allow a user to not be linked back to his or her IP address. The problem is the generally held belief that these technologies, if applied properly, are enough to counteract privacy threats. On the contrary, stylometry can present a threat to anonymity despite the best efforts of an individual to use these pseudonymity technologies. Rao and Rohatgi put out a call to action to the privacy community to take this threat seriously and present a series of experiments intended to demonstrate the threat and make the case that it is not something that should be ignored. Their research is reviewed in this section.

In order to demonstrate that existing privacy technologies are not sufficient given the threat of stylometry, Rao and Rohatgi constructed a study and performed a series of experiments. This study was performed on three data sets with large numbers of authors, and each author had a considerably large amount of available writing samples. These were posts from the sci.crypt newsgroup, the IPSec mailing list, and the RFC Database\(^7\). The stylometry technique used in all experiments was principal component analysis, considered one of the current state-of-the-art methods when the paper was published in 2000. The feature set for the study were the frequencies of function words from a specific list, a feature measuring the probability that a randomly selected word would not
be a function word. Authorship was chosen based on the Euclidian distance between an unknown sample and the potential authors.

The newsgroup persona clustering experiment was performed on five years of sci.crypt newsgroup data. Headers and signatures were removed, as were quotes and “non-language” words. The scenario under consideration is the attempt to break pseudonymity in the case of a newsgroup where multiple pseudonyms could be linked to a single identity. The data set was divided into authors and only used personas that met the minimum requirement for number of posts and words. Authors with very large amounts of data were split into two personas, and it is that set of authors that was used for testing with the goal of pairing the unknown authors together correctly.

There were two parts to this experiment. The first dealt with authors having a minimum of 50 posts and 5,000 words. Sixty-eight authors met the criteria for being split into two personas, and the point of the split was temporal, so the first pseudonym was the first half of their postings and the second pseudonym was the second half. Then a similarity-detection problem was constructed and evaluated using PCA resulting in 58.8% accuracy in correctly pairing the authors.

The second part increased the minimum number of words to 10,000. Forty-two authors satisfied this in a way that allowed for two individual pseudonyms to be generated in the same fashion as above. Once again it was treated as a similarity-detection problem, and in this case the success rate was 80.9%.

There is a big difference between the accuracies of the two parts so the authors did some more formal experiments and found that the threshold for protecting one’s identity in these experiments was about 6,500 words. That is, the change in accuracy between using authors with 6,500 words and 10,000 words was not significant.

The second experiment is an early look at cross-domain applications of stylometry. The authors felt it may not be realistic to pair pseudonyms in a newsgroup, as it is probably unlikely in reality that many people use multiple pseudonyms for the same forum. But the same can’t be said for using different pseudonyms in different domains. Rao and Rohatgi were able to identify seven authors who posted to the IPSec mailing list and the sci.crypt newsgroup at the same time, and participated
enough to generate sufficient data for testing (more than 10,000 words). This experiment tried to link the author from the mailing list to the author in the newsgroup. Five out of the seven authors were correctly paired and identified using PCA, demonstrating a privacy threat that is a more realistic scenario.

The authors attempted to also pair authors from the newsgroup and mailing list data sets to the RFC data set but were unable to do so effectively. They suspected it had to do with the different requirements for the domain—the RFC data set was much more formal than the others. They believe some sort of transformation could be done to link the two at some point in the future. That research has not been done, but it does lay the groundwork for the idea that an extended feature set identifying more nuances in writing styles could be more effective as was shown in the Writeprints technique.

The Impact of the “Pseudonymity” Paper  This paper laid the groundwork for approaching stylometry from an entirely new angle: as a threat to privacy. Many of the techniques and standard stylometry knowledge presented in this paper had been presented elsewhere in great detail and improved upon significantly since\textsuperscript{1,4,60}, but little work spawned from their call to action until recently.

They also make some interesting claims in the paper that may no longer hold weight. For one, they say that fairly short messages are safe from stylometry techniques. This is not necessarily the case anymore, as Writeprints demonstrated effectiveness on both eBay comments and short instant message communication, both of which are characteristically shorter in length than the data sets used here. Additionally, they claim that modifying the writing style of the document could potentially overcome the threat posed by stylometry, but doing so would change the meaning of the message. This also is likely not true, our work later in this document will show.

3.3.2 The Role of Stylometry in Privacy and Anonymity

A multidisciplinary approach to privacy that includes much more than traditional location-based circumvention and anonymity tools has long been suggested\textsuperscript{62}. Users of the Internet with the desire to publish anonymously may also desire to hide their writing style to circumvent stylometry.
techniques. The goal of these authors is to keep their identities private. With the growth of stylometry as an accurate means of determining authorship, the question of how it affects privacy and anonymity in the information age is becoming increasingly important. Privacy and anonymity are held in high regard by many activists, journalists, and law enforcement officers. The introduction of adversarial stylometry and the use of circumvention passages on stylometry are possible means of ensuring the privacy and anonymity of an individual. The largest example of the value of anonymous speech to these groups is the Tor Project, an anonymous communication tool originally developed by the Naval Research Laboratory that is utilized by hundreds of thousands of individuals, including law enforcement, journalists, activists, businesses, and more\textsuperscript{63,64}.

The privacy issues concerning stylometry can be summarized through an example scenario: Alice the anonymous blogger vs. Bob the abusive employer: Alice is an employee at Bob’s company, the Widget Design Corporation. Alice, a long-time employee, wishes to draw attention to the various systemic abuses that she has observed at the company under Bob’s management, such as harassment, unpaid overtime, and employees being encouraged to rip off competing widget designers. She decides to do this by publishing an open anonymous letter she has personally authored detailing these abuses on the web. She takes great care to use privacy-enhancing technologies as outlined by Rao and Rohatgi, so that it is very difficult, if not impossible, to trace her post back to her identity or her IP address\textsuperscript{17}.

The letter draws criticism of the company from the press, and Bob decides to discover the author, believing it came from within his company due to the details it revealed. In this case, Bob’s company has about 100 employees. If Bob has access to a stylometry system such as Writeprints that has very high accuracy for large numbers of potential authors\textsuperscript{1}, he can collect 6,500 words from each employee’s writing (probably though various reports and emails that they have written in their time at the company) and come to believe that Alice is the writer of the document with over 90\% probability. He can then take action against Alice that may compromise her job.

This hypothetical scenario represents a reasonable threat that is within the range of ability for current methods of stylometry. However, the threat presented to anonymity is not purely hypothet-
ical. In his 2011 book, “Inside Wikileaks”, former Wikileaks spokesperson Daniel Domscheitt-Berg discussed the potential impact of stylometry on the organization after attending a presentation on adversarial stylometry at the 26th Chaos Communication Congress:\(^3\):

“If someone had run Wikileaks documents through such a program, he would have discovered that the same two people were behind all the various press releases, document summaries, and correspondence issued by the project. The official number of volunteers we had was also, to put it mildly, grotesquely exaggerated.”\(^6^5\)

The reality of using stylometry to identify individuals who wish to remain anonymous can also be seen from the perspective of law enforcement. It is highlighted in the recently commissioned FBI report, *State-of-the-Art Biometric Excellence Roadmap* (SABER):

“As non-handwritten communications become more prevalent, such as blogging, text messaging and emails, there is a growing need to identify writers not by their written script, but by analysis of the typed content. Currently, there are some studies in the area of writers colloquial analysis that may lead to the emerging technology of writer identification in the blogosphere.”\(^6^6\)

Later in this chapter we examine the domain of deception and writing style in real-world scenarios like the “Gay Girl in Damascus” blog in which an American writer masqueraded as a Syrian woman. This fraudulent blog was uncovered through forensic means, but this research shows that stylometry demonstrates a strong correlation between the pseudonymous blog and writing samples by the true identity of the author\(^6^7\).

The technical ability for a method of stylometry to present such a threat to anonymity is explained in the next section. The methodology used in Writeprints in particular has demonstrated the potential for identifying a single author among up to 100 unique identities. As a result, bloggers and others may have reason to circumvent stylometry to protect their privacy and anonymity. Adversarial stylometry can be viewed in this way as a means for maintaining privacy or anonymity.

\(^3\)http://events.ccc.de/congress/2009/wiki/
The threat that large-scale authorship recognition can generally present to privacy and anonymity has been studied in Arvind Narayanan et al. in their 2012 work “On the Feasibility of Internet-Scale Author Identification” [16]. This work demonstrates the ability to discern an author 20% of the time when looking at a corpus of over 100,000 potential identities. The precision can be increased to 80% by halving the resulting recall and guessing the true author via confidence estimation. This work presents a new urgency for the anonymity and privacy concerns that are part of the motivation behind adversarial stylometry research.

3.3.3 Anonymous Speech and Public Discourse

The need for anonymity on the web has been well established. Individuals may require it for purchasing goods and services or communicating via email; corporations and militaries need trusted communication channels without revealing participants[14].

Systems and methods for anonymous communication are varied. Mixes were a popular early method of anonymous communication, and are the foundation of many subsequent systems[68]. A mix combines a series of encrypted messages into a batch and then decrypts and delivers some portion of that batch. The concept was introduced in 1981[69] and was later expanded into the concept of a mix network, which allows the sending of message through a series of mixes. Crucial components of a mix-based anonymous communication system include how to select the series of mixes in a mix network[70;71] and the algorithm used for dispensing messages from a batch, which are called the flushing algorithm or batching strategy[72;73]. There are a few particularly well-known deployments of mixes and mix networks including Mixmaster[74], and Mixminion[75].

Possibly the most widely used design for anonymous communication systems today is the onion routing approach, in which a series of routers relay client traffic instead of other clients[68]. This approach, similar to that of a mix network, was developed in 1996 by Goldschlag et al.[76] and was later improved upon with a formal method for the distribution of public keys[15].

All of these anonymous communication systems do not, however, offer any protection against the identification of users via linguistic means such as stylometry. The examination of the impact of linguistic authorship recognition has only arisen very recently. The first such tool for anonymization
of linguistic style is Anonymouth, a tool that identifies the most salient linguistic features attributed to an author and suggests ideal measurements as targets for adjusting linguistic style in order to preserve anonymity.

### 3.3.4 Detecting Deception in Stylometry

This effect can be seen in Figure 3.2, which compares both a modified ($D'$) and unmodified ($D$) unknown document against Hamilton ($H$) and Madison ($M$). As we know, the true author of the unknown document is Madison. The important comparisons here are $MvH$, $MvD'$ and $MvD$. When distinguishing between two authors who are in fact different, as in $MvH$, there is a slow decline in accuracy as features are removed. When distinguishing between two authors who are in fact the same, as in $MvD$, there is a consistently low accuracy. This makes sense because they are...
the same author, and thus it should be hard to distinguish between them. But when the \( MvD' \) case is looked at, the classifier is at first able to distinguish between them very accurately due to the modifications that make it look like \( D' \) is authored by Hamilton. But as the most salient features are removed the same features that were modified by Kacmarcik’s method–it becomes harder to distinguish between them very quickly, because they are in fact the same author.

The method that Kacmarcik and Gamon created for obfuscating documents can also be used to perform deeper obfuscation through modifying a larger number of features and performing more changes on each one. The effect this has on detection is presented in Figure 3.3. This graph shows the difference in accuracy between \( HvD' \) and \( MvD' \) based on the number of features modified (0, 14, or 35). As the number of modified features increases, so does the slope of the curve. This shows that it is possible to make it harder to detect obfuscated documents by increasing the amount of modification that is performed\(^4\).

