THE JOURNAL OF DIGITAL FORENSICS, SECURITY AND LAW

Journal of Digital Forensics, Security and Law

Volume 9 | Number 2

Article 14

2014

# “Time for Some Traffic Problems”: Enhancing E-Discovery and Big Data Processing Tools with Linguistic Methods for Deception Detection

Erin S. Crabb
*University of Maryland, College Park*

## Recommended Citation

Crabb, Erin S. (2014) "“Time for Some Traffic Problems”: Enhancing E-Discovery and Big Data Processing Tools with Linguistic Methods for Deception Detection," *Journal of Digital Forensics, Security and Law*: Vol. 9 : No. 2 , Article 14.
DOI: https://doi.org/10.15394/jdfsl.2014.1179
Available at: https://commons.erau.edu/jdfsl/vol9/iss2/14

# "TIME FOR SOME TRAFFIC PROBLEMS": ENHANCING E-DISCOVERY AND BIG DATA PROCESSING TOOLS WITH LINGUISTIC METHODS FOR DECEPTION DETECTION

Erin Smith Crabb
University of Maryland at College Park
College of Information Studies
College Park, MD 20740
ecrabb@umd.edu

## ABSTRACT

Linguistic deception theory provides methods to discover potentially deceptive texts to make them accessible to clerical review. This paper proposes the integration of these linguistic methods with traditional e-discovery techniques to identify deceptive texts within a given author's larger body of written work, such as their sent email box. First, a set of linguistic features associated with deception are identified and a prototype classifier is constructed to analyze texts and describe the features' distributions, while avoiding topic-specific features to improve recall of relevant documents. The tool is then applied to a portion of the Enron Email Dataset to illustrate how these strategies identify records, providing an example of its advantages and capability to stratify the large data set at hand.

**Keywords**: e-discovery, deception detection, anomaly detection, forensic linguistics, natural language processing

## 1. INTRODUCTION

The field of electronic discovery (e-discovery), which concerns itself with practices for identifying electronic evidence relevant to investigations or litigation, faces a variety of challenges, ranging from data sets which exponentially increase in size and complexity to aging text processing techniques that some have gone so far as to call "primitive" (Kroll Ontrack, 2013; Oard & Webber, 2013). Contemporary data sets are so large that processing by having a human review and categorize them is prohibitively expensive (Tingen, 2012). In order to more accurately and efficiently process these data, legal teams and courts need to be aware of what advanced technologies are available, how best to employ them and what their limitations are (Tingen, 2012). In addition, they must be prepared to defend their chosen course of action and justify its use (Kroll Ontrack, 2013; Tingen, 2012). Linguistic methods can provide theoretically rich, empirically valid methods to help support these decisions.

This paper explains how a linguistic text classifier can serve as an advanced search technique and how it may be used to detect records which are more likely to contain deceptive information. The proposed classifier benefits from theories of linguistic deception and techniques shown to identify relevant features, providing a tool which is cost effective and informative to improve search results. This linguistically-based approach is fundamentally different from many current search technologies because it draws attention and focus to an author's linguistic style, rather than relying on the words he or she uses. After all, the words found in an email may be misleading (Baron et al., 2007); for example, rather than intending a literal interpretation, an author may intend to communicate sarcasm or be making an allusion. In order to circumvent these issues, this paper provides a framework for summarizing language patterns and discourse quality within each example of an author's work, and then comparing them to find the most suspect texts. Section 2 will summarize the challenges and methods current in the e-discovery field as documented by recent survey works. Section 3 will address linguistic techniques for deception detection and detail how they are relevant to e-discovery. In section 4, I shall discuss preliminary results obtained from a proof-of-concept model exploring one author's email collection from the Enron Email Data Set (EDRM, LLC, 2014). Finally, section 5 concludes the paper with suggestions for further research.

## 2. LITERATURE REVIEW

As noted by Tingen (2012) and Belt, Kiker, and Shetterly (2012), some within the legal field have resisted adopting new technology for e-discovery efforts. Although some members of the legal profession have defended manual review, this is a time-consuming and expensive process, costing up to five hundred dollars an hour for review (Tingen, 2012). However, despite the price and time required for human review, previous studies such as Grossman and Cormack (2011) have shown that technology-assisted review not only requires less effort, but produces more accurate results. However, while the promise of reducing effort on the part of law firms certainly encourages the use of review-assisting products, the quality of returned data is dependent on a number of factors, including the quality of the original data (is it complete, with uncorrupted files), the methods the software employs (is it using basic, complex or a combination of search strategies), and their implementations (is the algorithmic interpretation of the search function stable and error-free) (Grossman & Cormack, 2011).

### 2.1 Modern Search Techniques

Tingen (2012) and the Sedona Conference (Baron et al., 2007) catalogue several of the most common e-discovery techniques and describe how they vary in intended application and sophistication (Grossman & Cormack, 2011; Tingen, 2012).

