The Impact of Message Length and Medium on Imitation Attack Creation and Detection by Humans

Sarah Lazbin
slazbin@uci.edu

Abstract

It's important to know who the true author of an electronic document is in an age where cybercriminals can steal a person's identity by impersonating them in electronic communication. This is known as an *imitation attack*. Current human ability to detect deception is generally poor, though automatic techniques relying on linguistic features fare better at identifying authorship deception like imitation attacks. However, it is unknown how accurate either these authorship techniques or humans are on messages of shorter lengths and with different communicative goals (e.g., tweets, texts, and emails). To evaluate authorship techniques in shorter messages, we first require a dataset of imitation attacks on these different message types. I asked Amazon Mechanical Turk participants to generate samples of their own writing style as well as imitations of previously generated samples. With these data in hand, I assessed (i) how well humans can detect imitations across different message length and medium types, and (ii) whether the imitations differ by 51 linguistic features used previously in automatic detection. In Aim 1, an investigation of human detection, participants were asked to read samples from a given target author and judge if an additional message was from the target author or an imitator. I found that shorter messages are easier to imitate and also make detection harder. People performed at about random chance when attempting to detect imitations in shorter mediums (texts and tweets) and only a little bit better in longer mediums (emails). Automatic detection on the other hand, historically performs better than humans at detecting imitations. In Aim 2, automatic linguistic feature analysis, I looked at linguistic feature usage to examine how much people were able to change their natural writing style to look more like that of the target when creating an imitation. This allowed me to determine which of these features are easier/harder to shift and if this varies across mediums. This research provides the foundation for future automatic linguistic-feature-based techniques that can detect imitation attacks robustly across different message types.

1. Introduction

a. Stylometry for Detecting Author Identity

In a day and age where the mere act of typing of someone's name is considered an "electronic signature" and therefore legally binding, it is extremely important to know who the true author of an electronic document is. While people often assume they are completely anonymous behind a computer screen, current computational techniques can identify authorship fairly accurately (Li 2006, Pearl & Steyvers 2012). More specifically, these computational techniques are quite accurate at detecting impersonations, sometimes called *imitation attacks*, for longer writing samples like an excerpt from a novel (Brennan & Greenstadt 2009, Pearl & Steyvers 2012). This statistical and computational analysis of linguistic features, known as stylometry, dates back to the 1880s when Thomas Mendenhall first analyzed word length in written letters (Tweedie 1996). Researchers interested in the true author of a message look at stylometric features in sets of writing samples that are thought to "reflect unconscious aspects of an author's style, appearing in the form of distinctive, quantifiable features that are salient, structural, and frequent" (Pearl & Steyvers 2012). Some examples of stylometric features one could choose to investigate include average length of word, proportion of exclamation marks to all punctuation marks, and average sentence length (Pearl & Steyvers 2012).

The Federalist Papers are perhaps the most famous example of determining authorship (Tweedie 1996). Originally published anonymously, they were 85 essays calling for support of the Constitution. It is now commonly known that two thirds were composed by Alexander Hamilton and the rest were composed by James Madison and John Jay. Notably, there were about 12 papers that both Hamilton and Madison claimed to have

written. Using stylometric techniques, Mosteller and Wallace (1964) distinguished who wrote which papers by using the papers with known authors as reference sets and comparing these against the disputed papers. Their conclusions for each of the 12 papers coincided with hypotheses of historians (Klarreich 2003).

b. Human Detection of Authorship Deception

Presumably, non-linguist humans do not use the same statistical approach in determining authorship. Naive analysis by humans might tend to rely on more targeted or holistic aspects to determine if an author's style has changed. For example, someone might get the feeling that a complex word is not within the vocabulary of the author (Klarreich 2003) -- such as if your ten-year-old sister described the mountain scenery as "placid". Another example is that the syntax just feels like it's wrong (Pearl & Steyvers 2012). For instance, suppose your friend typically speaks using run-on sentences with tons of dependent clauses; if she started using lots of short abrupt sentences, you would get the sense you are no longer talking to her. In both of these examples, the reader has to personally know the individual they are communicating with in order to determine that something has changed. That is, the reader has an internal, subconscious sense of the author's writeprint, and consciously notices something has deviated from that.

It is unknown if humans are good at imitation attack detection specifically, though, they typically aren't very good at deception detection in general (e.g., Vrij 2000; Newman, et al. 2003; Ott, et al. 2011; Ott, Cardie, and Hancock 2013; Fitzpatrick, Bachenko, and Fornaciari 2015). Because of this, if deception detection is pertinent to a job, humans need to be extensively trained to perform more accurately than someone who is

inexperienced (Zuckerman, DePaulo, and Rosenthal 1981; McCornack and Parks 1986; Ott et al. 2011; Levine 2014).

c. Computational Stylometry

Obviously, computers can be programmed to compute a multitude of different statistical analyses of linguistic features in seconds; thus, we might expect a stylometric computational analysis program to perform significantly better than a human would in identifying authorship.

Successful computational stylometry combines statistical analysis of linguistic features and machine learning approaches. To successfully do this, the field relies on a growing corpus of online writing samples to determine the identity of an author when the author is either anonymous, disputed, or uncertain (as in imitation attacks).

Computational stylometry can be thought of as analyzing a person's unique *writeprint* in order to determine authorship. A writeprint, which originates from the concept of identification through fingerprint (Pearl & Steyvers 2012, Pearl & Enverga 2015, Pearl, Lu, & Haghighi 2016), is a collection of linguistic features that are indicative of an individual's writing style. The idea is that the features in the writeprint are a subset of all possible linguistic features and are assumed to be unconscious and static across a person's writings (Escondo 2006). Writeprint features can include lexical features (e.g., richness of the vocabulary used or ratio of function words to stop words), syntactic features (e.g., passive versus present sentence structure or length of verb phrase), and structural features (e.g., indentation of paragraph), among many others.

Notably, most authorship detection studies use large collections of writing samples (e.g., novel length samples) as the basis of their analysis (Hirst 2007). The shorter the message,

the less certain the results have traditionally been (Hirst 2007, Burrows 2007). More importantly, the only existing, publicly available database of imitation attacks was developed by Brennan & Greenstadt in 2009 and involves essays of around 5000 words (Brennan & Greenstadt 2009). While the writeprint-based technique of Pearl & Steyvers (2012) had near perfect detection of imitation attacks for this dataset, it is unknown whether this technique would also succeed for shorter messages in mediums other than essay-writing.

d. Differences Across Length and Medium

While we know that humans are generally not good at deception detection, and computer programs perform quite accurately at imitation attack detection, it's currently unknown if each will perform similarly at detecting imitations attacks across different mediums and message lengths. Prior research by Kruger (2005) however, may lead one to assume that medium plays an important role in detection.