### 3.4 Circumventing Authorship Recognition

#### 3.4.1 Methodology

We collected writing samples from human subjects that represented both their unmodified writing style and attempts to create adversarial passages to circumvent stylometry methods. We identified three methods of creating documents to circumvent stylometry, two of which are written by human subjects and one of which is an automated method. We implemented three methods of stylometry to analyze the effect of these adversarial passages by attempting to recognize the original author.

The next section explains our method for creating a survey and collecting the data required for the unique adversarial corpus that this study was based on. We then detail the three types of circumvention passages that have been examined in this study. Finally we summarize the implementation of each method used in the study and their corresponding feature sets.

**Study Setup & Format**  While there is a virtually limitless number of corpora with known author information that can be used for stylometry research, we had to develop our own unique corpus for evaluating the effects of adversarial passages. This was necessary because in addition
to a large number of texts from unique authors we also needed adversarial texts written by those same authors. The collection of this data was performed through the use of a survey by voluntary participants of various backgrounds, though all participants held at least some college education. The Brennan-Greenstadt adversarial stylometry corpus is freely available\textsuperscript{4}.

There were three basic elements for participation in the survey. First, each author had to submit at least 6,500 words of pre-existing sample writing. Each writing sample had to be from some sort of formal source, such as essays for school, reports for work, and other professional and academic correspondence. This was intended to eliminate slang and abbreviations, instead concentrating on the consistent, formal writing styles of everyone involved. This also helped to limit possible errors that are not a result of the malicious attempts but nonetheless could have an effect on the accuracy of the authorship attribution. Participants submitted at least two sample documents, with a median submission of five documents per author, which were combined and split into approximately 500-word sample passages.

Second, each author had to write an obfuscation passage in which they tried to hide their identity through their writing style. This was accomplished by writing a new 500-word passage on a specific topic. The task given to them was to write a description of their neighborhood to someone who has never been there as part of a college application. This task was designed to encourage thought about the writing style and not weigh down the author with having to think about a complex topic or be too creative with their content.

Third, each author had to write an imitation passage in which they tried to imitate another author’s style. For this task the participants were given a 2,500 word sample from \textit{The Road} by Cormac McCarthy to model their passage after. This selection was made for a variety of reasons. Imitation and pastiche is a common writing task, as can be seen through contests such as the International Hemingway Imitation Competition and the Faux Faulkner contest\textsuperscript{67}. In addition, Cormac McCarthy has a distinct writing style:

\begin{quote}
\textquote{\textit{On the far side of the river valley the road passed through a stark black burn. Charred}}
\end{quote}

\textsuperscript{4}https://psal.cs.drexel.edu/tissec

\begin{quote}
\textsuperscript{67}Chapter 3: Managing Identity and Anonymity with Stylometry
and limbless trunks of trees stretching away on every side. Ash moving over the road
and the sagging hands of blind wire strung from the blackened lightpoles whining thinly
in the wind.”

Since the participants are not professional writers this would allow them to make a fair attempt
at emulating another style. The popularity of the book was also a factor in its selection for this
study as we wanted to engage participants with reading samples that had broad appeal to a variety
of readers. In engaging participants we sought more effort in writing passages and thus more robust
results. The writing task given to participants was to narrate their day from the point at which
they get out of bed, and to do so using a third-person perspective. This is also similar to the events
in the sample text. For testing purposes, an additional 2,500 words were taken from *The Road* and
used as training text for Cormac McCarthy along with the original sample. It should also be noted
that the excerpt distributed to readers is freely available as a promotional passage from the book.

Asking the participants to write the obfuscation passages before the imitation passage was in-
tentional. We were concerned that if participants chose to write the imitation passage first then all
of the obfuscation passages would simply read as a second Cormac McCarthy imitation attempt.

3.4.2 The Brennan-Greenstadt Corpora

We have published two freely available research corpora. The first is the Brennan-Greenstadt corpus,
which is based on a survey conducted through Drexel University and contains 12 authors who
volunteered their time and were not compensated for their efforts. This corpus was the basis for our
original work on adversarial stylometry\(^6\). The second is the Extended-Brennan-Greenstadt corpus
containing 45 authors solicited through the Amazon Mechanical Turk platform. Submissions were
vetted against a series of guidelines to ensure the quality of the content, as described below.

**Brennan-Greenstadt Corpus** Participants for the Brennan-Greenstadt corpus were solicited
through classes at Drexel University, colleagues, and other personal relationships. This provided us
with submissions from 12 authors. The Brennan-Greenstadt corpus used an earlier version of the
survey which had two relaxed requirements. Authors were only required to submit 5,000 words of
pre-existing writing and they were not required to fill out a demographic survey.

While this corpus was sufficient for preliminary results presented in earlier work\textsuperscript{60}, we desired a more robust corpus in order to confirm our original findings in a larger author space with a greater diversity of writers and tweaked survey requirements.

**Extended-Brennan-Greenstadt Corpus** We utilized the Amazon Mechanical Turk (AMT) platform to create a large and diverse corpus that could be used for more robust analysis. AMT\textsuperscript{5} is a platform that provides access to a large and diverse population that is willing to perform human intelligence tasks. Participants choose tasks that they would like to complete in exchange for a sum of money decided by a task creator.

A corpus equivalent to the Brennan-Greenstadt Corpus was creating using AMT, resulting in training, obfuscation, and imitation passages of 45 unique authors. The creation of this data set is discussed in detail in Section 4.2.

### 3.4.3 Circumvention Approaches

We developed three methods of circumvention against stylometry techniques in the form of obfuscation, imitation and machine translation passages. Two of these, obfuscation and imitation, were manually written by human subjects. These passages were very effective at circumventing attempts at authorship recognition. Machine Translation passages are automated attempts at obfuscation utilizing machine translation services. These passages were not sufficient in obfuscating the identity of an author. The full results and effectiveness of these circumvention methods are detailed in the evaluation section.

**Obfuscation** In the obfuscation approach the author attempts to write a document in such a way that their personal writing style will not be recognized. There is no guidance for how to do this and there is no specific target for the writing sample. An ideal obfuscated document would be difficult to attribute to any author. For our study, however, we only look at whether or not it successfully deters recognition of the true author.

\textsuperscript{5}https://www.mturk.com
Imitation  The imitation approach is when an author attempts to write a document such that their writing style will be recognized as that of another specific author. The target is decided upon before a document is written and success is measured both by how successful the document is in deterring authorship recognition systems and how successful it is in imitating the target author. This could also be thought of as a “framing” attack.

Machine Translation  The machine translation approach translates an unmodified passage written in English to another language, or to two other languages, and then back to English. Our hypothesis was that this would sufficiently alter the writing style and obfuscate the identity of the original author. We did not find this to be the case.

We studied this problem through a variety of translation services and languages. We measured the effectiveness of the translation as an automated method as well as the accuracy of the translation in producing comprehensible, coherent obfuscation passages.

We performed three language experiments in addition to the English baseline. In all cases the original and final language were English. We performed single-step translations from English to German and back to English as well as English to Japanese and back to English. We then performed two-step translations from English to German to Japanese and then back to English. German was chosen for its linguistic similarities to English and Japanese for its differences.

The two machine translation services we compared were Google Translate\(^6\) and Bing Translator\(^7\). Both services are free and based on statistical machine translation.

3.4.4 Methods and Feature Sets

We selected a series of stylometry techniques that represent a variety of potential approaches both in machine learning methodology and feature selection. The feature selections range from basic to comprehensive and the methods from simple and novel to robust and unique.

Neural Network and the Basic-9 Feature Set  The most straightforward stylometry techniques are those that use traditional machine learning methods with some set of linguistic fea-

\(^6\)http://translate.google.com
\(^7\)http://www.microsofttranslator.com
Table 3.5: The methods and feature sets examined in this study.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Network</td>
<td>Basic-9</td>
</tr>
<tr>
<td>SVM</td>
<td>Writeprints-Static</td>
</tr>
<tr>
<td>Synonym-Based</td>
<td>Vocabulary Frequency</td>
</tr>
</tbody>
</table>

The effectiveness of neural networks\(^4\) in stylometry established machine learning as an integral part of modern authorship analysis. We implemented a neural network with a simple, straightforward feature set. The purpose of the simple feature set and basic machine learning approach is to demonstrate the effectiveness of stylometry even with a limited representation of something as complex as writing style.

The features used for the neural network and SVM classifiers, which we will call the “Basic-9” feature set, include nine linguistic measurements: number of unique words, lexical density, Gunning-Fog readability index\(^8\), character count without whitespace, average syllables per word, sentence count, average sentence length, and the Flesch-Kincaid Readability Test\(^9\). The number of hidden layers in the Neural Network classifier was based on the number of features and the number of classes: 

\[
\text{(number of features + number of classes)}/2
\]

The feature extraction for this set was done with the JStylo tool\(^10\).

**Synonym-Based Approach** Explained in full detail in Section 3.2.

**Support Vector Machine and the Writeprints-Static Approach** As discussed in Section 3.2, Writeprints is one of the most successful methods of stylometry that has been published to date because of its high levels of accuracy on a range of data sets with large numbers of unique authors. Unfortunately, this accuracy comes at a high computation cost. In order to perform the robust experiments we designed for this study, we have created our own approximation of the Writeprints algorithm that performs comparably to the original but has much lower computation cost that allows us to run large numbers of experiments in a reasonable time frame. We will summarize the

\(^4\)http://en.wikipedia.org/wiki/Gunning_fog_index  
\(^8\)http://en.wikipedia.org/wiki/Gunning_fog_index  
\(^9\)http://en.wikipedia.org/wiki/Flesch-Kincaid_readability_test  
\(^10\)JStylo is available at https://psal.cs.drexel.edu
Writeprints method and highlight the feature sets created for this approach and how we merged those feature sets into our approach.

The Writeprints technique constructs a single classifier using feature sets that are specific to each individual author’s rather than being generalized across the set of potential authors. The method has two major parts: write print creation and pattern disruption. The write print creation step constructs $n$-dimensional hyperplanes that represent an individual author’s writing-style, where $n$ is the number of features in the feature set. The pattern disruption step identifies zero usage features and shifts the Writeprint representation further away from Writeprints that have non-zero values for the same features which in turn decreases the level of stylistic similarity between two separate authors. There are many nuances to this approach that we will not discuss here but are described in detail in the original research paper on Writeprints.

One of the most valuable pieces of research that has come from the creation of Writeprints is the baseline and extended feature sets\(^1\). The baseline data set has 327 features, whereas the extended set contains tens of thousands. The primary difference between these two data sets, however, is that the baseline set contains only static features in that the contents do not change with the addition and removal of documents. The extended feature set contains many elements that are based on the documents being classified, and is much larger as a consequence. Examples of these dynamic features are the most common misspellings and character bigrams in the corpus. We used many pieces of the feature set created by Writeprints.

To mitigate the issue of Writeprints’s high computation cost we have combined a hybrid version of the Writeprints feature sets with a support vector machine. This results in a faster method that has a higher, but comparable, precision on our corpus. We validated the effectiveness of this approach by comparing it to the original Writeprints approach for select data sets from the original Brennan-Greenstadt corpus and found that the precision of our approach is comparable to the precision of the complete Writeprints method, as can be seen in Figure 3.4. This is also in line with results from the original Writeprints research, which compared the approach with a variety of others, including SVMs using the same feature set.
The feature set we use combines the brevity and static nature of the baseline set with some of the more complex features of the extended set. We call this the Writeprints Static Feature Set. It contains 557 static features, detailed in Table 3.6. We applied this feature set to an SVM classifier in the form of a sequential minimal optimization (SMO) with a polynomial kernel using Weka machine learning software\footnote{http://www.cs.waikato.ac.nz/ml/weka/}.

