The Semantics of Persuasion: A Case Study Using Phishing Emails

by

Jacob J. van der Laan

Bachelor of Business Administration, UNB, 1988
Bachelor of Laws, UNB, 1991
Bachelor of Science in Computer Science, UNB, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Computer Science

In the Graduate Academic Unit of Computer Science

Supervisor(s):	Arash Habibi Lashkari, PhD, Faculty of Computer Science
	Christopher J.O. Baker, PhD, Department of Computer Science
Examining Board:	Saqib Hakak, PhD, Faculty of Computer Science, Chair
	Rongxing Lu, PhD, Faculty of Computer Science
External Examiner:	Mohsen Mohammadi, PhD, Faculty of Engineering

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

April, 2021

Abstract

As of 2021, phishing emails continue to be the primary means by which network breaches are facilitated. Notwithstanding the development of many tools to detect and block incoming phishing emails, many users continue to be plagued by them on a daily basis. In addition, the nature of phishing emails is changing as the incidence of more personalized forms, such as spear phishing and whaling, prove their effectiveness. These newer forms of phishing are harder to detect using traditional methods and emphasize the need for approaches which seek to enable detection based on persuasion based language features unique to phishing emails. To that end, this thesis draws insights from the phishing process, the applicable behavioural psychology research on persuasion, as well as linguistics, to inform an understanding of how phishing emails persuade. It then proposes a methodology for feature engineering of persuasion language related features for the phishing email domain, based on these insights. A proof of concept model is developed using persuasion based language features, and then implemented and tested using several machine learning algorithms. The performance of this model is as good, if not slightly better, than other more complex and labour intensive efforts which sought to capture semantic meaning using fewer detection features. The thesis concludes with a discussion of potential future work.

Dedication

I dedicate this work to Eileen M. Quinn, my partner in life, for her unwavering support and encouragement.

Acknowledgements

I wish to thank both of my supervisors, Dr. Christoper J.O. Baker and Dr. Arash Habibi Lashkari. Their persistent support and encouragement in the course of my research, as well as their invaluable guidance during my sometimes meandering walk through various phishing related disciplines and approaches to the problem, enabled me to finally arrive at the completion of this thesis. Their assistance in reviewing various drafts of this thesis is also very much appreciated.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

APWG	Anti-Phishing Working Group
API	Application Program Interface
BEC	Business Email Compromise
BSFS	Binary Search Feature Selection
CEO	Chief Executive Officer
CNN	Convolutional Neural Networks
CSV	Comma Separated Value
CxG	Constructions Grammar
DNN	Deep Neural Network
EML	Electronic MaiL - an email storage format
FBI	Federal Bureau of Investigation
FOMO	Fear Of Missing Out
GB	Gigabyte
IML	Informed Machine Learning
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MacOS	Apple's MacIntosh Operating System
MBOX	A generic email collection storage format
MSG	A Microsoft email storage format
MIME	Multipurpose Internet Mail Extensions
ML	Machine Learning
MWE	Multi-Word Expression
NLP	Natural Language Processing
NP	Not Phishing
OOV	Out Of Vocabulary
OST	Ontological Semantics Technology
P	Phishing
PCA	Principal Components Analysis
PCP	Phishing Campaign Perpetrator
PDTB	Penn Discourse Treebank
PMS	Persuasion Motivation Sequence
POS	Part of Speech
QR	Quick Response (code)

RAM	Random Access Memory
RF	Random Forest
RFC	Request For Comment
SME	Small and Medium Enterprises
SSD	Solid State Drive
SVD	Singular Value Decomposition
TF-IDF	Term Frequency - Inverse Document Frequency
TMR	Text Meaning Representation
URL	Uniform Resource Locator
VSM	Vector Space Model
XML	eXtensible Markup Language
ZIP	An archive file format

Chapter 1

Introduction

"You shall know a word by the company it keeps."

J.R. Firth, 1957

1.1 A Persistent Problem

Phishing is the deceptive practice of sending electronic communications, such as email, purportedly from a reputable person or organization, in an effort to induce the reader to click a link, open an attachment or perform some other action. The goal of phishing is usually to obtain account credentials or facilitate the installation of malware on the target's computer.

Phishing has been a nagging and persistent problem for decades to the point that today, the majority of cybersecurity breaches arise as a direct result of the successful phishing of a computer user [Blo19] [Ver19]. In both 2019 and 2020, the Federal Bureau of Investigation (FBI) also identified phishing as the number one cyber crime by victim count [FBoIF19] [FBoIF20]. This trend does not appear likely to abate, particularly with the COVID-19 pandemic and its compounding effects of increased

use and reliance on digital communications and remote work. The Fourth Quarter Report from the Anti-Phishing Working Group (APWG) noted a marked increase in phishing in 2020, particularly in the areas of Business Email Compromise (BEC) and spear phishing (discussed further below) [Gro20]. Phishing emails are also a critical component of ransomware attacks, which continue to plague many organizations, in particular small and medium businesses (SME's) in recent years.[1]

Phishing attacks are prevalent because they work, are cheap to execute, are repeatable, and can be targeted at many users at once.

Phishing is effective because it exploits the human condition. Absent any obvious signs of deception, humans are generally inclined to trust written communications such as email and thus predisposed to performing requested actions contained within them [SK19]. As noted in a recent paper[SARG20]:

> Phishing is one of the most successful forms of deception and persuasion in the cyber world, because it takes advantage of social engineering and psychological techniques that exploit human weaknesses. These human weaknesses include our almost inevitable tendency to rely on our own memory and experience to make decisions, our limited and often biased attention towards items that are "attention catching", and our tendency to believe that things that look similar have similar effects. These cognitive human factors result in human cognitive biases, which unfortunately, attackers seem to master quite well.

Phishing is now also increasingly present on social media networks [FF20], messaging apps used on mobile phones (sometimes referred to as "smishing") [CJ17], and even in QR codes (an attack referred to as "QRishing") [VOW+13].

The *sophistication* of phishing attacks is evolving in two primary ways: the use of

[1]For more insight into these developments, consult the author's cybersecurity resource page at `https://sites.google.com/view/cyber-resources/home` which contains a list of recent reports on cybersecurity trends.

behavioural science informed social engineering strategies, and more personalized versions of phishing.

1.1.1 Social Engineering

Much has been learned in the last decade about how humans make decisions and how those decisions can be influenced with a variety of social engineering strategies [KHHW15]. Phishing campaign perpetrators (referred to as "PCP's" in this thesis) are learning from this science and applying the lessons learned to improve the effectiveness of their phishing campaigns.

1.1.2 Personalized Phishing

PCP's have also observed there is great value in more tailored and *targeted* phishing attacks. Personalizing the attack leads to higher rates of success and the potential for a greater financial bounty. The approach is generally to conduct online background research on the target and then closely tailor the messaging to the victim's profile. This targeted strategy is generally known as *spear phishing* and is increasingly prevalent today[ACPZ19]. One particularly use of spear phishing is in the pursuit of Business Email Compromise (BEC) schemes, a growing area of online mediated fraud involving the manipulation of corporate employees to approve, process or generate fake invoices. BEC can produce significant financial losses to businesses [Aga20]. These newer strategies are increasingly employed where the target is of high value, such as accounting and financial employees in an organization, given that these types of individuals have access to more valuable information, or in the case of BEC, direct spending authority. When a spear phishing attack involves a CEO or other high level executive this type of phishing is often referred to as "whaling". An example spear

phishing email used to seek to execute an invoice based scam is set out in Figure 1.1[2]

These more personalized versions of phishing make automated detection much more difficult and as a result they usually do end up in the inbox of the targeted user [Wor08].

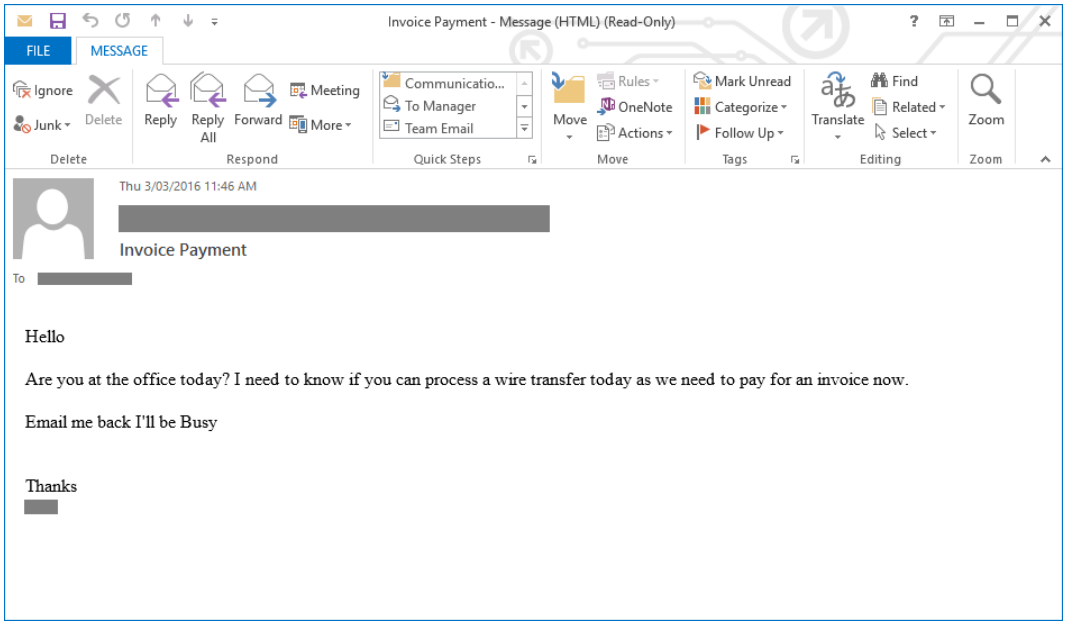

Figure 1.1: An Example Spear Phishing Email

1.1.3 Motivation

Even with various anti-phishing systems in place at many organizations and email service providers, the average email user must still deal with approximately sixteen phishing emails a month [C+18]. Personalized phishing strategies such as spear phishing and whaling make it even harder for humans to detect these attacks.

[2]Retrieved from `https://fraudwatchinternational.com/phishing/spear-phishing-targeting-organisations/`. Also see the Berkeley University Information Security Office website for a very good list of phishing email examples: `https://security.berkeley.edu/education-awareness/phishing/phishing-examples-archive`.

Exploring new strategies, particularly those which can help identify these increasingly sophisticated phishing attacks is important to try and whittle down this number.

1.2 Exploring Phishing Email Semantics

This thesis explores an aspect of phishing which plays an important role in the increasing sophistication of email based phishing attacks: the *language* of the messaging used, and more specifically, the *persuasive* aspects reflected in that language. To that end this thesis examines the phishing email problem from several different perspectives, all of which provide meaningful insights into how to build a detailed - domain specific - persuasion language based phishing detection approach. These perspectives are:

1. The operational aspects of the phishing event, and its constituent entities, relationships, and steps.

2. The human behavioural psychology dimensions of the persuasive purpose of phishing emails.

3. The linguistic make up of the phishing message itself and how persuasion is reflected in written language.

4. The prior research done with respect to using the text of an email to identify phishing, as well as the work pursued to extract forms of meaning from phishing emails.

5. The approaches to incorporating domain specific knowledge into the feature engineering process.

With the insights gained, an examination of various phishing email corpora is undertaken and a methodology derived for formulating persuasion specific language

features. This is followed by a review of the implementation of a 13 feature based proof of concept model and the results from testing the model using various machine learning algorithms. The thesis concludes with a discussion of potential future work.

The primary focus of this work is to build on the research efforts in [Fal16] and [Par18], two PhD dissertations which sought to investigate the effectiveness of conceptualizing lexical features, that is, the *meaning* of the language in a phishing email, with a view to designing targeted solutions.

1.3 Contributions

The contributions of this thesis are as follows:

1. A detailed summary of insights gained into the operational, behavioural, and linguistic structures underlying the persuasive purpose of phishing emails.

2. A comprehensive literature review of the research to date on modeling phishing emails by way of language based features and the applicable feature engineering principles.

3. A domain knowledge driven approach to deriving language based features which seek to capture the purpose of phishing emails, i.e. to persuade the reader to act in some manner.

4. A machine learning based proof of concept implementation of the model, showing classification results as good, if not better than, prior efforts relying only on more limited language features in phishing emails.

1.4 Summary

This chapter introduced email phishing as an evolving and pressing problem requiring ongoing study and the exploration of additional means of detection. The persuasion related semantics of the phishing email were identified as an important aspect of this domain which to date has not been deeply explored. A thorough understanding of the phishing "event", behavioural psychology and linguistics provides insights and understanding of how persuasion in email phishing works, and offers a pathway for better solutions to their detection.

Chapter 2

Perspectives on Phishing

2.1 The "Phishing Event"

At its operational level, email phishing is probably best understood as the interplay between several entities in the course of a "phishing event":

1. The **sender** of the email - who seeks to persuade the reader to take a certain action such as clicking a link or opening an attached document,

2. The **email message** itself - the written communication which embodies the appeal to the reader to act, and

3. The **reader** of the email - who brings a series of attributes to the event, which are usually appealed to and/or sought to be exploited in some way by the sender.

These entities and their interplay are set out in Figure 2.1.

All of these entities, actions and attributes have been the subject of academic study and there exists a good amount of research into various aspects of phishing such as:

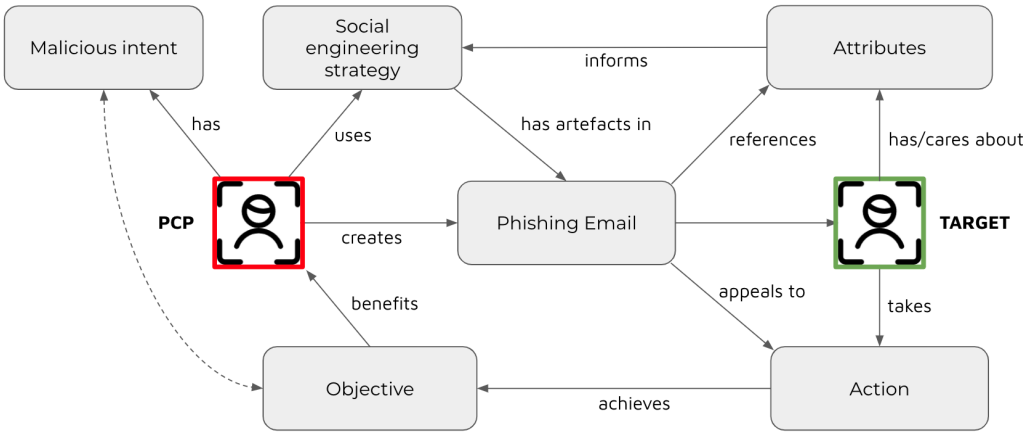

Figure 2.1: The Phishing Event

- The characteristics which influence the reader's susceptibility to being phished such as the work role of the reader [PHG04], his or her computer related activity [DF18], network access patterns [PCA16], browsing behaviour [CBB14], age [LCE+19] [PF20], level of perceived technical sophistication [VHC+11], personality [UQ14] [PJBC09] [MZP+17] [WHZ+15] [WLR16] [HMN15], and media use [VHC+11].

- The extent to which awareness training influences the reader's susceptibility and what makes for an effective awareness training strategy [HK17] [BSN19] [KCA+09] [CBD+19] [Aba14].

- The extent to which the context (or the "situation") within which the email is read and responded to affects phishing susceptibility. There is good evidence to suggest that such things as time pressure [JTRH19] [YG12], organizational culture [MA97], concurrent personal factors which match references in the email (such as for instance attending a college) [GWD17], health concerns [SW19], and emotional state [LHY20] all play a part. There is also research suggesting that phishing susceptibility is higher if the person has previously fallen victim to a phishing attack [CGR20].

This research illustrates the fact that the phishing event is a psychological battle between the persuasive efforts of the sender and the susceptibility of the reader. In general terms, where the persuasive efforts of the sender align with the susceptibility to persuasion inherent in the reader, the chance of a successful phish increases.

2.2 Continuous Innovation by Perpetrators

There is good evidence that some PCP's actively monitor efforts to defeat phishing attacks. For example, they often obtain keyword lists used in the matching systems in an effort to circumvent defence mechanisms that analyze the words of an email [Par18].

They also continually explore new ways to improve the efficacy of their efforts, by employing more sophisticated strategies such as spear phishing and whaling. The now ready availability of usable information in various social media problem enables these types of attacks[Cal13]. As a result, spear phishing attacks have continued to increase every year and some industry professionals now suggest that as of 2020, 95 percent of all attacks targeting enterprise networks are caused by some form of successful spear phishing.[1] Most notably, spear phishing emails were used in the compromise of the Democratic National Committee and presidential candidate Hillary Clinton's campaign leading up to the 2016 United States Election.

Spear phishing emails are usually crafted more carefully and are more likely to contain certain details of the target's environment (such as the name of a department or a supervisor) in order to create a sense of legitimacy in the reader.

[ACPZ19] identified the increasing importance of the *persuasive* elements in the text

[1] `https://www.kratikal.com/blog/staggering-phishing-statistics-in-2020/`

of spear phishing emails, which, as noted, are much more finely tuned to the circumstances of the targeted recipient.

It is worth noting that the study of spear phishing emails is evolving. One significant barrier is that due to privacy concerns, spear phishing detection approaches have not been conducted with adequate spear phishing samples. Collecting such data still remains a challenging task [Par18]. Hopefully methodologies can be developed to assemble anonymized spear phishing data sets which protect their victim's identities.

2.3 The Behavioural Dimensions of Phishing

The primary purpose of a phishing email is to *persuade*. It is thus no surprise that a reader's psychology plays a significant part in how that person responds to a phishing attempt [Jak07]. *Human behaviour* continues to be the key determinant of whether a particular phishing attempt succeeds or not [BDH+11].

Understanding how human persuasion works is thus useful. There is an abundance of research into persuasion strategies generally, as well as more recent research on its use in phishing emails specifically.

2.3.1 Persuasion Strategies

Persuasion strategies are generally classified into four broad categories [PBDL19]:

- **Commitment and Consistency** - The concept of completing an action which was previously initiated.
- **Liking** - Leveraging trust due to prior interaction or familiarity, such as for a known and recognizable person, group or brand.

- **Authority** - An authority figure mandating an action, usually with explicit or implicit consequences for failing to comply.
- **Scarcity** - A short and specific time frame to complete an action, often with a sense of urgency. Appeal to the "Fear Of Missing Out" (FOMO).

Other researchers add three additional strategies [CG02] [Cia16]:

- **Reciprocation** - A benefit or favour extended to the reader used to seek reciprocal action. "I scratched your back, you scratch mine".
- **Social Validation or Social Proof** - Others are doing this, you should too.
- **Unity** - An appeal for action on the basis of the benefits of unified action (e.g. "We are in this together").

[PT17] approaches the understanding of persuasion slightly differently, by identifying five distinct models or modalities of persuasion:

- **Appeal to authority** - Explicitly referring to an authority to convey and strengthen a message.
- **Compare and contrast** - Usually between two arguments, with the implication one is better than the other.
- **Problem** - Outlining a problem and its solution.
- **Hypothesis** - A rational outline of an argument with reference to evidence.
- **Association** - Relate something with favourably perceived things (logos, images, etc.).