Kruger (2005) investigated human ability to detect tone in emails, voice-recordings, and talking face-to-face. We know that certain stylistic elements corresponding to particular mental states (e.g., persuasive, brusque) (Pearl & Steyvers 2013). These linguistic signatures of tone can be described as *mindprints* and humans are generally quite good at determining them (Pearl & Enverga 2015). Tone is something that people easily recognize and are often confident they are good at detecting in others. With respect to actual human judgments, Kruger (2005) found that while people overestimated their ability to detect and transmit sarcasm over these mediums (on average they estimated they would detect sarcasm 97% of the time), participants were nonetheless fairly good at detecting sarcasm (62.8% of the time in emails, 73.9% in voice recordings, and 73.3% talking face-to-face). Participants performed only about 13% better than chance in emails,

as opposed to about 23% better than chance in talking face-to-face. Kruger mostly described this increase in understanding across the mediums as an "increase in paralinguistic cues such as gesture, emphasis, and intonation" done by the talker in later manipulations. He does not seem to believe that tone detection in emails is significant or reliable. These findings indicate that even when humans are doing something they are particularly good at, they are still not perfect. Additionally, this research demonstrates that different mediums are associated with different distinctive features that might aid detection; for example, how gesture aids tone detection when talking face-to-face.

e. Motivation

To summarize, more research needs to be done investigating human ability to detect imitation attacks. However, because humans are not very good general deception detection, we could expect them to not perform well at imitation attack detection either. Moreover, it is unknown if human ability to detect these imitations will be affected by differences in medium and length. Computational analysis, on the other hand, is fairly accurate on longer samples of writing like essays. Yet, it is unknown whether the same type of analysis will find the same features to be of importance in messages shorter than 5000 words, or, whether the different communicative goals of messages matter. There are some studies investigating appropriate authorship techniques for SMS texts (Ragel 2013) and emails (Chen 2011); however, I have yet to find a study pertaining to imitation in these mediums. Presumably, messages of different communicative goals may have different features they employ. Moreover, one person's writing can contain stylistic differences across mediums, with a tweet by one person likely having a different style than a short email or longer essay composed by that same person. Given these stylistic

differences, it is likely that different features are more salient or easier/harder to manipulate in different mediums.

f. Format

This investigation sought to determine if human judgement accuracy scores vary across medium and/or message length, features that are easier/harder to imitate in shorter messages, and if said features vary across medium and/or message length. The mediums used include texts which were defined as 90 characters, tweets defined as 120 characters, and emails defined as 300 characters. In the remainder of the paper, I first discuss the creation of the unique database used in both human judgement collection (Section 3) and linguistic feature analysis (Section 4). Then, I review the investigation of the impact of message length/medium on human judgements. Results suggest that imitations are harder to detect in shorter messages (text) than longer messages (emails). Afterwards, I review the linguistic feature analysis used with an emphasis on features that were changed more verses features that were changed less. Features investigated include 51 syntactic, structural, and lexical features useful in detecting imitations in prior research. Because the linguistic features analyzed are thought to be unconscious, I assume that features that are changed more on average can be considered easier to consciously change to match a target author. On average, it appears that number of word type (e.g. number of adjectives used, number of verbs, number of nouns) is easier to consciously change to look more like that of the target author. Finally, I discuss how these general results could mean that spam text messages are more believable than spam emails, which has the potential for misuse in the near future. I call for the creation of spam folders for text messages in light of this new discovery.

2. Database Creation

a. Database Format

The database was created in two unique segments. In the first segment (Section 2b), subjects responded to very general prompts in order to create sample messages of different lengths. These samples served as the basis of imitations in the second segment of the database (Section 2c). In the second segment (Section 2c), subjects performed the same task of responding to general prompts, but they were then asked to imitate the previously generated messages. These samples and imitations generated in the second segment (Section 2c) were then used as the stimuli in both the human judgement investigation (Section 3) and linguistic feature analysis (Section 4).

These different mediums were chosen for their average length (90 characters for texts, 120 characters for tweets, and 300 characters for emails) and their different communicative/structural styles. Texts for example are more casual than emails and tend not to have greetings or signatures. An example of a prototypical text reads, “Hamilton was amazing last night! I can’t believe I finally saw it after listening to the soundtrack for a year, thank you so much for the tickets!” Tweets on the other hand, have a very different and distinct structure. They tend to begin with an “@someone” and end with a “#intent-of-the-tweet”. For example, “@LinManuelMiranda great show last night! I can’t believe I was finally able to see it! #iloveHamilton” This style is not static across twitter users, as some use twitter for more blog-like purposes. Therefore it will be interesting to see how people interpret this stylistic direction. If the majority of participants do not use the typical punctuation marks associated with twitter (at-symbols and hashtags), it will be

harder separate texts from tweets and thus make drawing conclusions about the impact of medium alone difficult.

b. Collection of Samples for Basis of Imitations (Target Authors)

Three people were asked to write about 15 random prompts using one of three different styles; these styles were text messages, twitter replies, and email conversations. In other words, for each trial, participants responded to a random prompt in the style of a random medium which yielded 135 datasets (3 participants x 3 mediums x 15 prompts). Prompts used were very general, for example, "Explain why you were late" (see Appendix A for full list).

c. Collection of Writing Samples and Imitation Attacks (Imitators)

The messages from the target authors (Section 2b) were then used as basis for generating imitation attacks. Participants in this segment (the Imitators) were split into three groups. Groups were asked to generate different numbers of samples and imitations in an effort to normalize the amount of time it would take for a participant to complete the experiment. Group 1 was asked to write 10 samples of emails (300 characters in length) and 1 imitation of a message from a target author (Section 2b). Group 2 was asked to write 50 samples of tweets (140 characters in length) and 10 imitations of messages from target authors (Section 2b). Group 3 was asked to write 50 samples of texts (90 characters in length) and 10 imitations of messages from target authors (Section 2b). The samples of each of the groups served as a reference set for the imitator's personal writing style.

d. Datapoints

The number of datapoints per group also differed because this segment was separated into groups that produced differing numbers of imitations. A datapoint is defined as a single feature value in a given dataset. Because each message has a collection of 51 feature values per message, this collection of feature values is termed a dataset. The number of datasets generated in a group depends on the number of participants and the number of messages a participant was asked to generate. Subsequently, the number of datapoints for a given group is obtained by multiplying the number of datasets in a group by the 51 features. In group 1 (emails) there were a total of 6 datasets per participant (5 sample messages and 1 imitation of target author). There were 13 participants in group 1 which resulted in 3,978 datapoints (6 messages, 13 participants, 51 features). Similarly, in groups 2 and 3 (tweets and texts), there were a total of 60 datasets per participant for each feature of a tweet and for each feature of a text (50 sample messages and 10 imitations of target author). There were 11 participants in group 2 (tweets) and 12 participants in group 3 (texts) which resulted in 33,660 data points for group 2 and 36,720 data points for group 3 (60 messages, 11/12 participants, 51 features).

e. Participants

All participants were recruited using Amazon Mechanical Turk (MTurk). Both parts of the investigating, creating the database (Section 2) and collecting human judgements (Section 3), were IRB exempt therefore signatures of consent were not collected. Instead, participants were told that continuing with the experiment served as giving consent to use their data generated.