### 3.4.5 Evaluating Manual Circumvention Approaches

There are two ways to think of “success” when evaluating how stylometric methods respond to adversarial writing samples. One way is to measure the success of the method in identifying the true author of a document intended to circumvent stylometry and the other is to measure the success of the circumvention passage in preserving the anonymity of the author. We will examine the results from both angles.

To test the success we look at the performance on different sets of unique authors. Our data set consisted of a total of 45 unique authors. This is a larger number of unique authors than almost all of stylometry studies cited throughout this paper and is in line with the current state-of-the-art Writeprints, which looked at writing samples of 25, 50, and 100 unique authors.

In order to evaluate the corpus we set up test sets for 1,000 unique sets of 5, 10, 15, 20, 25, 30,
Table 3.6: Writeprints-Static Feature Set. adopted from the Writeprints approach.

<table>
<thead>
<tr>
<th>Group</th>
<th>Category</th>
<th>No. of Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>Word level</td>
<td>3</td>
<td>Total words, average word length, number of short words</td>
</tr>
<tr>
<td></td>
<td>Character level</td>
<td>3</td>
<td>Total char, percentage of digits, percentage of uppercase letters</td>
</tr>
<tr>
<td></td>
<td>Special Character</td>
<td>21</td>
<td>Occurrence of special characters</td>
</tr>
<tr>
<td></td>
<td>Letters</td>
<td>26</td>
<td>Letter frequency</td>
</tr>
<tr>
<td></td>
<td>Digit</td>
<td>10</td>
<td>Digit frequency (e.g., 1,2,3)</td>
</tr>
<tr>
<td></td>
<td>Character bigram</td>
<td>39</td>
<td>Percentage of common bigrams</td>
</tr>
<tr>
<td></td>
<td>Character trigram</td>
<td>20</td>
<td>Percentage of common trigrams</td>
</tr>
<tr>
<td>Vocabulary Richness</td>
<td></td>
<td>2</td>
<td>Ratio of hapax legomena and dis legomena</td>
</tr>
<tr>
<td>Syntactic</td>
<td>Function Words</td>
<td>403</td>
<td>Frequency of function words</td>
</tr>
<tr>
<td></td>
<td>POS tags</td>
<td>22</td>
<td>Frequency of Parts of speech tag</td>
</tr>
<tr>
<td></td>
<td>Punctuation</td>
<td>8</td>
<td>Frequency and percentage of colon, semicolon, qmark, period, exclamation, comma</td>
</tr>
</tbody>
</table>

35, and 40 authors out of a total sample pool of 45 authors from the Extended-Brennan-Greenstadt corpus. The precision measurements discussed for any author count throughout this section refer to the average across all 1,000 sets. All of the baseline results are based on ten-fold cross-validation. The precision for classifying obfuscation, imitation and translation passages is measured by training each classifier on the entire unmodified corpus for the authors in a test set and testing that classifier on the corresponding circumvention passages. The graphs in this paper refer to the precision because we believe that it is the most important and intuitive measurement when determining the authorship of an individual, unknown document. We have made additional graphs available reflecting recall and f-measure on our website.

The high number of combinations is uncommon in stylometry research but we believe it is important. Accuracy between different sets of authors can vary significantly depending on the specific authors chosen. By viewing the potential combinations of authors as the sample space and a specific combination of authors as a sample selection, we are able to make robust accuracy claims with minimal standard error. Standard error in this case is 3.1%.

12Precision, Recall and F-Measure graphs are available at https://psal.cs.drexel.edu/tissec
Baseline  Figure 3.5 demonstrates the effectiveness of the four methods we tested and the accuracy of random chance classification. The random chance line in all figures represents what the precision would be if the authorship of a document was determined by randomly selecting one of the potential authors. All of the results for the baseline precision measurements are statistically significant over random chance. All methods show a degradation of precision as the number of unique authors increases but the effectiveness is still quite substantial at even the largest author set. The Writeprints Static feature set utilizing an SVM demonstrates the highest precision overall. The synonym-based approach is also very effective. The Basic-9 feature set does poorly compared to the other two methods, but is still far above that of random chance. This is important and confirms our hypothesis
that even a very simple measurement of writing style is effective for small numbers of authors and still demonstrates significant ability for de-anonymization with larger numbers of authors.

Our SVM approach was evaluated using ten-fold cross-validation in the same manner as the rest of our experiments. This is more robust than the approach utilized by the Writeprints authors. In addition, the amount of training data available per author is much lower in our data set than most of the sets used in the Writeprints research. Our classification is based on approximately 6,500 words per author (we allow for error within 250 words of 6,500 in order to not break up sentences) as opposed to anywhere from 23,000 to 43,000 in most data sets used in the Writeprints study. These points are important because while our method does not achieve the same levels of precision at the Writeprints approach, it still approaches those numbers despite the relaxed constraints of our evaluation.

The Neural Network approach displays accuracies varying between 63% and 25%. This is not as effective as the other approaches but is high enough to be an effective form of authorship recognition and a legitimate threat to privacy and anonymity. The synonym-based approach, however, performs exceedingly well even at higher numbers of unique authors. The degradation in precision follows the same decline as the Writeprints Static feature set approach.

The synonym and Writeprints Static approaches follow a smooth degradation curve as the number of authors in each set increases. This confirms the hypothesis of our original research that the degradation curve we observed using the original Brennan-Greenstadt corpus would extend to larger numbers of authors.

**Obfuscation and Imitation Circumvention Approaches** Attempting to recognize the authors of the obfuscation passages results in a drop of accuracy to around that of chance classification as can be seen in Figure 3.6. Only the SVM (Writeprints-Static) approach displayed an effectiveness above that of random chance. This demonstrates the weakness of these methods in detecting writing style designed to be adversarial. The fact that the obfuscated passages track closely to random chance is exactly what we would expect for well-constructed obfuscation passages.

Imitation passages demonstrated a significant drop in effectiveness beyond the obfuscation ap-
Figure 3.7: Accuracy in imitation detection.

Figure 3.8: Success of imitation attacks.

approach for the neural network and SVM-based approaches, and maintained chance-level accuracy for Writeprints and the synonym-based methods as is seen in Figure 3.7. The lower-than-chance results for SVM and Neural Network approaches demonstrates a difference in how resistant different methods are to adversarial passages. This might indicate that the features modified in attempting to imitate Cormac McCarthy are well represented in the Basic-9 feature set but not so much by the extensive Writeprints Static feature set or the specific but unique synonym analysis.

Furthermore, the ability of the imitation passages to successfully imitate Cormac McCarthy was high. This is demonstrated in Figure 3.8. The synonym-based approach, despite very high precision in the baseline tests, was just as susceptible to misclassification of imitation passages toward Cormac.
McCarthy as the neural network approach that it greatly outperformed in the baseline evaluation.

It is important to note that the testing of imitation success required a modification to the evaluation process. In order to test for the success of an imitation passage in representing Cormac McCarthy's writing styles we needed to include training texts for Cormac McCarthy when building the classifier. This is only done for the measurement of imitation success—no Cormac McCarthy text is used in the training process for any other tests.

While stylometry methods that are resistant to adversarial passages have been hypothesized\textsuperscript{59,60}, we can see the first true instance of it here with the results from the SVM approach. The precision when classifying obfuscation passages never drops below that of random chance. Furthermore, it can be considered resistant to wrongly classifying imitation passages as their intended target as it generally is only half as likely as the other methods to classify an imitation passage as having been written by Cormac McCarthy. It is likely, however, that more sophisticated circumvention methods could pose a more significant threat to this approach.

### 3.4.6 Evaluating Circumvention By Machine Translation

Since publication of our initial 2009 research on this topic, circumvention via machine translation has been consistently questioned as a possible approach to maintaining anonymity. Our evaluation suggests that machine translation in its current form does not provide a reliable means of circumventing stylometry.

The translation evaluation was performed using passages from the Brennan-Greenstadt corpus. The single-step translations provided a drop in precision of anywhere from 10% to 30%. The two-step translation caused drops from 30% to 60%. Generally, Bing was better at identifying true authorship after Japanese translations and Google was better with German translations. Bing, however, performed very poorly with the two-step translations described below. We found that the synonym method was especially resistant to these passages and the neural network and SVM methods performed similarly throughout.
One-Step Translations One-step translations to and from German and Japanese were generally not effective at hiding the writing style of an author. While there was a reduction and it was sometimes statistically significant, such as Bing translations from English to German to Japanese to English on Neural network and SVM classifiers, the drop in accuracy is not large enough to grant much comfort to those looking to maintain their anonymity, as it generally only decreases the chance of identification from about 65% to about 50% in SVM and neural network approaches and only from 85% to 75% in the case of the synonym-based method. These numbers are not strong enough to warrant a claim that they are effective in providing an anonymizing effect on a document.

Figure 3.9: Google translation results.

Figure 3.10: Bing translation results.
Two-Step Translations  Two-step translations from English to German to Japanese and then back to English were generally no more effective at hiding the writing style of an author than a one-step translation with Japanese, except for the case of Bing Translator using SVMs and neural networks as explained below.

Translation Service Comparison  Single-step translation approaches to and from German were less effective using Google Translate and Japanese translations were less effective on Bing Translator. Bing seemed to produce very effective two-step translation passages for the NN and SVM methods. Overall, it appears as though Bing’s ability to construct adversarial passages well from an accuracy standpoint is greater than Google Translate’s. Bing’s average accuracy across correctly identifying adversarial translation passages is 12 points lower than Google and when the especially effective synonym-based method is removed, the difference increases to 17 points. This would indicate Bing is better for attempting a machine translation-based circumvention approach but overall the accuracies are still not low enough to suggest it would be sufficient to protect privacy and anonymity.

Classifier Method Comparison  The most effective of the three for all experiments as well as the baseline was the synonym-based method. This method was demonstrated to have high accuracy in past work but there is no previous work indicating that accuracy persists when looking at larger numbers of unique authors until now. The baseline of 84.5% for the synonym-based method was the highest of the three classifiers. The SVM and Neural Network using the Basic-9 feature set had baseline accuracies of 65.1% and 65.9% respectively.

Using Google Translate, the synonym-based method maintained an accuracy of 84.8% for Japanese and 80.5% for German one-step translations. The accuracy dropped only to 75% for the two-step translation. Similar results were found with Bing Translator. The SVM and neural network methods dropped to 46% and 44.4% for Japanese, respectively, and 50.1% and 58.7% for German. They also saw accuracies similar to Japanese one-step translations for the two-step translation experiment.

Overall these results are, for the most part, not statistically significant in favor of the translation having an anonymizing effect on the writing style of an author, and we believe the reduction in
accuracy is not enough to warrant calling this an effective approach to circumventing stylometry.

**Effectiveness of Translated Documents**  Even if we were to accept a drop in accuracy by 15 to 35 points as sufficient for aiding the anonymization of a document, would the resulting translated passage be acceptable for communication purposes or publication? We observed that the answer to that question depended heavily on the complexity of the language being translated. Here is an example sentence from Cormac McCarthy that appeared in his novel *The Road*, along with each translation:

(Original)

Just remember that the things you put into your head are there forever, he said.