With respect to task-related communications, i.e. attempts to compel a person to act in some way, the perceived *power* of a communicator positively affects the success

that the communicator may have [BL76] [FCK+03]. Power is usually demonstrated by way of a threat or suggestion of a looming consequence, some indication that the target's situation will be worse if the communicated task is not performed. Perhaps somewhat surprisingly, the simple act of *using* powerful acts (such as a threat) creates the perception that such an actor is *in fact* more powerful [SL87]. The expression of power is thus, in a sense, a self fulfilling promise.

2.3.2 Persuasion Strategy Effectiveness in Phishing Emails

The research shows that the type of persuasion strategy or combination of strategies used in phishing emails correlates with differing rates of success.

As it relates to strategies which employ some form of implicit or explicit threat, research has shown that the effectiveness of threats is associated with "the size of the threatened punishment, the perceived likelihood that the threat will be enforced, and the magnitude of the resources that enable the threat to be enforced" [FCK+03].

A 2014 study determined that phishing messages using either a Scarcity or Liking strategy were the most successful [WJT+14].

Two later studies in 2017 and 2018 concluded that the use or inclusion of an Authority strategy was the most effective strategy in convincing users to take the desired action in a phishing email [WHJ18] [BPP+17].

A further 2019 study found that reader susceptibility was highest with Scarcity and "legal" emails and lowest for Social Validation/Proof and "financial" emails [LCE+19].

Another 2019 study, which focusing on spear phishing specifically, found that Authority and Scarcity based persuasion strategies appeared to be the most effective in tricking targets to respond to those types of phishing e-mails [dK19].

2.3.3 Persuasion Strategy Prevalence in Phishing Emails

Several studies have analyzed the frequency of use of the various persuasion strategies in phishing emails.

Given the above research on the effectiveness of Authority based messaging, it is not surprising that a 2019 study, which examined emails sent between 2008 and 2017, found that Authority based persuasion strategies were the most prevalent persuasion strategy in phishing e-mails [FT19].

A relatively small and earlier (2014) study of 207 phishing emails suggested that the two most used strategies at that time were Authority and Scarcity based approaches [Akb14]. This study also pointed out that the artefacts of multiple persuasion strategies are often found in phishing emails, with evidence of an Authority based approach being present in nearly all. See Figure 2.2.

Cialdini's principles	Frequency	Percent
Authority	199	96.1
Scarcity	85	41.1
Likeability	45	21.7
Consistency	36	17.4
Reciprocation	20	9.7
Social proof	11	5.3

Figure 2.2: Phishing Email Persuasion Strategy Frequency (from [Akb14])

2.4 The Language of Persuasion

The components of an attempt to persuade a reader to act on a phishing email must *by necessity* be embedded in the message - the written language - in the phishing email. It is invariably the only communication between the sender and the reader. An appreciation of how language "does this" is thus an important aspect of looking for ways to model persuasion in phishing emails. Insights from various areas in the field of Linguistics are instructive.

2.4.1 The Structure of Language

The study of linguistics can be separated into the following general categories [Ben13] [JM12]:

- **Phonetics**: The study of the sounds of human language.
- **Phonology**: The study of sound systems in human languages.
- **Morphology**: The study of the formation and internal structure of words - the meaningful components of words.
- **Syntax**: The study of the formation and internal structure of sentences - the structural relationships between words.
- **Semantics**: The study of the meaning of sentences.
- **Pragmatics**: The study of the way sentences - with their semantic meanings - are used for particular communicative goals. That is, how they are related to the goals and intentions of the speaker.
- **Discourse**: The study of knowledge from "linguistic units larger than a single utterance".

As it relates to the written language of phishing emails, we are primarily concerned with **syntax** and **semantics** and less so with pragmatics. The relationship between these three is generally that:

> "Syntax is what we use to do our best to communicate on the most basic level. Semantics helps us determine if there's any meaning to be found. Pragmatics enables us to apply the correct meaning to the correct situation."[Bre20]

2.4.2 Language Patterns in Phishing

With respect to the syntax and semantics of persuasive speech, [Bro18] conducted an in-depth linguistics based study of the nature of language in phishing emails. This study also recognized the importance of Authority in persuasion, in fact it is central to the analysis in this work.

[Bro18] explains that, in order to successfully persuade, there needs to be *organization* in the way persuasive information is communicated - it needs to flow in a particular manner. To that end, she cites [Wes15] which provides a useful organizational framework known as "The Motivation Sequence" (see Figure 2.3).

In the case of phishing emails [Bro18] identifies the Motivation Sequence as being animated as follows (referred to hereafter as the "Persuasion Motivation Sequence" or PMS):

> "[E]mails need to follow through a particular framework that *grabs the attention* of the recipient, *maintains* his attention, *explains what is expected* of him, *explains why* there is no other possible route to take, and *reinforces* the action necessary."(emphasis added)

Within this persuasive framework, the actual persuasive language in a phishing email is reflected in "speech acts". Speech acts are *actions* performed through language,

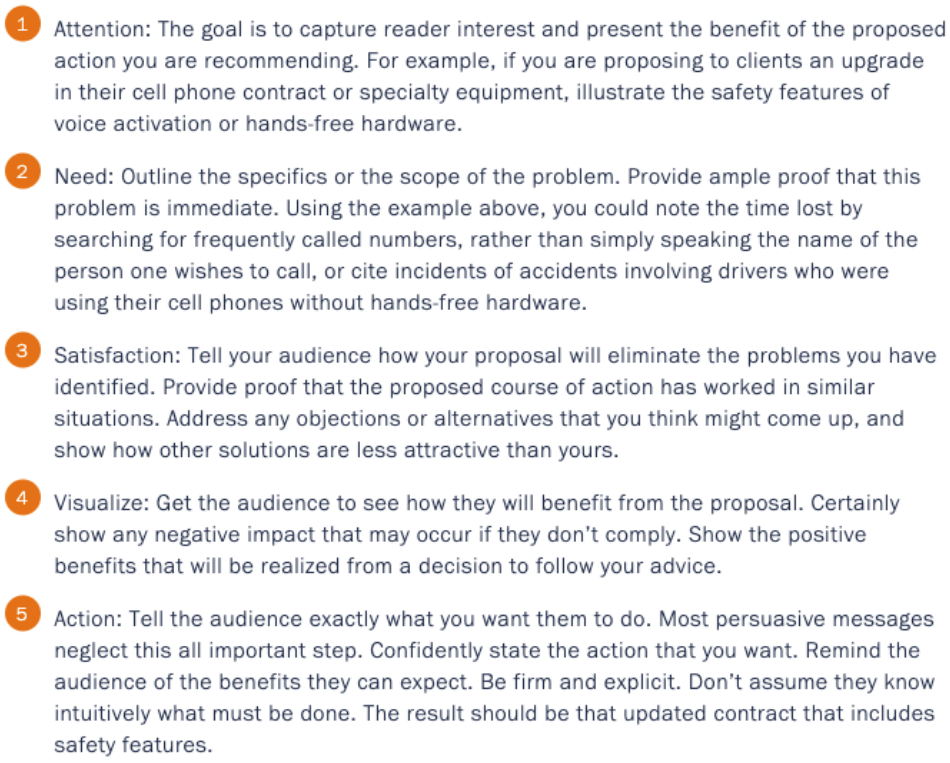

Figure 2.3: The Motivation Sequence (from [Wes15])

which can embody a certain "attitude" of the speaker (or writer).

To achieve this effect, a speech act has three components: the words spoken (the *act*), the intention behind the words (the *force*), and the actual realized effect of the words (the *effect*)[2]. The second component, the force, is particularly important in persuasion, and it is parseable into five categories [GM12]:

1. **Assertive**: stating how things are in the world.
2. **Commissive**: committing to doing something.
3. **Directive**: seeking that the reader do something.
4. **Declaratory**: the speaker doing something currently.
5. **Expressive**: transmitting an attitude about facts or objects.

[2]These three components are referred to as "locution", "illocution" and "perlocution".

Beyond these five categories, speech acts can also generally be divided into two types: direct and indirect speech acts. A direct speech act is declarative and unambiguous. An indirect speech act is more circumspect and used when politeness is in order. An example provided by [Bro18] is illustrative: "I invoke my right to counsel" (direct) versus "Maybe I should speak to a lawyer?" (indirect).

Direct speech is more persuasive [Alt16] and is often used to make unambiguous assertions or statements about the actuality of a particular state of affairs.

[Bro18] systematically analyzed a corpus of phishing emails and made a number of important observations, including:

- The language used in phishing emails was drastically different from that used in ordinary emails. Phishing emails had a clear "agenda" in seeking out specific action. They embody a form of the Motivation Sequence.
- Themes of authority, threat, expressions of power, and superiority all played a persistent role in the persuasive components of phishing emails.
- Many of these themes were expressed using direct declarative statements about the sender, using such phrases as "I am" and "We are".
- Many contained "If ... then ..." and "I will" type conditional or consequential statements.
- Imperative sentences starting with, or containing a key verb with a strong "force" dimension (e.g. "change your password") were often used.
- Language referring to the reader directly, particularly with the word "your", was prevalent and used to engage the reader's interest as well as make factual assertions about matters of presumed interest to, or directly tied to the reader.

- Many phishing emails were thematically focused on money, usually by reference to some aspect of the use of money, such as accounts, transfers and payments.
- Verbs indicative of some sort of movement from one party to another, such as "transfer", "send" and "receive" were highly prevalent in phishing emails.
- Non phishing emails might also seek to persuade, but did so, almost invariably, through politeness based speech acts.
- Phishing emails were generally structured as a persuasive process towards a goal, whereas regular emails had no such common structure and were more "informational" in nature.

2.4.3 Word Meaning Dimensions

In seeking to interpret the role of language in phishing emails, and specifically the import of the use of specific words, it is helpful to appreciate that the meaning of a word may be captured along three general dimensions : **evalution** (i.e. evoking a good or bad emotive reaction), **potency** (strong or weak), and **activity** (active or passive). These provide a helpful lens in reviewing words used and how they might relate to the persuasive purpose of a particular phrase or sentence component. Models have been built scoring words along these three dimensions, albeit with mixed success [OST57] [Hei70] [FCK+03].

2.4.4 Modeling Persuasion

Some research has sought to develop models to detect the presence of persuasion in text.

[AKBG+11] explored the computational detection of perlocutionary (e.g. flattering, insulting, and scaring) speech acts. They used a corpus of blog posts annotated ac-

cording to the presence of 14 different persuasion related *tactics* (see Figure 2.4, and fed simple n-gram as well as Latent Dirichlet Analysis (LDA) generated representations of the posts to various machine learning algorithms. The results were mixed but did demonstrate the ability to model persuasion *tactics* using these approaches.

[IS19] sought to explore this area further, and employed a language structure focused methodology, relying on a parse tree representation of the text as input to various machine learning algorithms. They used the same persuasion tactics as [AKBG+11] and built prototype parse trees representing an "average" representation of each particular tactic. A specific text instance was then classified on the basis of the shortest edit distance to a particular prototype's parse tree. The corpora used were posts on a popular Reddit[3] community, U.S. Supreme Court oral argument transcripts, a large collection of various blog posts, and political speeches by Donald Trump and Hillary Clinton. Some of these data sets were annotated[4] on the basis of the presence of one or more of 14 identified persuasion tactics (see Figure 2.4).

This modeling approach demonstrated the ability to classify each of the various tactics with varying degrees of success, but generally confirming their thesis that language structure is a useful way of detecting persuasion in text.

2.5 The Grammar of Language

There are two general theories of linguistic knowledge, i.e. how language works. The first is the Generative Grammar approach, or the "Dictionary and Rules" model, which relies on a large lexicon of words with a set of grammatical rules on the basis of which language is constructed. This model is based on the fundamental work of

[3]Reddit is a popular topic based discussion forum. See `http://reddit.com`

[4]Some using Mechanical Turk, `https://www.mturk.com/mturk/`

Broad Categories				
Outcomes	**Generalizations**	**External**	**Interpersonal**	**Other**
Outcome. Mentions some particular consequences from uptake or failure to uptake	**Good/Bad Traits.** Associates the intended mental state with a "good" or "bad" person's traits.	**VIP.** Appeals to authority (bosses, experts, trend-setters)	**Favors/Debts.** Mentions returning a favor or injury	**Recharacterization.** Reframes an issue by analogy or metaphor.
Social Esteem. States that people the persuadee values will think more highly of them	**Deontic/Moral Appeal.** Mentions duties or obligations, moral goodness, badness	**Popularity.** Invokes popular opinion as support for uptake	**Consistency.** Mentions keeping promises or commitments	**Reasoning.** Provides a justification for an argumentative point based upon additional argumentation schemes e.g., causal reasoning, arguments from absurdity
Threat/Promise. Poses a direct threat or promise to the persuadee			**Empathy.** Attempts to make the persuadee connect with someone else's emotional perspective	
Self-Feeling. States that uptake will result in a better self-valuation by the persuadee			**Scarcity.** Mentions rarity, urgency, or opportunity of some outcome	

Figure 2.4: List of Persuasion Tactics used in [AKBG+11] and [IS19])

Noam Chomsky in the 1950's [Cho55].

The second is the Constructions Grammar approach, which views language as a large hierarchical "inventory" of constructions (a "construct-i-con") which are linguistic pattern generalizations [Gol95]. An examination of the differences between these two approaches, or their respective merits, is beyond the scope of this thesis and unnecessary for the purpose of aiding the effort of identifying persuasion in phishing emails. There are, however, some core observations and insights, particularly from the area of Constructions Grammar, which are helpful.

2.5.1 Grammar Basics

Verbs and nouns are the core components of language. They describe actions and the subjects of action, which comprise the vast majority of what we (functionally) express with language.

The interaction of verbs and nouns is usually articulated along certain semantic roles.

Every verb has a certain set of compatible roles. See Table 2.1.

Table 2.1: Semantic Roles (adapted from [Hil14])

Role	Description	Example
Agent	The initiator of an action	**Pat** ate a pie
Patient	The participant undergoing an action or state change	Pat ate a **pie**
Theme	The participant which is moving	Pat threw **the rope** over
Experiencer	The participant who is aware of a stimulus	**Pat** heard a sound
Stimulus	The participant that is experienced	Pat heard **a sound**
Beneficiary	The participant who benefits	Pat sang for **me**
Recipient	The participant receiving an item	Pat gave **me** a waffle

Verbs have a *syntactic valence* which describe the number of "participants" which need to be expressed in the case of a particular verb. For example, the word "yawn" has a syntactic valence of 1, given that only one participant is required: "I yawn". "Send" on the other hand, has a syntactic valence of 3, given that it requires someone to do the sending, something being sent, and a (perhaps implicit) destination: "John sends Jane a package". Having an appreciation of syntactic valency is helpful in identifying required arguments of a verb in a particular sentence, and in separating different types of verbs according to their valency.

2.5.2 Constructions Grammar

Constructions Grammar (also sometimes referred to as CxG) is a more recent linguistic theory (first proposed in 1995) which postulates that rather than a "dictionary and rules" based approach, one should look at language as a set of classes of expressions or grammatical (form/meaning) patterns [GC08]. In CxG, the core concept is the "construction", which is defined as the basic unit of language: "the network of constructions captures our grammatical knowledge of a language *in toto*, i.e. it's constructions all the way down." [Gol95]. Everything in language is either a con-

struction or a combination of constructions.[5]

Whereas "Chomskyan" approaches to modeling language focus on rules, Construction grammar based approaches focus more on patterns (mapping form to meaning) and templates for these patterns.

CxG has been successfully used in various text parsing methodologies within Natural Language Processing (NLP) implementations, and is particularly effective in describing the patterns of more nuanced language concepts such as, for example, causation. In one such application, [DLC17] created a set of core "causal" language components, and then "offloaded" the articulation of more complex constructions to machine learning, with very good success. See [MRDM20] for another exploratory study into the merits of using a constructions based approach to natural language processing. Some researchers are also exploring algorithms to try and map constructions patterns in an automated manner [Dun17].

As noted, constructions, and in turn their meanings, may be combined to articulate more complex meaning structure patterns. As a result, there are many such differing patterns. The following example, which is often cited in the literature, is instructive [MRDM20]:

> The sentence "She sneezed the foam off the cappuccino" ... is an instance of the *caused-motion* construction. The verb "sneeze" on its own cannot be interpreted as a motion verb, nor is it usually used as a ditransitive verb, i.e. it does not normally take any complements. It is the caused-motion construction that activates or highlights the motion dimension of the verb "sneeze". Other examples of the caused-motion construction

[5]There is some debate within the constructionist linguistic academic community around the question whether everything in language is by necessity a "construction", with some taking the view that they are limited to only patterns which create meaning beyond the meaning of the language parts of which the pattern is composed (i.e. "idiomatic" meanings). That distinction is not material as it relates to this thesis.

include "She pushed the plate out of the way" or "They moved their business to Oklahoma". All these constructions share a similar syntactic pattern (form) and a meaning of caused motion.

Another example, of an "unaware of the process" type of construction, is presented in Figure 2.5.

She didn't understand how I could do so poorly.

Kiedis recalled of the situation: "He had such an outpouring of creativity while we were making that album that I think **he really didn't know how** to live life in tandem with that creativity."

We didn't know how or why.

One day she picked up a book and as she opened it, a white child took it away from her, saying **she didn't know how** to read.

In a 1978 interview, Dylan reflected on the period: "**I didn't know how** to record the way other people were recording, and I didn't want to.

And it can be on my album, too, **I just didn't realize how** it worked... At first when I got this, people didn't know that I was an artist, so it was, like, 'Oh, this songwriter BC.'

Figure 2.5: Examples of the Construction *Personal Pronoun + didn't + Verb + how* (from [MRDM20])

The importance of simple terms ("I, we", "how"), verb tense ("sneezed, moved") and negation ("didn't") in articulating the meaning of a particular type of construction is illustrated by the above two examples. This is an important observation, as in many instances these types of elements are removed in more traditional NLP based language modeling approaches.

It is also notable to point out that there is a degree of "self-disambiguation" in multi-part constructions. The context created by its components clarifies the meanings of its constituent words or components (which is often a problem with very fine-grained, i.e. word based, meaning extraction approaches, particularly where stemming or lemmatization processes have been applied to the text).

There are a large number of constructions in English, but four of them are very common and important [JM12]. They are referenced here to illustrate the general

approach:

1. Declarative Constructions: a subject noun followed by a verb phrase, used for many purposes. ("We want to move to New Brunswick").

2. Imperative Constructions: a verb phrase without a subject, almost always used for commands or suggestions. ("Show me your passpost").

3. Yes-No Constructions: an auxiliary verb followed by a subject noun phrase, followed by a verb phrase. ("Does New Brunswick have trees?").

4. Wh- Constructions: a wh- word followed by a subject-question or non-subject question ("Which provinces have trees?").

Simply put, constructions are a way to map syntactical *patterns/form* to meaning [Dun17].

2.5.3 Computational Linguistics

Computational linguistics is the scientific and engineering discipline concerned with "understanding" language from a computational perspective. Much of the work in this area continues to rely on the "dictionary and rules" based model of context-free grammars (CFGs) defined by Noam Chomsky, because this approach has a closed set of word types, is logically simple and efficient to parse with a machine [CFL13] [HH01].

Perhaps as a by-product of the "dictionary and rules" based approach, much of the challenge in this area continues to revolve around resolving *ambiguity*. The simple fact is that a sentence can have multiple different meanings, which can only be differentiated with contextual information. An entertaining example provided in [JM12] is instructive. Take the statement "I made her duck". As the author notes, this statement could mean any of the following:

- I cooked waterfowl for her.
- I cooked waterfowl belonging to her.
- I created the (plaster?) duck she owns.
- I caused her to quickly lower her head or body.
- I waved my magic wand and turned her into undifferentiated waterfowl.