The most basic method is one many internet users are familiar with: keyword search. A keyword search strategy typically involves developing a list of potentially relevant terms and phrases, and then searching text data for occurrences of those words (Tingen, 2012). While its ease of use has led to its widespread implementation, not all relevant documents necessarily include one or more of the keywords searched for. The reverse is also true: containing a keyword is not a guarantee of a document's relevance (Tingen, 2012). A further complication arises from the fact that keyword searches often only match exact strings; words with spelling errors or inflected words, such as *walked* instead of *walk*, may not be returned (Tingen, 2012).

Utilizing Boolean operators and fuzzy search technologies permits more flexibility in what is still essentially keyword searching. For example, the Boolean operator wildcard, often represented as *, indicates to the computer that any character may take the wildcard's place and that any pattern match should be a returned result (Tingen, 2012). Therefore, using a Boolean operator search for *read** may return documents containing not only *read*, but also *reader*, *reading*, and *ready* (Tingen, 2012). Similarly, fuzzy search techniques attempt to account for human error by weighting letters in a word according to their position and allowing for low-weight letters to differ (Tingen, 2012).

Other search tools include ontologies and taxonomies, Bayesian classifiers, document clustering methods, such as topic modeling, and recently, sentiment analysis (Tingen, 2012; Baron et al., 2007). Ontologies and taxonomies also serve to expand the usefulness of keyword searching, allowing the computer to retrieve keyword synonyms (Baron et al., 2007; Tingen, 2012). For example, if one were to use a taxonomy and search for pen, results may include ballpoint, rollerball or felt tip. Bayesian classifiers use probability algorithms to determine document relevance by weighting particular words or phrases, taking into account their frequency, and also by examining the document's proximity and similarity to others in the set (Tingen, 2012). Document clustering involves statistically analyzing the words in each document in the data set and grouping it with others with similar statistical measurements (Tingen, 2012). Sentiment analysis involves classifying words within a text according to their emotional value (Pang & Lee, 2008). The measurements of emotion can vary greatly; some sentiment classifiers try to determine whether or not a given sentence is negative or positive by averaging emotion scores across all the words in a sentence, while others go so far as to classify sentences according to the "six "universal" emotions [...]: anger, disgust, fear, happiness, sadness, and surprise," (Pang & Lee, 2008).

The techniques discussed in the previous paragraph possess an advantage over simple keyword searches: as discussed by (Baron et al., 2007), when used effectively, they can reveal unexpected results by identifying related concepts that human operators may not have associated before. Each of these methods is effective at returning some proportion of relevant records, but they all share a similar limitation.

## 2.2 The Limitation of Looking at Words

As can be seen, all of these techniques rely on the presence of words, whether they are the particular words being looked for in a keyword search, or related concepts being united through document clustering. However, authors do not always use words in the way they were intended (Baron et al., 2007), and sometimes, the absence of words is far more important than a keyword search may indicate. This is an acknowledged problem in the realm of sentiment analysis: developing methods for identifying sarcastic text, for example, is still a current challenge (González-Ibáñez, Muresan, & Wacholder, 2011).

Zhou et al. (2003) defines deception as, "the active transmission of messages and information to create a false conclusion." The authors further specify that there must be an intent to deceive; messages that a transmitter does not know are false are not considered deceptive under this definition (Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). These messages can deceive in different ways. While some involve the transmission of misleading information, it is also possible to deceive with one's writing style to hide the identity of the author (Afroz, Brennan, & Greenstadt, 2012; Brennan, Afroz & Greenstadt, 2012; Juola, 2012). Although the transmission of misleading information is the focus of this study, stylistic deception will be briefly addressed as well (Afroz, Brennan, & Greenstadt, 2012; Brennan, Afroz & Greenstadt, 2012; Juola, 2012).

Past studies have shown that during deceptive interaction, there are a variety of cues which can indicate the deception (Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). However, due to the relatively recent rise in computer-mediated communication (CMC), the majority of research identifying these deceptive indicators has focused on face-to-face, human-to-human interaction (Enos, Shriberg, Graciarena, Hirschberg & Stolcke, 2007; Fitzpatrick & Bachenko, 2009; Zhou, Burgoon, & Twitchell, 2003; Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). Further research into linguistic methods of detecting deception is needed; due to the fact that many of the indicators previously studied, such as pupil dilation and pressing lips together, are unavailable when the interaction's participants are separated by time, space and technology (Keila & Skillicorn, 2005; Lee, Welker, & Odom, 2009; Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). The following section will discuss deception in CMC and how it may be measured.

## 2.3 Linguistic Techniques for Deception Detection

The scope of our search can be broadened by examining categories of words rather than just words themselves. This section examines previous literature on deception, focusing on studies centered on CMC, and identifies features which will be implemented in the model classifier (see section 4).