(English ⇒ German ⇒ English)

Remember that the things that you are dead set on always there, he said.

(English ⇒ Japanese ⇒ English)

But things are there forever remember what you put in your head, he said.

(English ⇒ German ⇒ Japanese ⇒ English)

You are dead, that there always is set, please do not forget what he said.

The original sentence was reasonably complex and did not fare well through the translation process. While the translated sentences were coherent, the meaning was fundamentally changed in each one. But when we look at a simpler sentence from that same passage we find more consistent results.

(Original)

They passed through the city at noon of the day following.

(English ⇒ German ⇒ English)

They crossed the city at noon the following day.

(English ⇒ Japanese ⇒ English)

They passed the city at noon the following day.
They crossed the city at noon the next day.

The translations of the simpler sentence are more effective but lack obfuscation. The goal of the translation approach is to alter the writing style while retaining the meaning. There are many examples of this that can be found in the translated passages, such as “Fighting was tough, with each house and factory fiercely contested” being translated to “The fight was hard, fought hard with every home and factory”. But these are outweighed by the number of significantly altered meanings, incoherent translations, and very good but non-obfuscated translations.

3.4.7 Measuring Writing Style Modification

In informal discussions with participants after completing the study, we found that many of them tried to obfuscate their style by “dumbing down” their writing by using shorter sentences and less descriptive words. When imitating the writing style of Cormac McCarthy, the participants described attempting to use descriptive and grim language. We can verify at least some of these claims through analysis of the feature frequency changes in the obfuscation and imitation passages.
We compare the normalized frequencies of the features to understand which ones people change to hide their writing style. The changes made to a selection of features for both obfuscation and imitation passages can be seen in Figures 3.11. This graph illustrate the changes in frequencies for each feature. The $y$-axis contains a list of features that have been adjusted in the passages and $x$-axis of the graph denotes the change in each feature. We compute the change in feature using the following formula:

\[
\text{Change in Feature } f, C_f = 100 \times \frac{(f_{\text{adv}} - f_{\text{reg}})}{(f_{\text{reg}} + 1)}
\]

where,

- $f_{\text{adv}}$ = Average values of feature $F$ in adversarial documents.
- $f_{\text{reg}}$ = Average values of feature $F$ in regular documents.

We add one with $f_{\text{reg}}$ in the denominator to avoid divide-by-zero error, as $f_{\text{reg}}$ can be zero for some features. The amount to the right of the $y$-axis represents the increases in a feature and the amount to the left represents the decreases.

In our experiments, the most changed features are average syllables, average word length, sentence count, average sentence length, usage of personal pronouns, adjectives, verbs, and readability index. We do see hints that authors are “dumbing down” their writing style with shorter sentences, lower readability scores, and less complex words. Most of these are the features in our Basic-9 Feature set which may explain why that feature set can be effective despite its relatively small size.

In imitation passages, all the participants use more personal pronouns and verbs and shorter sentences than the regular cases. The personal pronouns can likely be attributed to the direction of the writing survey which asked participants to describe their day in the third person. The shorter sentences, however, are likely a result of imitating Cormac McCarthy, who often uses short sentences in his prose. In both imitation and obfuscation passages, participants use shorter and simpler words (those with only one or two syllables) and shorter sentences. As a result, adversarial writings are
3.4.8 Detecting Deception in Writing Style

This work was completed by Sadia Afroz with support from Michael Brennan.

When an American male blogger, Thomas MacMaster, posed as a Syrian homosexual woman Amina Arraf in the blog “A Gay Girl in Damascus” and wrote about Syrian political and social issues, several news media including The Guardian and CNN thought the blog was “brutally honest” and published email interviews of Amina\(^\text{13}\). Even though no one had ever spoken to or met her and no Syrian activist could identify her, Amina quickly became very popular as a blogger. When Amina’s cousin announced that she had been abducted by the Syrian police, thousands of people supported her on social media and made the US state department investigate her fictional abduction\(^\text{14}\). This scrutiny led to the discovery of the hoax.

Can this deception be detected in writing style with stylometry? This section presents a brief analysis of this question. A full examination of this issue is outside of the scope of this document, and a more in-depth explanation of these results can be found in “Detecting Hoaxes, Frauds, and Deception in Writing Style Online”\(^\text{67}\). We show that linguistic cues can detect stylistic deception in the Extended-Brennan-Greenstadt adversarial corpus. Also, we can detect indication of masking in the documents collected from two additional data sets, the Ernest Hemingway and William Faulkner imitation contests. We also show how long-term deceptions such as the blog posts from “A Gay Girl in Damascus” are different from these short-term deceptions. We found these long-term deceptions to be more robust to our classifier but more vulnerable to traditional stylometry techniques.

**Approach and Results** The three approaches we used are the Writeprints approach, as described in Section 3.2, the Basic-9 feature set as described in Section 3.4.4, and a new lying-detection feature set. This feature set includes features that were known to be effective in detecting lying type deception in computer-mediated communications and typed documents\(^\text{79,80}\). These features are:

1. Quantity (number of syllables, number of words, number of sentences)


2. Vocabulary Complexity (number of big words, number of syllables per word)

3. Grammatical Complexity (number of short sentences, number of long sentences, Flesh-Kincaid grade level, average number of words per sentence, sentence complexity, number of conjunctions)

4. Uncertainty (Number of words that express certainty, number of tentative words, modal verbs)

5. Specificity and Expressiveness (rate of adjectives and adverbs, number of affective terms)

6. Verbal Non-Immediacy (self-references, number of first-, second- and third-person pronoun usage)

We use the list of certainty, tentative, and affective terms from LIWC 2007\textsuperscript{31}.

To evaluate our approach, we perform a three class-classification, where the three classes are Regular, Imitation, and Obfuscation, with three feature sets. We use 10 fold cross-validation with Support Vector Machine (SVM) and J48 Decision Tree classifier. In $k$-fold cross-validation the original data set is randomly partitioned into $k$ equal folds or subsets. Then the classifier is trained on $k - 1$ subsets and tested on the remaining one subset. The cross-validation process is then repeated $k$ times (the folds), with each of the $k$ folds used exactly once as the validation data. The final estimation of the classifier is the average of the $k$ results from the folds. The results are best understood by examining the driving questions behind the experiments.

**Can we detect stylistic deception in documents?** The results show that a classifier trained on sets of adversarial and non-adversarial documents can detect deceptive documents with 96.6% accuracy on our best feature set, as is shown in Table 3.7.

**Which linguistic features indicate stylistic deception?** We examined the Information Gain Ratio (IGR)\textsuperscript{81}. The top features are mostly function words, as shown in Table 3.8. Other than function words, some syntactic features such as personal pronoun, adverbs, adjectives, and average word length were some of the most discriminating features.

**Are some adversarial attacks more difficult to detect than others?** In our experiment, obfuscation was harder to detect than imitation. In imitation, participants followed one specific
### Table 3.7: The table shows performance of different feature sets in detecting regular and adversarial writing samples. The Writeprints feature set with SVM classifier provides the best performance in detecting deception.

<table>
<thead>
<tr>
<th>data set</th>
<th>Feature set, Classifier</th>
<th>Accuracy by Type</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extended-Brennan-Greenstadt</td>
<td>Writeprints, SVM</td>
<td>Regular: 98%</td>
<td>96.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 89.5%</td>
<td></td>
</tr>
<tr>
<td>Extended-Brennan-Greenstadt</td>
<td>Lying-detection, J48</td>
<td>Regular: 95.7%</td>
<td>92%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 75.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 59.9%</td>
<td></td>
</tr>
<tr>
<td>Extended-Brennan-Greenstadt</td>
<td>9-feature set, J48</td>
<td>Regular: 94.5%</td>
<td>89%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 48%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 43%</td>
<td></td>
</tr>
<tr>
<td>Amazon Mechanical Turk</td>
<td>Writeprints, SVM</td>
<td>Regular: 97.5%</td>
<td>95.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 77.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 86.9%</td>
<td></td>
</tr>
<tr>
<td>Amazon Mechanical Turk</td>
<td>Lying-detection, J48</td>
<td>Regular: 95.2%</td>
<td>90.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 57.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 61.8%</td>
<td></td>
</tr>
<tr>
<td>Amazon Mechanical Turk</td>
<td>9-feature set, J48</td>
<td>Regular: 94.3%</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 40%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 36%</td>
<td></td>
</tr>
<tr>
<td>Brennan-Greenstadt</td>
<td>Writeprints, SVM</td>
<td>Regular: 96.9%</td>
<td>94.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 90.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 66.7%</td>
<td></td>
</tr>
<tr>
<td>Brennan-Greenstadt</td>
<td>Lying-detection, J48</td>
<td>Regular: 91.4%</td>
<td>85.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 87%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 9.5%</td>
<td></td>
</tr>
<tr>
<td>Brennan-Greenstadt</td>
<td>9-feature set, J48</td>
<td>Regular: 91.5%</td>
<td>84%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imitation: 25%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obfuscation: 55.6%</td>
<td></td>
</tr>
</tbody>
</table>

writing style, the writing style of Cormac McCarthy. Different people followed different linguistic aspects in imitating him; for example, some participants used short sentences, some used descriptive adjectives, and some used a conversational format with dialogs. But the overall writing style was limited to the style of Cormac McCarthy. Obfuscation is different than imitation as in obfuscation an author can choose to imitate more than one author’s writing style or develop a new style different from his own. However, when we include multiple imitated authors it becomes correspondingly more difficult to detect imitation attacks.

**Can we generalize deception detection?** We check whether our deception detection approach that can detect imitation and obfuscation on the Extended-Brennan-Greenstadt can detect imitation samples from the Ernest Hemingway and William Faulkner imitation contests. We performed a 10 fold cross-validation on the Hemingway-Faulkner imitation corpus. We used Writeprints
Table 3.8: This table shows the features that discriminate deceptive documents from regular documents. The top discriminating features according to Information Gain Ratio are mostly function words.

<table>
<thead>
<tr>
<th>Top 20 features</th>
<th>Imitated documents</th>
<th>Obfuscated documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>whats</td>
<td>alot</td>
<td></td>
</tr>
<tr>
<td>atop</td>
<td>near</td>
<td></td>
</tr>
<tr>
<td>lately</td>
<td>up</td>
<td></td>
</tr>
<tr>
<td>wanna</td>
<td>theres</td>
<td></td>
</tr>
<tr>
<td>underneath</td>
<td>thousand</td>
<td></td>
</tr>
<tr>
<td>anymore</td>
<td>ours</td>
<td></td>
</tr>
<tr>
<td>beside</td>
<td>shall</td>
<td></td>
</tr>
<tr>
<td>she</td>
<td>thats</td>
<td></td>
</tr>
<tr>
<td>herself</td>
<td>cuz</td>
<td></td>
</tr>
<tr>
<td>beneath</td>
<td>whats</td>
<td></td>
</tr>
<tr>
<td>like</td>
<td>havent</td>
<td></td>
</tr>
<tr>
<td>he</td>
<td>Frequency of comma</td>
<td></td>
</tr>
<tr>
<td>till</td>
<td>lots</td>
<td></td>
</tr>
<tr>
<td>her</td>
<td>tons</td>
<td></td>
</tr>
<tr>
<td>onto</td>
<td>anyway</td>
<td></td>
</tr>
<tr>
<td>soon</td>
<td>plus</td>
<td></td>
</tr>
<tr>
<td>Frequency of dot</td>
<td>other</td>
<td></td>
</tr>
<tr>
<td>Personal pronoun</td>
<td>maybe</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.9: Imitated document prediction result: A classifier trained on imitation and obfuscation data set can detect imitated articles from the International Hemingway Imitation contest and the Faux Faulkner contest.

