Detecting Deception in a Phishing Corpus using NLP Techniques

John Schriner
Computational Linguistics Program, Graduate Center, CUNY
jschriner@qcc.cuny.edu

Abstract

We hypothesize that by using Natural Language Processing techniques and tools, we can detect deception in a phishing corpus. By contrasting this *known-bad* corpus with a *known-good* corpus *EnronSent*, devoid of phishing emails, tools such as LIWC, AntConc, and Coh-Metrix will provide features common for deception. A second study explores a *Nigerian Letters* corpus for automatic detection of deception.

Introduction

The literature shows that the LIWC tool is popular among researchers working towards detecting deception in text. Franklin (2015) noted some theoretical assumptions with LIWC that ought to be considered before conducting research. Duran et al. (2010) examined the findings of Hancock (2007) and compared them to their own findings using Coh-Metrix, another tool for natural language processing, though with its own limitations[1]. Coupling Duran et al.'s (2010) methods with Franklin's (2015) insight into supplementing/auditing with a concordancer, we will be able to account for context and arrive at accurate findings.

The hypothesis is that NLP tools will accurately detect deception, especially deception via the persuasion principle of authority. By examining a phishing corpus, it is likely that we'll find evidence of authority in the texts: this can be shown in LIWC with the *Clout* feature (Kacewicz et al. 2013, Xu 2018).

Related Work

In a recent meta-study on social engineering and deception, authority, and with it obedience, was found to have the largest compliance score at 62.5% compliance in 23 studies, with other techniques from moderately to significantly high: conformity at 28.8%, reciprocity at 41.1%, and commitment at 45.2%. By designing original research in the area of information security around the findings of meta-analyses and replication studies, Bullée et al. (2018) were confident they could expect to find that some persuasion principles were more prevalent than others in social engineering attacks: their focus would be the social influences used by the offender.

After collecting transcripts of deceptive and truthful conversations, Hancock et al. (2007) employed LIWC to show linguistic features. As these were conversations, the "dynamics of deception" could be tracked in real time. LIWC was designed for cohesion and has 72 words characteristics. LIWC, developed in the 1990's and updated/audited regularly after, is word-counting software that's "blind to context" (Newman et al. 2003), and

[1] Except for a Portuguese port of Coh-Metrix, the tool is ostensibly End of Life as its only instance is web-based and only intermittently functional: http://tool.cohmetrix.com/

thoroughly critiqued in Franklin (2015). Indeed, Franklin (2015) notes the theoretical assumptions implicit in using LIWC: word frequency can tell us something about a person, words have meaning in isolation, inaccuracies due to context-specific factors are negligible, among others. The takeaway from Franklin (2015) is that the researcher using LIWC ought to prioritize context by using a concordancer (e.g. AntConc) alongside LIWC. A researcher ought to know the theoretical underpinnings, potential bias, miscategorization, and account for nuanced context of their texts.

The Corpora

For the purposes of testing LIWC and AntConc for deception-detecting, we use a *known-bad* phishing corpus of over 2000 emails first collected by Jose Nazario and now hosted on GitHub (Ocampo 2019). For contrast and a baseline, the *known-good* corpus is the EnronSent corpus that is devoid of phishing emails. It a subset of the Enron Corpus collected in 2002 and the EnronSent corpus "contains 96,107 messages from the 'Sent Mail' directories of all the users in the corpus"(Styler 2011). Contrasting the two corpora is similar to the methods described in Stone (2007) and Ocampo (2019).

Coh-Metrix, given that it's limited to fifteen thousand characters, is clearly not the correct tool to use for large corpora like either the phishing corpus or the EnronSent corpus due to volume. Coh-Metrix is used used for representational emails selected from the *CLAIR collection of fraud email* (Radev 2008). All corpora are sanitized and the email headers, HTML, CSS, MIME-type information, and attachments removed.

LIWC and AntConc aid in detecting phishing emails by detecting deception and lexical clues. Using best-practices described in Franklin (2015), LIWC and AntConc compliment each other by removing misleading findings.

The representational emails and the findings from Coh-Metrix on the CLAIR collection are contrasted with the findings on LIWC for deception features in phishing emails as well as social engineering in fraudulent emails. An examination of each corpora's unique context will add depth to the findings.