In the above example "duck" can be a verb or a noun, "made" can refer to "her" or "duck" (if "her" is interpreted in a possessive sense), etc.

Computational linguists have developed a number of approaches to modeling knowledge in language to resolve ambiguity (as well as other knowledge extraction challenges). These approaches primarily revolve around state machines, rule systems, logic, probabilistic models and vector-space models. In most cases any automated processing of language using these approaches, involves the general theme of *a search through a space of states representing hypotheses about an input*[JM12]. With respect to disambiguation in particular, part-of-speech (POS) tagging processes are helpful in resolving questions like whether "duck" is a verb or a noun. Word sense disambiguation tools can help determine if "made" refers to creating or cooking.

An approach focused less on individual words, and more on phrasal patterns, like that pursued in a constructions based model, would appear to be less susceptible to the ambiguity effects of this traditional approach to computational linguistics.

2.5.4 Some Other Language Related Observations

The English language boils down to a list of semantic "building blocks" not much larger than 20,000 words, when names and acronyms are excluded [BSMK16].

The English language is also a highly *analytic* language, requiring a rigid word order from its users [Fal17].

The frequency distribution of terms in a language corpus generally follows the rule that the frequency of a word is inversely proportional to its rank. This rule is known as Zipf's Law. A thorough analysis is available in [Pia14]. The effect of this rule is that frequency distributions of words are very much top heavy, which result in the effect where, for example, the first 2,000 words in the frequency distribution of words in articles in Time Magazine account for close to 75 percent of the total words used.[6]

Age and education influence the size of one's vocabulary [BSMK16]. We have little insight into the age or education of PCP's but given the sophistication of the language generally seen in phishing corpora, it is unlikely that most phishing emails are being crafted by 50+ year old academic linguists. And even if they were, given the PCP's goal of wanting to communicate effectively, clearly and unambiguously, he or she should *actively seek* to use simple language, using a more limited vocabulary, fully available to its intended targets.

In summary, it appears likely that the (persuasive) language used in phishing emails is subject to an upper bound both in vocabulary size as well as syntactic and semantic complexity. Given that phishing emails have limited unique and distinct language features in light of their single persuasive purpose, and given the existing insight into how persuasion is recognizable in text, it should thus be possible to derive a set of persuasion related constructions for each of the various organizational components of the persuasive effort which reveal themselves in the vast majority, if not all, phishing emails, and use those to build a detection model.

[6] See `https://www.wordfrequency.info/samples.asp`

In order to assess the extent to which such an approach might comport with prior work, a literature review was conducted with respect to computer science research into automated detection of phishing emails.

2.6 Literature Review

The research into computationally driven phishing detection can be divided into two general categories: research focusing on phishing *websites* (the target to which a phishing email is often seeking to steer the reader) and the phishing *email* itself. This literature review focuses on the latter. A few papers from the first category are referenced where the use of semantic components of phishing websites provide useful insights.

Many anti-phishing strategies focus on extracting features from an email and then applying some processing methodology to that data to classify the emails as either phishing or non-phishing.

These features can be divided into three general categories:

1. **Non text-based features** such as whether the email has an attachment, the type of domain from which the email was sent, or whether the email has multiple recipients or not.

2. **Structural text-based features** like the presence of keywords/n-grams in the body text of the email.

3. **Semantic text-based features** which seek to represent the meaning of the text in a set of features, based on a set of general or specific linguistic rules.

Of these three categories, there is a good deal of research in the first two categories. The academic work in the third category is growing and evolving. Machine learning based approaches to classifying phishing emails are prevalent in all three categories.

A review of language focused email phishing research follows. References to non-phishing based implementations are also included where a particular approach is helpful to a similar issue in the email phishing domain. In the same vein, some papers on the semantics of emails in general are referenced as well.

Many researchers have built models with features from both of the first two categories set out above. For ease of reference these are referred to as using "hybrid" approaches. A table summarizing and comparing specific language based approaches (where sufficient detail was available in the applicable papers) is provided at the end of this section.

2.6.1 Text-based Approaches

[CNU06] explored, in addition to several non text-based structural features, the presence and frequency of 18 keywords in the subject and body segments of the email. The testing corpus was quite small (only 400 emails) and the classification accuracy varied greatly once non-word based features were removed.

A somewhat similar approach was followed in [BCP+08] where the presence of 10 words and word stems were tracked and counted. The model also included a combination of several other non text-based structural features such as the presence of any URL's, the web technologies used in the email, and whether the SpamAssassin API[7] classified the email as spam.

[7] `https://spamassassin.apache.org/`

[MOWB09] pursued an approach combining a large number of hybrid features extracted from phishing emails to create several models for classification. Using an information gain feature selection process, the best features were sought to be selected. Of these key features, the "words in the subject" and "words in the body" categories achieved the highest rank (with a frequency distribution of 0.3177 and 0.2281 respectively), followed closely by a feature based on the presence of links in the body of the email (at 0.2266). It is notable that the fourth ranked feature scored 0.01, demonstrating clearly the importance of text-based features in phishing email classification. The authors used a very large corpus (659,673 emails of which 45,525 were phishing emails (about 7 percent)) and were able to generate good results using a number of different algorithms. Notably, this research suggested that decision tree based algorithms performed best in modeling phishing emails on the basis of text-based features.

[PR12] used 23 most prevalent keywords extracted from the email body, reduced to 12 features via a t-statistic based feature selection process, to test a number of machine learning algorithms for classification. They used a corpus of 2,500 emails made up of equal numbers of phishing and non-phishing emails. The research concluded that the feature selection process did not detrimentally affect the machine learning based detection, compared to the full feature set.

[ST$^+$14] employed a model using 46 hybrid features with a large proportion of text-based features. A high rate of accuracy (in excess of 99 percent) was achieved using a Random Forest classification approach.

[MCB17] employed a feature set comprised of the number of web links in the email

body, whether the email was HTML or simple text, the presence of JavaScript, and, after stop word and special character removal, a vectorization of all the words in the email. A Neural Network based modeling approach generated good results.

[BNBW19] employed a vectorized full text, top-down NLP approach to seek to capture inherent characteristics of phishing email text, which were then used to classify emails as phishing or non-phishing using machine learning and deep learning strategies.

[Son20] employed 41 features in four categories (email subject, email body, links present, and readability) to drive a Binary Search Feature Selection (BSFS) approach to phishing email classification.

2.6.2 Language Based Approaches

Several researchers have explored the use of the semantic make up of the text in phishing emails to build a classification model.

[KTTZ05] outlined a methodology for engineering language driven ontology-based knowledge systems to detect various email mediated scams including phishing. The ontological model consisted of identifying *concepts* (e.g. a user's "account") and (action based) relations (e.g., "open") as syntactic patterns within the text of emails. The proposed methodology had four tracks: system engineering; terminology engineering; knowledge engineering; and language engineering.

[LHWR10] used an approach which extracted word clusters constituting an inferred "significant meaning" from phishing emails by applying a version of Latent Semantic Analysis (LSA), a k-means clustering technique, to the phishing emails corpus used.

The authors then used a Latent Dirichlet Allocation (LDA) algorithm to extract "topics" reflected in the clusters. From the paper:

> "LDA is a model where latent topics of documents are inferred from estimated probability distributions over the training data set. The key idea behind LDA is that every topic is modeled as a probability distribution over the set of words represented by the vocabulary, and every document as a probability distribution over a set of topics. By using this text-mining method, different topics can be extracted and used as input features for the phishing classification task."

The top 10 keywords for each topic were then used to mark up the emails. The total number of useful keywords was 405.

The "topics" generated by this approach, although generally recognizable as phishing related, are not easily further classifiable/parseable by humans in any sort of "meaning" sense. See Figure 2.6 illustrating the word content from "topics" derived in this research.

TEN MOST RELEVANT WORDS FOR FIVE TOPICS EXTRACTED BY USING THE LDA TOPIC-MODEL OVER THE PHISHING CORPUS.

Topic 1	Topic 2	Topic 5	Topic 15	Topic 20	Topic 25
paypal	account	account	grupo	bank	click
account	messag	fraudul	imagen	account	visa
secur	suspend	bank	cuenta	bankof	card
password	inform	thank	para	america	receiv
protect	termin	suspend	click	wellsfargo	free
inform	warn	fraud	cliente	well	credit
verifi	legal	login	googl	fargo	usernam
click	agreement	secur	bancaria	barclay	success
access	liabil	notif	nuestro	huntington	want
assist	resolv	regard	dato	client	wish

Figure 2.6: Example Topic Words Extracted Using LDA (from [LHWR10])

The research did demonstrate that this LSA/LDA based approach outperformed those based on purely keyword based features.

An interesting poster, from what appears to be the early 2010's, explored the subject and object of verbs in their usage between phishing emails and legitimate emails. It concluded that features could be created with respect to some verbs, and that more study was required with respect to the utility of others [PT15a].

[VSH12] sought to separate emails on the basis of whether they were "informational" or "actionable" in nature, by way of a number of hybrid features. With respect to the text-based features the researchers used an NLP based methodology which involved extracting, after stemming and stop word removal, four different types of keywords from the email text. Features were then formulated by way of Wordnet sense disambiguation[8] and a SenseLearner tagging process [MF04]. It is of note that the keywords and stems used were personally curated by the researchers by reviewing what appears from the paper to be a relatively small set of emails (20). Several other structural features were also extracted. After disambiguation, term frequency was used to assign a score using a feature weighting formula developed by the authors. This formula also incorporated the other structural features (such as whether a URL link is present). The model also sought to include a broader context focused feature in the classification methodology, by modeling the user's *other* emails to aid in the identification of an "out of the ordinary" (and thus more likely to be phishing) email.

[VH13] employed a (stop word reduced) body text only approach, which among other features used a semantic "action detector", to classify phishing emails. This action detector component was based on finding "property" n-grams, defined as any sequence which matched the word "your", followed by a reference to a "property" term (identified by way of a match with a term from a term list) followed by a URL link.

[8]This API is passed a word and the context in which it occurs (i.e. its surrounding text) and returns a string describing the most likely meaning of the word.

[HAK13] used a combination of structural and text based features to achieve a classification accuracy of 94 percent, using several machine learning algorithms, including Random Forest and Support Vector Machine. The text based features were based on a "black list" of terms not further particularized in the paper.

A 2014 study recognized the importance of integrating human-accessible semantics into computational solutions like phishing email detection [PSTR14].

[AA14] pursued a Random Forest algorithm exclusive approach using 15 features representing various text-based and structural aspects of phishing emails. They were able to achieve an classification accuracy above 99 percent.

[AKS14] sought to extract the presence of semantic components such as the presence of a "reply inducing sentence" and a "sense of urgency" from phishing emails using a Part of Speech(POS) and word stemming based approach.

[PM15] presented a system which used generative grammars to create dynamic e-mail contents for use as test cases for anti-phishing research, demonstrating that there are reproducible inherent grammatical structures which are consistent with phishing emails.

[PT15b] reported on a syntactic sentence similarity experiment comparing phishing and non-phishing emails. The experiment examined the subject and object of verbs in the email text. The results indicated that the syntactic structures of sentences driven by verbs was not enough to play a role in a definite differentiating between phishing and non-phishing. The researchers felt this was due to the fact

that verbs can have multiple meanings. They did identify interesting language structures around the expressed intention of the sender, as well as the use of the word "your" followed by some attribute of the reader, and the need to explore the effect of meaning as future work. One of the researchers pursued this further in [Par18], discussed below.

[YA16] used, in additional to several non-text features, a frequency based weighting of phishing terms extracted from the email header and body (after text stemming, stop word removal, and synonym supplementation using WordNet) to try to estimate the semantic meaning of an email.

[ZZJ+17] demonstrated an approach which extracted a series of semantic features from phishing websites (as opposed to phishing emails) with a Word2Vec based machine learning approach, with impressive results. The authors opined that the majority of phishing websites are effectively identifiable by *only* mining the semantic features of word embeddings.

[PHS18] explored a methodology to detect social engineering attempts using the presence of malicious question/command verb-object pairs, "urgent tone", a generic greeting and a malicious URL link. This work also concluded that semantic information is valuable in phishing detection.

[GDSV+20] proposed a multi-stage approach to email phishing detection, using the text portion of emails only. Text vectors were lemmatized and stop words were removed, and then subjected to various vectorization methodologies, automated feature selection methodologies, and finally, machine learning. The classification accuracy was excellent - in excess of 99 percent, using several machine learning algorithms.

[LZW20] pursued an effort to detect three types of persuasion in emails, namely Authority, Reciprocation and Scarcity, using short term lists for each.

[BFSS20] offers a very recent systematic literature review of research into the classification of phishing attack solutions with deep learning strategies, with an emphasis on URL related data. The review indicates that with respect to that domain, the most often used approaches involved the use of a Deep Neural Network (DNN) or a Convolutional Neural Network (CNN).

A similar literature survey was done, also on machine learning based detection approaches to phishing websites, in [KA19].

Details of the specific text based and other features used, where available, are summarized in Tables 2.2 - 2.4.

Much of the above research focused on capturing *some* semantic aspects of the language in phishing emails using automated approaches or with relatively little evident exploration of the psychological or linguistic aspects of the phishing problem domain.

There are a few researchers who have explored the idea of more systematically using the semantics of phishing emails to model them.

2.6.3 Systematic Semantic Modeling of Phishing Emails

Two PhD dissertations from 2016 and 2018 respectively, [Fal16] and [Par18], sought to more systematically explore the question of capturing the semantic structure of phishing emails. Their work confirms that there is significant value in seeking a more

Table 2.2: Feature Details for Some of the Papers Cited

Citation	Text Features	Other Features	Classification
[CNU06]	Presence of: account, access, bank, credit, click, identity, inconvenience, information, limited, log, minutes, password, recently, risk, social, security, service, suspended. Number of unique words.	URL analysis, presence of embedded forms	Support Vector Machine
[BCP+08]	Presence of: account, update, confirm, verify, secur, notif, log, click, inconvenien. Latent topic modeling using LDA	Presence of MIME email components, URL's, Web technologies, SpamAssassin classification	Support Vector Machine
[MOWB09]	Presence of: account, update, confirm, verify, secur, notif, log, click, inconvenien, bank, urgent (among others - no complete list provided)	Presence of links in the the email, invisible links, non-matching URL's, forms, scripts	Decision Tree, Random Forest, Multi-layer Perceptron, Naïve Bayes, and Support Vector Machine
[PR12]	Presence of: Account, Response, Member, Offer, Access, Transaction, Email, Agreement, Address, Registration, Update, Person, Price, System, Market, Process, Online, Service, Information, Request, Work, Message, Credit	None	Multilayer Perceptron, Decision Tree, Support Vector Machine, Group Method of Data Handling, Probabilistic Neural Net, Genetic Programming, and Logistic Regression
[ST+14]	Presence of: dear, verify your account, verify, suspension, login, click, reply, debit, bank. Number of unique words, number of words	Structural features including the presence of URLs, images, periods in the URL, scripts.	Naïve Bayes, Random Forest and AdaBoost
[AA14]	Presence of: update, Confirm, user, customer, client, suspend, restrict, hold, verify, account, notif, login, username, password, click, log, SSN, social security, secur, inconvinien	URL's containing IP addresses, disparities between "href" attribute and link text, the words click, here, login, update,and link in the link text, number of dots in domain name, HTML or text email, Javascript, number of links, among others	Random Forest
[MCB17]	Word2Vec vectorization of all the words in the email	Number of the web links, HTML/text email, JavaScript, number of the email's parts	Neural Network

Table 2.3: Feature Details for Some of the Papers Cited - continued

Citation	Text Features	Other Features	Classification
[BNBW19]	All the text of the email (converted to one-hot encoded vectors)	None	Naïve Bayes, Support Vector Machines, Decision Tree, Long Short Term Memory (LSTM) and Convolutional Neural Networks
[Son20]	In the subject line: account, update, security, important, resent, notice, verify, please, verification, credit, bank, online, and in the body: account, update, information, transfer, post, credit, priority, user, resent, security, status, address, access, time	Presence and nature of URL's, URL visibility and match with visible text, IP based URLs, URL length, hyphens and dots in URL, image tags, absence of unsubscribe link, etc.	Binary Search Feature Selection
[LHWR10]	Keyword extraction using k-means clustering with subsequent Latent Dirichlet Allocation	None	Logistic Regression, Naïve Bayes, Support Vector Machine
[VSH12]	(VERBS): click, follow, visit, go, update, apply, submit, confirm, cancel, dispute, enroll (ADVERBS): here, there, herein, therein, hereto, thereto, hither, thither, hitherto, thitherto (WORDS CONVEYING A SENSE OF URGENCY): now, nowadays, present, today, instantly, straightaway, straight, directly, once, forthwith, urgently, desperately, immediately, within, inside, soon, shortly, presently, before, ahead, front (DIRECTION WORDS): above, below, under, lower, upper, in, on, into, between, besides, succeeding, trailing, beginning, end, this, that, right, left, east, north, west, south. The word "money",	Context feature based on consistency of email with other emails of user. Email header analysis, link analysis	Pattern matching classifiers developed by the researchers
[VH13]	Presence of "property" n-grams, defined as any sequence which matched the word "your", followed by a reference to a "property" term (identified by way of a match with a term from a term list) followed by a URL link, extracted from the body text only.	None	Pattern matching classifiers developed by the researchers

Table 2.4: Feature Details for Some of the Papers Cited - continued

Citation	Text Features	Other Features	Classification
[AKS14]	The presence of a "reply inducing sentence" and a "sense of urgency" from phishing emails using Part of Speech (POS) and word stemming. Mention of "money"	Presence of names	Scoring methodology developed by the researchers
[PT15b]	The subject and object of the verbs: access, click, confirm, enter, follow, protect, update and use, in the email text	None	Experimental comparison. No classification attempted.
[YA16]	Phishing terms extracted from the email header and body (after text stemming, stop word removal, and synonym supplementation using WordNet) to try to estimate the semantic meaning of an email. The words account, dear, paypal, login, bank, verify, agree and suspend were constituted as independent features.	Presence of HTML content in the email body, any hexadecimal characters in URL's in the email, the number of domains referenced in the email, and the number of dots in referenced URL's	Random Forest, J48
[LZW20]	Presence of: paypal, verify, fraud, management, identity, benefits, bank, customers, accounts, updates, limited, services, suspension, suspended, terminated	None	K-nearest Neighbour, Decision Tree, Bayes

structured model of the *meaning* of the email text and that *semantically meaningful data as input* to a detection system in turn makes for *more meaningful output*.

2.6.3.1 The Falk Research

[Fal16] is the first research project which sought to methodically examine the importance of semantic structures in the context of phishing emails. He explored whether or not machine learning algorithms perform better when provided *semantic* structures as opposed to simple *lexical* structures. The author coined the term Meaning Based Machine Learning (MBML) to represent the first category.

The specific semantic structure chosen to reflect meaning to be found in a phishing email was the "TMR" or Text Meaning Representation, which is a concept from the Ontological Semantics Technology (OST) theory of natural language processing [Tay10].