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Lying-detection features, J48</th>
<th>Writeprints features, SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imitation</td>
<td>73.9%</td>
<td>73.2%</td>
</tr>
<tr>
<td>Regular</td>
<td>62.4%</td>
<td>72.2%</td>
</tr>
</tbody>
</table>

and lying-detection features with SVM and J48 classifiers respectively from the WEKA tool. Our classifier can distinguish imitated articles from the original writings of Ernest Hemingway and William Faulkner with 88.6% accuracy (Table 3.9). We also performed an experiment where a classifier trained on the Extended-Brennan-Greenstadt data set was tested on Hemingway-Faulkner Imitation corpus. Only 57.1% of the imitated documents were considered as imitation in that case.

How does all of this apply to long-term deception cases, such as “A Gay Girl in Damascus”? We performed an authorship attribution test on the posts Thomas MacMaster created as himself and as Amina that existed in a history-focused Yahoo! group. We found that he consistently maintained a separate writing style as Amina in the Yahoo! group as none of Amina’s posts were attributed to Thomas in the authorship attribution test. Thomas’s writing style as
Amina was different than his regular writing style. The use of upper-case letters and modal verbs was noticeable in Amina's posts, whereas Thomas used longer sentences and more adjectives and adverbs.

Moreover, all of the posts Thomas wrote as Amina and as himself and posts by his wife Britta, an early suspect for the true identity of Amina, were considered as regular when tested on an SVM classifier which was trained with the Extended-Brennan-Greenstadt corpus. Deception classification of the posts from A Gay Girl in Damascus also did not show any indication of masking. In our test, only 14% of the blog posts were considered deceptive, which is less than the error rate, suggesting a random effect. Thirteen blog posts were classified as obfuscated documents, 22 were classified as imitated documents. During authorship attribution, 57.14% of the deceptive documents were attributed to Amina.

But maintaining an alternate writing style consistently for a long time is hard, which was evident in the Thomas-Amina case. When people started questioning Amina’s existence, Thomas and his wife Britta were suspected as possible writers of the blog based on various pieces of evidence. For example, Thomas’s address was used in Amina’s account, and photos from Britta’s Picasa album were used in Amina’s blog. In the end, Thomas admitted that he was “Amina.” Regular authorship recognition also supports this fact. More than half of the blog posts (54.03%) were attributed to Thomas during authorship attribution with an SVM classifier and the Writeprints feature set. Only 10 posts were attributed to Britta and the rest were attributed to “Amina.” Native language detection, age, and other demographics analysis are other possible ways to detect this form of deception; they are not explored in this paper.

3.4.9 Discussion and Future Work

Authors, Topics, and Skill The amount of training text for each author is not exceptionally large, but the total number of authors in the study is significantly larger than most studies of stylometry methods and on par with new methods such as Writeprints. Having such a large number of authors to work with allowed us to do more extensive testing by choosing random groups of authors and averaging the accuracies across them. This also allowed us to see interesting patterns
in the study, such as which authors did a better job than others in creating successful adversarial passages. Through our anecdotal observations it was clear that certain authors did a poor job in writing their adversarial passages.

Additionally, certain authors also had a style that seemed to share a disproportionately large amount of features with many obfuscation attempts. Authorship of obfuscation passages were often attributed to these authors when they were a member of a test set. An interesting avenue of research would be to determine if it is possible to create a generic writing style by automated means.

The domain of possible content in this study was fairly open. Participants were allowed to present samples from a variety of subjects so long as they were scholarly in nature. It could be beneficial to study the effects of adversarial attacks in stricter domains, where there is less room for maneuvering and thus fewer options for how an author could hide his or her identity. Stylometry used in restricted domains may prove less susceptible to our attacks.

In general, the participants in this study were unskilled in the field of stylometry and were not professional writers. Despite this lack of expertise they were able to consistently circumvent the authorship attribution methods that were put in place. This strengthens our findings, as it would be reasonable to expect that authors with expertise in such areas could do a better job at attacking the system.

**Open Problems in Adversarial Stylometry**  Given the evidence in this research that hiding one’s writing style is an effective means to circumventing authorship recognition, one of the next logical steps is to develop end-user software that can assist users in modifying their writing. This is being addressed with the release of Anonymouth\(^{15}\). Anonymouth augments the writing style modification process with intelligent suggestions driven by implementations and analysis of stylometry techniques outlined in this research and elsewhere. Open research avenues include identifying the most effective structured approach for writing style modification, resolving the trade-off between comprehensive modification and over-fitting the changes for specific recognition methods, and identifying which features may be heavily automated and which must rely greatly on manual input.

\(^{15}\)https://psal.cs.drexel.edu/anonymouth

---

Chapter 3: Managing Identity and Anonymity with Stylometry
Another important part of continued research in this area is larger and more defined corpora in different languages. Our general corpus satisfies a number of reasonable demands for consistency, length, and focus, but there are many other, more specific domains that could produce different results when examining adversarial passages. For example, will writing style modifications be more effective in a highly restricted domain such as complex scientific research papers or in a very broad domain such as fictional short stories? Are the most salient features for identification in other languages similar to those in English?

3.5 Stylometry in Social Media

This work was completed by Diamond Bishop with support from Michael Brennan and Bob Grimshaw.

The majority of authorship recognition research has been done on documents of formidable size. With the advent of social networking and the widespread ownership of mobile devices, however, people communicate and express their thoughts in increasing proportion through short messages that are much smaller than the texts studied by past researchers in the field.

One popular form of communication that exemplifies the rise of the short message format is Twitter. Tweets are a text-based message format limited to 140 characters posted to the social networking and micro-blogging website Twitter. Twitter allows users to keep others updated on real time posts regarding anything from news to what their cat is doing\textsuperscript{82}. The rise in tweeting as a means of communication has resulted in recent messaging rates of 340 million/day\textsuperscript{83}. There are similarities between Twitter and Short Message Service (SMS), a textual messaging medium with even more pervasive adoption. Annualized totals for 2010 are expected by some to total 1.8 trillion messages in the U.S. alone\textsuperscript{84}.

SMS has become an integral part of electronic communication in the developed world. The standard limit on the size of an SMS message is 160 characters, just enough to get a quick and informative message to someone. With 160 characters at your disposal, texting can be a viable form of communication for conveying almost any type of information. Through the use of almost any mobile phone, someone can be reminded to pick up groceries, informed of a disaster, or learn of an engagement.
These platforms also contain many challenges. Criminals can use the anonymity of a prepaid mobile phone for the planning of drug sales, murders, and any number of surreptitious activities. Similarly, through the use of 140 characters in a tweet, a potential criminal can send a message to others to gather for something harmful such as a violent flash mob. Some promise has been shown in narrowing the scope of criminal investigations involving SMS messages by using authorship recognition techniques and harvested SMS messages, but research is limited by the lack of availability of public corpora.

Anonymity is also an integral component to public discourse. Means for maintaining one’s anonymity on the web are widespread but have focused primarily on pseudonyms, location-based privacy, and encryption. Authorship recognition is an effective means for discovering the author of an anonymous document. This has significant implications for individuals who rely on anonymity as a component of their public speech. Authorship recognition has been actively researched for decades but developments in machine learning combined with the massive text data publicly available on the web have led to the ability to identify individuals amongst hundreds or even thousands of potential authors. It is important to understand the degree of identifiability that comes with a tweet or a Twitter feed.

To meet this challenge, we test the accuracy while varying different techniques for authorship recognition of Twitter messages. The large size and public nature of this data makes it feasible to develop and test a tweet authorship recognition system. The same public nature that puts Twitter within reach of a researcher may give an investigator the chance to train a classifier against a corpus reaped from public information in his or her pursuit of a suspect, making this a potentially profitable area of investigation for law enforcement. Furthermore, the small amount of research of authorship recognition systems in Twitter messaging makes our work novel in its area of use and interesting to explore.
3.5.1 Related Work: Short-Form Message Analysis and Authorship Recognition

Some work has been done in the analysis of short text messages that has shown there are differences in the classification and processing of short messages. Healy, Delany, and Zamolotskikh have done work in case-based reasoning-focused classification on multiple short message types. They have focused on using spam filtering applied to SMS messages and customer comments from guests of a large hotel chain. Their results show that the short text message classification models require different feature representation and types. Our work similarly exploits feature frequency models for classification of short message types, as opposed to binary feature existence models. Authorship recognition of SMS messages has utilized character based N-grams with an SMS corpus to recognize message authors. This approach identified 65-72% of their test authors when using a trigram character-based approach.

Natural language processing has been applied to Twitter as well; however this work has not focused on authorship recognition. One example of Twitter classification is the electronic new processing system TwitterStand. TwitterStand uses a Nave Bayes classifier that is trained on a corpus of tweets that have already been marked as news or junk to classify tweets as news or junk. The interesting aspect of TwitterStand’s approach to text classification is that they use a static corpus and a dynamic corpus for training their model. This dynamic corpus is gathered from trusted new tweeters, or what they refer to as seeders. This use of seeders allows for the inherently dynamic nature of news-following. Similarly, Sririam, Fuhry, Demir, Demirbas, and Ferhatosmanoglu have used Twitter-specific feature selection and known information about authors to classify new tweets into a number of different areas. There has also been work on Twitter-related authorship recognition in the past, but they utilize methods not tied exclusively to linguistic styles such as the source-code author-profile method, which relies on n-grams that could be affected by content and do not link short-form text with long-form blog data.

There has been extensive research on stylometry, including a broad overview of the field by Dr. Patrick Juola. The majority of stylometry research has focused on larger sets of text, with new
approaches being able to distinguish among up to 100 authors through the use of various supervised and unsupervised learning techniques\(^1\). There has also been work examining author identification across domains\(^67\), but not across different communication mediums as this research does.

### 3.5.2 Methodology and Data

**Twitter Data**  We began our research using a corpus of tweets collected by Ritter et al at the Microsoft Research Group\(^92\). This data was collected using the Twitter API to capture all public tweets sent between July 1 and August 27, 2009. The resulting, filtered data set contained approximately 9.4 million tweets, posted by approximately 296,000 users. The data was anonymized by replacing all user IDs by an increasing number sequence, and redacting URLs and long number sequences (possibly phone numbers, SSNs, or credit card info). All other content was left unmodified, according to the authors. A similar corpus containing an even larger number of Twitter messages (97 million), composed by Petrovic et al. at the University of Edinburgh in late 2009, has been pulled from their site per a request from Twitter\(^93\).

The Microsoft corpus contains only English-language tweets that received at least one reply (along with the replies). Kelly et al.\(^94\) claims conversational tweets make up only 37% of all Twitter traffic, so our results will necessarily not represent all Twitter authors.

**Blogger Data**  To run the tests identifying individuals across different domains we looked for authors with blogs as well as active Twitter accounts. While there are many such authors, we chose to use a list of the 100 top entrepreneur-focused bloggers with Twitter accounts\(^16\). These authors were chosen to introduce some control for the differences in subject matter that can increase authorship recognition results.

