Detecting Deception and Masked Accountability in Enron Emails

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
| **DRAFT/MILESTONE** |
| Colby Carter, Maura Cullen, and Clay Miller  W266: Natural Language Processing with Deep Learning  UC Berkeley School of Information  colby.carter, mcullen, cmiller11 @berkeley.edu |
|  |
|  |
|  |
|  |
|  |

Abstract

In this paper, we attempt to bring together several semi-supervised learning methods to evaluate the potential for deception within the now-public Enron emails. Our motivation is the not-infrequent attempts by individuals and companies alike to mislead and cover up harmful information, whether those be insureds trying to make money off of fraudulent claims or executives misleading investors in order to line their pockets. We draw upon several studies on deceptive language and expand on the use of N-gram features and machine learning algorithms in order to draw clusters of similarly-suspicious emails to those identified by targeted heuristics.

Introduction

Individual acts of fraud and corporate corruption can hurt companies, employees and customers alike, while at times guiding investors toward placing faith in a dishonest company. The discovery of fraud within a successful company can even affect the entire economy as was the case with Enron Corporation in 2001.  Enron was an American energy, commodities and services company which had to file for bankruptcy in 2001 after willful accounting fraud and corruption was discovered within the company.

Monitoring and combing through emails or other written statements to detect fraud or unlawful business practices within companies is expensive and tedious to do by hand. Automating the fraud detection process is difficult because codewords, hints, and obfuscations in emails and verbal statements are used with the intent of not being discovered. We attempt to mimic the human process of reviewing emails to be able to flag those suspicious communications with modest precision. To learn from a real occurrence of fraudulent behaviour masked within company communications, we will be using the public emails gathered from the Enron case. We hope to be able to identify emails that appear to obscure potentially-damning information, which could be at least correlated with larger fraudulent activity. We would like this process to lead to more generalized fraud detection using written statements in other contexts, such as insurance and credit fraud.

Background

In our research, we have looked at a few other papers which have attempted similar goals of identifying suspicious or deceptive behavior in written language, emails or speech.

Bachenko et al. (2008)[[1]](#footnote-2) illustrated that many types of deception can be identified through verbal and non-verbal behavior. Using police transcripts, they identified linguistic indicators of many types of deception seen in criminal investigations. We consider use the idea of these linguistic groups to create features for our models. However, because the methods they used were focused on the context of criminal deception in police reports with verifiable details, we will have to get more creative when considering the broader topic of suspicious content in emails.

In an effort to bridge the gap between criminal statements and emails we will also borrow ideas from Cohen et al.[[2]](#footnote-3) for classifying “speech acts” in emails. The idea is generally to separate statements in emails to focus on words that imply the sender is or will be taking action on these statements. They suggest using the noun and verb etymology. As you will see, we will use TFIDF-weighted bag-of-words as a baseline for processing context-free text to be used in the model.

We have also researched other prior studies such as Feng et al. (2012)[[3]](#footnote-4) to understand how to process context-free grammar and extract meaningful features from statements. These features, in addition to uni- and bi-grams, can include context-free grammar trees, part-of-speech tags, and lexico-syntactic patterns within the emails. With added features, we will hopefully allow our model to pick up on context within emails that do not inherently have any context and could very possibly have slang, misspellings and improper grammar.

Methods

Our approach to identifying emails with suspicious phrases begins with parallel efforts toward developing semi-supervised models: (1) with learnings from prior research on deceptive language, we explore the various types of emails in the Enron dataset (e.g. financial, personal, news reporting, hiring, etc.) and conduct targeted heuristics to identify and positively label emails with suspicious or information-masking language; and (2) we develop a simple, baseline approach for clustering emails using uni- and bi-gram features in a K-means algorithm, then evaluate cluster quality and the types of emails with the strongest similarity to our labeled data.

Our exploration began with a review of several hundred randomly drawn email bodies from the first 50,000 examples in the dataset (which totals over half of a million, including a substantial number of redundancies with email threads and forwarded messages). Topics can include anything from legitimate attempts at buying and selling energy assets and other innocuous business dealings, to personal exchanges about birthday parties and fantasy sports leagues, all with varying lengths and content type, such as email addresses, links and image file names. Identifying attempts to obfuscate information requires a more targeted and nuanced approach, particularly when those attempting to hide information are not doing so obviously. Targeted phrases can include a range of subtlety, from the obscure “*no recollection of*”/“*I don’t recall*” to more damning word choice such as “*manipulate*” in conjunction with “*price*.” Similar leads were derived from Enron-specific keywords such as “*condor*” and “*raptor*.”