OST (or OntoSem as it is sometimes also referred to) seeks to "meaningfully connect the overlapping concepts and relationships used in ... text descriptions" on the basis of the underlying idea that the state of something can be described in many different ways, but there are common *properties* and *attributes* for all descriptions. For example the sentences "a ball is over a cube" and "a cube is under a ball" both reflect a common spatial attribute about the ball and the cube. The pattern based philosophy underlying OST is not dissimilar to that of the Constructions based approach discussed above. The basic components of an OST based system are:

1. An Ontology;
2. A Lexicon, or vocabulary;
3. An Ontomasticon (a vocabulary of proper names); and

4. A Fact Repository (to store learned instances of concepts and TMR's).

In OST all the applicable properties and attributes are stored in the Ontology, where entities (usually subjects in sentences) are reflected by way of "case roles". See Figure 2.7.

Case Role	Definition
AGENT	Entity responsible for an action.
BENEFICIARY	Entity deriving benefit from an action.
DESTINATION	Ending point of an action.
INSTRUMENT	Object used to carry out an action.
LOCATION	Place where an action occurs.
MANNER	Style in which an action is done.
PATH	Route taken during an action.
PATIENT	Entity affected by an action.
SOURCE	Starting point of an action.
THEME	Entity manipulated by an action.

Figure 2.7: List of Case Roles as used by Ontological Semantics (from [Fal16])

A TMR is a graph based structure which (ambitiously) seeks to represent the "totality of knowledge available" by connecting abstract concepts, such as the case roles, via "slots" (properties). For the purpose of feeding a TMR to a machine learning algorithm, it is embodied in a parseable S-expression (a "symbolic expression" - a way to represent a nested list of data, not unlike XML).

[Fal16] pursued a limited implementation of an OST based model for phishing emails. This appears to have been a laborious exercise as no TMR parser was available. As a result he prepared two small data sets manually, one using a TMR interpretation, and another which was constituted as unigram representations of the email text. He then tested these against three different machine learning algorithms: Naïve Bayes,

Support Vector Machines and a J48 Decision Tree approach. A cross validation[9] approach to guard against overfitting was also employed.

The TMR based data set performed markedly better, see Figure 2.8.

Test	**TP**	**FP**	**Prec**	**Rec**	**F_1**	**ROC**
Uni-NB	0.800	0.404	0.673	0.800	0.723	0.741
Uni-SVM	0.754	**0.207**	**0.813**	0.754	0.762	**0.773**
Uni-J48	0.607	**0.271**	0.716	0.607	0.628	0.673
TMR-NB	**0.864**	**0.261**	**0.776**	**0.864**	**0.813**	**0.905**
TMR-SVM	**0.832**	0.286	**0.771**	**0.832**	**0.786**	**0.773**
TMR-J48	**0.882**	0.393	0.701	**0.882**	**0.773**	0.745

Figure 2.8: Test Results From 3 ML Algorithms for a TMR and non-TMR Data Set (from [Fal16])

A number of limitations were imposed on this research, while others flowed from the approach itself:

- It only focused on having emails with PDF attachments, on the premise that a particular type of attachment would limit the number of action verbs related to that file type.
- As noted, the process of creating TMR's was done manually and was very time intensive. As a result the data set for testing was very small.
- While the end goal of the research would be a system that could automatically distinguish between phishing and non-phishing emails based on their semantic parses, building such a system was outside the scope of the dissertation.

[9]Cross validation is a randomized re-sampling strategy which repeatedly selects a different subset of a corpus for training and comparison to the remaining (smaller) test subset.

Iit is of note that in [Fal17] Falk explored the development of a parser to transform natural language into TMR's. He was successful with a heuristics based approach using a genetic algorithm.

2.6.3.2 The Park Research

Two years later, another PhD student, Park, further examined the issues explored by Falk [Par18]. This research, again towards a "meaning based" approach to email phishing detection, was driven by his key observation, citing [LSY+16], that:

> ... fraudsters have found ways to bypass phishing detection measures. One possible explanation of this is that most security defenses are based on superficial features of emails, which has been susceptible to continuously newly-crafted phishing emails. For instance, once phishers find out a list of addresses in blacklists or literal keywords registered as dangerous in email servers, they can easily re-forge emails to infiltrate the systems.

For the purpose of his analysis, Park separated the prior approaches to phishing into three categories:

1. List-based approaches: using a list of identified phishing websites URL's to identify phishing emails which contain these links.

2. Heuristic based approaches: using a heuristics driven model to identify phishing emails or websites as being phishing.

3. Machine Learning based approaches: using various algorithms to identify phishing.

He succinctly summarized the various limitations of these non-semantic based approaches, reproduced here as Figure 2.9.

Park noted that "it is imperative that a rigorous study be conducted to identify what email body text means and conveys. Specifically, body text of emails needs

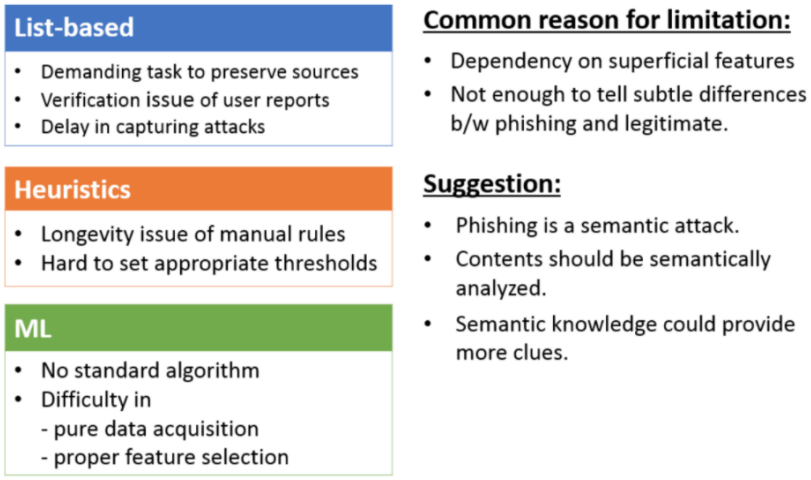

Figure 2.9: The Drawbacks of Technical Approaches (from [Par18])

to be analyzed semantically, and meaning of words and sentences should be used as features to determine the email's legitimacy". His research sought to further confirm that the absence of seeking to model *semantic factors that humans might be able to catch*, but which are simply not considered in structural feature driven ML approaches is a significant weakness, particularly as new more personalized versions of phishing emails are being used. As succinctly put by Park: "the intention of phishing does not change over time, and thus, it is expected that clustering words in semantic domains can create distinguishable features between phishing and legitimate emails".

In the research pursued by Park, he also approached the issue using an ontology driven approach, focusing exclusively on the body text of the emails and no other features.

The primary focus of his research was to find a difference in the nature of verb driven "action requests" between phishing and legitimate emails using various verbs and "their dependents" in the sentences of the emails. Park recognized that the same verbs may appear in both phishing and non-phishing emails, but that their meaning

would likely be different based on their interplay with objects in the text. Park was confronted, as was Falk, with the reality that manually creating syntactic structures from phishing email instances is very time consuming. This again limited the scope of meaning modeling pursued.

Park approached the research question by first identifying 50 verbs with the highest frequency in his testing corpora, which were the following:

`access, activate, approve, attach, bring, build, buy, change, check, choose, click, complete, confirm, consider, create, deliver, deposit, earn, enter, file, find, give, kill, lose, make, open, pay, protect, provide, raise, receive, reconfirm, reduce, register, release, remove, review, save, select, send, set, share, sign, submit, transfer, update, use, verify, visit, win.`

He then followed the following methodology:

- He looked for the presence of the identified verbs and their related objects in the emails, so as to capture "action" based meaning. He used the Stanford typed dependency parser [DMM08], to parse sentences to identify the sentence components indicating initiators of actions (verbs) and targets of the actions (objects).

- He designed an ontology to disambiguate the "action" meaning for combinations of these verbs and objects into "concepts".

- He then tested the modeling value of various representations of the phishing emails (i.e. the email itself, the sentence containing the verb, the *lexeme* (the actual verb and object construction), and its representative concept) using several machine learning algorithms. See Figure 2.10. He also tested the impact of replacing certain verbs or objects with their synonyms.

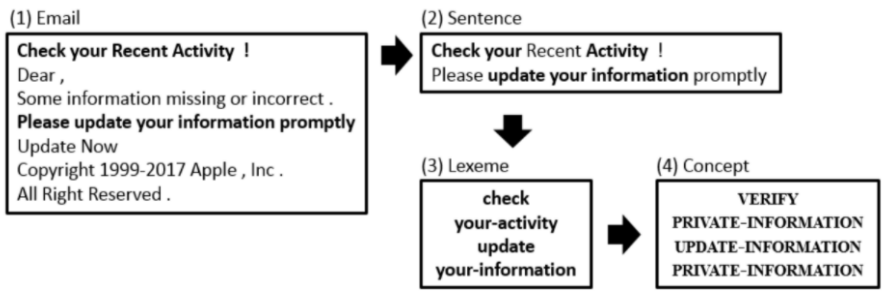

Figure 2.10: The 4 Input Types for ML Testing Pursued in [Par18]

The results indicated that as the machine learning algorithms were fed more data, the accuracy improved. When he tested the machine learning algorithms against *only* the lexemes and the concepts, the scores were as set out in Table 2.5.

Table 2.5: Lexeme and Concept based ML Accuracy from [Par18]

Algorithm	Input	F1 Score
Logistic Regression	Lexeme	0.832
Support Vector Machine	Lexeme	0.833
Logistic Regression	Concept	0.843
Support Vector Machine	Concept	0.843
Random Forest	Concept	0.888
Random Forest	Lexeme	0.891

[Par18] drew several conclusions from his study:

- The inputs to teach a machine phishing detection need to go beyond superficial clues. Semantic features are useful resources to improve machine based phishing detection capability.

- Lexical feature-based models became more vulnerable to unseen data than the

models that leveraged conceptualization. Semantic models are more durable.

- The effectiveness of conceptualization decreases as the number of uncontrolled features increases. We need to analyze semantic aspects of emails in a broader spectrum.

Future work included exploring additional semantic elements (beyond verb and object), conducting semantic analysis in a broader scope beyond sentence units, and looking for ways to integrate more human knowledge in designing phishing detection models.

2.6.4 Insights from the Literature Review

Several observations stand out from the review of various approaches to modeling phishing email language features generally, as well as the semantics focused work specifically.

2.6.4.1 Meaning is a Valuable Feature Set

Every effort to give voice to meaning seems to improve performance. Finding forms of meaning representation are also generally viewed as providing a potential means to more robustly model evolving spear phishing emails.

2.6.4.2 The Struggle To Capture Meaning

Extracting exact meaning from text is hard. LSA approaches like the one pursued in [LHWR10], despite offering improved performance, are difficult to evaluate from the perspective of true meaning. The approach is not clearly "understandable". In turn, the highly formalized OST approach is very labour intensive, and likely too ambitious in that it seeks to capture the "totality of knowledge".

These two differing approaches beg the question whether there is an approach somewhere between the two. Is there room for *not* overthinking it? Is it possible to expand the scope of captured meaning without the intimidating overhead of an OST? Can we achieve results with less formalization? Can we employ simpler versions of an approach like OST if the language in a particular domain is of markedly lesser complexity? Can the approach to reflecting meaning be made more "understandable"?

2.6.4.3 Shallow Pre-Processing

A great proportion of the research seeking to classify phishing emails attack the problem from the traditional text mining approach, which incorporate stemming and stop word removal usually as a given, without appearing to consider the impact of such measures upon the richness of the language data and the concomitant potential loss of available detection information.

As will be explored below, in the specific context of phishing emails, these processing steps can rob the text of critical semantic nuances. An example derived from a phishing email used in this work proves the point. See Figure 2.11. It is quite easy to find both "action" as well as "informational" (to follow the approach in [VSH12]) expansions of a post processing derived root sentence, demonstrating the potential loss of significant meaning.

2.6.4.4 Emphasis on Term Frequency

Where specific term features were identified by way of term lists, these seem to have been curated on the basis of term frequency alone and in most cases, without much thought of the *functional* semantic aspects of email phishing, i.e. the persuasion driven functional meaning which might be available to be derived from the terms. Given that phishing emails are generally shorter than normal emails, term frequency

1. Original

After the last annual calculations of your fiscal activity we have determined that you are eligible to receive a tax refund of **$109.30.** Please submit the tax refund request and allow us 6-9 days in order to process it.

2. Non word and stop word removal

annual calculations fiscal activity determined eligible receive tax refund submit tax refund request days order process

3. Stemming

annual calcul fiscal activ determin elig receiv tax refund submit tax refund request dai order process

Note that this is also a derivative of the following text, which has a markedly different meaning:

Annual calculation of your fiscal activities may determine if you are eligible to receive a tax refund. If so, you can submit a tax refund request yourself. Daily orders will be processed.

Figure 2.11: An Example of Loss of Meaning due to Text Pre-processing

(as opposed to presence and context) would appear to be of less importance.

2.6.4.5 Lack of In-depth Domain Analysis

Very few of the research papers explored the operational, behavioural or persuasion focused underpinnings of phishing emails, and when they did it was in a superficial manner. The importance of doing this is emphasized by a number of the best practices in feature engineering, which is explored in the next section.

2.7 Feature Engineering

Given the intent of deriving language based features, a review of feature engineering best practices and recent research in that area was undertaken.

As noted by a popular blogger in this area:

> ... feature engineering is [a] topic which doesn't seem to merit any review papers or books, or even chapters in books, but it is absolutely vital to

> ML success. [...] Much of the success of machine learning is actually success in engineering features that a learner can understand.[Loc14]

2.7.1 Approaches

Feature engineering is generally pursued by way of one of three broad approaches [DL18]:

2.7.1.1 Classical (hand-crafted) Feature Representation

In this approach, features are (able to be) carefully designed by domain experts with knowledge about the domain specific data properties and the problem sought to be solved. Given that this approach integrates human real world knowledge in the design process, it generates features which are easy to understand and interpret.

An example of a classically engineered feature is the shape of an object in visual recognition systems, articulated by way of its dimensions in three dimensional space.

2.7.1.2 Latent Feature Representation

This, usually automated, approach is used when it is more difficult to readily identify features, for instance, where features are sparse and of low dimensionality. Various feature selection algorithms and approaches exist, usually around an objective function being optimized. Obtaining features with this approach can be difficult and may require extensive reformulation and/or optimization.

An example of the use of latent features is Principal Component Analysis, or PCA or the LSA approaches discussed previously. Eigenfaces [TP91] is a good example of the use of PCA for use in real-time face recognition systems.

2.7.1.3 Deep Learning Feature Representation

This approach achieves automated feature generation with minimal pre-processing and without the need for any domain expertise, through the use of Convolutional Neural Networks (CNN) and similar strategies. A certain amount of design is still required, but this relates more to the appropriate architecture of the CNN (the types and number of layers, the number of neurons, etc.) The downsides of this approach are that it requires large data sets and computational resources, and is susceptible to learning errors due to (latent) biases present in the data, exacerbated by limitations in the ability to explain how the derived information is generated.

An example of the application of these types of deep learning generated features is diagnostic image based lung cancer identification, where particular shapes and densities in images are identified as correlating with disease [SZQ17].

2.7.2 Feature Engineering for Text Data

Techniques for engineering features from text data have been developed by researchers in the areas of information retrieval, natural language processing, and data mining [DL18].

Early techniques depended on term frequency [Luh58] and evolved over time to general "bag of words" representation and Term Frequency Inverse Document Frequency (TF-IDF) weight based approaches, which continue to see persistent use today.

Term frequency based strategies tend to be quite shallow and are not able to capture the syntactic nuances of text where more complex goals such as capturing a sense of "meaning" are pursued. Where that is important, strategies using syntactic phrase, parse-tree, and entity-relation identification are used. More recently, machine learn-

ing enabled Latent Semantic Indexing (LSI) based approaches, have also been found to be useful.

The current cutting edge of text feature engineering is through "word embeddings". A word embedding is a type of word representation that allows words with similar meaning to be understood by machine learning algorithms. This is achieved by mapping words into vectors of real numbers based on how they "co-occur" with other words in a particular text corpus. Word2Vec is one such embeddings approach which places words with similar meanings closer together, and more dissimilar words farther away, in a (multi-dimensional) vector space.[10]

2.7.3 The Importance of Domain Knowledge

Where it is possible, incorporating domain knowledge in the feature engineering process is very beneficial. Simply put, it *avoids the model having to learn something that we already know*, enhances the precision of the model, and reduces the risk of faulty correlation [BBM20].

Domain knowledge should incorporate different dimensions of the problem space. In the realm of text classification in particular, features which interpret sentence data from different perspectives improve performance [DL18] [JZ07].

The purposeful inclusion of domain knowledge is sometimes referred to as *Informed Machine Learning* (IML) [vRMB+19]. See Figure 2.12. A helpful history and taxonomy of approaches to incorporating domain knowledge into feature engineering is provided in [vRMB+19].

[10] see `https://medium.com/@zafaralibagh6/a-simple-word2vec-tutorial-61e64e38a6a1` for more detailed explanation.

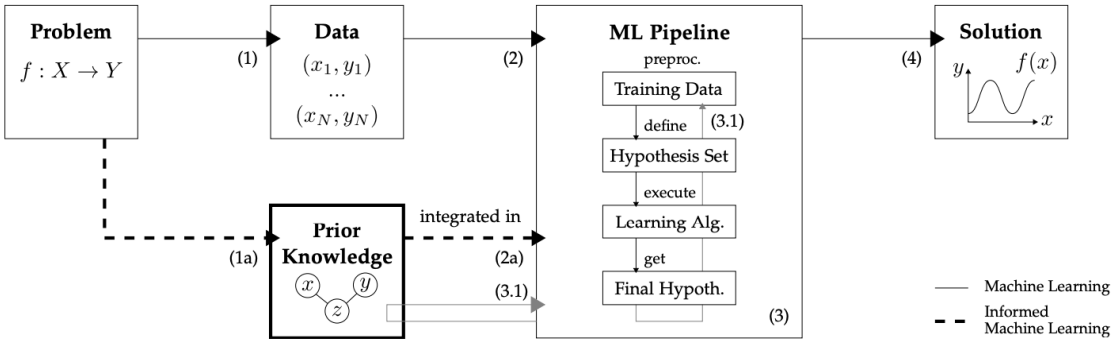

Figure 2.12: Informed Machine Learning (from [vRMB+19])

In the context of IML, "knowledge" is defined as "validated information about relations between entities in certain contexts". This knowledge often stems from natural or social sciences or is a form of expert or world knowledge [vRMB+19].

Incorporating domain knowledge in the feature engineering process also improves the explainability of machine learning solutions which is an increasing area of concern for some of its users [RBDG20].

2.7.4 The Process of Feature Engineering

As a whole, the quest for great features is met by finding insightful ways to *meaningfully* describe the structures inherent in the domain specific data which - together - best embody the problem sought to be solved. In this process, as the author of an oft cited article in feature engineering states: "intuition, creativity and 'black art' are as important as the technical stuff" [Dom12].

One effective methodology to this end is the development of a Concept Vocabulary for the problem domain. This approach was effectively demonstrated in [SS05] with respect to the domain of classifying incoming emails for a customer center. The method sought to specifically leverage domain-dependent knowledge through the cu-

ration of a key concept dictionary manually provided by human experts which was then used to create a concept relation dictionary via an inductive learning algorithm. This dictionary had three layers, a Concept class, a Key Concept and Expressions. An example of this from [SS05] is reproduced in Figure 2.13.