Twenty authors were randomly chosen out of the list of 100 to run experiments on. When selected, their Twitter accounts and blogs were checked to see if they were active (> 10 blog entries, > 100 tweets). If they were active, they were added to our list of chosen authors until we reached a limit of 20. Tweets were scraped from these 20 authors and the most recent 10 blog entries were taken.

\(^{16}\)http://www.solobizcoach.com/small-business-twitter/

Chapter 3: Managing Identity and Anonymity with Stylometry
**Approach**  
Our feature set consists of 46 lexical, syntactic, and structural features. These were selected to gain a quick look at a wide range of possibly useful features. Context-specific features like n-grams were avoided to help control for classification due to subject rather than linguistic style. The set of 46 features and their descriptions can be seen in Tables 1 and 2. All attribute values were discretized using the WEKA toolkit’s supervised discretization, which makes use of the class information to turn numeric attributes into nominal ones through Fayyad and Irini’s MDL method.

Experiments were run using the WEKA toolkit. A Support Vector Machine (SVM) with a linear kernel was used for classification. Results for the Twitter-specific features were confirmed using 10 experimental iterations of 10 fold cross-validation. The cross-domain experiments constituted 10 iterations of training and testing on two separate data sets. We used information gain calculations as the basis for ranking features throughout our experiments.

### 3.5.3 Evaluation

**Twitter Evaluation**  
For each experiment, authors were chosen at random from the Microsoft data corpus. When an author was chosen, 300 tweets were randomly selected from the author’s tweet set and agglomerated into sets of $m$ tweets, where $m$ is the current number of tweets that are placed together and treated as one document.

One of the problems with authorship attribution in short messages is the lack of features in a short message. Since each tweet is so small, there is a distinct lack of information to gather. We agglomerate different numbers of tweets into larger bundles of messages to understand if this is a structural problem, or purely a data size limitation. During these tests we performed our classification tasks using the same set of 300 tweets for each of 10 users, bundling them in sets of 1, 5, 10, 20, 25, and 30.

Using the best size of tweet agglomeration from the previous set of experiments, these experiments take a look at the affects of the number of authors on Twitter authorship recognition. By varying the number of authors we can see if the authors of tweets are distinguishable as the number of authors increases. This set of experiments checks 2, 10, 20, 40, and 50 sets of authors.
Table 3.10: Feature Set Part I

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>characters_nospace</td>
<td>Number of characters, not counting spaces</td>
</tr>
<tr>
<td>characters_space</td>
<td>Number of characters, counting spaces</td>
</tr>
<tr>
<td>uniqueWords</td>
<td>Number of unique words</td>
</tr>
<tr>
<td>sentencecount</td>
<td>Number of sentences</td>
</tr>
<tr>
<td>avg_sentence_length</td>
<td>Average sentence length</td>
</tr>
<tr>
<td>avg_syllables</td>
<td>Average number of syllables per sentence</td>
</tr>
<tr>
<td>complexity</td>
<td>Unique words/total words</td>
</tr>
<tr>
<td>largeWords</td>
<td>Number of words larger then 6 characters</td>
</tr>
<tr>
<td>percentLargeWords</td>
<td>Percent of large words</td>
</tr>
<tr>
<td>insightVal</td>
<td>Number of words that are considered insight related</td>
</tr>
<tr>
<td>spPronoun</td>
<td>Number of singular pronouns</td>
</tr>
<tr>
<td>fpPronoun</td>
<td>Number of first person pronouns</td>
</tr>
<tr>
<td>objPronoun</td>
<td>Number of objective pronouns</td>
</tr>
<tr>
<td>tpSingPronoun</td>
<td>Number of 3rd person singular pronouns</td>
</tr>
<tr>
<td>tpPluralPronoun</td>
<td>Number of third person plural pronouns</td>
</tr>
<tr>
<td>certaintyVal</td>
<td>Number of words that are considered certainty related</td>
</tr>
<tr>
<td>tentativeVal</td>
<td>Number of words that are considered tentative related</td>
</tr>
<tr>
<td>timeframeVal</td>
<td>Number of words that are considered time frame related</td>
</tr>
<tr>
<td>percentInsightVal</td>
<td>Percent of words that are considered insight related</td>
</tr>
<tr>
<td>percentSpPronoun</td>
<td>Percent of singular pronouns</td>
</tr>
<tr>
<td>percentFpPronoun</td>
<td>Percent of first person pronouns</td>
</tr>
<tr>
<td>percentObjPronoun</td>
<td>Percent of objective pronouns</td>
</tr>
<tr>
<td>percentTpSingPronoun</td>
<td>Percent of 3rd person singular pronouns</td>
</tr>
<tr>
<td>percentTpPluralPronoun</td>
<td>Percentage of words that are third person plural pronouns</td>
</tr>
<tr>
<td>percentCertaintyVal</td>
<td>Percent of words that are considered certainty related</td>
</tr>
<tr>
<td>percentTentativeVal</td>
<td>Percentage of words that are considered tentative related</td>
</tr>
<tr>
<td>percentTimeFrameVal</td>
<td>Percent of words that are considered time frame related</td>
</tr>
</tbody>
</table>
### Table 3.11: Feature Set Part II

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>allPunc</td>
<td>Number of punctuation markers</td>
</tr>
<tr>
<td>semicolon</td>
<td>Number of characters that are semicolons</td>
</tr>
<tr>
<td>qmark</td>
<td>Number of characters that are question marks</td>
</tr>
<tr>
<td>comma</td>
<td>Number of commas</td>
</tr>
<tr>
<td>period</td>
<td>Number of characters that are periods</td>
</tr>
<tr>
<td>colon</td>
<td>Number of characters that are colons</td>
</tr>
<tr>
<td>paren</td>
<td>Number of parenthesis</td>
</tr>
<tr>
<td>exclaim</td>
<td>Number of exclamation points</td>
</tr>
<tr>
<td>percentAllPunc</td>
<td>Percentage of characters that are any punctuation</td>
</tr>
<tr>
<td>percentSemicolon</td>
<td>Percent of characters that are semicolons</td>
</tr>
<tr>
<td>percentQmark</td>
<td>Percent of characters that are question marks</td>
</tr>
<tr>
<td>percentComma</td>
<td>Percent of characters that are commas</td>
</tr>
<tr>
<td>percentPeriod</td>
<td>Percent of characters that are periods</td>
</tr>
<tr>
<td>percentColon</td>
<td>Percentage of characters that are colons</td>
</tr>
<tr>
<td>percentParen</td>
<td>Percentage of characters that are parenthesis</td>
</tr>
<tr>
<td>percentExclaim</td>
<td>Percentage of characters that are exclamation points</td>
</tr>
<tr>
<td>readability_gf</td>
<td>Gunning Fog index</td>
</tr>
<tr>
<td>readability_experimental</td>
<td>Flesh Kincaid ease of reading formula</td>
</tr>
<tr>
<td>percentSpPronoun</td>
<td>Percentage of words that are singular pronouns</td>
</tr>
</tbody>
</table>

Evaluating the accuracy over different sizes of tweet bundles was done with 10 users using sets of 1, 5, 10, 20, and 30 tweets for the bundles. The results can be seen in Figure 3.12. We can see that as tweets are bundled in larger sets, the accuracy increases and error decreases. To understand exactly what affects the accuracy of our classifier, as the bundle sizes changed we calculated the information gain for our features and took a look at the top 10. By looking at these features we saw that the features that were percentages instead of absolute counts fared better in smaller bundles, while the count features and features relying on more data, such as sentence length, increased with bundle size.

Since the bundling of 30 tweets was the most accurate, we moved forward with the next experiment using 10 sets of 30 tweet bundles for each author. We then varied the number of authors to give us the results that can be seen in Figure 3.13.

These results did not produce a clear trend, but they were able to confirm that we can continue to be successful as the number of Twitter authors increases. We do see a general drop-off in accuracy as the number of potential users increases. This is in line with stylometry research in general.
Figure 3.12: Accuracy when varying the bundle size.

Figure 3.13: Accuracy when varying the number of users.
Cross-Domain Analysis  We gained an insight into the features of blogs vs. tweets by training on tweets and testing on blogs, and testing on tweets and training on blogs. We trained on blogs and tested on tweets, then did the opposite. We repeated this experiment three times: one with all of the features, one with the 10 most useful features from the blog data and one with with the 10 most useful features from the Twitter data. These features can be seen in Figure 3.14.

During the cross domain analysis experiments, we used sets of 10 users to verify results. Our first experiment was to use all the features from our Twitter experiments to train on the blog data and test on the Twitter data. This resulted in an accuracy of 21% ($\sigma \approx 3\%$). Using these features once more, we trained on tweets and tested on blogs, receiving a low average accuracy of 11% ($\sigma \approx 3\%$).

Testing the effects of different features, we used the 10 best tweet features to train on the blogs and test on the tweets. This produced better accuracy of 38% ($\sigma \approx 4\%$). Alternatively, training on tweets and testing on blogs with these features produced an accuracy of 16% ($\sigma \approx 3\%$).

Using the 10 best blog features we trained on tweets and tested on blogs, once again producing better results than the initial use of all features, an accuracy of 15% ($\sigma \approx 2\%$). Using these features to train on blogs and test on tweets produced a comparable 16% ($\sigma \approx 3\%$) accuracy.

<table>
<thead>
<tr>
<th>Blog Features</th>
<th>Tweet Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>characters_space</td>
<td>characters_nospace</td>
</tr>
<tr>
<td>uniqueWords</td>
<td>avg_syllables</td>
</tr>
<tr>
<td>percentAllPunc</td>
<td>sentencecount</td>
</tr>
<tr>
<td>PercentPeriod</td>
<td>readability_experimental</td>
</tr>
<tr>
<td>readability_experimental</td>
<td>allPunc</td>
</tr>
<tr>
<td>avg_sentence_length</td>
<td>largeWords</td>
</tr>
<tr>
<td>avg_syllables</td>
<td>percentLargeWords</td>
</tr>
<tr>
<td>largeWords</td>
<td>percentPeriod</td>
</tr>
<tr>
<td>complexity</td>
<td>percentExclaim</td>
</tr>
<tr>
<td>percentTimeFrameVal</td>
<td>percentAllPunc</td>
</tr>
</tbody>
</table>

**Figure 3.14:** Blog Features vs. Tweet Features.
3.6 Conclusions: How Does Stylometry Affect Online Identity and Anonymity Management?

This work demonstrates the effectiveness of adversarial writing against modern methods of stylometry. It also demonstrates the ability of stylometry to effectively identify individuals based on writing in the first place. From the perspective of stylometry, analysis of current techniques and their weaknesses to adversarial writing demonstrates that we must test stylometry methods for their resistance to adversaries in situations where their presence is likely. From the perspective of anonymity, we advocate a stronger stance of not relying on stylometry in sensitive situations where the authorship of an unknown document must be known with a high degree of certainty unless the possibility of a modified writing style is negligible. But we also see that it is possible to maintain one’s anonymity in the methods utilized above.

The obfuscation approach weakens all methods to the point that they are no better than randomly guessing the correct author of a document. The imitation approach was widely successful in causing authorship to be attributed to the intended imitation target. Additionally, these passages were generated by participants in very short periods of time by amateur writers who lacked expertise in stylometry. Translation with widely available machine translation services does not appear to be a viable mode of circumvention. Our evaluation did not demonstrate sufficient anonymization and the translated document has, at best, questionable grammar and quality.