The average age of participant was 36 years old, though, participant ages did range from 21 years to 66 years. The average amount of education was observed as "some high school" education.

Participants in segment 1 (Section 2b) were awarded $0.75 after the completion of the Human Intelligence Task (HIT) because it took around 25 minutes. Participants in segment 2 (Section 2c) who were asked to generate emails (Group 1) were awarded $1.55 after the completion of the HIT because it took around 60 minutes. Participants in segment 2 (Section 2c) who were asked to generate texts or tweets (Groups 2 and 3) were awarded $0.75 after the completion of the HIT because it took around 30 minutes.

f. Proper Participant Response

9 participant datasets of the 45 participants tested were not included due to participants not responding properly to the prompts given. With 39 participant datasets (3 target authors (Section 2b), 36 imitators (Section 2c)) and 51 features analyzed, there were a total of 74,358 data points analyzed to determine authorship.

Table 1: Length of Participant Response Across Mediums

	Mean	**Standard Deviation**
Emails (300 character)	345.65	141.85
Tweets (120 characters)	122.9	10.6
Texts (90 characters)	89.27	7.6

Table 1: Mean and standard deviation of number of characters in imitator sample messages across mediums. Email has a very different mean and standard deviation than tweets and texts because an adjustment was made to the code to add a minimum length requirement after reviewing participant responses to emails.

For the most part, participants followed directions in terms of character length for tweets and texts. As can be seen in Table 1, there is a much larger standard deviation in message length for emails. This is probably because in the groups following (tweet and text) a character minimum to continue to the next trial was added in an effort to collect more appropriate messages.

3. Experiment 1 - Human Judgement

a. Participants

10 MTurk workers participated and were awarded $1.00 after completion of the Human Intelligence Task (HIT).

b. Stimuli & Procedure

Statistics were calculated using Python Scipy Toolkit.

Participants were asked to read 4 target author messages from the database. Then they pressed a "next" button to see a 5th message and were instructed to judge if the additional message was composed by the same author or a different author. Of the target samples, there was a 75% chance that the participant would read target samples of the same medium and 25% chance that the participant would read mixed medium samples (see Table 2). In the same medium trials, the 5th message matched in medium to the target author samples. In the mixed medium trials, the medium of the target author samples and 5th additional message was random. Participants indicated their confidence on a slider labeled from "definitely not same author" to "definitely same author". The slider was scaled 0-100 and judgements less than or equal to 50 were labeled as "definitely not same author" and judgements greater than 50 were labeled as "definitely same author". Note

that there are other ways to interpret this confidence level data and future research could explore the easier/harder features correlated with high confidence ratings.

Table 2: Types of Trials

Name of Trial	Chance of Getting Trial	Example
Mixed Trial / Imitation Attack	25% chance for mixed, 50% chance for imitation = 12.5% chance for mixed imitation trial	1st message = text (target author) 2nd message = tweet (target author) 3rd message = email (target author) 4th message = text (target author) 5th message = tweet (imitation author)
Mixed Trial / Same Author	25% chance for mixed, 50% chance for same author = 12.5% chance for mixed same trial	1st message = text (target author) 2nd message = tweet (target author) 3rd message = email (target author) 4th message = text (target author) 5th message = tweet (target author)
Same Medium Trial / Imitation Attack	75% chance for same medium, 50% for imitation author = 37.5% chance for same medium imitation author	1st message = text (target author) 2nd message = text (target author) 3rd message = text (target author) 4th message = text (target author) 5th message = text (imitation author)
Same Medium Trial / Same Author	75% chance for same medium, 50% for same author = 37.5% chance for same medium same author	1st message = text (target author) 2nd message = text (target author) 3rd message = text (target author) 4th message = text (target author) 5th message = text (target author)

Table 2: Participants were asked to read 4 messages from a target reader and then an additional message that was varied by medium (same or different as target author samples) and author (target or imitator). Column 2 displays the chances of getting each of the 4 different trial types and column 3 displays an example of what said trial type would look like.

Table 3: Accuracy Definitions

	Same Author	**Different Author**
Judge as Same Author	True Positive	False Positive
Judge as Different Author	False Negative	True Negative

Table 3: The signal detection theory distinctions used in analysis of human judgements. According to this definition, true positive and true negatives would be considered correct

judgements of authorship, while false positives and false negatives would be considered incorrect judgements of authorship.

c. Results

Figure 1: Proportion of Accurate vs. Inaccurate Judgements of Imitation Attacks

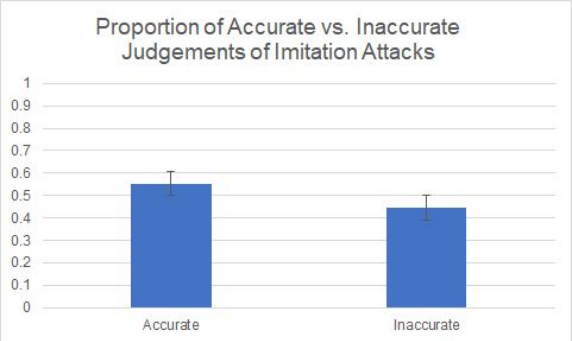


Figure 1: The Rand Index equation (sometimes referred to as "accuracy") was used to determine if humans were more or less accurate in general. Humans appear to be accurate at around chance (55% of the time), and inaccurate a little bit less of the time (45% of the time). A T-test found this difference between accurate and inaccurate scores to be significant ($t(15) = 2.48$, $p = 0.013$).

$$\text{Rand index} = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + True\ Negative + False\ Positive + False\ Positive)}$$

In Figure 1, a Rand Index (RI) equation was used to determine if human judgements tend to be correct or incorrect more of the time. For the accurate measure, the RI was calculated by adding the true positive and true negative scores across mediums and dividing this by the total sum of true positives, true negatives, false positive, and false negatives. Similarly, the inaccurate measure was calculated using the sum of false positive and false negative as the numerator, divided by the total sum of the judgements.

Both the accurate and inaccurate score were about chance, 55% of the time and 45% of the time respectively. A T-test found this difference to be significant ($t(15) = 2.48$, $p = 0.013$). Therefore participants were slightly more accurate than not.