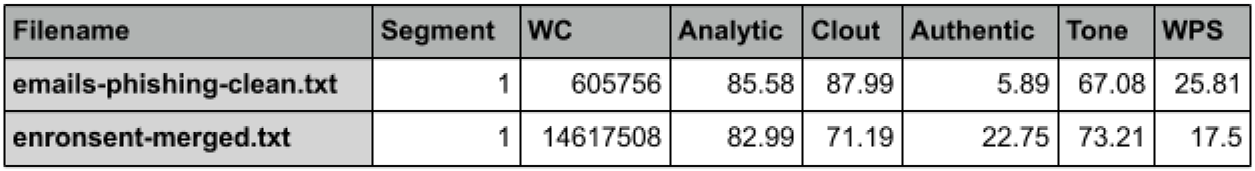
This research, as opposed to informal spoken dialog (Hancock et al. 2007, Duran et al. 2010), examines premeditated text that was written to deceive the reader. A feature such as *complexity* found using both LIWC and Coh-Metrix, has far less to do with demands on working memory than they did with spontaneous conversational text, but traits like redundancy and the LSA given-new value, we hypothesize, will be found well-represented in the fraudulent email corpus as the writer desires greater coherence: emphasizing the monetary reward for correspondence several times, the writer entices the reader to write back.

Findings

Although we would expect that clout/authority would be high in both corpora, there is significantly more clout in the phishing corpus: 87.99 clout compared to the 71.19 in the EnronSent corpus (Figure 1)[2]. Clout signifies an expertise and confidence. A corpus sourced from non-hierarchical text would undoubtedly have less clout. Indeed, the feature 'power,' which points to references of social hierarchy and dominance, is higher in the EnronSent corpus.

[2] Full table can be found at: https://github.com/johnschriner/NLP/blob/master/LIWC2015%20Results%20(phishing%20and%20enronsent).csv

Authenticity, expressed as a composite (i.e. *black box*) feature in LIWC based on prior research, shows that the EnronSent corpus is over four times more *authentic* or *personal*: the phishing corpus with its low authenticity points to "distanced discourse" (Pennebaker et al. 2007) which aligns with the style of emails sent by a corporation to its users, as opposed to colleagues in a workplace.



Filename	Segment	WC	Analytic	Clout	Authentic	Tone	WPS
emails-phishing-clean.txt	1	605756	85.58	87.99	5.89	67.08	25.81
enronsent-merged.txt	1	14617508	82.99	71.19	22.75	73.21	17.5

Fig. 1 - LIWC findings (excerpt).

The phishing corpus contains several types of phishing methods and sites. For phishing attempts in spoofing PayPal or bank account suspension notifications, the tone is formal. For eBay phishing emails, on the other hand, the tone is often personal and uses the word "I" because the phisher is appealing to emotion and an explanation as to why they haven't received payment. Even accounting for the eBay phishing messages, the use of "I" first person pronoun is still very low in the phishing corpus at .6% as compared to the Enron corpus at 2.39% of all text.

Another interesting and illuminating feature found is *risk*. Risk is found 1.08% in the phishing corpus as opposed to .39% in EnronSent. Risk is a feature for "references to dangers, concerns, things to avoid" (Pennebaker et al. 2007). This aligns with our impressions of warning and negative valence words like *compromise, penalty,* or *suspension*.

As LIWC is dictionary-based and focusing on common words, misspelled words, such as those found in the following examples, would not be accounted for, while they could possibly set off alarms for a human reader:

> This email is to inform that we had to block your PayPal account because this ip 64.12.117.14 **tryed** to access your account 3 times.We **apologise**[3] for the **inconvinience** but the safety of your account is our main priority.
>
> **You're** Billing Information
>
> If you choose to ignore our request, you leave us no **choise** but to **temporaly** suspend your account.
>
> [emphasis mine]

Using a python tool for readability called *textstat*, the phishing corpus was found to have a Flesch Reading Ease Score of 16.56 (i.e. *very confusing*), and the EnronSent

[3] This is the British-English spelling of *apologize*.

corpus has a -18.4, which is valid with the tool (worse than very confusing). This could be attributed to longer, more complex sentences containing more syllables per word in the EnronSent corpus. It could also be because of corporate-speak, spreadsheet data, or simply too much detritus from the original corpus.