## 2.4 Deception in Explicit Experimental Space

While the literature on deception as a speech act is quite rich, deception in CMC is a sparsely studied field, in part due to the difficulty of collecting authentic data. Lee, Welker, and Odom (2009), Hancock, Curry, Goorha, and Woodworth (2008) and Zhou, Twitchell, Qin, Burgoon, and Nunamaker (2003) together with Zhou, Burgoon, and Twitchell (2003) employ similar experimental methods, which involve encouraging some participants to deceive a randomly assigned, anonymous communication partner. (Incentivizing experimental participants to deceive has also been undertaken in face-to-face communication studies, such as Enos, Shriberg, Graciarena, Hirschberg, and Stolcke (2007).) The language used by the deceptive and non-deceptive participants was then compared in order to identify whether or not there were language patterns endemic to deception, and if so, what they were (Hancock, Curry, Goorha, & Woodworth, 2008; Lee, Welker, & Odom, 2009; Zhou, Burgoon, & Twitchell, 2003; Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). In each study, students were anonymously paired with others, and each of these communication dyads was then instructed to communicate about some outside entity (Lee, Welker, & Odom, 2009; Zhou, Burgoon, & Twitchell, 2003; Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). In Zhou, Twitchell, Qin, Burgoon, and Nunamaker (2003) and Zhou, Burgoon, and Twitchell (2003), students were told to reach agreement on a ranking of items which would be most useful in surviving in the desert using asynchronic CMC. The other study involved one student describing the state of real estate properties to an inquiring party via a specially-designed email system (the other student) (Lee, Welker, & Odom, 2009). Unlike the others, the Hancock, Curry, Goorha, and Woodworth (2008) study involved synchronic communication between participants through computer terminals; they were still deprived of face-to-face interaction, but responses were potentially much more immediate.

In all three cases, at least one of each pair of students was presented with some level of motivation to deceive their partner. However, as noted above, these interactions are not what one could consider "real-world" as they would not have occurred outside the experimental sphere. Although the authors attempts to mitigate this by informing some participants assigned to the deceptive group that their partner's belief was important, there was little to no real risk involved with being unsuccessful with their deception attempts (Hancock, Curry, Goorha, & Woodworth, 2008); as Lee, Welker, and Odom (2009) note, "The kinds of messages that people construct defensively when lying to a superior who can exact harm on them may differ from the kinds of messages that individuals construct when lying to unknown students..."

The linguistic features examined in the articles described thus far varied, but each one identified some cues which were significantly associated with the act of deception. Zhou, Twitchell, Qin, Burgoon, and Nunamaker (2003) determined that use of ellipsis, wordiness, passive voice, second person address and possessive forms were significantly indicative. The authors also found that deceptive study participants used more words (especially verbs, modifiers and noun phrases) and yet also displayed less lexical and content diversity (Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). Zhou, Burgoon, & Twitchell (2003) confirmed the significance of these features, but also shed light on the importance of chronology in deception analysis. They found that at the beginning of a deception, the cues available for analysis are "fair [in] number", and as time progresses, they shift in number, increasing and then decreasing until the end of the communicative event, at which time very few cues are exposed; furthermore, not all cues were significant over all time periods (Zhou, Burgoon, & Twitchell, 2003). To mitigate this chronological challenge, the authors suggest either merging all messages or selecting certain messages toward the middle of a communicative event for further scrutiny (Zhou, Burgoon, & Twitchell, 2003). Following this guidance, the model classifier proposed in section 4 below functions on a merged set which is not chronologically separated.

The increase in wordiness found by Zhou, Twitchell, Qin, Burgoon, and Nunamaker (2003) was corroborated by Hancock, Curry, Goorha, and Woodworth (2008). Additionally, it was determined that deceptive participants used fewer causal words (such as because or effect) when lying (Hancock, Curry, Goorha, & Woodworth, 2008). Deceptive participants were also more likely to use sense-related words (see or listen) when they were not telling the truth.

According to Afroz, Brennan and Greenstadt (2012), function words are the most useful, content-independent features for detecting deceptive writing. Additionally, they state that these function word features can be used to identify style deceptions; that is, although an author can alter their style to hide their identity, function words can be used to determine that style obfuscation has taken place (Afroz, Brennan, and Greenstadt, 2012). Juola (2012) also successfully utilized character trigrams (series of three characters, such as *cha* or *har* from *character*) to detect imitation and stylistic obfuscation.

Lee, Welker, and Odom (2009) elaborated on the difference between deceiver and truth-teller content words, stating that deceivers' messages included more causation words, first-person singular pronouns, present-tense verbs and tenacity verbs. This finding regarding first-person singular references is somewhat different from one of the findings reported in Zhou, Twitchell, Qin, Burgoon, and Nunamaker (2003), where the authors found that group references were significant, but first-person singular references were not. In comparison, Hancock, Curry, Goorha, and Woodworth (2008) found that first-person references fell in number during deceptive communication. In light of the disagreement, the model classifier (see section 4) will examine these features (first-person singular and first-person plural references) separately. It is possible these observed differences may be due to the context of the conversation: for example, when discussing desert survival such as in the Lee, Welker, and Odom (2009) study, it may be more prudent to refer to the survival of the group rather than oneself.