Figure 2: Judgements from Same Author vs. Judgements from Different Author

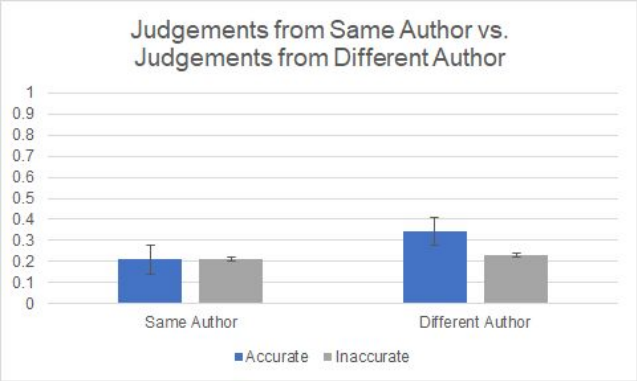
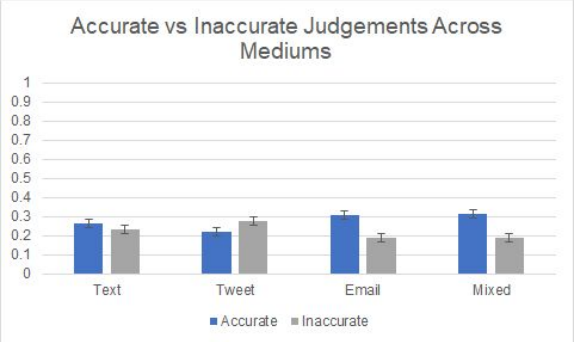


Figure 2: Average RI scores were investigated to see if participants are better at determining when something is composed by the same author or a different author. The results of multiple T-tests imply that there is no significant difference between judgements from same author vs judgements from different author ($t(7) = 1.281$, $p = 0.202$), nor a significant difference between the correct judgements and incorrect judgements for same author ($t(4) = 0.334$, $p = 0.739$) or different author ($t(4) = 1.666$, $p = 0.101$).

Then, the RI scores of conditions were the same author is presented was compared to the scores of conditions where a different author was presented to determine if participants were better at determining if something was composed by the same author versus determining if something was *not* composed by the same author. Even though in Figure 2 there visually looks to be a difference between judgements from the same author verses from different authors, a T-test failed to find this difference to be significant ($t(7) =$

1.281, p = 0.202). This means that people are not better at determining an additional message is from a target sample than not from a target sample.

Figure 3: Accurate vs Inaccurate Judgements Across Mediums



Across mediums, judgement scores for texts and tweets are less accurate. Both these mediums have lower accurate scores and higher inaccurate scores compared to the email medium. This appears to be more pronounced in the tweet medium. A T-test revealed that there is not a significant difference between accurate and inaccurate scores in the text medium ($t(3) = 1.454$, $p = 0.151$) or the tweet medium ($t(3) = -1.066$, $p = 0.291$), however there was a significant difference in accuracy scores in the email medium ($t(3) = 2.526$, $p = 0.014$) and mixed medium condition ($t(3) = 2.219$, $p = 0.030$).

As Figure 3 shows, on average people detected imitations in texts and tweets less well than in emails or mixed medium trials. In the text medium, participants made a correct judgement about 27% of the time, compared to the 23% of the time participants made an incorrect judgement. In the tweet medium this effect is even more pronounced; participants were incorrect more often (28% of the time) than they were correct (22% of the time). T-tests reveal that these differences in accuracy scores are not significant in texts ($t(3) = 1.454$, $p = 0.151$) or tweets ($t(3) = -1.066$, $p = 0.291$).

The opposite effect is seen however in the email and mixed medium trials where people are more accurate more of the time. In the email medium, participants were accurate 31% of the time, compared to the fact that they were inaccurate only 19% of the time. Similarly, in the mixed medium condition, people were accurate 31% of the time compared to inaccurate about 19% of the time. A T-test reveals this difference in accuracy scores to be significant in both the email condition ($t(3) = 2.526$, $p = 0.014$), and the mixed medium condition ($t(3) = 2.219$, $p = 0.030$). It is interesting to note that the mixed medium condition results look very similar to the email condition results, probably because at least one email was typically included in the mixed medium condition.

d. Discussion

In this investigation, participants were presented with 4 messages generated by one target author then 1 additional message. After, participants were asked to determine if the additional message was composed by a target author or an imitator. The target samples were either the same medium as the questionable additional message (same medium condition), or a different medium (mixed medium condition). Afterwards judgement accuracy scores were obtained where true positive and true negative judgements were defined as correct and false positive and false negative judgements were defined as incorrect judgements. We found that on average people were a bit more accurate than inaccurate (55% of the time compared to 45% of the time), though, when looking at accuracy scores across medium length, participants were more accurate in shorter text lengths (texts and tweets) than longer text lengths (emails) meaning that it is harder to detect an imitation in these shorter text lengths.

Judging by Figure 1, it appeared that on average participants were more accurate than inaccurate. This was found by using the RI equation (Section 3c). A T-test confirmed that this difference in accuracy scores was indeed significant. Upon further investigation however, it appears that medium greatly influenced this finding.

Before investigating the impact of medium, the impact the different types of judgements were separated in order to determine if it is easier to judge something as from the same author as a target set or judge something as *not* from the same author as a target set. Though in Figure 2 it looks like people are more accurate when determining if an additional message was composed by a different author, T-tests fail to find this difference to be significant. This is interesting because it is easy to image a scenario where the target set looks completely different than an imitation (for instance, if the target author used lengthy run on sentences and the imitation was short and to the point) making this type of decision fairly easy. Because there is no difference in accuracy when samples are from the same author or a different author, it implies that the features participants employ when imitating, and also look at in detection, are very nuanced. Further research should look into features that are more likely to be detected correctly or incorrectly.

Lastly, the impact of medium was investigated. In Figure 3, it is clear that shorter mediums such as texts and tweets appear to be incorrect more of the time than the email and mixed medium conditions. A T-test fails to find the difference between accuracy scores in the shorter mediums as significant. Even though it looks like participants were more incorrect than correct when looking at tweets, because the difference in accuracy scores is insignificant, it implies that participants were just guessing in the shorter mediums.

Not only are participants correct more of the time in the email condition, a T-test found this difference to be significant. Participants were likely not guessing at random when judging emails because there was more information to base their decisions off of. When there are more characters, there are presumably more features and said features can appear more often which is helpful in detection. Participants are likely looking for patterns in feature usage to help determine true authorship. However, humans probably could not tell you why they are more accurate when there is more information, assuming that they are not actually counting occurrences of features. In addition to investigating the which features aid in accurate detection, it would be interesting to investigate if people are aware that they are paying attention to these features.

It is interesting that in Figure 3 as well as the results of the T-test from Figure 3, the mixed medium results look more similar to the email condition than the short medium conditions. This effect was probably because the mixed medium condition was likely to have at least one email in it (because there were 4 messages that were each one of the 3 mediums). It must be significantly easier to detect imitations in longer messages if the effect of the longer message is seen even when surrounded by shorter messages, as in the mixed medium condition.

These results should be taken with a degree of caution however. Even though humans are decidedly better at determining imitations in longer mediums, they are only accurate about 9% more of the time than in the lowest scoring medium (in this case the tweet medium which was 22% of the time). Even though humans are better at detecting authorship in longer mediums, we are still not that great at imitation attack detection.