In parallel to the heuristic labeling approach, we utilized a simple clustering approach using uni- and bi-gram features, still on the first 50,000 examples in the dataset. Feature generation was performed by using NLTK’s sentence tokenizer to split emails into sentences of individual tokens, then preprocessing the messages for start and stop characters, punctuation, numeric digits, and limiting our vocabulary to 10,000 distinct tokens while replacing uncommon tokens with a string for “unknown.” The N-gram counts were then vectorized into features, keeping only those N-grams that were present in at least five email bodies and no more frequently than 20% of the emails, then transformed with TF-IDF to properly weigh both the more and less common words within documents. The resulting feature set is just over 75,000 weighted uni- and bi-grams counts, a dimensionality likely too large for successful clustering but sufficient for a simple baseline.

With these simple features and labeled examples, we trained a K-means model with five clusters in order to test how well the labeled examples tend to be grouped together. While the suspicious emails make up a very small fraction of the overall dataset, clusters that purely contain our labeled examples can give us optimism that the set of potentially suspicious emails can be narrowed down significantly to just those most similar to the labeled examples. We evaluate cluster quality by measuring the cosine similarities of our labeled examples and in-cluster emails (a method for normalizing by document length, for which the emails vary greatly), and examine the closest emails for suspicious content.

Results and discussion

Our initial results from the simple clustering on N-grams suggest both promising potential improvements as well as remaining challenges. With a subset of about ten percent of the overall data and a simple K-means producing five clusters, we see our labeled examples consistently captured in the same cluster, but with cluster sizes being largely unbalanced and still significant variation within clusters. We suspect this is due to the rough bag-of-words approach that does not capture the variety of contexts in and across emails. Looking within the cluster in which our labeled examples fall, we then evaluated the cosine similarities of nearby vector examples: while there are occasionally emails that show phrases that border on obscurity, they do not tend to meet the threshold of “suspicious”--albeit a subjective one. Correspondingly, we see cosine similarities on the order of 0.16, which are very low and likely due to the sparsity of our N-gram feature set; we also experience slow model training due to the large dimensionality. We hope that the following improvements will reduce our feature set and greatly improve the quality of our clustering.

Next Steps

Because this simple model suffers from a dimensionality problem with little context being captured, we intend to next test a couple of refined approaches to topic clustering, while also expanding our heuristic approach of searching for suspicious phrases. Motivated by the deep learning methods of word-vectorizing (e.g. word2vec), we have encountered similar uses at the paragraph level on the Enron emails with doc2vec. Like word2vec, it uses a similar continuous bag-of-words model to estimate a vector representation of an entire document, regardless of the document length. With the reduced-dimension representation of the emails, we can train K-means clustering with more examples and clusters to improve both the volume of training data as well as the purity of our clusters, for evaluation similar to our base model’s using N-grams only.

Along a similar vein, we intend to implement and evaluate the performance a different type of model, latent Dirichlet allocation (LDA), which has been frequently cited and performed well on similar tasks. Specifically, the model attempts to identify multiple unobserved topics within each document. Given our set of labeled examples, we can evaluate other emails that share predicted topics and estimate what percentage show potentially suspicious content.

Additionally, most research up to this point has largely focused on word usage and there has not been much research on sentence structure. Given time, we would like to explore this feature in our data set to determine if there is any difference in the way deceptive phrases are structured.

1. Bachenko, Joan, Eileen Fitzpatrick, and Michael Schonwetter, “Verification and Implementation of Language-Based Deception: Indicators in Civil and Criminal Narratives,” *ACL*, 2008, <http://www.aclweb.org/anthology/C08-1006> [↑](#footnote-ref-2)
2. Cohen, William W., Vitor R. Carvalho, and Tom M. Mitchell, "Learning to Classify Email into ‘Speech Acts’," *EMNLP*, 2004, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.329.4178&rep=rep1&type=pdf> [↑](#footnote-ref-3)
3. Feng, Song, Ritwik Banerjee, Yejin Choi, “Syntactic Stylometry for Deception Detection,” *ACL*, 2012, <http://www.aclweb.org/anthology/P12-2034> [↑](#footnote-ref-4)