### 2.5 Deception in Real-World Data

While these findings are an excellent place of departure, it is also important to verify their findings with research on real-world data, if and when available (Feng, Banerjee & Choi, 2012; Fitzpatrick & Bachenko, 2009; Lee, Welker, & Odom, 2009; Zhou, Burgoon, & Twitchell, 2003; Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003). Any data not produced as the prompting of a study could be considered real-world; for example, Feng, Banerjee and Choi (2012) use reviews collected from several review websites while Fitzpatrick and Bachenko (2009) collect both spoken and written narratives from sources such as Court TV and police case files.

The Enron Email Data Set is a large collection of real world emails released to the public domain. It is comprised of the mail folders of 150 former Enron employees, including some of the top executives who were later prosecuted, some of whom were convicted of deception based crimes (EDRM, LLC, 2014). What is important is that the data were not motivated by a study, and it is reasonable to assume that some of the emails within the set contain evidence of deception (Keila & Skillicorn, 2005).

Keila and Skillicorn (2005) examine the Enron Email Data Set in detail. The authors built a classifier which ranked emails by the likelihood that they contained deception based on the measurement of a variety of linguistic features (Keila & Skillicorn, 2005). This approach, in essence, identifies emails which deviate most from a model of language, rather than trying to predict whether each individual text is deceptive (Keila & Skillicorn, 2005). Although the authors' tool does not discriminate the identities of senders or receivers of email, they state that their approach can also be applied successfully to the emails of a single person to identify their most unusual messages (Keila & Skillicorn, 2005).

Gupta (2007) proposes the use of the Pennebaker deception model to find deceptive emails within the Enron Email Data Set (EDRM, LLC, 2014). The applied model relies on changing frequencies of personal pronouns, exclusive words, negative-emotion words and action words (Gupta, 2007). However, the sparsity of the created model (due to the lack of most words not occurring in most of the texts) results in the need for specialized normalization procedures to prevent zero values from unduly influencing the model (Gupta, 2007). While Gupta (2007) suggests that deception word features can be customized to suit any domain, the danger of data sparsity remains. It also does not take into account syntax (such as the noun *drive* as compared with the verb *drive*), context or literary devices, which can all affect the meaning of a used word in a text (Gupta, 2007).

Louwerse, Lin, Drescher, and Semin (2010) follow in the footsteps of Keila and Skillicorn (2005) and Gupta (2007) and examine the Enron Email Data Set (EDRM, LLC, 2014). Using linguistic categories and several derived features, they analyze the data set to determine if certain features could be linked to fraudulent events (Louwerse, Lin, Drescher, & Semin, 2010; EDRM, LLC, 2014). They conclude that abstractness (which was calculated using a formula to compare usage of verbs and adjectives) is most indicative of fraudulent events in the corpus (Louwerse, Lin, Drescher, & Semin, 2010). However, they qualify their findings by stating, "By no means are we arguing that by using the LCM model we can predict whether an email consists of fraudulent information or not," (Louwerse, Lin, Drescher, & Semin, 2010).

A similar model to the one proposed in this paper was reported by Feng, Banerjee and Choi (2012). The authors built a classifier which utilized both lexical and syntactic rule features to uncover deception in several different corpora of reviews (Feng, Banerjee & Choi, 2012). However, this requires a deep syntactic parser for the language being studied, which may not available and may have varying accuracy rates (Feng, Banerjee & Choi, 2012). Their process also relies on lexical features, which they state have been identified as useful within a specific genre, such as reviews (Feng, Banerjee & Choi, 2012).

## 3. MODEL

In this section, two methods of modeling and identifying unusual texts within a single author's texts will be outlined, along with their results. Both models will rely on the same set of generalized linguistic as opposed to lexical features; this avoids overfitting the models to specific lexical items as seen in Gupta (2007). Although Feng, Banerjee and Choi (2012) identify that word-based features can be useful in genre-specific deception detection, the current study does not employ them. This is because the genres of emails included cannot be guaranteed, and this model is intended to be run over potentially genre-mixed data; for example, some emails may be formal business electronic letters, while others may be very personal and informal, incorporating conventions such as emoticons and numeric substitutions.