Despite the structural changes that come with short messages, we successfully identified the authors of sets of tweets. Given a large number of tweets it seems to be best practice to use bundled tweets to increase classifier accuracy. These results also show promise for further work in cross-document analysis using readily available data of one type (blogs) to classify unknown works of a different document type (tweets). Future work would study the best way to choose features that can cross-document domains and increase classifier accuracy. Overall, short message classification seems to be an area of study that can be expanded successfully for real-world usage.

There has long been a case to be made for a multi-disciplinary approach to privacy and anonymity. This research shows both the necessity of considering writing style analysis as a component of that
approach and demonstrates the possibility for privacy conscious-individuals to take steps to maintain their anonymity in the face of advanced stylometric techniques.

This work provides further evidence that learning techniques used in adversarial settings need to be tested with adversarial test sets. This research also has implications for machine translation research through the use of stylometry as a method for testing the effectiveness of machine translation. If a machine-translated data set shows comparable accuracy in an adversarial stylometry setting, then the results may be used to validate the translation method.

This work also strengthens the original claims of high accuracies by validating the methods on a large set of new data produced for a variety of purposes. When these methods are used in situations where adversaries are not considered to be a threat, they perform quite well.
Chapter 4: Building Data Sets and Obtaining Baselines for Adversarial Research

The work in this section demonstrates the strengths and weaknesses of crowdsourcing platforms in obtaining data for adversarial analysis. It suggests that obtaining new adversarial data can be done effectively by using a crowdsourcing platform, and that establishing baseline quality measurements on pre-existing data with crowdsourcing platforms has mixed results. The crowdsourcing system we use is Amazon Mechanical Turk (AMT), a platform that provides access to individuals willing and able to complete human intelligence tasks.

A major challenge in analyzing the effect of adversaries in a system is acquiring a quality data set with labeled results and baseline measurements. This challenge occurs when analyzing and managing identity with stylometry because adversaries attempt to obfuscate their true identity. Despite the wealth of text data on the web, little of it has the author identity attached to it. We refer to the true identity of a document’s author as the ground truth. This requires new data sets to be constructed, and this section demonstrates how they may be created using AMT.

This challenge also occurs when analyzing quality in online discourse. Developing new approaches to handling adversaries in commenting systems requires a ground truth of the identity of adversarial content. Many domains that may benefit from such research do not have a ground truth with which to build such approaches. The content exists but baseline measurements of adversarial content must be created. This section identifies how adept AMT workers are adept at creating such baseline measurements.

This work is important because establishing baseline data sets can be an exhaustive process that drains time and money. Taking into consideration the weaknesses of crowdsourcing systems and building a process that leverages them effectively in order to generate data sets for evaluating adversarial content has the potential to increase the amount of quality research emerging from adversarial research.
4.1 Challenges in Obtaining Adversarial Text Data

This work addresses two primary challenges in obtaining adversarial text data for analysis.

1. **Inaccurate or Incomplete Baseline.** In order to evaluate the effectiveness of a system, a complete baseline measurement must exist from which to base comparisons. When identifying the accuracy of machine learning in identifying adversaries in a community such as Slashdot, we must know the true number of adversaries that are present. If we accept the crowdsourced ratings as a baseline, we have the issue of comments that go unmoderated due to factors such as time or length. This creates an inaccurate or incomplete baseline.

2. **Lack of Known Data.** In some contexts, such as adversarial stylometry, the intent of the adversary is to mask the presence that an adversary exists at all. This can be seen in Section 3.3.4, where the author of “A Gay Girl in Damascus” attempted to mask the fact that Amina was a false identity.

Humans can resolve both challenges. In the case of inaccurate or incomplete baselines, in order to examine the ability of machine learning to effectively address the problem that is normally solved by humans, we need a human measurement to compare against. In the case of lack of known data, we are comparing against human adversaries and thus need data sets manually generated by humans. We acquire the human participants for these tasks through the AMT platform.

4.1.1 Related Work: The Mechanical Turk Approach

As discussed in Section 3.4.2, AMT provides access to a large and diverse population of participants that can perform human intelligence tasks. These participants, referred to in this section as “AMT workers”, choose tasks on their own and complete them in exchange for financial compensation. Each task is called a “Human Intelligence Task”, commonly known as a HIT.

AMT has been used extensively for data set generation and analysis. This is particularly true for creating and analyzing language data\textsuperscript{96,97}. There are limitations in the granularity of participant selection. The basic means of exclusion are based on geography and success of past HITs. This
requires the creation of quality control mechanisms as a component of the data collection process. This has been done successfully with data analysis based on AMT\textsuperscript{98}.

The question of utilizing AMT for adversarial research has been addressed in the past through security applications. These efforts have involved collecting data on human use of network security games\textsuperscript{99,100}. But adversarial research concerning AMT in general has been somewhat limited. There is a general recognition of the challenges of AMT research in this field, however, and the need for effective quality control and methods obtaining objective data for machine learning tasks and social media analysis\textsuperscript{101,102}.

There has also been a great deal of work in the area of spam detection as a means of detecting quality. Relevant work to this dissertation includes research on the use of machine learning to co-opt existing spam filters\textsuperscript{103}. Social spam has been defined and researched in the form of detection frameworks\textsuperscript{104,105}, and studying specific cases of it such as fake reviews for goods and services\textsuperscript{106}. None of this work, however, has resolved whether or not it is possible to crowdsourced a baseline for data sets that are difficult to obtain or don’t yet exist.

4.2 Creating a Stylometry Corpus with AMT: The Extended-Brennan-Greenstadt Corpus

The original adversarial stylometry corpus was created by enlisting 13 trusted individuals to participate in generating adversarial text. These were colleagues, students and other individuals with whom the researchers had a trusted relationship. They were not given any guidance on how to change their writing style. Generating a corpus in this manner was time-consuming and tedious, and a broader approach was required in order to create the sizable data set required to deliver robust results, so the Extended-Brennan-Greenstadt corpus was developed using AMT with the original Brennan-Greenstadt corpus used as a basis for comparison in quality and effectiveness of stylometry approaches. One hundred submissions were obtained, of which 45 were accepted into the final data set based on quality control measures. This new corpus was nearly identical in accuracy across a range of adversarial stylometry evaluations: baseline, obfuscation, and imitation passages.

As in the original Brennan-Greenstadt Corpus, the data acquired from each author was at least
6,500 words of unmodified training data, a 500-word obfuscation passage, and a 500-word imitation passage. For full a full understanding of obfuscation and imitation approaches to circumventing stylometry, please see Section 3.3.

4.2.1 Quality Control in the Extended-Brennan-Greenstadt Corpus

Submission quality is a serious consideration when using the AMT platform, as the completion of a task does not necessarily indicate that the worker has followed the directions and completed it correctly. In order to ensure that the submissions were acceptable, we reviewed every submission and scrutinized them according to the guidelines and requirements listed on the submission form. We only removed authors from the data set who did not adhere to the directions of the survey. We did not remove authors because of the quality of their writing, demographic information, or anything other than their ability to follow directions.

In addition to the existing requirements, we published four guidelines that submissions should adhere to:

1. The submitted pre-existing writing should be “scholarly” in nature (i.e: a persuasive piece, opinion paper, research paper, journals, etc.).

2. Anything that is not the writing content of the work should be removed (i.e: citations, urls, section headings, editing notes, etc.).

3. The papers/samples should have a minimal amount of dialogue/quotations.

4. Please refrain from submitting samples of less than 500 words, laboratory and other overly scientific reports, Q&A style samples such as exams, and anything written in another person’s style.

As an added incentive for authors to take care with their submissions, we offered a bonus payment of two dollars on top of an original payment of three dollars if their submission adhered to the quality guidelines. Of the 100 submissions we received, 45 satisfied the requirements of the survey. These 45 submissions make up the Extended-Brennan-Greenstadt adversarial stylometry corpus and are
the basis of the evaluation for this research. It took about one hour on average for a participant to finish the complete task. The entire instruction set for participation is available online\(^1\).

### 4.2.2 Evaluating the AMT-based Corpus

In order to substantiate the results we present in this paper as being in line with our previous work, we evaluated all of the methods presented in this work on our original data set, the Brennan-Greenstadt corpus. We utilized the author counts available in our original paper given the smaller data set. We found that the precision for each approach is comparable on all data sets. The Basic-9 neural network approach saw a slight drop as can be seen in Figure 4.1. The others were nearly identical as seen with the synonym-based approach in Figure 4.2 and the Writeprints-Static approach in Figure 4.3.

This similarity in accuracy on the different corpora with the same methods, experimental approach, and data-generation questions indicates that the new data set is at least as robust as the original Brennan-Greenstadt corpus. This gives us confidence that enlisting AMT workers to generate new data as part of an adversarial corpus is feasible, albeit with the need for strong quality control and adherence to the unbiased standards set forth. A valuable path for future research is to take a more nuanced approach to the quality control measures and identify ideal financial rewards and automated approaches to preventing poor submissions.

### 4.3 Creating a Crowdsourced Ratings Baseline with AMT

Using the Slashdot data set presented in Section 2.2, we have conducted experiments to understand the strengths and limitations of crowdsourcing a baseline for social spam with AMT. The Extended-Brennan-Greenstadt corpus shows that AMT workers are effective in generating adversarial data, given the right quality control. This experiment examines whether the workers were effective at determining quality of existing data and found that they had high precision in identifying adversarial content, but low precision in identifying high quality content.

---

\(^1\)Online appendix, including participation instructions: [https://psal.cs.drexel.edu/tissec](https://psal.cs.drexel.edu/tissec)
Figure 4.1: Brennan-Greenstadt and Extended-Brennan-Greenstadt with the Basic-9 Neural Network approach.
Figure 4.2: Brennan-Greenstadt and Extended-Brennan-Greenstadt with the Synonym-Based approach.
**Figure 4.3:** Brennan-Greenstadt and Extended-Brennan-Greenstadt with the Writeprints Static SVM approach.
4.3.1 Data Set Prep and Quality Control in the Slashdot Data Set

To remove any ambiguity in the data set and the baseline ratings within it, only a subsection of the original Slashdot data set was used: any comment that was submitted by a registered user and that received a final rating of -1, 0, 3, 4, or 5. This ensures that any comment used in this data set has either a positive or negative rating as decided by a member of the Slashdot community. This is because comments receive a default score of either 1 or 2 for registered users based on their posting history on the site. It is possible that some comments that received ratings are not included in this subset (for example, comments that started at a score of 1 and were moderated to a score of 2)–but every comment in this data set received a rating from at least one member of the Slashdot community.

There were two components of unbiased quality control: average precision and number of ratings performed. The minimum number of HITs required was 25, and that was communicated to the AMT workers at the beginning of each HIT. The minimum threshold required was 55%, but this was not communicated to the workers. Instead, the AMT workers were given a goal of 80% accuracy. Of the 3,110 HITs that were registered by AMT workers, 2,071 met the quality control requirements. Of the 32 workers that completed 25 or more HITs, 24 had accuracies above 55%.

AMT workers were also incentivized to perform well by offering a 60% bonus for quality assessment accuracies of 80% or more. Only two workers met this threshold. The full list of instructions, which included quality control requirements, were communicated to AMT workers as follows:

You will be shown a series of comments, the articles they come from. You are being asked to evaluate these comments under the assumption that they are part of a larger discussion which may be of interest to the readers of the article.