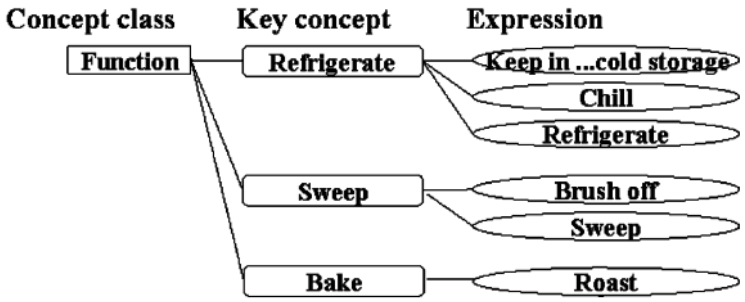

Figure 2.13: The Three Layers of the Concept Vocabulary in [vRMB+19]

The implementation took the subject and body text of an email as input and then classified the email into one of several categories. The researchers noted that this approach corresponded well with the intuition of operators in the customer center and as a result gave highly precise ratios in the classification. The methodology was also highly understandable for everyone involved.

2.8 Summary

This chapter summarized three key dimensions of the email phishing problem:

- The nuances of the phishing event itself;
- The behavioural aspects relating to the fact that emails are efforts in persuasion; and
- The aspects of language and linguistics dealing with understanding how persuasion and related concepts such as meaning are reflected in language.

Next, the various approaches to modeling phishing emails were reviewed. This modeling has to date predominantly focused on relatively shallow text-based and non-text based feature sets and more recently on more "meaning" or semantic driven approaches. The literature review provided a number of insights, including the clear value of meaning based strategies, the continuing challenge of capturing meaning effectively, and what appeared to be a general lack of considering the domain specific, and admittedly more nuanced, behavioural and linguistic aspects of email phishing in classification solutions.

The chapter concluded with an overview of the process of feature engineering and identified a useful Concept Dictionary based approach to incorporating domain knowledge into feature engineering.

Chapter 3

Feature Engineering Methodology

3.1 Introduction

In an effort to build on the work of Falk and Park, this thesis aims to explore semantic aspects of phishing email language beyond merely the "action" verb + object combination. This is sought to be achieved by identifying language components in the text of emails reflective of the various parts of the Persuasion Motivation Sequence as explored in the previous chapter, using a Constructions based Concept Vocabulary focused feature engineering methodology. A summary of the various steps towards that goal follows.

3.2 Problem Domain Theory

The review summarized in the previous chapter enabled the formulation of the following theory towards a broader persuasion language based modeling of phishing emails:

1. Phishing emails are primarily an effort at persuasion.

2. Persuasion is achieved using recognizable persuasion strategies and modalities.

3. Written persuasion uses a particular kind of language and has identifiable persuasion components (the PMS).

4. These persuasive elements are constituted as templatable combinations of language snippets (constructions) of distinct types.

5. These types of language snippets are identifiable in the text of the phishing emails.

6. A Concept Vocabulary based model can be crafted to classify phishing emails on the basis of the presence of these various constructions in an email.

Based on the research set out in the previous chapter, the following specific language statements were identified as candidate construction types:

- Statements which seek to capture the attention of the reader.
- Statements to establish authority, or to build familiarity or trust.
- Declarative statements about the sender.
- Declarative descriptions of some problem, state, completed action, benefit or favour.
- Conditional statements referring to potential consequences.
- Imperative statements which direct some form of action.
- Statements seeking to create a sense of scarcity.
- Statements referring to the reader directly.
- Verbs which embody action.

- Descriptions or references to things or concepts which are important to the reader.
- Descriptions of positive or negative outcomes.

3.3 Developing the Concept Vocabulary

The Concept Vocabulary in [SS05] was comprised of three iterative layers: a Concept Class, a Key Concept and Expressions. In the domain of phishing emails, this approach is constituted as a layered vocabulary of Constructions, Features and n-grams. A visualization of the proposed Concept Dictionary is presented in Figure 3.1.

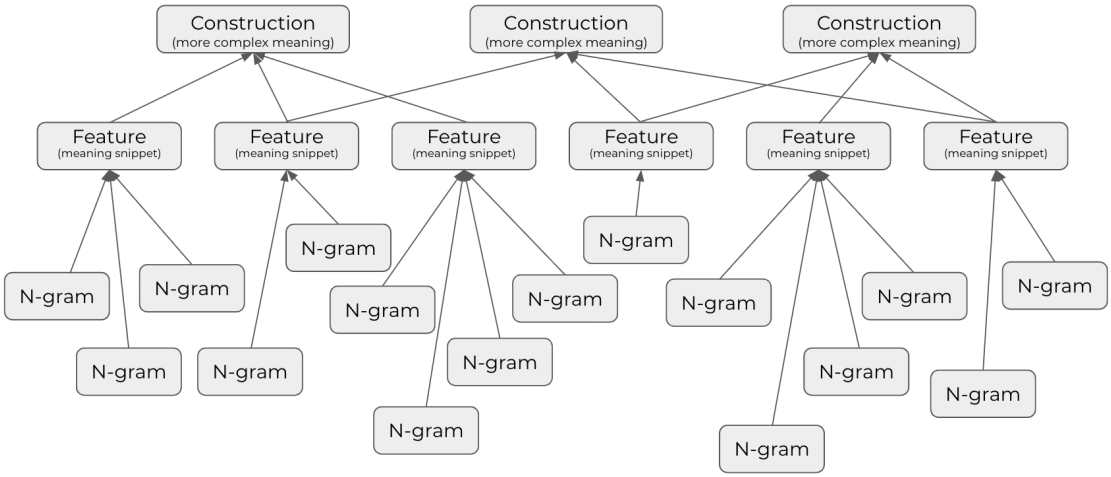

Figure 3.1: The Concept Vocabulary

The methodology towards populating the layers of the Concept Vocabulary is based on the key engineering "tracks" set out in [KTTZ05] and the linguistic methodology followed in [Bro18], and is as follows:

1. Conduct an inspection of the domain corpora and gather candidate language

snippets reflective of persuasive language and the Persuasion Motivation Sequence.

2. Create key concepts by gathering expressions of similar form, with the same or similar meaning/purpose in the PMS as potential feature sets.

3. Test the utility of identified feature sets and combinations of feature sets in classifying phishing emails.

4. Revise as appropriate until there is satisfaction that a sufficient number of important components of the PMS are being captured and classification is effective.

A summary of each of these steps follows.

3.4 Corpora Examination

A review of several email phishing corpora, more particularly set out in Appendix A, was conducted. To contrast these corpora with "regular" emails, the "Podesta emails", a large (57,000+) corpus largely free from phishing emails was also retrieved and examined.[1]

The review sought to identify instances of the various types of (templatable) language snippets outlined above. This was pursued through a visual inspection of a large number of emails, as well as a number of searches for word combinations reflective of PMS related language.

In an effort to enable an effective comparison of both the phishing and non-phishing corpora, two representative samples from each corpus was prepared with a size of

[1]Ironically, this corpus was introduced into the public domain after its owner, John Podesta, fell victim to a phishing email.

3,757 and 3,859 emails respectively (the "Sample Corpora"). The non-phishing set was reviewed carefully to remove any obvious duplicates or highly similar emails, as well as to reduce the number of emails reflecting lengthy threads, which were highly prevalent.[2]

The key observations and insights gleaned from this process are summarized below.

3.4.1 Sender Identity

At a high level, the senders of emails in the various corpora can be segregated into two categories:

- Emails from corporate senders, such as banks, retailers, news agencies, service providers etc.
- Emails from personal contacts of the reader, such as family, co-workers, colleagues outside the organization, social network connections, etc.

Emails incorporating Authority based persuasion strategies tend to largely reflect senders in the first category.

3.4.2 Subject Line Content

In the phishing email corpora, the subject line of the email is often used to draw attention to the reader. This is done in a variety of ways, but a largely predictable set of terms is readily discernable. In many instances the exclamation mark is also used to emphasize the need for attention.

[2]It is worth noting that the corpora used represented mostly "business emails" rather than personal emails. This is helpful given that the primary targets of email phishing are businesses and their employees.

3.4.3 Body Length

Most phishing emails appeared to have 3 to 10 substantive sentences. The body length of a phishing email appeared, on average, to be less than that of the average non-phishing email. A plot of the body length frequency distribution of the Sample Corpora (both reduced to n=3,744), confirmed this observation. In general, most phishing emails had a body character length less than 2,000 characters, with very few outliers beyond that size. The non-phishing sample included a large segment of emails having a size well beyond 2,000 characters. See the histogram in Figure 3.2.

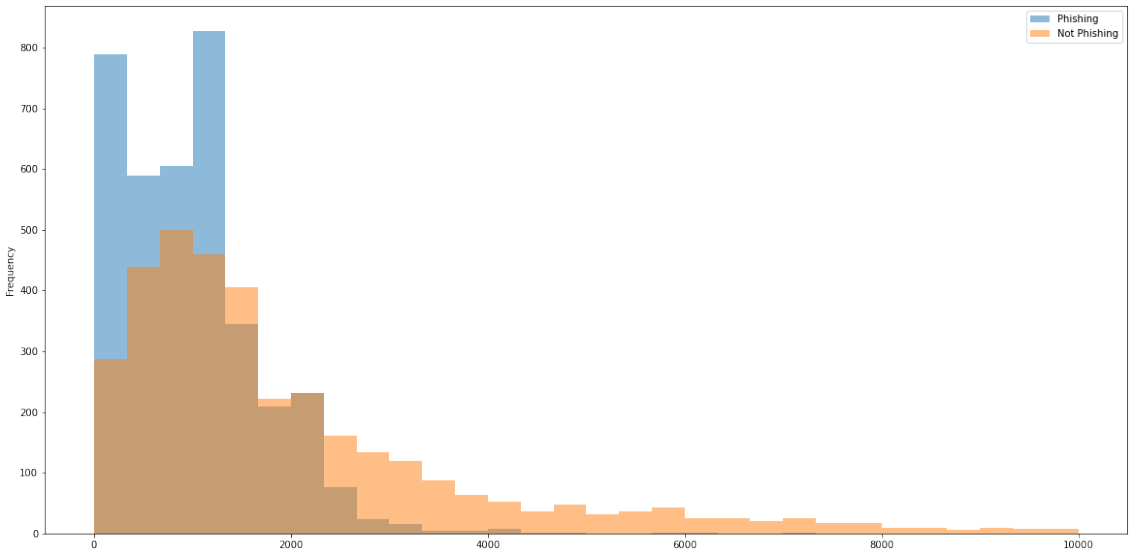

Figure 3.2: Sample Corpus Body Length Frequency Distribution (n=3,744 for each of Phishing and Non-phishing)

3.4.3.1 Very Short Emails

Both corpora contained several very short emails, usually containing only a short sentence like "*I have attached the document we discussed*", "*Here is a link to that article*", "*You have received a fax*", and "*Please see the document*". These types of very short emails are easily discernible as either phishing or not phishing. As a result,

most researchers will exclude very short emails from any model. Likewise, emails with bodies containing short sentences were removed from the Sample Corpora.

3.4.4 Term Frequency

Term frequency lists were generated from the Sample Corpora (without any stemming) and examined to identify language snippets consistent with the Domain Theory. Over 600 language snippets were identified as potential candidates. The top 20 most frequent snippets are set out in Table 3.1.

As a whole, the distribution of these types of phishing related language snippets between the two corpora was markedly different, with the phishing corpus having a much richer overall presence, at almost twice the global incidence rate of the non-phishing corpus (116,440 versus 62,966).

3.4.5 Bi-gram Analysis

A more fine-grained bi-gram term frequency analysis was also conducted. The goal of this exercise was to determine if there were any differing types of combinations of language snippets present in the corpora and if there was a difference in frequency of such terms. To that end numerous bi-grams were identified and searched for with regular expression matching searches such as words following "*your*", "*my*", "*our*", "*if you*", "*has been*", etc. The top 20 results for two such searches appear in Table 3.2.

Table 3.1: The Top 20 Phishing Related Language Snippets in the Sample Corpora

	Phishing Emails		Non-Phishing Emails	
Rank	Term	Frequency	Term	Frequency
1	your	16171	is	14532
2	account	9594	your	3796
3	ebay	9531	our	3577
4	is	7051	my	3038
5	paypal	5308	sent	2950
6	our	3694	was	2742
7	bank	2952	i am	1432
8	sent	2718	will be	1297
9	update	2112	you are	1193
10	access	2051	we are	1122
11	online	1853	you have	875
12	was	1839	we have	868
13	privacy	1655	change	803
14	you have	1506	i have	724
15	unauthorized	1328	would be	645
16	my	1289	could be	608
17	notification	1206	we can	603
18	respond	1161	is not	554
19	id	1104	health	520
20	you are	920	update	519

3.4.6 Possessive Terms

In the English language we express the direct state of possession between two persons or entities (as opposed to ownership by a third person) using only three words: "*my*", "*our*" and "*your*". These words have no synonyms and are thus very useful in email phishing feature engineering. These terms were highly prevalent in the email phishing corpora, and were often used to refer to a particular action or state of something, as part of one or more of the components of the Persuasion Motivation Sequence.

Table 3.2: The Top 20 Results for two Sample Bi-Gram Searches in the Sample Corpora

"your" + any word				**words ending in "ed"**			
Phishing		Non-Phishing		Phishing		Non-Phishing	
Term	Frequency	Term	Frequency	Term	Frequency	Term	Frequency
your account	4177	your notification	132	registered	792	need	1224
your paypal	1090	your rsvp	116	limited	714	united	719
your email	478	your address	110	reserved	510	forwarded	673
your notification	376	your browser	85	designated	440	attached	551
your ebay	343	your profile	79	located	414	received	515
your records	307	your help	78	updated	366	wanted	447
your business	269	your support	74	suspended	307	asked	412
your identity	259	your thoughts	73	automated	302	intended	345
your online	245	your email	69	generated	294	wed	312
your personal	244	your name	58	included	288	subscribed	263
your information	224	your time	54	listed	281	used	261
your item	184	your system	53	need	277	based	194
your transactions	174	your inbox	40	unauthorized	272	called	192
your prompt	171	your calendar	29	received	251	needed	181
your patience	162	your family	29	required	251	worked	174
your protection	154	your review	28	united	242	updated	174
your password	149	your mobile	28	originated	241	changed	173
your access	142	your account	27	accessed	236	privileged	170
your billing	134	your own	25	committed	234	interested	158
your preferences	116	your donation	24	answered	234	invited	153

3.4.7 Verbs in the Past Tense and "State"

Phishing emails often refer to the "state" of some reader attribute, usually in order to raise a level of concern in the reader as part of one or more of the persuasion strategies discussed previously.

Many of these statements are in the past tense[3], and identifiable by way of a "*your + has been + past tense verb*" construction, where in most cases the past tense verb ends in "*ed*". See Table 3.3 for the results of a "*has been + any word*" search within the Sample Corpora.

It is notable that this "*ed*" ending in this "state" construction is valid irrespective of

[3]There are some exceptions, such as "*your account is at risk*", but these appear to have a low frequency.

the specific past tense used (e.g. Past Simple Passive: *The account was compromised*; Present Perfect Passive: *The account has been compromised*; Past Perfect Passive: *The account had been compromised*; or the Past Continuous Passive: *The account was being compromised.*)

Table 3.3: The Top 20 Results for "has been" + any word in the Sample Corpora

"has been" + any word			
Phishing		Non-Phishing	
has been resolved	107	has been a	57
has been reported	64	has been changed	45
has been limited	38	has been canceled	13
has been suspended	26	has been in	12
has been randomly	22	has been the	10
has been used	20	has been an	10
has been sent	16	has been doing	9
has been locked	15	has been responsive	7
has been temporarily	13	has been approved	7
has been flagged	13	has been on	6
has been selected	12	has been so	5
has been restricted	9	has been by	5
has been done	8	has been about	5
has been violated	8	has been sent	5
has been successfully	6	has been reported	5
has been placed	5	has been fighting	4
has been receiving	4	has been to	4
has been temporary	4	has been approached	4
has been finalised	4	has been set	4
has been connecting	3	has been working	4

There is a relatively small set of irregular verbs which do not end in "ed", for example "*withheld*", "*lost*", "*spent*" and "*withdrawn*", so these were reviewed for relevancy and frequency manually.[4]

[4]There are approximately 200 irregular verbs in the English language. See `https://www.englishpage.com/irregularverbs/irregularverbs.html`

With the high prevalence of simple and common words in these snippets, as well as the importance of the past tense, it became quite apparent that stop word removal and stemming of text were potentially detrimental to the goal of capturing at least some of the various artefacts of the Persuasion Motivation Sequence in phishing emails.

3.4.8 Use of the Future Tense and "State"

There was a prevalence of language snippets indicating some future state, usually in the context of some future impact on the reader if the desired action sought in the email was not completed by the reader. See Table 3.4 for a comparison of matches to the search for "*will be*" + any word. The presence of "negative potential impact" type constructions in the phishing corpus was quite evident. These types of statements were generally absent in the non-phishing data set. The use of the past tense of verbs to indicate a future state was again observed.

3.4.9 Action Language

"Action language", for the purpose of this thesis, is language referencing the specific action sought by the sender. This is clearly a subset of the much larger set of English verbs. Not every language snippet reflecting an action is "phishing action" language.

Identifying PMS specific action verbs was done with the assistance of a few revealing searches. Many of the action verbs found in the phishing corpora followed identifiable patterns, such as "*please + verb*" or "*you must + verb*" and searches of these patterns was insightful. See Table 3.5 for the top 50 results from a "*please + word*" search in the Sample Corpora.

Table 3.4: The Top 20 Results for "will be" + any word in the Sample Corpora

"will be" + any word			
Phishing		Non-phishing	
will be provided	116	will be in	89
will be suspended	95	will be a	86
will be limited	58	will be able	37
will be deactivated	37	will be held	34
will be required	32	will be on	29
will be forced	31	will be at	26
will be able	26	will be the	26
will be stored	26	will be there	26
will be terminated	25	will be sorry	16
will be restricted	19	will be served	15
will be upgrading	14	will be an	15
will be checking	13	will be leaving	14
will be deleted	12	will be automatically	14
will be allowed	12	will be running	13
will be closed	11	will be here	12
will be effective	10	will be out	11
will be locked	9	will be very	11
will be done	8	will be to	10
will be covered	8	will be available	10
will be notified	7	will be sending	8

A by-product of the "*please*" search, was the observation that it occurred slightly more often in phishing emails (1,388 instances in phishing emails (n=3,757) and 1,047 for non-phishing (n=3,859), but which much less *variability* than non-phishing emails. In the phishing corpus there were only 51 unique "*please + word*" pairs, whereas in the non-phishing sample, there were 130 such variations.

There was a significant contrast in the presence of action related language between the phishing and non-phishing corpora. The emails in the phishing corpora, largely without exception, sought some sort of action on the part of the reader. The Podesta emails were much more "informational" in nature, with most emails not containing any explicit overture to some sort of desired action like that invariably sought in

Table 3.5: The Top 50 Results for "please" + any word in the Sample Corpora

"please" + any word			
Phishing		Non-phishing	
please click	186	please log	131
please be	163	please notify	84
please help	155	please let	70
please take	140	please send	63
please visit	129	please contact	61
please understand	97	please add	60
please contact	76	please click	49
please review	53	please do	45
please follow	44	please visit	38
please do	35	please https	36
please use	32	please advise	28
please access	24	please delete	19
please go	23	please find	18
please sign	19	please go	16
please login	18	please call	16
please disregard	16	please email	15
please call	16	please review	15
please ignore	16	please get	15
please update	15	please rsvp	13
please confirm	15	please use	13
please become	13	please reply	10
please tell	11	please take	10
please verify	10	please feel	9
please let	9	please immediately	8
please provide	9	please follow	7
please check	7	please be	7
please read	5	please tell	6
please immediately	5	please note	6
please notify	4	please respond	6
please mail	4	please keep	6
please reconfirm	4	please http	5
please enter	4	please see	5
please send	4	please know	4
please try	4	please e	4
please delete	3	please make	4
please pay	2	please reach	4
please email	2	please help	4
please don	2	please consider	4
please log	2	please shout	3
please please	1	please sign	3
please start	1	please treat	3
please www	1	please say	3
please know	1	please forward	3
please spare	1	please meet	3
please launch	1	please join	3
please authenticate	1	please destroy	3
please act	1	please stay	3
please begin	1	please unsubscribe	3

phishing emails. Where such overtures were present, they usually occurred as the result of an information exchange or a desire to share information (which in turn was usually not directly related to an attribute of the reader). In general, observations were consistent with those of [Bro18] that requests for action in regular emails constitute themselves in a more polite and indirect manner.