### 3.1 Data Set

Both classification methods were run on a portion of the Enron Email Data Set, due to the fact that it is publicly available and is comprised of real-world data (EDRM, LLC, 2014; Keila & Skillicorn, 2005). As discussed previously, this model seeks to find interesting records within a specific author's work (as opposed to drawing comparisons across a population) and demonstrate the value of linguistic feature-based email ranking as an investigative tool. Fornaciari and Poesio (2012) conclude from their study of deception in court transcripts that, "increasing homogeneity [of subsets] is effective provided that the remaining set is still sufficiently large." Therefore, the data set is composed of emails sent or drafted by one author within the Enron Email Data Set, who will be unnamed in the interest of privacy (EDRM, LLC, 2014). The data set for the present experiment was built by exporting the mailbox of this single author from the full collection, and then removing all messages not written by him; for example, all emails received were removed, as were forwarded portions of those he sent. As he was one of the first Enron executives prosecuted, it is reasonable to assume that his emails may contain items of interest, although they may not necessarily be deceptive (Keila & Skillicorn, 2005). It is important to note that although no definitive identification of deceptive emails is possible, as no ground-truth is available, the purpose of the model is to identify potentially relevant or interesting documents which may be unresponsive to topic searching or predictive coding techniques.

### 3.2 Methods

To begin, a list of relevant features, derived from those identified in the deception detection literature, was assembled. Other features were then added to create a more robust language model for each email, and the complete list of features can be seen in Table 1 below (EDRM, LLC, 2014). For each of the part of speech and word count features, emails were first split into sentences using the Natural Language Toolkit sentence tokenize function (Bird, Loper, & Klein, 2009). Each sentence was then tokenized on the word level and passed through the Penn Treebank part of speech tagger (Bird, Loper, & Klein, 2009). Frequencies for relevant part of speech tags were calculated within each sentence, and some tag values were condensed; for example, one condensed tag set is comprised of nouns and noun phrases, as lexical diversity was the feature of interest, not types of noun phrases. Finally, these tag-per-sentence frequency values are averaged across the sentences within the email to form the email-level feature.

Two analytic methods for reviewing the emails were then employed. First, WEKA, a tool comprised of machine-learning algorithms published originally by Hall et al. (2009), was used to create a cluster-based classification. The mean feature values for each cluster were then examined to identify which had higher means for parts of the speech associated with deception. The goal of employing this method was to identify groups of emails with similar patterns of language variation, on the basis that the most interesting clusters could then be extracted for analysis.

The second method involved calculating the mean values for each feature in isolation, identifying what emails had statistically significant differences for deception-relevant features, and subjecting these to human review.

For comparison, a list of keyword terms were put together and run on the data set as well. As discussed in section 2.1 above, there are many "predictive coding" tools available on the software market (Grossman &

Cormack, 2011). It is my assumption that the results obtained from one tool’s use may be very different from another. As it is not the goal of this paper to judge the efficiency of particular industry tools, I have chosen to employ a traditional keyword and Boolean operator search of the data using Python. In order to facilitate this, emails were converted to plain text.

### 3.3 The Five Clusters

The WEKA Expectation Maximum (EM) algorithm was used to identify clusters within the created feature matrix (Hall et al., 2009). The algorithm was analyzed with 10-fold cross validation, and was set to generate its own number of clusters (Hall et al., 2009). The table below outlines the clusters built by WEKA. The mean values for each feature, as well as their standard deviations, are included. It is important to note that the third singular feature includes words for third singular subjects and objects, as well as possessive pronouns. First person pronouns were not weighted in the clustering algorithm, because of the conflicting results from previous works as discussed in 2.4 above, and were therefore not reported.

As can be seen, a variety of features were examined, not only those previously determined as significant with regard to deception detection. As discussed briefly above, this was done in order to give a more thorough outline of the style of each email. Furthermore, the additional features may reveal other potentially interesting features, and they contribute to an analysis of lexical diversity. As discussed above, lower lexical diversity occurs with a lower frequency of content words. This study assessed this by taking into account the frequency of the remaining weighted word types, including repeated and unique nouns, adjectives and adverbs.

Table 1 summarizes the features utilized in the cluster-based classifier. Mean values for each feature’s frequency, as calculated across a cluster, are in standard font while standard deviations are in italics.