- Please also answer all questions.
- Your bonus pay is based on the quality and accuracy of your answers. If your answers are at least 80% correct, you will receive an additional $.03 per HIT.
- To be eligible for bonus pay, you MUST complete at least 25 HITs.
If you participated in a similar HIT in the past, you may still participate in this HIT.

The range of accuracies for AMT workers can be seen in Figure 4.5, along with a line denoting the threshold for passing quality control.

### 4.3.2 Performance of AMT Workers in Measuring Quality

The data set that was created by AMT workers was analyzed in a variety of ways to determine the ability of the workers to effectively generate baseline classifications for comment quality. The results were mixed, showing a strong ability to detect good comments while detecting bad comments was difficult. Each subsection below explains a piece of analysis that was done.

**Assessing the Ability of AMT Workers to Assess Quality**  There were two primary ways in which the AMT workers assessed quality. First, they looked at the quality of a comment: does a
given comment have a positive or negative effect on the quality of the discussion? AMT workers were, on average, much better at identifying quality content than bad content. The task of identifying quality comments, those with baseline ratings of 3, 4, or 5, was performed with 85.7% precision and 82.6% recall. The task of identifying negative comments, however, was performed with 33.9% precision and 34.2% recall. The results for precision and recall based on quality assessment can be seen in Figure 4.6.

Assessing the Quality of AMT Workers’ Social Spam Ratings The second way workers assessed quality was in whether or not they believed a comment was “social spam” according to a the given definition. The definition used in the AMT task for the term “social spam” is “unwanted content of any sort that is experienced by users of social media or other user-generated content based sites”, which is based on a definition by the anti-spam company Imperium\(^2\). The baseline metric for social spam was the set of comments that were rated -1 or 0 by the Slashdot community. The

\(^2\)http://impermium.com/social-spam.php

Figure 4.5: Assessing the accuracy of quality assessments by AMT workers.
results for precision and recall based on social spam can be seen in Figure 4.7. This graph shows 80% precision and 95% recall for positive comments and 60% precision and 19% recall for negative comments.

**The Effect of Funny Comments.** Earlier we saw that funny comments had a negative effect on accuracy for a machine-learning-based approach to replicating ratings on Slashdot. We found the same thing to be true when AMT workers were assessing quality regarding negative comments, and a slightly negative change for positive comments.

The precision of accurately identifying negative comments rises from 34% to 47.2% when examining the “quality” metric, and just slightly improves from 59.5% to 61.3% for the “social spam” metric. Positive comment precision went down slightly from 82.6% to 79.1% on “quality” and from 79.8% to 78.3% for “social spam.”
Assessing Worker Accuracy Based on Confidence  AMT workers were also asked about their confidence in regards to the topic at hand. The hypothesis behind asking this question was that content on Slashdot, a tech-based news site, may be unfamiliar to the average AMT worker. The exact question was as follows: “Are you confident in your ability to determine the quality of comments related to this article?” with YES, NO, and UNSURE being the possible answers. Workers decided that they were in fact confident 97.9% of the time. This means we only saw a slight deviation in accuracy (71.4% instead of 70.9%) and could not draw robust conclusions as to the effect of self-determined confidence on quality measurements. The results were similar for both the HITs that passed quality control and the raw data set.

Can we identify negative comments?  The case looks grim for being able to accurately identify a range of negative comments. The best that can be done is to look at the per-comment averages of social spam ratings and exclude funny comments.
The challenge with this approach is that workers were less likely to say something was “social spam” rather than simply saying it took away value from the discussion. In fact, workers decided comments were “social spam” less than half as often as they determined comments would have a “negative impact” on discussion. As a result we looked at a “social spam” threshold. A comment was determined to be classified as “social spam” if a certain percentage or more of workers who rated a comment decided they felt it was social spam. This was examined at a range of different thresholds, which can be seen in Figure 4.8 along with their corresponding precision and recall values.

If we take only comments that have a social spam average rating of 50% or more, we see that the precision for detecting "bad" comments is 80%. This accuracy decreases as we lower the threshold to 76.5% for a 40% threshold, 76% for 30% threshold, 66.7% for 20% threshold, and 53.3% for a 10% threshold. While these precision levels are reasonable, the recall is very poor for all cases. Given that the purpose of these experiments is to determine the possibility of establishing a baseline measurement of a data set, recall is important and the poor performance here is concerning.

If we look at the recall for comments that were negative and rated as BOTH social spam and negative in quality, our precision goes up a bit to 44.1%, but our recall goes down to 13.6% because of how few HITs actually meet that criteria.

Given the poor ability of AMT workers to identify negative comments, we suggest a new series of experiments be run to acutely understand the causes of this behavior. These experiments should vary the questions asked to understand if this is an unavoidable trend or if the wording and arrangement of questions are a volatile mix that can drastically sway the answers of the workers.

Interestingly, these results conflict a bit with the results of an initial smaller study to scope the problem. We selected about 50 comments at random to be rated and while we found the same level of precision in positive and negative comments, we found that the accuracy of identifying negative comments with a 15% threshold was 73%. When we looked at those same comments alone in the larger study, we obtained 73.3% precision. And, like the results above, removing the “funny” comments in that list raises our precision to 84.6%. This means that the ability of the workers was quite consistent between the smaller study and the larger study, and suggests that the larger and
Figure 4.8: Assessing the precision and recall of AMT workers in identifying social spam when varying the decision threshold.

more diverse data set is the culprit - the original small set was not a large enough sample of the data set.

Why is this the case? Are the news stories that were related to the earlier set somehow easier to assess in relevance and quality? Was the selection of comments particularly clear? These are important questions, though they fall outside of the scope of this work. The data set has been created and is publicly available

4.3.3 Criticisms of AMT Analysis

The complexity of the experimental setup and diversity of research questions mean that there is a myriad of possible approaches to determining the ability of AMT workers to effectively generate baselines for comment quality. Some of the concerns that may arise have been addressed in this work, and others have been left out. This section addresses a number of potential concerns and

---

3 All data sets are available at [www.mbrennan.net](http://www.mbrennan.net)
either provides an explanation of the considerations taken or explains why they fall beyond the scope of this work.

**Structuring of Question One.** The context of the question that prompted these ratings is important. The question on the questionnaire that prompted this question was “Do you believe this would have a negative impact on the value of a larger discussion surrounding the article” and the choices given were:

- **YES** (This comment would have negative impact on the value of the discussion.)
- **NO** (This comment would have no effect or a positive effect on the value of the discussion.)
- **UNSURE** (I am not sure what effect this comment would have on the value of the discussion.)

**The “Random-Chance” Threshold** Throughout this section the basis for comparison is said to be that of random chance, or 50%, suggesting that there were just two options for a worker to choose from in any given question. There were three choices for each question, however, not two. The reason for choosing the baseline threshold of 50% rather than 33.3% is that very few workers selected “UNSURE” for any given question. For question one, for example, “UNSURE” was selected just 1.5% of the time. We felt, therefore, that choosing a random chance threshold of 33.3% would be misleading. As a result, a choice of “UNSURE” was considered to be incorrect regardless of the true answer.

**Uneven Data Set** The data set contained 69 negative comments and 243 positive comments. It is possible that this lopsided data set altered the ability of AMT workers to effectively judge quality due to a larger number of positive comments coming through. Further experiments that modified the number of negative and positive comments an AMT worker sees are warranted.

**Removing Anonymous and Neutral Comments** A large portion of comments on Slashdot are submitted by anonymous users and/or receive no moderation and remain at their default level. As the purpose of this study was not to analyze Slashdot but rather to analyze the ability of AMT
workers to establish accurate baseline measurements of quality, these comments were removed from the data set as they can add ambiguity to the results.

4.4 Conclusions: Can AMT Create Data Sets for Adversarial Research?

This section demonstrates that in the case of generating original content for an adversarial data set, AMT can perform quite well. In the case of analyzing existing content, however, AMT has mixed results. Further research is needed in order to draw more wide-ranging conclusions, but the evidence in this section is strong in suggesting that generation of adversarial text for stylometric analysis can be done effectively, whereas analyzing comments on a website is much more challenging—succeeding in the ability to highlight good content but failing to accurately identify adversarial content.
Chapter 5: Conclusions

The work presented here demonstrates the ability of machine learning to affect the way we process the massive amount of discourse on the web from the perspectives of quality, relevance, and identity. It also illuminates the effect that it would have on anonymity and adversaries.

Information overload is a huge problem on the Internet and moderation requires significant amounts of human labor. Automatic metrics that can help sort through online discourse for insightful or informative content would be useful—even in large communities like Slashdot with complex moderation schemes where many comments do not get tagged or rated. Our results show that the work of moderators can be amplified by machine learning techniques and that finding interesting discourse automatically is an interesting and likely achievable goal for natural language processing.

This work demonstrates that machine learning can be a valuable tool for gaining an objective understanding of how values are embedded in technologies, how communities develop reputations and norms, and how socio-technical communities can combine human and machine computation. The work we have done thus far with the Slashdot data set has shown that author past performance (reputation) is a good proxy for future results. However, the linguistic feature results suggest that there are interesting and unexpected features to be found that can provide insight into the workings of these community filtering mechanisms. Even in an irreverent community like Slashdot, “I-statements” are indicators of good content, and civility matters.

The work on topic identification on Twitter demonstrates both the benefit and feasibility of incorporating machine learning into social media in order to ease information overload and organize discourse. Twitter is an example of how machine learning is underused in social media. It also provides a fruitful testbed for machine learning approaches to be implemented with great effect. It also shows that despite limited use in modern machine learning research, efficient Bayesian approaches like TWCNB can be implemented easily and are effective enough to be useful in text-based domains that are a large part of social media platforms.
Stylometry is a deep field that has provided a great deal of opportunity for important research in linguistics and computer science, and will not stop doing so anytime soon. Given the large amount of groundwork laid over the past 50 years, it is clearly time to take the next step in a few directions.

The impact that stylometry has on privacy and anonymity, specifically on the Internet, must continue to be looked at seriously. Privacy has become an increasingly important issue given the amount of information available and the constantly improving means of automatically synthesizing it. Stylometry is a clear example of this, and technology should be available to those who desire to preserve their anonymity when publishing documents while still being mindful of the difficult balancing act between privacy and security that comes with it.

Adversarial stylometry is an important subject for future research because of its implications to both the effectiveness of new large-scale stylometry techniques and its potential for protecting individual privacy. This is an especially recent development in stylometry and will require significant work to discover the impact that adversarial attacks may have on the field. There is a potential for an arms race between increasingly effective stylometry techniques and similarly effective attacks.

Continued research in these areas underscores the need for robust data sets with accurate baselines. The work presented here regarding Amazon Mechanical Turk demonstrates that it is possible to generate data sets for adversarial research quite effectively, but that analyzing adversarial data in order to obtain a solid baseline is a mixed bag. Further research may shed light onto the reasons behind the challenge of identifying social spam content using massive crowdsourced platforms such as AMT.

All of the data used to generate the work is open and available on the web. Future work in these areas will bring us closer to a reality where machine learning can effectively join human analysis in combined systems that solve the individual weaknesses of each approach.

---

1 All data is available at [www.mbrennan.net](http://www.mbrennan.net)
Bibliography


[47] Robert Matthews. Linguistics on trial: Forensic scientists have fiercely condemned a technique used in court to show that confessions have been tampered with. New Scientist, 1887, 1993.