Where action was sought in the emails in the phishing corpus, these overtures can be divided into three distinct categories:

- Requests to click a link in the document, usually to access information referenced in the email, connect to a login page, or access a document stored in the cloud.
- Requests to open an attachment attached to the email.
- Requests to respond to the email. This category, although much less prevalent, and less immediately risky to the user's computer or network, is also a serious threat as these types of requests are in many cases the first step in building a relationship for exploitation in a more intricate social engineering effort such as a spear phishing based attack or BEC.

3.4.10 The Word "please"

As noted above, the word "*please*" is highly prevalent in both phishing and non-phishing corpora. It is used to express a polite wish or request and thus may be used as part of the underlying Persuasion Motivational Sequence.

3.4.11 Reader Attributes

There was a high prevalence of reader attributes: terms identifying things or concepts of import to the reader. As noted previously, in phishing emails these are often

referenced using direct "*your*" based statements. Prevalent nouns following "*your*" are set out in Table 3.2. The high frequency of "*your*" generally is evident in Table 3.1. It was the most prevalent language snippet identified.

As is evident from the searches for words following "*will be*" (see Table 3.4) and "*has been*" (see Table 3.3), many of these are verbs indicating some sort of state change with respect to what appeared to be similar types of attributes. To explore in more detail what was being referred to, a number of searches were done to identify words *preceding* those and similar terms. The top 50 search results for two such searches are set out in Table 3.6.

3.4.11.1 Deriving the Attribute Feature

Language relating to the attributes of interest to the email reader appeared important in the persuasion based analysis of phishing emails. These types of attributes are not always explicitly stated, but when they are, it is what gives force to the other components in the Persuasion Motivation Sequence. It and some referenced state change or "danger" with respect to it, in many cases provide the "why" behind the action sought by the phishing email.

In order to curate a list of n-grams for this feature, the following process was adopted:

1. A general term frequency list was generated and nouns which constituted attributes and had a relative high frequency (compared to non-phishing), were added to a candidate list.

2. Various searches were conducted for noun presence in language patterns where attributes were likely to be referenced, such as matches for "*your*" + word, word + "*will be*", word + "*has been*", word + "*have been*", etc.

Table 3.6: The Top 50 Results for Any Word + "has been" and "will be" in the Sample Corpora

any word + "has been"				**any word + " will be"**			
Phishing		Non-phishing		Phishing		Non-phishing	
account has been	211	event has been	60	account will be	158	it will be	135
issue has been	107	it has been	36	you will be	112	i will be	133
access has been	36	she has been	28	information will be	104	we will be	104
item has been	9	that has been	26	we will be	72	there will be	67
it has been	8	he has been	24	features will be	54	you will be	57
this has been	8	there has been	16	it will be	51	she will be	45
instruction has been	7	who has been	15	and will be	35	and will be	40
email has been	5	clinton has been	14	ebay will be	33	this will be	39
message has been	4	sanders has been	12	below will be	19	he will be	24
department has been	4	department has been	12	verification will be	8	that will be	23
you has been	4	and has been	11	listings will be	6	they will be	21
union has been	3	hrc has been	9	funds will be	6	who will be	19
ebay has been	3	this has been	8	transaction will be	5	which will be	17
profile has been	3	which has been	7	i will be	5	collected will be	16
block has been	3	what has been	7	money will be	4	event will be	14
data has been	2	life has been	6	of will be	4	meeting will be	13
process has been	2	group has been	5	accounts will be	4	shipping will be	11
banking has been	1	work has been	5	that will be	4	session will be	9
services has been	1	response has been	5	case will be	3	lunch will be	9
identity has been	1	outcry has been	4	upgrade will be	3	hrc will be	8
door has been	1	approach has been	4	online will be	3	gabe will be	8
acount has been	1	campaign has been	4	services will be	3	emily will be	8
and has been	1	amanda has been	4	strike will be	3	team will be	7
bank has been	1	staff has been	4	system will be	3	prices will be	7
tfcu has been	1	email has been	4	card will be	2	transportation will be	6
card has been	1	biden has been	4	us will be	2	message will be	5
number has been	1	nancy has been	4	br will be	2	classes will be	5
		katherine has been	4	which will be	2	clinton will be	5
		recovery has been	3	database will be	2	information will be	5
		draft has been	3	access will be	2	names will be	5
		schedule has been	3	patience will be	2	hillary will be	5
		say has been	3	order will be	2	call will be	5
		committee has been	3	item will be	2	guests will be	5
		nothing has been	3	amazon will be	2	but will be	4
		benghazi has been	3	australia will be	2	group will be	4
		data has been	3	update will be	1	candidate will be	4
		activities has been	2	then will be	1	tickets will be	4
		hillary has been	2	this will be	1	calculations will be	4
		vote has been	2	one will be	1	status will be	4
		requirement has been	2	address will be	1	discussion will be	4
		p has been	2	platform will be	1	polls will be	4
		everyone has been	2	mail will be	1	refreshments will be	4
		proposal has been	2	request will be	1	what will be	4
		successor has been	2	business will be	1	hope will be	4
		trump has been	2	provide will be	1	so will be	4
		degree has been	2	banks will be	1	center will be	4
		christine has been	2	limit will be	1	us will be	4
		one has been	2	payment will be	1	myself will be	4

3. The candidate list was then reviewed and obvious non-attribute terms removed.

3.4.12 Authority Language

Phishing emails contain a good deal of language which is authoritative in nature. Use of the word "*we*" together with some verb with pejorative potential ("force) is prevalent. For example "*we have determined*". These types of constructions were usually found early (usually in the first sentence) in the body of emails, and appeared to be employed to immediately state some sort of state of concern or interest to the reader, from a power based perspective.

3.4.13 Other Observations

As a whole, phishing emails contain various degrees of other language features which also contribute to some degree to the persuasive effort at hand:

- Negative directive statements, such as "*do not*", "*don't ignore*" etc. are often present.
- There are often references to "temporal" language such as "*immediately*", "*act now*" etc., usually in an apparent effort to support a sense of urgency.
- Imperative language, such as "*you must*", "*you will*", as referred to above, are generally prevalent.
- References to some detrimental consequence to the reader (or a reader attribute) of some sort are usually included, or implied.

3.5 Differences Between the Corpora

The review summarized in this Chapter also revealed some more general differences between the phishing and non-phishing corpora, which deserve being mentioned.

3.5.1 "Transmitted Meaning Complexity"

Differences were noted in the language complexity between phishing and non-phishing emails which deserve some discussion. One might use the term "transmitted meaning complexity" to describe the *depth* of expressed language in the emails, i.e. the extent to which the reader needs to corral a "semantic" or even "pragmatic" lens in order to fully capture the meaning expressed by the sender. As a whole, the contrast in the scope of transmitted meaning communicated, as between phishing and non-phishing emails, was notable. Where the Podesta emails varied greatly in their subject matter, topic depth, and as a result sentence structure, the phishing emails were generally quite predictable in the various forms of transmitted meaning employed. This observation was bolstered by the results from the various searches. Phishing emails had a high prevalence with low variability of terms, whereas the non-phishing emails has lower prevalence and higher variability of results.

Interpreting the language in phishing emails is generally not a complex literary exercise. This observation provided comfort that a model based on the more rudimentary aspects of language was appropriate for the phishing email classification problem, and that gathering a large set of persuasion related language snippets would adequately cover the persuasion language used in phishing emails generally.

3.5.2 Conversation Threads

Many emails in the Podesta corpus include threads of prior emails, as the result of a conversation carried on by email. This pattern was completely absent in the phishing emails reviewed.

3.6 Feature Identification and Curation

The corpora review confirmed that phishing emails contained identifiable language snippets of the types set out at the beginning of the Chapter. In excess of 600 instances of these language snippets were gathered by way of examining the available phishing email corpora, and initially grouped as candidate feature sets, based on these types. These candidates were then reviewed and pruned using the following criteria.

3.6.1 Criterion for Feature Curation

The articulation of a particular feature was driven by a single criterion: is this feature set an atomic component having a pattern and common meaning, which is an identifiable and *distinct* segment of the PMS of a phishing email?

3.6.2 Criteria for Language Snippet (n-gram) Selection

The inclusion of a particular n-gram in a feature set was determined using two criteria:

1. Is the snippet functionally consistent with the "role" (form/meaning) of the function set?

2. Is the snippet sufficiently unambiguous to belong to the identified feature set?

3.6.3 Iterative Refinement

At various points versions of feature sets and their constituent n-grams, as well as various combinations of them, were tested to assess their ability to classify between phishing and non-phishing emails. This resulted in the elimination of several candidate feature sets and the removal of a number of candidate n-grams in others.

3.7 Summary

This chapter summarized the feature engineering methodology implemented and key insights gained from an examination of various corpora towards that goal. The next chapter describes the persuasion focused features derived by that effort.

Chapter 4

A Semantic Persuasion Model

4.1 Introduction

The outcome of the feature selection process summarized in the previous chapter was the identification of 13 features suitable for modeling persuasion in phishing emails. Each feature has an associated curated n-gram term list, ranging in size from three words for one feature to 121 n-grams for the largest feature. The total number of n-grams used in the model is 441.

The features are summarized below.[1]

4.1.1 AlertTerm

These are terms which seek to imbue the reader with a sense that the message deserves priority attention, sometimes because the state of a certain something is not as it should be, artefacts of which may be found close to an AlertTerm. Many instances of these terms have an appended "!" in an apparent effort to amplify the need for attention prioritization. Most of these are constituted as single words and

[1]The names of the features follow the naming conventions for Classes in the Java programming, given that Java was used to implement these features in Phishalyzer, the processing pipeline.

appear in the subject line of a phishing email, but are also often repeated again in the body as well. Examples are:

- attention
- alert
- urgent
- critical
- action required

AlertTerm are often accompanied by an AttributeTerm and a PossessiveTerm.

4.1.2 SenderActionVerb

These short bi-grams reflect a tense agnostic personal action or state on the part of the sender. Examples are:

- I have
- We advise
- I regret

This feature is often found with an ActionVerb in a sentence.

4.1.3 TenseStateVerb

These verb based n-grams reflect or refer to some recent action - sometimes by, on behalf of, or observed by the sender of the email - in relation to some object. In phishing emails, the object of the n-gram is often some attribute of importance to the reader (see AttributeTerm below). Examples are:

- has been
- has not been
- could not be

4.1.4 ActionStateVerb

These are past-tense versions of verbs depicting some state change, which in the case of phishing emails is often associated with an attribute of interest to the reader. This feature has a large list of n-grams. Examples are:

- accessed
- cancelled
- identified
- locked

4.1.5 PersonalActionTerm

These are personalized n-grams intended to specifically articulate that certain action is required, or appropriate, on the part of the reader. Examples are:

- you must
- you should
- could you

The term "please" is very often placed immediately before or after a PersonalActionTerm.

4.1.6 ActionVerb

These are specific, instructive, and often infinitive verb form based n-grams, identifying the specific action desired from the reader in phishing emails. Some of these include observed modifiers such as "on" and "the", as the simple form appeared too general and ambiguous. Examples are:

- click on
- open the

- update
- log on
- view the

This feature may be followed by an AttributeTerm.

4.1.7 PossessiveTerm

This feature is only comprised of the three terms "*your*", "*my*" and "*our*". As previously noted, none of these terms have a synonym and their use is limited to the purpose of articulating ownership of something. Given that many phishing emails make appeals based on some undesirable state of some attribute, these terms are often found in phishing emails with an instance of AttributeTerm.

4.1.8 AttributeTerm

These usually single word terms reference some attribute of importance to the reader, with respect to the state of which some action on the part of the reader is sought in phishing emails. Attributes are diverse and include material things such as "*package*" as well as more intangible things commonly referenced or implied with such words as "*spending*" or "*financial*".

Examples are:

- account
- financial
- ticket
- package
- identity
- date of birth
- spending

4.1.9 AuthorityTerm

The concepts of authority, threat and fear are semantically connected. There is no fear without a perceived threat, a threat loses much of its gravitas if there is no fear, and a threat without authority (i.e. ability to follow through) has no potency. In light of this, various language snippets which seek to embody these various dimensions of the same idea were combined into one feature. This set comprises n-grams which explicitly state or imply/infer (weakly or strongly) some consequence for the reader for not following the desired action articulated or implied by the sender.
Authority terms are often found in the concluding parts of phishing emails in an effort to (re-)emphasize to the reader the importance of acting. Examples are:

- unless you
- if we do not
- you fail to
- is required
- be advised that
- if you choose to ignore

4.1.10 TemporalTerm

Many phishing emails seek to convey the need for immediate action as soon as the reader has seen and read the email. Many thus use language which has temporal dimensions, usually referring to deadlines for action, correction or consequences, as well as to emphasize scarcity. This feature seeks to capture n-grams which reflect this time-pressure purpose. Examples are:

- as soon as possible
- on or before
- immediately

4.1.11 FutureImpactTerm

There are a number of distinct n-grams conveying the potential or reality of some impact or effect in the future. A threat may be conveyed on the basis of these n-grams alone, or in combination with a further state based term. The FutureImpactVerb below, although similar, does not have this property. Examples are:

- can result in
- may impact
- may cause
- will halt

It appeared impractical to reduce some of these to merely "will" or "may" given the high prevalence of those words in English language use generally. The key use of this feature seemed to be its combination with an instance of the ActionStateVerb feature.

4.1.12 FutureImpactVerb

This feature reflects a threat agnostic reference to some state change of something which precedes it by way of the infinitive verb that follows it. Examples are:

- can be
- would be
- should not be

The feature is usually found with a ReaderAttribute and/or an ActionStateVerb.

4.1.13 PleaseTerm

As discussed in the last chapter, the word "*please*" plays a role in persuasion and there are some differences in its presence between the phishing and non-phishing corpora.

4.2 Interplay with Persuasion Motivation Sequence

A graphical representation of the features and their primary interplay with the components of the Persuasion Motivation Sequence is presented in Figure 4.1.

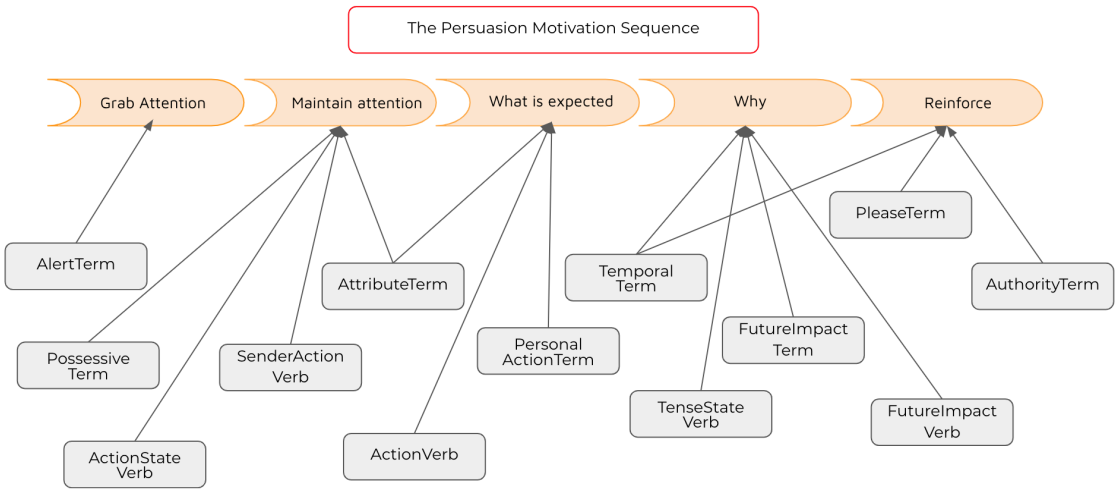

Figure 4.1: Primary Feature Interplay with the Persuasion Motivation Sequence

4.3 A Related Language Feature: EmailSize

As noted there was a notable difference in the average length of phishing and non-phishing emails. Although not a direct indication of persuasion in the language, this aspect might be helpful in separating between phishing and non-phishing emails in certain cases and might also serve to help explore the utility of combining persuasion based features with other features in emails.

The feature is represented as an integer value denoting the character length of the body of the email.

4.4 Discarded Candidate Features

As noted in the previous chapter, a number of candidate features which appeared promising at first, were discarded as a result of the iterative feature engineering process. A brief discussion with respect to a few of the more interesting candidates follows.

4.4.1 AlertTermInSubject

There appeared to be a notable use of AlertTerm instances in the subject line of phishing emails. In most cases, the same term was also repeated in the body of the email. The separation of the presence of these terms between the subject line and body text, by way of distinct features, appeared to not matter. They were thus combined to provide a stronger AlertTerm signal as a whole.

4.4.2 SocialMediaTerm

A long list of social media related terms was curated. These were very specific and not readily conformable to the persuasion based approach pursued, and were hard to accommodate into the PMS. In addition, this feature might end up confusing the model by mis-classifying email notifications from social media providers (which also use persuasive strategies to entice their users to check their profile, etc.) as phishing. Given that, as well as the general focus on "business" emails, these were excluded.

It may be possible to conduct additional analysis into how phishing email attacks are perpetrated along the social media vector in the business environment, to inform the curation of better features for that use case.

4.4.3 CorporateTerm

A long list of corporate related language, including corporate entity names, was substantially trimmed and folded into the AttributeTerm and AuthorityTerm features, after realizing that the persuasive aspects of these language snippets were more effectively covered under those features.

4.4.4 EmploymentAttributeTerm

Given the high prevalence of employment and income related terms generated by the initial review, a category for these had initially been kept separate. These terms were later combined with the general AttributeTerm feature to enable better modeling on the basis of the general reader attribute role of these terms in the PMS.

4.4.5 EmotiveTerm

In light of some of the "softer" persuasion strategies which do not rely on an authority or scarcity based paradigm, a number of terms expressing emotive language ("happy", "pleased", etc.) were initially identified as candidates. These were discarded as these terms were also widely present in the non-phishing corpus, and therefore did not appear to add any meaningful horsepower to the model.

4.4.6 Persuasion Strategy Candidates

An effort was undertaken to seek to distinguish among differing persuasion strategies based on the language used with the goal of developing some identifiable taxonomy. After an extended effort to find patterns in the emails concordant with such an approach, as well as an effort to frame out a potential ontology, it became clear that this approach was not workable. This was for two primary reasons. First, the candidate artefacts identified were too specific. Phishing emails tend to combine various

persuasion strategies in their efforts to persuade, creating a "persuasion artefact patchwork". It is easy to see the persuasion artefacts in the emails, but it is difficult to clearly trace these back to *separate* persuasion strategies to enable classification on such a fine-grained basis.