| Feature | Cluster 1<br>(74) | Cluster 2<br>(35) | Cluster 3<br>(137) | Cluster 4<br>(74) | Cluster 5<br>(85) |
|---|---|---|---|---|---|
| **Metadata** | | | | | |
| Subject length | 19.8934 | 20.2261 | 17.1884 | 21.3642 | 22.5450 |
| | *13.2477* | *11.8727* | *8.7631* | *11.9969* | *10.3724* |
| Email length | 1.0148 | 4.3292 | 7.5018 | 3.5584 | 10.4032 |
| | *0.1209* | *2.2240* | *4.6405* | *2.8894* | *7.3635* |
| Emoticons | 0.0540 | 0 | 0.0237 | 0.0007 | 0 |
| | *0.2794* | *0.1252* | *0.0779* | *0.0089* | *0.0001* |
| Raw punctuation | 3.3141 | 1.6032 | 1.6197 | 1.4976 | 1.7495 |
| | *20.8164* | *0.7167* | *0.7265* | *1.6505* | *0.5947* |
| ‘@’ or ‘#’ | 0 | 0.0303 | 0 *0.0002* | 0 | 0.0156 |
| | *0.0454* | *0.0997* | | *0.0454* | *0.0541* |
| **Grammatical features** | | | | | |
| Modal verbs | 0 | 0.0090 | 0.0031 | 0.0008 | 0.0041 |
| | *0.0041* | *0.0057* | *0.0028* | *0.0016* | *0.0031* |
| Prepositions | 0.0001 | 0.0167 | 0.0148 | 0.0139 | 0.0178 |
| | *0.0008* | *0.0092* | *0.0075* | *0.0107* | *0.0053* |
| Coordinators | 0.0001 | 0.0039 | 0.0046 | 0.0008 | 0.0039 |
| | *0.0006* | *0.0034* | *0.0037* | *0.0016* | *0.0027* |
| Determiners | 0 | 0.0116 | 0.0156 | 0.0143 | 0.0153 |
| | *0* | *0.0061* | *0.0072* | *0.0103* | *0.0056* |
| Adjectives | 0.0021 | 0.0049 | 0.0092 | 0.0060 | 0.0071 |
| | *0.0067* | *0.0046* | *0.0076* | *0.0090* | *0.0043* |

| Feature | Cluster 1<br>(74) | Cluster 2<br>(35) | Cluster 3<br>(137) | Cluster 4<br>(74) | Cluster 5<br>(85) |
|---|---|---|---|---|---|
| Adverbs | 0.0011 | 0.0039 | 0.0062 | 0.0053 | 0.0072 |
| | *0.0066* | *0.0040* | *0.0063* | *0.0086* | *0.0061* |
| ‘To’ | 0 | 0.0028 | 0.0057 | 0.0031 | 0.0050 |
| | *0* | *0.0031* | *0.0049* | *0.0043* | *0.0033* |
| **Pronouns** | | | | | |
| 2nd | 0.0009 | 0.0029 | 0.0047 | 0.0015 | 0.0023 |
| | *0.0073* | *0.0044* | *0.0060* | *0.0030* | *0.0027* |
| 2nd possessive | 0 | 0.0001 | 0.0010 | 0.0006 | 0.0007 |
| | *0.0015* | *0.0005* | *0.0021* | *0.0018* | *0.0015* |
| 3rd sg. | 0 | 0.0006 | 0.0003 | 0.0012 | 0.0008 |
| | *0.0026* | *0.0019* | *0.0009* | *0.0039* | *0.0018* |
| 3rd pl. | 0 | 0 *0.0003* | 0 | 0 | 0.0019 |
| | *0.0010* | | *0* | *0.0010* | *0.0016* |
| 3rd pl. possessive | 0 | 0 | 0.0001 | 0 | 0.0006 |
| | *0.0005* | *0.0005* | *0.0003* | *0.0005* | *0.0011* |
| **Verb types** | | | | | |
| Base | 0 | 0.0125 | 0.0090 | 0.0042 | 0.0082 |
| | *0.0001* | *0.0064* | *0.0065* | *0.0058* | *0.0046* |
| Present tense | 0.0001 | 0.0091 | 0.0190 | 0.0150 | 0.0165 |
| | *0.0011* | *0.0067* | *0.0095* | *0.0124* | *0.0066* |
| Past tense | 0 | 0.0029 | 0.0065 | 0.0071 | 0.0079 |
| | *0* | *0.0035* | *0.0049* | *0.0082* | *0.0049* |
| **Nouns** | | | | | |
| Repeated nouns | 0.0015 | 0.0006 | 0.0001 | 0.0004 | 0.0007 |
| | *0.0081* | *0.0012* | *0.0003* | *0.0016* | *0.0009* |
| Unique nouns | 0.0235 | 0.0811 | 0.0653 | 0.0870 | 0.0601 |
| | *0.0533* | *0.0234* | *0.0209* | *0.0314* | *0.0135* |
| **Wh- words** | | | | | |
| No pronouns | 0.0002 | 0.0005 | 0.0022 | 0.0004 | 0.0014 |
| | *0.0020* | *0.0014* | *0.0039* | *0.0013* | *0.0022* |
| Wh- pronouns | 0 | 0.0001 | 0.0025 | 0.0004 | 0.0013 |
| | *0.0021* | *0.0005* | *0.0038* | *0.0010* | *0.0027* |

Cluster 2 is the most interesting of the clusters above, for several reasons. It has high frequencies of modal, base and present tense verbs, medium to high frequencies of second person pronouns and high frequencies of function words (most notably, prepositions and coordinators). This is combined with medium to very low frequencies of unique and repeated nouns (although the means are relatively high, the standard deviation is as well), medium to low frequencies of adjectives and adverbs, and medium to long email bodies, this cluster is a candidate for human review. Cluster 5 too has high frequencies of these features, and given the small numbers of emails placed into these clusters (120), human review for both is plausible, although this may not always be the case.