As mentioned above, Franklin (2015) suggested best-practices of using LIWC alongside AntConc to provide context and to avoid misleading findings. AntConc confirms our findings in LIWC of the frequency of the second-person “you.” In LIWC, “you” was found at 3.67% of the text in the phishing corpus, and 1.79% in the EnronSent corpus even given the conversational nature of workplace email. The high prevalence in the phishing corpus could be attributed to the frequent “your account has been compromised” and it could be also used as a method of authority or instruction as in, for example, “you must click the link.” This is confirmed in AntConc, as “your” and “you” are among the top 5 most frequent words.

Study on the Nigerian Letter corpus using Coh-Metrix

This second part of the study complements the first study but without the real benefit of a *known-good* corpus. The NLP tool Coh-Metrix (Graesser, 2004) shares many features with LIWC. Coh-Metrix was originally designed to explore “cognitive constructs of cohesion in written text” (Duran 2010). As mentioned above, it’s not a practical tool for computational linguistics research given its limit of fifteen thousand characters. Duran et al. (2010) introduce their use of Coh-Metrix by re-examining the texts and findings of Hancock et al. (2008). Coh-Metrix, like LIWC, tracks word-level features but also “incorporates modules and algorithms that assess collocation of words” as well as advanced indices for referential overlap, syntactic complexity, easability, and others that number to over 700 indices (Duran et al. 2010). Deceptive linguistic behavior is particularly context-dependent so the researchers sought to first ascertain the conversational context and then decide on indices that are most relevant to the context: this would help to prevent overfitting. As opposed to their study on conversational text, we are using text that was written to deceive. We employ Coh-Metrix using some of the features identified by Duran et al. in conversational deception including quantity and redundancy.

The corpus for this study is the CLAIR collection of “Nigerian Letters” or “419” emails[4] collected by Radev (2008). This corpus contains 2500 emails of a very different style from the phishing or EnronSent corpora. We can compare some of these features in Nigerian Letters with those in the previous two corpora: we have to keep in mind the fact that the corpora content lengths are very unequal and the analytical and programmatic differences between LIWC and Coh-Metrix.

	Nigerian Letters	Phishing	EnronSent
WPS	**23.857**	**25.810**	**17.500**

Figure 2 - LIWC and Coh-Metrix shared features

When we look to word per sentence (Figure 2), the Nigerian Letters fall nearer to the amount of WPS in the phishing corpus.

[4] https://www.fbi.gov/scams-and-safety/common-fraud-schemes/nigerian-letter-or-419-fraud

Levitan et al. (2018) found that “more words per sentence were significant indicators of deception” in spoken interviews. Clearly these are very different modalities and this needs to be researched further. When we look next to redundancy we use the features *argument overlap* and *latent semantic analysis* (LSA) *given new values*. These indices are “broad indicators of between-sentence conceptual redundancy” (Duran et al. 2010). This is a feature we would suspect having high values in the Nigerian Letter samples. The goal is not, as in the phishing emails, to trick the reader into clicking, but rather to be engaged with the story and write an email back. It is expected that there is great repetition and concreteness in the narrative as well as repetition of the money value just waiting to be sent. In fact, in most Nigerian Letter emails, the money value is written in dollars and then explicitly spelled out in words:

> (US$30,000,000.00) Thirty Million United States Dollars Only

High values in both argument overlap and LSA given-new signify cohesion between sentences (Duran et al. 2010). We find high levels of concreteness (as opposed to abstract concepts), argument overlap, and LSA given-new in the Nigerian Letters corpus (Figure 3).

	Nigerian Letters
Concreteness	**384.29 (mean)**
Argument Overlap	**0.593 (mean)**
LSA given-new	**0.291 (mean)**

Figure 3 - Coh-Metrix indices

If Coh-Metrix supported larger samples or whole corpora, comparing results would be illuminating for these features, but confidence is low that such a small sample can yield meaningful results.