The better approach is the general persuasion based feature engineering approach reflected in this thesis and the use of machine learning to demonstrate its efficacy.

4.5 Constructions

Using the 13 features described above, several email phishing domain specific, and more complex constructions can be readily identified. Three simple examples are set out below. This provided comfort that the selected features could reflect more complex persuasion constructions and that a machine learning algorithm could assist further in finding signature combinations of these features in phishing emails, in a manner similar to that achieved with identifying causation in [DLC17].

4.5.1 Example 1

A SenderActionVerb followed by an ActionStateVerb, such as "*we have suspended*", occurs quite often in phishing emails particularly those with an authoritative bent, to clearly indicate to the reader that the sender has identified a certain state, with which the reader needs to be concerned for one reason or another.

4.5.2 Example 2

A PersonalActionTerm followed by an ActionVerb, such as "*you must update*", constitutes a clear imperative directive statement to the reader.

4.5.3 Example 3

PersonalActionTerm is often followed by TemporalTerm in phishing emails. For example "*respond as soon as possible*" or "*download immediately*".

4.5.4 Feature Co-occurence

The curated features are postulated to be the building blocks of the persuasive components of the Persuasion Motivation Sequence. They should thus naturally occur together if they are used towards a persuasive purpose. In order to gain some insight into how the proximity of certain features within an email's text might correlate with its "phishiness", and thus confirm that these features provided a usable "phishiness signature", a feature co-occurrence analysis was undertaken on the Sample Corpora.

A first analysis examined whether a feature co-occurred with another feature immediately before it or after it, once extracted from an email. A 3-D rendering of this co-occurrence for phishing emails is set out in Figure 4.2. A similar rendering for non-phishing emails is set out in Figure 4.3.

It is evident that PossessiveTerm and AttributeTerm have significant co-occurrence with each other, as well as with ActionVerb. In the non-phishing corpus, the more generic TenseStateVerb has a high co-occurrence with itself as well as several of the other features. This is likely the result of these types of n-grams being found several times in large segments of normal text. A co-occurrence analysis of the much larger Podesta corpus generated a very similar signature. See Figure 4.4.

A second analysis explored the prevalence of 3 feature combinations together in the corpora. This analysis revealed a marked difference in the feature sequences prevalent in the phishing emails versus the non-phishing emails. Within the top sequences

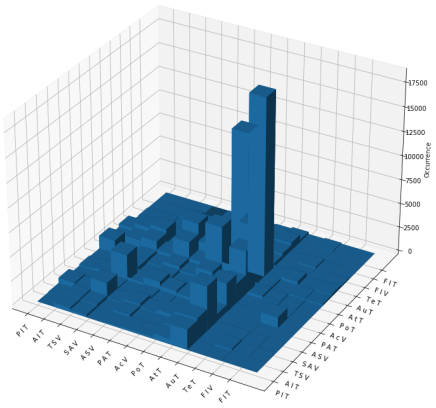

Figure 4.2: Sample Corpus Phishing Email Feature Co-occurrence

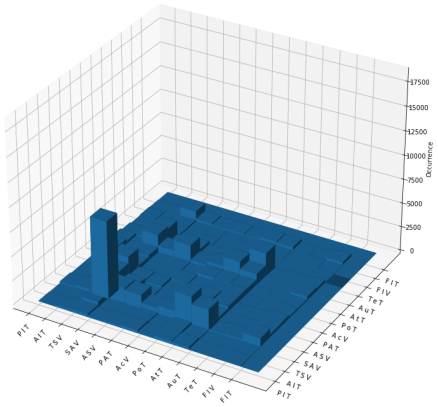

Figure 4.3: Sample Corpus Non-Phishing Email Feature Co-occurrence

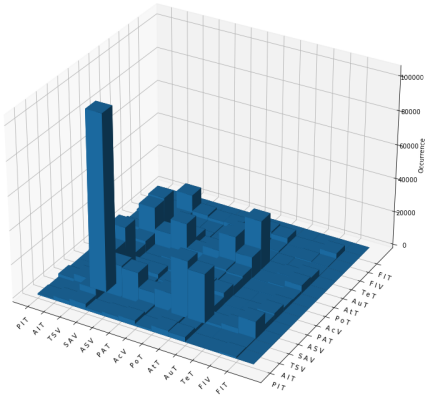

Figure 4.4: Podesta Corpus (Non-Phishing) Email Feature Co-occurrence

found in the phishing corpora, a number of clear constructions are evident. 4.1 sets out the top 10 3-feature sequences found in the phishing corpora, together with some examples of the types of constructions these feature combinations embody. The frequency of these in the phishing and non-phishing parts of the Sample Corpora is also set out. The contrast in frequency is evident.

This provided comfort that the identified features would enable classification between phishing and non-phishing emails.

Table 4.1: Top 10 Constructions Evident in the Sample Phishing Corpora

Rank	Feature Combination	Phishing	Non-Phishing	Example
1	ActionVerb + PossessiveTerm + AttributeTerm	4027	242	[you can] update your password
2	AttributeTerm + ActionVerb + PossessiveTerm	1761	97	[to access your] account restore your . . .
3	PossessiveTerm + AttributeTerm + TenseStateVerb	1661	317	your account has been . . .
4	ActionStateVerb + PossessiveTerm + AttributeTerm	1634	209	. . . cancelled your creditcard
5	AttributeTerm + TenseStateVerb + ActionStateVerb	1491	272	payment has been blocked
6	PossessiveTerm + AttributeTerm + ActionVerb	1211	101	your password [is there to] protect you
7	PossessiveTerm + AttributeTerm + PersonalActionTerm	1003	78	[to view] your reimbursement you can
8	ActionStateVerb + AttributeTerm + PossessiveTerm	883	38	login [to the] account [with] your
9	TenseStateVerb+ PossessiveTerm + AttributeTerm	861	206	[this] could be your information
10	AlertTerm + PossessiveTerm + AttributeTerm	855	21	attention, your document [is ready]

4.6 Implementation

The implementation of the model was done in two phases. Phase 1 involved the conversion of the corpora into data sets of email instance feature information for further processing, using the Phishalyzer pipeline developed for this thesis project.

Phase 2 involved additional data preparation and testing of various machine learning based approaches on the test data. A high level visual representation of the implementation work flow is set out in Figure 4.5.

Additional technical details of the implementation are set out in Appendix B.

4.7 Summary

This chapter summarized the 13 persuasion focused features curated using the feature engineering methodology set out in the previous chapter. It also briefly outlined

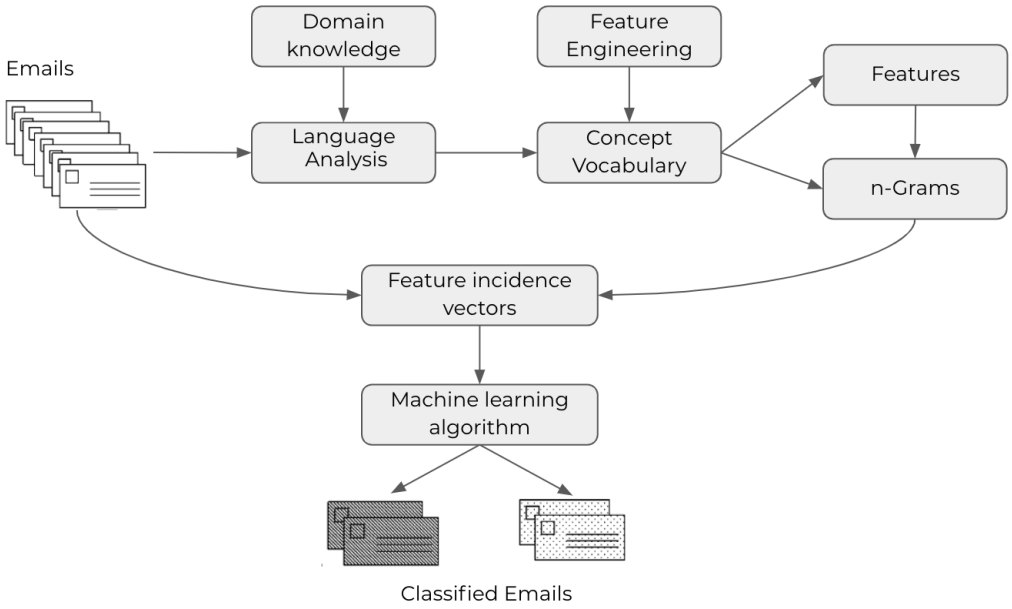

Figure 4.5: High Level Workflow Overview

the implementation process, more fully described in Appendix B. The next chapter summarizes and offers an analysis and discussion of the results obtained.

Chapter 5

Results and Discussion

5.1 Introduction

The efficacy of the identified features in classifying phishing emails was tested using a machine learning approach.

Machine learning can be defined as the process of solving a practical problem by gathering a data set and algorithmically building a statistical model based on that data set which answers the question posed by the problem [Vem20].

The problem of distinguishing between a phishing and non-phishing emails is a *binary classification* type problem where a yes/no answer to the question "is this a phishing email" is sought to be produced by the model.

The usual approach in binary classification is the generation of a data set of labeled instances, with each instance's features reflected by way of a numeric feature vector. An example email, marked up with identified features, from which such a feature count vector is derived, is presented in Figure 5.1.

Figure 5.1: An Example Phishing Email Marked Up With Features

The data set is then split into a training set and a test set, usually at a ratio of 80/20 or 70/30. The machine learning algorithm is applied to the learning set and the efficacy of the generated model determined by its success in classifying the test set.

As is evident from the literature review, several supervised machine learning algorithms are often used in binary classification problems generally, and phishing email classification specifically. The ones most prevalent in the papers cited were selected for testing, and are briefly summarized below.

5.2 Classification Algorithms Selected

5.2.1 Naïve Bayes

Naïve Bayes is a fast classification algorithm based on Bayes' Theorem[1] which describes the probability of an event given prior knowledge of conditions that might be related to the event. In the context of machine learning a Naïve Bayes classifier considers every feature as contributing *independently* to the probability of an event, that is, without any consideration of any correlations between them. This is of course not always the case, hence the "Naïve". It is thus generally less effective in generating good models where inter-feature connections are important to the problem being modeled.

5.2.2 Logistic Regression

A Logistic Regression algorithm seeks to find a regression function (a sigmoid or "S" curve) which optimally separates the feature data along a binary outcome (e.g. yes/no), by assigning weights to each of the various features. Logistic Regression works well if there are no outliers in the data and where there is little correlation among the features.

5.2.3 J48 Decision Tree

In a decision tree the training data is split (or branched) along the feature values using a tree structure. Features which seem to split the data the best (a property which is usually referred to as the "split quality"), appear at the top and those which add less to the mix are lower in the tree. At a branch, a threshold value for how the feature contributes to the classification is determined. New nodes are added in this manner to refine the classification and once a node resolves to one of the classification

[1] `https://en.wikipedia.org/wiki/Bayes%27_theorem`

states, no further nodes are added. Decision trees are easy to interpret but are more susceptible to the influence of outliers.

5.2.4 Random Forest

A Random Forest algorithm uses multiple decision trees (hence the forest) that operate as an *ensemble*. In essence each of the trees generates a prediction based on the feature values, and the prediction with the highest prevalence constitutes the classification. Different trees are generated by allowing each tree to sample differing feature data from the data set. This approach of using many trees to generate a result as a "committee" outperforms single instances of decision trees in many instances.

5.2.5 Support Vector Machine

The Support Vector Machine (SVM) is widely used in pattern recognition and classification problems. It seeks to find a "hyper plane" of a dimension less than the dimensions of the feature set (i.e. the number of features) which optimally separates feature values according to the classification outputs. SVMs are memory-intensive, hard to interpret, and difficult to tune.

5.2.6 Metrics

Machine learning model efficacy for classification problems is usually assessed using the Precision and Recall metrics, as well as the F1 Score[2], all of which are calculated from the following result categories:

1. **True Positives** (TP): The number of correct positive classification predictions.
2. **True Negatives** (TN): The number of correct negative classification predictions.

[2] `https://en.wikipedia.org/wiki/Precision_and_recall`

3. **False Positives** (FP): The number of incorrect positive classification predictions that should have been a negative.

4. **False Negatives** (FN): The number of incorrect negative classification predictions that should have been a positive.

The TP, TN, FP and FN numbers from a particular prediction effort are generally displayed as a Confusion Matrix:

$$\begin{vmatrix} TP & FP \\ FN & TN \end{vmatrix}$$

The obvious goal is to reduce FN and FP as much as possible.

5.2.6.1 Precision

The Precision metric is calculated as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{5.1}$$

Given that the numerator is the total number of instances predicted as positive, improving Precision is helpful if the cost of false positives is high.

5.2.6.2 Recall

The Recall metric is calculated as follows:

$$Recall = \frac{TP}{(TP + FN)} \tag{5.2}$$

Given that the numerator is the total number of both positive and negative instances predicted correctly, improving Recall is useful when the cost of false negatives is high.

5.2.6.3 The F1 Score

The F1 Score seeks to find a balance between Precision and Recall, particularly in cases where the real world incidence rate for the problem has a large number of actual negatives. The F1 score is calculated as follows:

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)} \tag{5.3}$$

5.2.6.4 Accuracy

A general overall Accuracy score is also sometimes helpful. Accuracy is calculated as follows:

$$Accuracy = \frac{(TN + TP) * 100}{(TP + TN + FP + FN)} \tag{5.4}$$

5.2.7 Test Corpus

For the purpose of testing, a representative sample corpus (n=7,616) was prepared, made up of a cross section (n=3,757) of all the phishing emails, combined with a random sample (n=3,859) of non-phishing emails from the Podesta Emails. Feature count vectors were generated for all individual email instances in the sample corpus which were then also labeled as being either phishing or non-phishing, using the Phishalyzer application.

5.3 Classification Based on Feature Cardinality

A set of modeling experiments were completed by importing the data sets into several Jupyter Notebooks and applying the selected machine learning algorithms. A 70/30 training/test split of the data set was employed.

5.3.1 Test Corpus Feature Correlation

Given that several of the machine algorithms perform better if features are not substantively correlated, a feature correlation analysis of the data set was performed. This confirmed that there were no directly or significantly correlated features. See the correlation matrix in Figure 5.2.

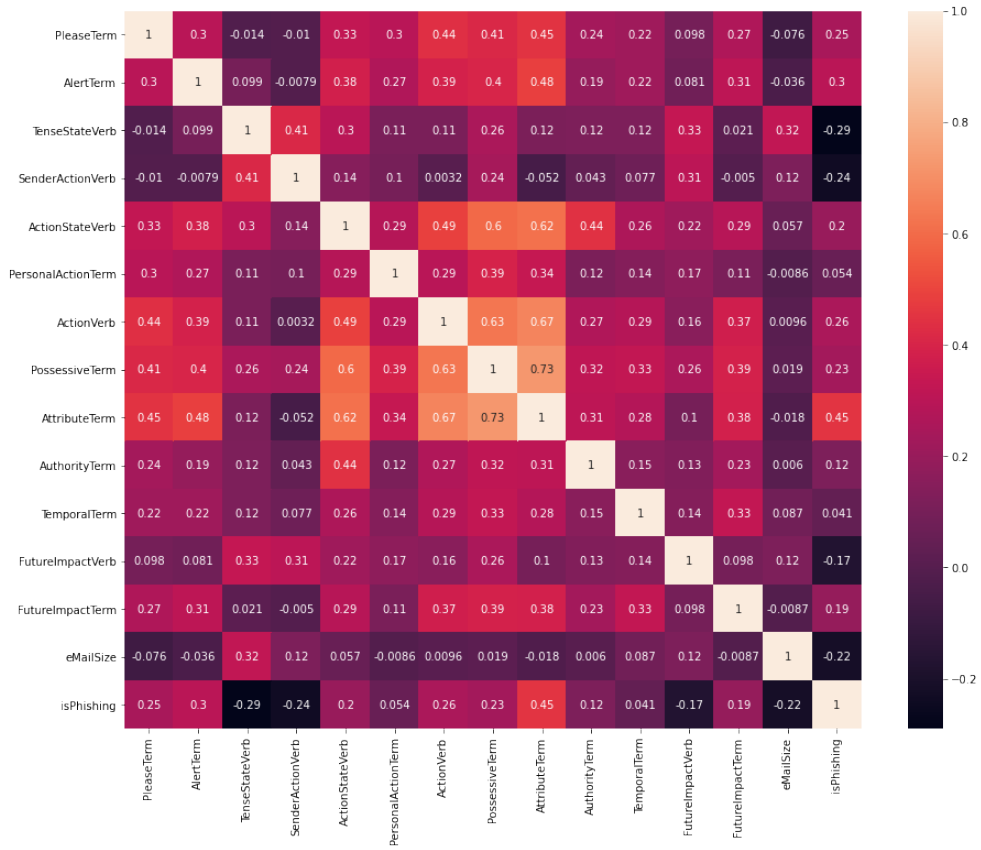

Figure 5.2: Test Corpus Feature Correlation Matrix

5.3.2 Classification with All Language Features

The combination of all 13 language specific features, i.e. excluding EMailSize, generated accuracy results between 76.890 (Naïve Bayes) and 87.462 (Random Forest) for the Test Corpus. The J48 Decision Tree generated significantly more False Positives, where Naïve Bayes generated significantly more False Negatives. The overall results

of these tests are summarized in Table 5.1

Table 5.1: Metrics for "All Language Features" Models

Algorithm	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
Naïve Bayes	902	73	303	349	0.925	0.749	0.828	76.890
J48 Decision Tree	820	155	151	501	0.841	0.844	0.843	81.192
Logistic Regression	892	83	184	468	0.915	0.829	0.870	83.589
Support Vector Machine	923	52	184	468	0.947	0.834	0.887	85.495
Random Forest	917	58	146	506	0.941	0.863	0.900	87.462

5.3.2.1 Feature Importance

In order to gain some insight into the relative importance of the features in the context of the Random Forest algorithm, a feature importance analysis was performed. This revealed that AttributeTerm, TenseStateVerb, and PossessiveTerm play a major role in the classsifation. See Figure 5.3.

This was consistent with the results of the co-occurrence analysis previously conducted, and provided support for the conclusion that the identified features are capable of constituting themselves as the potential constructions identified in that analysis.

In addition, they appear to provide a degree of "explainability" in that features representing components of the Persuasion Motivational Sequence are useful in classifying between phishing and non-phishing emails.

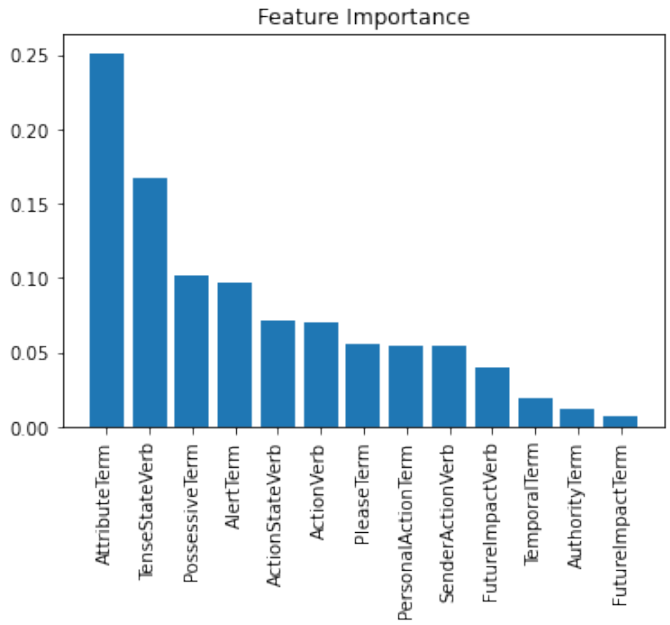

Figure 5.3: RF Feature Importance Analysis Results (All Language Features)

5.3.3 Classification without AttributeTerm

In an effort to gain additional insight into the relative role of AttributeTerm, modeling was pursued without that feature for all algorithms. This removal did degrade performance slightly for Naïve Bayes and by about 3 to 4 points for the other algorithms. The prediction results for this test are set out in Table 5.2.