## 3.4 Individual Feature Analysis

Table 2, which summarizes the results of the second method discussed in 4.2 above, can be found below. Mean values for each feature were calculated across the entire data set. The statistical significance values are standard deviations from the mean. The final column is the number of emails with values which are statistically significant in the direction under scrutiny; that is, higher frequencies of features are more important. Even in the case of lexical diversity this is true, because the model should not be drawing attention to emails with high unique content.

Table 2 shows the features examined, their corpus-wide means and standard deviations. The final column identifies the number of emails in the corpus with significant values for each feature.

| Feature | Mean value | Standard deviation | Number of emails with significant values (out of 405) |
|---|---|---|---|
| **Metadata** | | | |
| Subject length | 20.50617 | 11.67137 | 75 |
| Email length | 4.935802 | 5.053753 | 55 |
| Emoticons | 0.0136596 | 0.1252213 | 7 |
| Raw punctuation | 1.905513 | 9.005654 | 5 |
| '@' or '#' | 0.0070054 | 0.0454155 | 13 |
| **Grammatical features** | | | |
| Modal verbs | 0.0027279 | 0.0041302 | 60 |
| Prepositions | 0.0125777 | 0.0101776 | 64 |
| Coordinators | 0.0021978 | 0.0030283 | 69 |
| Determiners | 0.0116533 | 0.0093803 | 60 |
| Adjectives | 0.0057991 | 0.0075239 | 52 |
| Adverbs | 0.0047896 | 0.0072721 | 45 |
| 'To' | 0.0031837 | 0.0040969 | 71 |
| **Pronouns** | | | |
| 1$^{st}$ sg. | 0.0044918 | 0.0061031 | 27 |
| 1$^{st}$ sg. possessive | 0.0002954 | 0.0013422 | 23 |
| 1$^{st}$ pl. | 0.00092 | 0.002454 | 36 |
| 1$^{st}$ pl. possessive | 0.000301 | 0.001155 | 28 |
| 2$^{nd}$ | 0.0021805 | 0.0048731 | 39 |
| 3$^{rd}$ sg. | 0.0006766 | 0.0026263 | 28 |
| 3$^{rd}$ pl. | 0.0003241 | 0.0009732 | 43 |
| 3$^{rd}$ pl. possessive | 0.0001186 | 0.0005079 | 25 |
| **Verb types** | | | |
| Base | 0.006042 | 0.0066367 | 79 |
| Present tense | 0.0122779 | 0.0111024 | 61 |
| Past tense | 0.0052184 | 0.0064449 | 54 |
| **Nouns** | | | |
| Repeated nouns | 0.0006551 | 0.0036768 | 10 |
| Unique nouns | 0.0668139 | 0.0396788 | 55 |
| **Wh- words** | | | |
| No pronouns | 0.0008137 | 0.002271 | 37 |
| Wh- pronouns | 0.0006929 | 0.0021356 | 35 |

As can be seen, there are fewer emails found significant on this per-feature basis (the maximum of which was 79) than those identified in clusters 2 and 5 (*n=120*). This is likely because of the flexibility of the clustering algorithm and weighting system. Not every email in the relevant clusters had significant values for every feature; they simply had a high enough proportion of significant features as to pattern together.

## 3.5 Analysis

Overall, these two methods pointed to similar potentially-deviant collection subsets, with the latter suggesting fewer in total. Although this second, smaller subset returned more precise results, its recall may be too low and the cluster model may be more fitting for smaller data sets more suited to human review. The smaller subset is problematic for the collection size being tested (about 400 emails in total from the author in question) in this instance, because its overlap with keyword search is fairly large (for example, 22 of the 41 emails which had statistically significant modal and base verb features were also returned by keyword searching).

The emails in clusters 2 and 5 contained more disparate data, including a number of personal contacts (discussing topics such as golf plans and his job dissatisfaction) and business related emails in a more informal linguistic style (discussing and justifying several notable business practices which he was later prosecuted for) that were not identified by keyword search.

One interesting pattern which emerged was that a few of his emails in cluster 5, especially those in a more informal style, contained descriptions and unflattering characterizations of those involved in business situations he disagreed with, along with information about what those situations were in general terms. While these records are undoubtedly relevant (both from an e-discovery and investigative perspective), he does not refer to the people he is describing by name; sometimes he does not even refer to them by the company they work for. This, combined with a general vagueness, lack of nouns he usually associates with his workplace and no use of notable keywords resulted in some of these emails not being returned by keyword searching. Additional emails in this style were returned by searching some of the most general keywords (such as trade and trading), although any more sophisticated filtering mechanism, such as requiring more than one keyword, would have eliminated the record as irrelevant.