This result of humans being bad at deception detection was expected, as they generally aren’t good at detection in general. Confidence ratings that the additional message was

"definitely the same" or "definitely not the same" were recorded but not evaluated due to time. It would be interesting to see if participants in general had more confidence ratings towards the middle range or on the extremes, and if these confidence levels vary depending on which features are present. This type of information would be useful in both creating imitation attacks (know which features to put more effort into changing) and detecting imitation attacks (pay attention to the features that are harder to change).

Though it is clear that imitations are easier to detect in longer messages, humans are still not that great at detecting imitations and more research needs to be done on the features that affect this detection in order to know how to help humans to improve their ability.

4. Experiment 2 - Linguistic Feature Analysis

a. Procedure

Linguistic features were analyzed using the Natural Language Toolkit (NLTK) and Python.

Features used were a subset of the features in Pearl et al. 2016 (shown in Table 4) that were easy to extract using NLTK. The majority of the features were collected in terms of a ratio. For instance, the feature "number of nouns" was calculated by taking the number of nouns present in the sample and dividing it by the total number of words in the sample.

Table 4: Features Analyzed

Feature	Description	#	Implementation
alphabetic characters	all letters	1	$\frac{\text{\# of character tokens}}{\text{total \# of tokens}}$
digits	all digits 0-9	1	$\frac{\text{\# of digit tokens}}{\text{total \# of tokens}}$

punctuation	all punctuation marks	1	# of punctuation tokens / total # of tokens
average word length	average length of words	1	# of character tokens / # of word tokens
punctuation	period, at, comma, hyphen, question, exclamation, semicolon, apostrophe, colon	9	# of punctuation tokens / total # of punctuation tokens
multiple punctuation	multiple periods, exclamation, question	3	# token + space / # token
total characters	total number of characters	1	# character tokens
foreign words	foreign words (FW)	1	# of foreign words / # of word tokens
lexical diversity	measure of how many different word types are used in a text	1	total # of word types / # of word tokens
total words	# total words	1	total # of word tokens
content fraction	number of words contentful words (e.g. “dog”, “cat”) versus number of stop words (e.g. “the”, “a”)	1	total # of content words / total # of content + stop words
vocab size	measure of how many new words are used in a text	1	# of new words per message / # of words per message
word type	all adjectives (JJ), comparative adjectives (JJR), superlative adjectives (JJS) ,all adverbs (RB), adverbs comparative (RBR), adverb superlative (RBS), wh-adverbs (WRB), determining words (DT), predeterminer (PDT), Wh-determiner (WDT), interjections (UH), proper noun singular (NNP), proper noun plural (NNPS), possessive ending (POS), possessive pronoun (PRP$), possessive wh-pronoun (WP$), personal pronoun (PRP), possessive pronoun (PRP$), wh-pronoun (WP), possessive wh-pronoun (WP$), verb base form (VB), verb past tense form (VBD), verb gerund or present participle (VBG), verb past participle (VBN), verb non-3rd person singular present (VBP), verb 3rd person singular present (VBZ), coordinating conjunction (CC)	32	# of word type / total # of words

Table 4: The features analyzed were pulled from Pearl et al. 2016 and were easily extractable using NLTK. Column 1 lists the name of the feature group, column 2 describes the feature group purpose or if the feature group is composed of many features,

it lists the individual features (in the word type feature group, column 2 lists the different word types found and their part of speech tag for the NLTK word tagger), column 3 states the number of said features in said feature group, and column 4 explains how the feature value was calculated.

b. Rationale

Recall that a writeprint is a collection of features that comprise a person's writing style. For instance, if someone tends to use lots of semicolons or dashes in the middle of their sentences, this would be a feature of their writeprint. To get a sense of if an imitator's writing style has changed when imitating, we can investigate the feature values of the imitation. If the feature values look more like they would come from the target author's writeprint than the imitator's writeprint, then it can be considered a "good" imitation with respect to that feature. This analysis can be used to determine the features that are easier or harder to consciously imitate. Further research could potentially incorporate these results into their machine learning classifier to aid in accuracy.

c. Calculations

Before calculating the distributions of feature values that comprise a writeprint, because the number of natural writing samples collected from the imitator differed between the email group (5 samples) and the text and tweet groups (50 samples), I first normalized the amount of input data used to calculate the distributions. To do so, instead of using every datapoint from a sample to calculate the distribution of features, the average of every 10 data points in the text and tweet groups was used to calculate the mean and standard deviation for every feature distribution. Therefore, distributions for imitator's natural

writing style were calculated based on 5 data points across mediums and assume a normal distribution.

Equation 1

$$Pr(\textit{feature value} \mid \textit{target author sample}) \textit{ vs. } Pr(\textit{feature value} \mid \textit{imitating author sample})$$

To determine if an imitator mimicked a feature from a target author well, we determine the probability of said feature value within each author distribution (target author sample distribution and imitating author sample distribution). A higher probability in the target author's distribution than the imitation author's distribution indicates that the imitator was able to replicate that certain feature well. This is described as the likelihood of a feature value, given the sample. For example, suppose a target author's distribution of commas has a mean of 10 and a standard deviation of 1. Now, suppose an imitating author's distribution of commas has a mean of 5 and a standard deviation of 0.5. If the imitation contained 9 commas, the probability of it being from the target author's distribution is higher than the probability of it being from the imitating author's distribution, making it a "good" imitation with respect to comma usage.

Equation 2

$$\frac{Pr(\textit{feature value} \mid \textit{target author sample})}{Pr(\textit{feature value} \mid \textit{imitating author sample})}$$

In order to determine if the probability of the feature value coming from the target author's distribution is higher than the probability of the feature value coming from the imitating author's distribution, Pearl & Steyvers (2012) took the ratio of the probabilities. In this instance, a positive number would indicate that the probability of the numerator is

higher, which in turn means the feature value is more likely to come from the target author's writeprint. Relating back to the comma example, if we took the probability of 9 commas in target author's distribution (mean=10, std=1) divided by the probability of 9 commas in the imitating author's distribution (mean=5, std=0.5), it would yield a large number, because the numerator is larger, indicating that the 9 commas are more likely to have come from the target author's distribution.

Equation 3

$$shift\ score\ = log(\frac{Pr(\ feature\ value\ |\ target\ author\ sample)}{Pr(\ feature\ value\ |\ imitating\ author\ sample)})$$

I follow Pearl & Styvers (2012) and take the log of this probability, which I call a *shift score* (Equation 3). In this investigation, the shift score is used to represent the amount of change an imitator produced from their natural writing style to form an imitation. Because the features analyzed are thought to be unconscious, the amount of change in said feature from natural style to imitation style can be seen as the amount of effort the writer put into matching the style of the target author. Features that are changed more on often average are most likely easier to change.