Further Research

The phishing corpus, although cleaned of headers, HTML, CSS, and remnants of MIME-type and attachments, still contained *word salad*, the method of adding nonsense text to emails to counteract naive Bayesian spam filters further described in Spense (2011). Two examples of word salad left in the corpus:

> Best to save it for later, then. chinquapin bricklaying That's why I went in the first place.He did not just pass this beneath her nose but pressed it briefly against her lower face. She hadn't asked them, but they had gone in there anyway. Never. It was horrible, but also sort of funny. Fair enough? This was shortly, after he had asked the traditional when-the-sleeper-wakes question and she had told him he was in the little town of Sidewinder, Colorado. Your phone has to ring at least once a day or Mountain Bell comes and takes it out? apollonian
>
> "Annie†ó "You know what they want? blackmail cherubim Instead of pulling back, they jerked a little and lay still.You were the tough young gunsel looking to make a rep off the tired old turd of a sheriff, right? Yes, here they were. ""Oh boy,ªshe said. Let that be the end of her. ""Oh? The prosecution wove its net as well as it

> could, but he handprint with the mark of the ring was really the most damning bit of evidence it could come up with. The panic was yammering more loudly now, asking what was he going to do, what was he going to do, for Christ's sake, this might be his last chance†ó What I'm going to do first is a thorough job of checking this situation out, he told himself grimly. agrimony

The phishing corpus could be cleaned of these to see whether it affects the findings in LIWC. We hypothesize that the words are unique enough that it will have little effect besides adding to the personal pronouns and signifying a more conversational or narrative tone.

Moving forward, the phishing types could be separated in the phishing corpus so that researchers could focus on, for example, the first person pronoun and emotion in eBay phishing attempts. eBay emails as well should be divided into formal corporate notifications on one hand, and personal phishing messages on the other. This would make the new corpora more valuable by satisfying unique phishing email types and the features they provide.

References

Bullée, J., Montoya, L., Pieters, W., Junger, M., & Hartel, P. (2018). On the anatomy of social engineering attacks—A literature‐based dissection of successful attacks. *Journal of Investigative Psychology and Offender Profiling*, 15(1), 20-45.

Duran, N. D., Crossley, S. A., Hall, C., McCarthy, P. M., & McNamara, D. S. (2009). Expanding a catalogue of deceptive linguistic features with NLP technologies. In *Twenty-Second International FLAIRS Conference*.

Duran, Nicholas D., Hall, Charles, McCarthy, Philip M., & McNamara, Danielle S. (2010). The Linguistic Correlates of Conversational Deception: Comparing Natural Language Processing Technologies. Applied Psycholinguistics, 31(3), 439-462.

Franklin, E. (2015). Some theoretical considerations in off-the-shelf text analysis software. In *Proceedings of the Student Research Workshop* (pp. 8-15).

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*(1), 1-23. doi:10.1080/01638530701739181

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun use reflects standings in social hierarchies. Journal of Language and Social Psychology, 33, 125-143.

Levitan, S. I., Maredia, A., & Hirschberg, J. (2018, June). Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1941-1950).

Ocampo, D. (2019). This project will determine which of the five supervised classification machine learning algorithms performs best in detecting phishy emails: diegoocampoh/MachineLearningPhishing. Retrieved from https://github.com/diegoocampoh/MachineLearningPhishing (Original work published 2017)

J. W. Pennebaker, R.J. Booth, and M. E. Francis. 2007. Linguistic inquiry and word count: LIWC2007 operators manual. University of Texas.

Radev, D. (2008), CLAIR collection of fraud email, *ACL Data and Code Repository*, ADCR2008T001, http://aclweb.org/aclwiki

Spence, C. (2011). I'll have the word salad, please. Retrieved from http://www.mxpolice.com/spam-trends/ill-have-the-word-salad-please/

Stone, A. (2007). Natural-Language Processing for Intrusion Detection. Computer, 40(12), 103-105.

Styler, Will (2011). The EnronSent Corpus. Technical Report 01-2011, University of Colorado at Boulder Institute of Cognitive Science, Boulder, CO., http://wstyler.ucsd.edu/enronsent.html

Xu, Weiai and Zhang, Congcong, "Sentiment, richness, authority, and relevance model of information sharing during social Crises—the case of #MH370 tweets" (2018). *Computers in Human Behavior*. 64.

n.b. Corpora and files pertinent to this project may be found here:

https://github.com/johnschriner/NLP