It is important to note that both of these methods in essence divide texts by their writing style, and then select subsets based on the concentration of frequencies of deception-associated features. Brennan, Afroz and Greenstadt (2012) demonstrated that it may be possible for authors to obfuscate their writing style purposefully. While Brennan, Afroz and Greenstadt (2012) focuses on concealing an author's identity, the ability for participants to confuse stylometric tools by attempting to write differently suggests that it may be possible to consciously alter one's writing style to change the distribution of certain features. These stylistic countermeasures may be detectable, however (Afroz, Brennan, & Greenstadt, 2012; Juola, 2012).

## 4. CONCLUSION

As demonstrated, linguistic methods for deception detection can contribute to e-discovery search by stratifying the data and highlighting records which may otherwise go undetected during a search phase. Although this classifier method does not replace current search technologies and some results may be overlapping, it can be effectively utilized alongside them to bring attention to unusual and interesting emails. Further investigation is needed to determine what search methods best compliment the proposed model. Additional verification is also recommended, preferably with a data set in which deceptive texts are identified, so that empirical, ground-truthed testing may be undertaken. Research into the ability of an author to purposefully conceal deceptive indicators would also be highly recommended.

## REFERENCES

Afroz, S., Brennan, M., & Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. In 2012 IEEE Symposium on Security and Privacy (SP), 461-475.

Baron, J. R., Braman, R., Withers, K., Allman, T., Daley, M., & Paul, G. (2007). The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in e-Discovery. *The Sedona Conference Journal*, *8*, 189-223.

Belt, W., Kiker, D., & Shetterly, D. (2012). Technology-assisted document review: Is it defensible? *Richmond Journal of Law and Technology*, *XVIII*(3), 1-43.

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*, O'Reilly Media Inc.

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Transactions on Information and System Security (TISSEC), *15*(3), 12:1-12:22.

EDRM, LLC (2014). Enron Email Data Set v2. Retrieved from http://www.edrm.net/resources/data-sets/edrm-enron-email-data-set

Enos, F., Shriberg, E., Graciarena, M., Hirschberg, J., & Stolcke, A. (2007). Detecting deception using critical segments. *INTERSPEECH*, 2281-2284.

Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, *2*, 171-175. Association for Computational Linguistics.

Fitzpatrick, E. & Bachenko, J. (2009). Building a forensic corpus to test language-based indicators of deception. *Language and Computers*, *71*(1), 183-196.

Fornaciari, T., & Poesio, M. (2012). On the use of homogenous sets of subjects in deceptive language analysis. Proceedings of the Workshop on Computational Approaches to Deception Detection, 39-47. Association for Computational Linguistics.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 581–586.

Grossman, M., & Cormack, G. (2011). Technology-Assisted review in e-Discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, *XVII*(3), 1-33.

Gupta, S. (2007). Modelling Deception Detection in Text (Master's thesis). Retrieved from http://qspace.library.queensu.ca/handle/1974/922

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1), 10-18.

Hancock, J., Curry, L., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, *45*(1), 1-23.

Juola, P. (2012). Detecting stylistic deception. In Proceedings of the Workshop on Computational Approaches to Deception Detection, 91-96. Association for Computational Linguistics.

Keila, P., & Skillicorn, D. (2005). Detecting Unusual and Deceptive Communication in Email. Technical report.

Kroll Ontrack (2013). 5 Daunting Problems Facing EDiscovery: Insights on EDiscovery Challenges in the Legal Technologies Market. Technical report. Retrieved from http://www.krollontrack.com

Lee, C., Welker, R., & Odom, M. (2009). Features of computer-mediated, text-based messages that support automatable, linguistics-based indicators for deception detection. *Journal of Information Systems*, *23*(1), 5-24.

Louwerse, M., Lin, K. I., Drescher, A., & Semin, G. (2010). Linguistic cues predict fraudulent events in a corporate social network. Proceedings of the 32nd Annual Conference of the Cognitive Science Society, 961-966.

Oard, D., & Webber, W. (2013). Information retrieval for E-Discovery. *Foundations and Trends in Information Retrieval*, 7(2-3), 99-237.

Pang, B., & Lee, L. (2008). Opinion mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135.

Tingen, J. (2012). Technologies-that-must-not-be-named: Understanding and implementing advanced search technologies in E-Discovery. *Richmond Journal of Law and Technology*, *XIX*(1), 1-49.

Zhou, L., Burgoon, J., & Twitchell, D. (2003). A longitudinal analysis of language behavior of deception in e-mail. In Chen, H. Miranda, R., Zeng, D., Demchak, C., Schroeder, J., Madhusudan, T. (eds). *Intelligence and Security Informatics*, LNCS 2665, 102-110. Springer Verlag, Berlin Heidelberg.

Zhou, L., Twitchell, D., Qin, T., Burgoon, J., & Nunamaker, J. (2003). An exploratory study into deception detection in text-based computer-mediated communication. Proceedings of the 36th Hawaii International Conference on System Sciences, 1-10.