A large positive shift score indicates that there was a lot of change from the imitator's natural style to form an imitation that looks more like a sample from target author. A negative shift score indicates the opposite, in that there was little shift from imitator's natural writing style in forming an imitation. A machine learning classifier would likely determine that an imitation with lots of negative shift scores for various features likely came from the true author, making it a "bad" imitation. In the comma example, if we took the log of that ratio of probability from the target author (mean=10, std=1) and

probability from imitator's sample (mean=5, std=0.5), we would find that the shift score is a high positive number (30 to be exact), meaning that the imitator shifted a lot from their baseline to produce a imitation that looks more like the target author's writing than their own.

We can compare these shift scores on average to tell us which features imitators generally use to imitate another person's writing style. For instance, if the shift score of "number of commas" was found to be high and positive on average and the imitation score for "number of exclamation marks" was found to be lower positive numbers on average, we can assume that people generally are better at imitating number of commas more than the number of exclamation marks when they are attempting to write like another person.

Comparison among imitation scores on average informs us which features are easier or harder to shift. Then, we can compare scores across mediums to determine if medium plays a role in a feature being easier/harder to shift.

c. Results

The log of the probability of a feature in target author's distribution divided by the probability of a feature in the imitating author's sample distribution was calculated and resulted in a positive or negative imitation score. These scores were separated by inspection: scores above +20 were labeled "super positive features", scores between 20 and 0 as "positive features", and scores below 0 as "weak features". Note that there were no scores near the -20 range that needed to be labeled.

After scoring each participant for each feature, the number of times the feature was found to be super positive, positive, or weak was found depending on medium to determine if certain features are used to imitate more often than others in certain mediums.

Table 5: Classification of Imitation Scores Across Mediums

	Texts	**Tweets**	**Emails**
Super Positive - Shift Score Above 20	# adverb # multiple question marks # question marks # apostrophes # cardinal numbers # possessives	# multiple question marks # question marks content fraction # adverb # multiple exclamation marks #multiple question marks # question marks	# multiple question marks average length of word
Positive - Shift Score 0-20	content fraction lexical diversity # adjectives # adverbs # apostrophes # cardinal numbers # commas # coordinating conjunctions # determiner words # exclamation marks # multiple exclamation marks # nouns # possessives # pronouns # question marks # verbs average length of word # multiple periods	content fraction lexical diversity # adjectives # adverbs # apostrophes # cardinal numbers # commas # coordinating conjunctions # determiner words # exclamation marks # multiple exclamation marks # nouns # possessives # pronouns # question marks # verbs average length of word # multiple periods vocab size number of words	average length of word content fraction lexical diversity # adjectives # adverbs # apostrophes # periods # cardinal numbers # commas # coordinating conjunctions # determiner words # exclamation marks # multiple exclamation marks # nouns # possessives # pronouns # question marks # verbs # multiple periods vocab size number of words
Weak - Shift Score Below 0	# comma	# comma	# cardinal numbers # commas

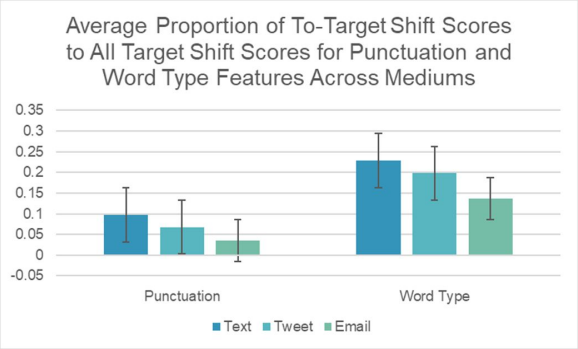
Table 5: List of features that were observed to have super positive shift scores, positive shift scores, and weak shift scores for at least one imitation. A feature could be listed as both a super positive feature and a weak feature because they were features of different imitations from different authors. Weak features to not appear to depend on medium however super positive features do change depending on length.

Weak features, defined as an imitation score less than 0, appear to be static across mediums. In these instances, the probability that the feature value came from the imitator's natural writing sample distribution was higher than the probability of the feature value coming from the target author's distribution. In other words, these features

were not changed much from their natural writing style. These weak features are presumably harder to imitate than features that were shifted more. 0 was selected as a "weak" classifier because there were no shift scores less than -20 -- in fact, there were not even scores less than -5. Because there were few scores below 0, it means that for the most part people changed their writing style for the imitation. The email medium was the only medium that had an additional weak feature other than number of commas, namely cardinal numbers; this is likely from participants writing out the name of numbers instead of the symbol (e.g. "one" vs. "1") when there is space to do so. Note that for other imitations, number of commas and number of cardinal numbers both had positive shift scores.

Super positive features however tend to vary across mediums. Number of multiple question marks and average length of word were the only features that were consciously changed a lot in emails, as opposed to a much longer list of punctuation and word type features that appear to be changed drastically in texts and tweets. This might be because when a space is small for a feature to appear, it is more noticeable and therefore more likely it will be imitated.

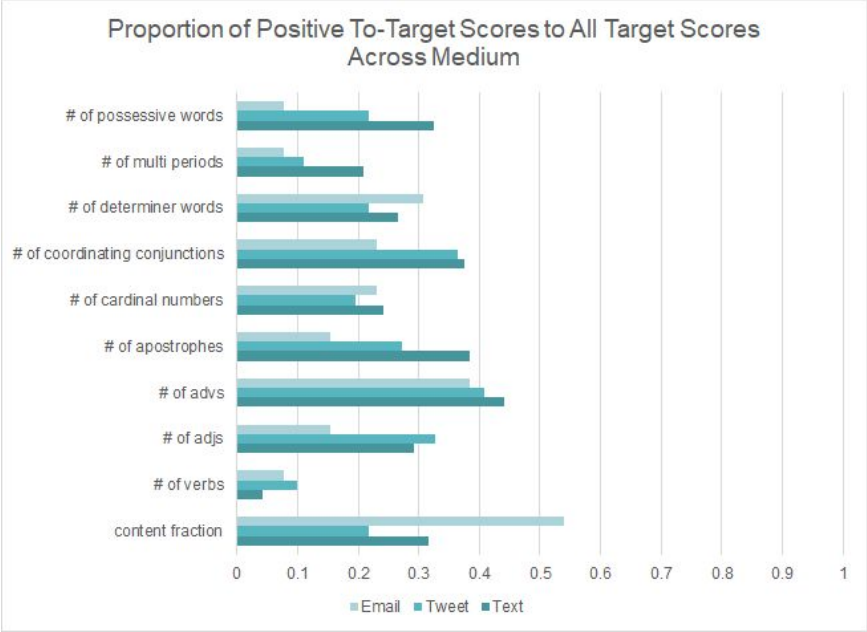
Figure 4: Average Proportion of To-Target Shift Scores to All Target Shift Scores for Punctuation and Word Type Features Across Mediums



Number of word type used appears to be easier to change than number of punctuation. This result however neglects a lot of features and it is more interesting to look at these results on a feature by feature basis.

As far as features that were easier/harder to shift, aside from the amount of shift, we looked at features that had positive shift scores across mediums. Figure 4 displays the two categories that had the highest shift scores on average demonstrates that on average and regardless of medium, number of word type (e.g. number of nouns, calculated by counting the nouns in a sample divided by the total number of words in a sample) appears to be a feature that is easier to imitate than other features, namely number of punctuation marks. It is also clear from Figure 4 that medium plays a role in feature shift scores, however we will investigate this further across all features in Figure 5. This averaging in Figure 4 misses a lot of important information. In order to determine which features exactly are easier or harder to imitate, we need to analyze the impact of medium/length on a feature by feature basis, as seen in Figure 5.

Figure 5: Proportion of Positive To-Target Scores to All Target Scores Across Mediums



Positive shift scores were obtained by calculating the feature value probability in the target author's distribution and the imitator's sample distribution and taking the log of the ratio of the two. Positive shift scores indicate that the feature value was moved significantly towards the distribution of the target author making it a better imitation. Across mediums the frequency of positive shift scores is seen to increase indicating that certain mediums utilize certain features more than others

Figure 5 demonstrates how shift scores are drastically different even within feature categories. For instance, number of verbs and number of possessive words both concern word type; Figure 4 showed that number of word type is easier to change, however Figure 5 shows that number of verbs is successfully shifted significantly less often than number of possessive words. Additionally, Figure 5 demonstrates that on average, as

message length decreases (from emails to tweets to text), the frequency of successful shifting increases. That is to say that shorter messages are easier to imitate. Number of possessive words clearly demonstrates this finding, in the email medium participants only shifted their number of possessive words to look more like the target author about 8% of the time, about 22% of the time in tweets, but about 32% of the time in texts.

d. Discussion

In this investigation, we were looking to see if medium impacts the features that humans employ in generating imitation attacks. We replicated linguistic analysis for a subset of Pearl, Lu, & Haghighi 2016's features and found that people can drastically change a subset of features from their personal writing style to imitate another depending on medium.

It is interesting to note that there were no super weak features, which would have had been shift scores around -20. There were not even shift scores below -5. This means that in general, people changed their writing style enough to make their imitations look more like the writing style of the target author. For this reason, I assume that participants tried their best to do the task at hand. However, it would be interesting to see if participants would respond differently if motivated in a more realistic way. Future researchers could make the platform for data collection look more like the medium participants are generating, or they could pay participants each time a machine learning classifier determines that their imitation is more likely to have come from the target author, making it a "good" imitation. These manipulations might result in different or more clear data that show that medium alone plays a role in detecting imitation attacks.

I begin with analysis of weak features, or features that were not shifted very much to match the target. In Table 5 it is clear that number of commas appears to be a harder-to-shift feature that people did not consciously change from their natural writing style across mediums. This result might lead one to think there is something unique about commas that make them inherently different from other punctuation marks. However, when you look at the data, there was on particular participant in segment 1 (Section 2b) of database generation that started out every entry, regardless of medium, with something along the lines of "Hi Sarah, …" It is possible that participants felt like this type of greeting was not medium appropriate and consciously decided not to imitate this aspect. Texts and tweets do not usually start with a greeting along these lines. In texts people will typically jump right into what they have to say, and typical tweets will usually start with an at-symbol (@) but no greeting. Because instructions were intentionally vague, in efforts to not influence the way in which participants imitated the samples, it is possible that participants thought that the task was to both imitate the language of the sample and change the sample to be more medium specific. One thing that future studies might want to account for is the quality of the samples used as the basis of imitations because this participant's dataset might have skewed our results.

Features that were drastically changed however seem to be dependent on medium. Table 5 displays very similar super positive features for texts & tweets and much less for emails. This is in contrast to the long list of features with positive shift scores for emails. It is possible that people have an easier time shifting their imitation to look more like the target author when the space to generate that imitation is smaller. When the space is smaller the features are likely more apparent and therefore you are more likely to reproduce this feature when making an imitation.

It is interesting to note that average length of word appears to be a unique super positive feature for emails. When imitating emails, people change their length of vocabulary to look more like that of the target author. One interpretation of this is that because of the formal nature of emails, people change their vocabulary to match. People writing emails tend to have "maximum verbosity engaged" (Livingstone) in an effort to sound like they know what they are talking about. Tweets and texts, on the other hand, are seen as less formal and therefore people use more short, colloquial words. Average length of word appears to be the only structurally unique feature that was observed as a super positive feature.

I did not find certain medium specific punctuation marks to be important indicators of style. I expected to find higher numbers of hashtags (#) and "at" symbols (@) in target author tweets and consequentially see imitators shift their usage to match a target author. However, none of the target authors employed these punctuation marks, meaning that this feature was not available for imitators to imitate. It is interesting to note that a few participants included these punctuation marks in their natural writing samples but did not include them in their imitations because they were following the target author's style. Furthermore, we also expected semicolons or colons to appear at least a few times in the effort of making an emoji face in either the texts or emails. Again, none of the target authors made any emoji faces and thus no imitations were made of emoji faces. Further studies might want to encourage participants to use more medium specific features such as hashtags or emojis as they are becoming increasingly common in online communication and carry information about the author's natural writing style (Hogenboom 2013).

Figure 4 shows that on average, features like word type are a lot easier to imitate than features like punctuation. It is possible that this is because it is common to use a lot of different word types and it is less common to use lots of punctuation marks; therefore, there is less room for normal change within these features. It would be normal to use a lot more adjectives and adverbs in an effort to be more descriptive, but people do not typically use a ton of exclamation marks, for example, to further their point. If a person used 10 exclamation points at the end of their sentence you would probably question their age or maturity level.

Taking the average of the features really loses a lot of information, however. Figure 5 informs us that even though number of verbs and number of possessive words are both word type features, it is clear that number of possessive words is changed much more often. It is important to do this kind of analysis on a feature by feature basis because features within the same category are not always changed the same amount.

Additionally, figure 5 provides more evidence that shorter message lengths are easier to imitate. As message length decreases from emails to tweets to texts, the frequency positive shift scores increase. Again, this is probably because features are more obvious in smaller spaces and therefore more likely to be imitated.

We are unable to draw conclusions about medium or length independently for a multitude of reasons. One reason was that we were unable to control for medium-specific features such as punctuation or structure. The only real thing we could control was participant response length, and they did respond within the requested message length. However, because the requested message length for texts and tweets were so similar (90 characters and 120 characters), and there were little structurally significant features unique to tweets

observed, we can really only conclude that message length and/or medium play a role in imitation attack creation. Future studies should make adjustments in their data collection stage so that they are able to conclude which one.

5. General Discussion

In this investigation, we were looking to determine whether or not human judgements of imitation attacks vary across mediums, features that are easier or harder to imitate, and if these features vary across mediums. Evidence from experiment 1 (Section 3) suggests that people are better at detecting imitations in longer messages (emails) than shorter messages (texts and tweets), likely because there is more space for a feature to appear multiple times. This is contrasted with the second main finding of experiment 2 (Section 4). While the evidence suggests that number of word type and number of punctuation appear to be the most easily shifted, it is important to note that medium plays a large role in the frequency of the feature being shifted; in other words, evidence suggests that shorter messages are easier to imitate. This is presumably because when a space is small for a feature to appear that feature is more noticeable and you are more likely to change it.

Recall that in *generating* imitations (Section 2) participants were only given one sample from the target author, as opposed to when *detecting* imitations (Section 3) participants were given four samples of a target author. When you are given multiple messages, it makes it easier to look for a pattern of some sort for feature usage (whether this process is conscious or unconscious is still unknown however), and in pattern matching the more information the better. In generating imitations however, it is not likely to see a pattern in just one message. Therefore, participants latch on to whatever feature they see and hope

that said feature is an aspect of the target author's writing style and not just an accident. When searching for just any feature that exists, the less information the better because said feature is more prominent than if it was hidden among other features. For these reasons, it is easier to imitate shorter messages but easier to detect imitations in longer messages.

This finding could have drastic implications in the wrong hands. If people are better at imitating shorter messages, one could imagine hackers and cybercriminals intentionally imitating your texts to scam your trusted friends and family members out of their hard-earned money. They are extremely susceptible to this kind of deception and wouldn't even realize that they aren't talking to you because the messages are so short. For example, how is my mom supposed to know that it is not really me asking for money when the message simply reads, "Need money for rent." Humans do not stand a chance at detecting this type of imitation with so little input information. My results suggest that a spam text folder (as we do for emails) might be a valuable application. If humans aren't very good at this kind of detection, we can rely on stylometric computational analysis to detect these deceitful cybercriminals.

5. References

Brennan, M. R., & Greenstadt, R. (2009, July). Practical attacks against authorship recognition techniques. In IAAI.

Chen X., Hao P., Chandramouli R., Subbalakshmi K.P. (2011) Authorship Similarity Detection from Email Messages. In: Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2011. Lecture Notes in Computer Science, vol 6871. Springer, Berlin, Heidelberg

Clifford, B. R. (2001, 09). Detecting lies and deceit: The psychology of lying and the implications for professional practice. *Applied Cognitive Psychology, 15*(5), 581-583. doi:10.1002/acp.743.abs

Fitzpatrick, Eileen, Bachenko, Joan & Fornaciari, Tommaso 2015. Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies* 8(3), 1–119.

Fornaciari, Tommaso & Poesio, Massimo 2014. Identifying fake Amazon reviews as learning from crowds. *In Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Hirst, G., & Feiguina, O. (2007, 09). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing, 22*(4), 405-417. doi:10.1093/llc/fqm023

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013, March). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp. 703-710). ACM.

Kendall, M. G., Mosteller, F., & Wallace, D. L. (1966, 03). Inference and Disputed Authorship: The Federalist. *Biometrics, 22*(1), 200. doi:10.2307/2528232

Klarreich. (2004, 01). Correction: Bookish Math. *Science News, 165*(3), 47. doi:10.2307/4014939

Kruger, J., Epley, N., Parker, J., & Ng, Z. (2005). Egocentrism over e-mail: Can we communicate as well as we think? *Journal of Personality and Social Psychology, 89*(6), 925-936. doi:10.1037/0022-3514.89.6.925

Levine, Timothy R 2014. Truth-Default Theory (TDT) A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology* 33(4), 378–92.

Li, J., Zheng, R., & Chen, H. (2006, 04). From fingerprint to writeprint. *Communications of the ACM, 49*(4), 76-82. doi:10.1145/1121949.1121951

Livingstone, S. (n.d.). Size doesn't matter: Why short words are better than long ones. Retrieved from https://www.articulatemarketing.com/blog/short-words

McCornack, Steven A & Parks, Malcolm R 1986. Deception detection and relationship development: The other side of trust. *Annals of the International Communication Association* 9(1), 377–89.

Newman, Matthew L, Pennebaker, James W, Berry, Diane S & Richards, Jane M 2003. Lying words: *Predicting deception from linguistic styles*. Personality and social psychology bulletin 29(5), 665–7.

Ott, Myle, Choi, Yejin, Cardie, Claire & Hancock, Jeffrey T 2011. Finding deceptive opinion spam by any stretch of the imagination. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 309–319. Association for Computational Linguistics.

Ott, Myle, Cardie, Claire & Hancock, Jeffrey T 2013. Negative deceptive opinion spam. In HLT-NAACL, pp. 497–501.

Pearl, L. & Enverga, I. 2015. Can you read my mindprint? Automatically identifying mental states from language text using deeper linguistic features. Interaction Studies, 15(3), 359-387

Pearl, L., & Steyvers, M. (2012, March 7). Detecting authorship deception: A supervised machine learning approach using author writeprints. Literary and Linguistic Computing. Retrieved from http://www.socsci.uci.edu/~lpearl/papers/PearlSteyvers2012_AuthorshipDeception.pdf

Pearl, L., & Steyvers, M. (2012). "C'mon – You Should Read This": Automatic Identification of Tone from Language Text. International Journal of Computational Linguistics, 3(1). Retrieved December 6, 2016, from http://www.socsci.uci.edu/~lpearl/papers/PearlSteyvers2012_MentalStatesLangText.pdf

Ragel, Herath, Senanayake, "Authorship detection of SMS messages using unigrams," *2013 IEEE 8th International Conference on Industrial and Information Systems*, Peradeniya, 2013, pp. 387-392. doi: 10.1109/ICIInfS.2013.6732015

Tweedie, F., Singh, S., & Holmes, D. (1996). Neural Network Applications in Stylometry: The "Federalist Papers" *Computers and the Humanities, 30*(1), 1-10. Retrieved from http://www.jstor.org/stable/30204514

Vrij, Aldert 2000. *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley.

Zuckerman, Miron, DePaulo, Bella M & Rosenthal, Robert 1981. Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology* 14, 1–59.

Appendix A

Prompts Used For Eliciting Short Messages of Different Mediums and Lengths:

1. Explain why you're going to be late for a meeting
2. Thank your (best) friend for a present
3. Congratulate a friend on an accomplishment
4. Explain to a friend why you have to cancel plans
5. Write a message thanking your parents for something
6. Suggest to go on a vacation somewhere to a significant other
7. Confront a friend about an annoying habit
8. Comment on the culture of the city you visited to a friend
9. Comment on your favorite season to a friend
10. Comment to a friend about a terrible day you're having
11. Ask your boss for a raise
12. Tell a friend what you want for your birthday this year
13. Say where you see yourself in 5 years
14. Tell your friend something about your favorite movie
15. Write something to your favorite author/actor