Table 5.2: Metrics for "All Language Features" Models Without AttributeTerm

Algorithm	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
Naïve Bayes	882	61	309	258	0.935	0.741	0.827	75.497
J48 Decision Tree	782	161	171	396	0.829	0.821	0.825	78.013
Logistic Regression	845	98	226	341	0.896	0.789	0.839	78.543
Support Vector Machine	892	51	226	341	0.946	0.798	0.866	81.656
Random Forest	858	85	163	404	0.910	0.840	0.874	83.576

5.3.3.1 Feature Importance

The relative feature importance for this feature set (without AttributeTerm present) within the Random Forest model was again determined. A graph displaying the rankings of the remaining features is presented in Figure 5.4. Five of the 12 features, TenseStateVerb, PossessiveTerm, ActionStateVerb, AlertTerm and ActionVerb now contribute well over 60 percent to the classification.

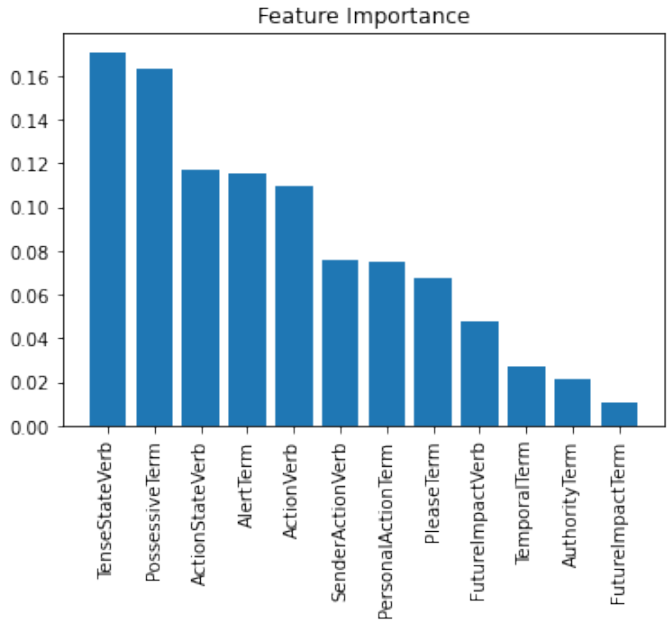

Figure 5.4: RF Feature Importance Analysis Results (All Language Features without AttributeTerm)

5.3.4 Adding the EmailSize Feature

As previously discussed, the EMailSize feature was added as a non-language feature with the expectation it might be useful to separate edge cases on the premise that longer emails are more likely to be legitimate, and as a way to test the utility of combining a persuasion language based model with non-text features.

Adding the EmailSize feature to the model did improve results by a few additional points, compared to the data set comprised of all language features (including AttributeTerm). The prediction results for these 14 features together is set out in Table 5.3.

Table 5.3: Metrics for the All Language Features Models and EMailSize

Algorithm	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
Naïve Bayes	1009	86	308	461	0.921	0.766	0.837	78.863
J48 Decision Tree	965	130	137	632	0.881	0.876	0.878	85.676
Logistic Regression	954	141	122	647	0.871	0.887	0.879	85.891
Support Vector Machine	1053	42	174	595	0.962	0.858	0.907	88.412
Random Forest	1045	50	113	656	0.954	0.902	0.928	91.255

5.3.4.1 Feature Importance

A feature importance analysis for this data set was also conducted and confirmed *AttributeTerm* as still the primary driver, with *EMailSize* and *TenseStateVerb*, and to a lesser degree the single word "*please*", being also important features. The word "*please*" appears to be somewhat more important in the presence of EmailSize. See Figure 5.5.

It is notable that the interplay of AttributeTerm and TenseStateVerb is consistent with the types of constructions anticipated in the co-occurrence analysis at the end of Chapter 4.

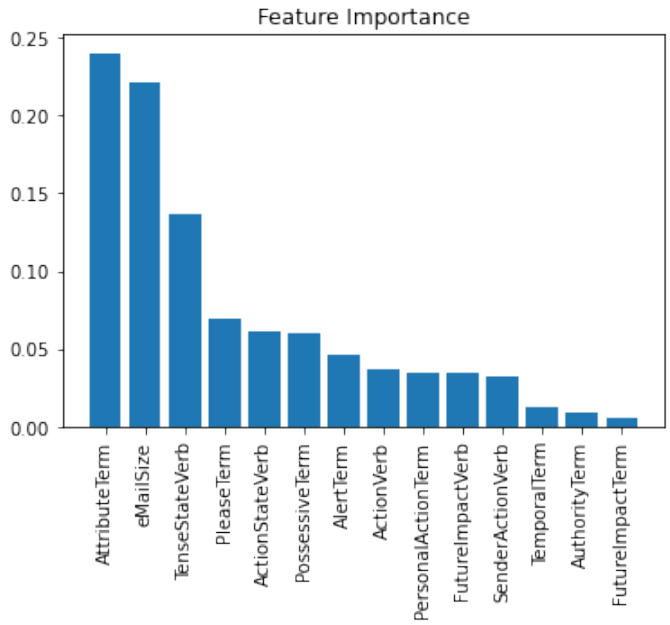

Figure 5.5: Random Forest Based Feature Importance Analysis Results (All Features + EMailSize)

5.3.5 Optimizing the Random Forest Classification

The above efforts pointed to the Random Forest algorithm as the best candidate for modeling based on the selected feature sets. An effort to optimize the implementation for this algorithm was thus pursued.

Optimizing the results from a particular algorithm strategy can be achieved by seeking to tune the various parameters for the model provided by the API, in this case SciKitLearn. This can be automated with a script which tries a large set of random permutations within a range of various input parameters. See Figure 5.6 for a short Python script to that effect.[3]

This strategy improved Accuracy to 91.577 using the following algorithm parameters:

[3]Adapted from `https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74`

```
# import the RF Random Search Cross Validation model
from sklearn.model_selection import RandomizedSearchCV
# set the number of trees in the random forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
# set the number of features to consider at every split
max_features = ['auto', 'sqrt']
# set the maximum levels in each tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# set the minimum number of samples required to split a node and at each leaf node
min_samples_split = [2, 5, 10]
min_samples_leaf = [1, 2, 4]
# set the method of training each tree
bootstrap = [True, False]
# create the random grid object
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
# display it so we can see what things look like before we pull the trigger
pprint(random_grid)
```

Figure 5.6: Example Random Forest Parameter Optimization Script using the SciKitLearn Cross Validation Model

```
'bootstrap': True, 'max_depth': 20, 'max_features': 4,
'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 1500
```

.

A summary of the results from the optimization effort for the three different data sets is set out in Table 5.4.

Table 5.4: Metrics for Optimal Random Forest Models for Various Data Sets

Data set	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
No AttributeTerm	862	81	162	405	0.914	0.842	0.876	83.907
All Language	920	55	143	509	0.944	0.865	0.903	87.830
All and EmailSize	1045	50	107	662	0.954	0.907	0.930	91.577

5.4 Classification with Feature N-grams

To gain insight into how well the identified 13 features constituted an effectively classified abstraction of the various types of n-grams identified, a classification based solely on the presence of the individual feature n-grams was conducted. To that end, another data set comprised of a concatenated sequential string representation of each email's n-grams was generated using the Phishalyzer application. A sample n-gram from each type of email is provided for illustration:

```
phishing_1220,1,account alert discovered unusual activity
your account unauthorized unauthorized your we have temporarily
suspended account verify you will will not be debit account you
need to please as soon as possible your let us know account was
your find online online review verify account your account number

podesta_006250,0,is would be sent my you have please please i
am sent update will be closed your regularly sent you are currently
assistance please your as soon as possible will be regularly
please form your is will be chance enrolled may result in your
your patience your sent
```

These n-gram sequences were imported into a Jupyter Notebook and then converted to numeric input, usable by the machine learning algorithms, using three different vectorizers:

- A *count* vectorizer, which counts the number of times a token shows up in the document and uses this value as its weight (essentially the same approach used in vectorizing the features for the testing in the previous section).
- A *hashing* vectorizer, which uses a hash function to directly map n-grams to their frequency.
- A *TF-IDF* vectorizer, which assigns a weight to an n-gram depending on its frequency in the email and its prevalence in the data set generally.

A 70-30 training/test distribution was again used. The results from these tests are set out in Table 5.5. The classification accuracy improved by 4 to 6 percent for all algorithms used.

Table 5.5: Metrics for Vectorized N-gram based Models

Algorithm	Vectorizer	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
Naïve Bayes	Hashing	1039	90	277	875	0.920	0.790	0.850	83.911
Naïve Bayes	TF-IDF	931	222	121	1007	0.807	0.885	0.844	84.963
Support Vector Machine	Count	1161	7	335	778	0.994	0.776	0.872	85.007
Naïve Bayes	Count	1108	59	210	904	0.949	0.841	0.892	88.207
Logistic Regression	TF-IDF	1057	102	161	961	0.912	0.868	0.889	88.470
Logistic Regression	Count	1015	133	109	1024	0.884	0.903	0.893	89.391
Support Vector Machine	Hashing	1058	92	141	990	0.920	0.882	0.901	89.785
Random Forest	Count	1029	100	129	1023	0.911	0.889	0.900	89.961
Support Vector Machine	TF-IDF	1085	50	162	984	0.956	0.870	0.911	90.706
Random Forest	TF-IDF	1122	65	120	974	0.945	0.903	0.924	91.890
Logistic Regression	Hashing	1095	52	137	997	0.955	0.889	0.921	91.714
Random Forest	Hashing	1094	59	114	1014	0.949	0.906	0.927	92.416

5.5 Discussion

The persuasion language based features identified in this thesis are able to predict whether an email constitutes phishing or not with a good deal of accuracy. This supports the conclusion that persuasion language is detectable as an important, and distinguishing, component of a phishing email.

As a whole, the set of persuasion language related features performed better than

the limited verb-object based feature sets assessed in [Fal16] and [Par18], without the overhead of needing to manually curate the feature representation into a complex Text Meaning Representation (TMR).

The scope of phishing emails sought to be detected in the approach of this thesis has also not been limited as it was in ([Fal16] specifically limited the examination to emails having PDF attachments, and [Par18] limited his examination to the semantics surrounding 50 identified verbs).

With respect to the machine learning algorithms used, Random Forest performed best. Support Vector Machine based results were also very good and demonstrated the highest Precision metrics throughout, i.e. generated the lowest number of false positives. Random Forest's Accuracy was, however, the highest, primarily because it consistently generated the *lowest* number of false negatives. The question of why a Support Vector Machine based approach had higher Precision than that of a Random Forest algorithm was identified but not further investigated. Given that the cost of wrongly classifying a phishing email as non-phishing is greater than classifying a non-phishing email as phishing, Random Forest appears to be the preferred algorithm for implementation of the classification model set out in this thesis, at least among the algorithms used.

Although not reflected in the results above, it was observed during testing that the execution run time for the various algorithms was very reasonable and easily managed on a consumer grade computer.

5.5.1 Feature Specific Observations

A number of features stood out as more important than others.

First, AttributeTerm has significant importance in the Random Forest driven models. This is consistent with the general focus of phishing emails around some aspect of the reader which is sought to be accessed or is referenced to seek to coerce the reader into action. Without some *object of interest* to the reader to compel action, persuasion is unfocused. The prevalence of these (necessary) types of terms in emails is reflected in the importance of the feature which represents them. The accessibility of domain informed explainability of the importance of this feature is a pleasant effect of having sought to embed domain knowledge in the features of the model.

Second, ActionStateVerb is also an important feature. This is particularly interesting in light of the fact that, with the exception of a handful of n-grams, these are all constituted as past tense versions of verbs ending in "ed". The importance of this feature emphasizes the relevance of step 2 of the Persuasion Motivation Sequence - the need to establish some sort of recently occurred "problem state".

The past tense dimension of this feature set appears to be important. This supports the conclusion that stemming without forethought may detrimentally affect the richness of certain language based features like this. The temporal-spatial aspect of this feature, articulating an act or state change in the (immediate) past would have been lost with stemming or lemmatization.

Third, the addition of the EMailSize feature improved the overall accuracy of every algorithm used. Notwithstanding the goal of this thesis to focus exclusively on persuasion language based features, the significance of using one additional contextual

feature bodes well for improving performance by adding others. This is also consistent with the findings by [Par18] that more diverse information mediates toward improved results.

Lastly, the alignment of the importance of certain features with the prevalence from the co-occurrence analysis summarized in Chapter 5, particularly those incorporating AttributeTerm, PossessiveTerm and TenseStateVerb features, would appear to support the conclusion that the machine learning algorithms are identifying persuasion related construction patterns in the text of phishing emails, consistent with the Persuasion Motivation Sequence.

5.5.2 Feature N-grams

The testing of the n-grams themselves as individual features generated a modest additional increase in performance. This suggests that abstraction of these n-grams into the 13 features might perhaps be improved with additional tuning of these features, such as splitting certain features into sub-categories or perhaps combining others. A deeper exploration of the science of linguistics would benefit the ability to generate additional insights into this issue.

5.6 Summary

In this chapter the results from several machine learning algorithm based implementations was summarized. The results demonstrated the efficacy of modeling Persuasion Motivation Sequence based language features in emails with the Random Forest algorithm.

As a whole, prediction results were consistent with, if not slightly better than, those

from [Fal16] and [Par18], without limiting the scope of emails able to be tested, or any of the data preparation overhead.

A number of specific observations were made about the importance of certain features and the alignment of the co-occurrence analysis with the importance of these features in the predictability of the model used.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This work confirms the efficacy of curating *Persuasion Motivation Sequence* informed language features from the text of phishing emails for use in phishing email detection strategies. The benefits of a thorough domain exploration of the various dimensions of the phishing email domain towards that goal was demonstrated.

As a whole, performance of this limited set of language only based features is not in the realm of many of the existing structural feature driven approaches, but that was not postulated nor expected.

The key contribution from this work is the identification of a combination of persuasion based language features as an additional arrow in the quiver of phishing email detection strategies.

6.2 Future Work

Although the set of features is broader than those used in [Fal16] and [Par18], they are unlikely to be linguistically comprehensive. Additional study of the linguistic foundations of persuasive language, and perhaps potential collaboration with linguists who are experts in written persuasion is likely to enable better, or perhaps different or additional, persuasion driven features.

The development of a more structured modeling of persuasion by way of an ontology (such as the OST referenced in the thesis) may provide a more robust framework for parsing the linguistic components of persuasion in written text. The construction of an effective persuasion text parser to match such an ontology would be beneficial.

The development of a comprehensive "persuasiveness" feature implementation to add to existing feature sets used for detection may improve the detection capacity of those approaches. The potential implementation of this as a web service would likely be very usable for researchers working on phishing detection strategies.

Study of the specific types of persuasion language used in spear phishing attacks would be beneficial and data sets for such work need to be developed.

There are also a number of potential refinements to the approach set out in this thesis which might improve performance further. For example, there may be benefits in exploring the presence and removal of domain specific stop words which have no bearing on model efficacy, in a manner consistent with the work in [AAH$^+$20]. In addition, the development of richer n-gram lists for the various language concept features is likely possible by examining additional and more diverse corpora.

There may also be further opportunity for improving classification performance by way of developing an *ensemble* Random Forest - Support Vector Machine based classification approach, with the aim of relying on the Random Forest approach to reduce false negatives and the Support Vector Machine approach to reduce false positives. There appears to be some precedent for this approach in the field of land-cover classification using remote sensing image data [ZZMIV17].

Bibliography

[AA14] Andronicus A Akinyelu and Aderemi O Adewumi. Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014, 2014.

[AAH+20] Farah Alshanik, Amy Apon, Alexander Herzog, Ilya Safro, and Justin Sybrandt. Accelerating text mining using domain-specific stop word lists, 2020. arXiv:2012.02294.

[Aba14] Jemal Abawajy. User preference of cyber security awareness delivery methods. Behaviour & Information Technology, 33(3):237–248, 2014.

[ACPZ19] Luca Allodi, Tzouliano Chotza, Ekaterina Panina, and Nicola Zannone. The need for new antiphishing measures against spear-phishing attacks. IEEE Security & Privacy, 18(2):23–34, 2019.

[Aga20] Alessandro Ecclesie Agazzi. Business email compromise (bec) and cyberpsychology, 2020. arXiv:2007.02415.

[Akb14] Nurul Akbar. Analysing persuasion principles in phishing emails. Master's thesis, University of Twente, 2014.

[AKBG+11] Pranav Anand, Joseph King, Jordan L Boyd-Graber, Earl Wagner, Craig H Martell, Douglas W Oard, and Philip Resnik. Believe me-we

can do this! annotating persuasive acts in blog text. In Computational Models of Natural Argument, 2011.

[AKS14] Shivam Aggarwal, Vishal Kumar, and SD Sudarsan. Identification and detection of phishing emails using natural language processing techniques. In Proceedings of the 7th International Conference on Security of Information and Networks, pages 217–222, 2014.

[Alt16] Sahar Altikriti. Persuasive speech acts in barack obama's inaugural speeches (2009, 2013) and the last state of the union address (2016). International Journal of Linguistics, 8(2):47–66, 2016.

[BBM20] Andrea Borghesi, Federico Baldo, and Michela Milano. Improving deep learning models via constraint-based domain knowledge: a brief survey. arXiv preprint arXiv:2005.10691, 2020.

[BCP+08] Andre Bergholz, Jeong Ho Chang, Gerhard Paass, Frank Reichartz, and Siehyun Strobel. Improved phishing detection using model-based features. In CEAS, 2008.

[BDH+11] Michael W Boyce, Katherine Muse Duma, Lawrence J Hettinger, Thomas B Malone, Darren P Wilson, and Janae Lockett-Reynolds. Human performance in cybersecurity: A research agenda. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 55, pages 1115–1119. SAGE Publications Sage CA: Los Angeles, CA, 2011.

[Ben13] Emily M Bender. Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. Synthesis Lectures on Human Language Technologies, 6(3):1–184, 2013.

[BFSS20] Eduardo Benavides, Walter Fuertes, Sandra Sanchez, and Manuel Sanchez. Classification of phishing attack solutions by employing deep learning techniques: A systematic literature review. Developments and advances in defense and security, pages 51–64, 2020.

[BL76] Samuel B Bacharach and Edward J Lawler. The perception of power. Social Forces, 55(1):123–134, 1976.

[Blo19] The Phishlabs Blog. Phishing Number One Cause of Data Breaches: Lessons from Verizon DBIR. "https://info.phishlabs.com/blog/phishing-number-1-data-breaches-lessons-verizon", 2019. Accessed: 2020-12-23.

[BNBW19] Sikha Bagui, Debarghya Nandi, Subhash Bagui, and Robert Jamie White. Classifying phishing email using machine learning and deep learning. In 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), pages 1–2. IEEE, 2019.

[BPP+17] Marcus A Butavicius, Kathryn Parsons, Malcolm R Pattinson, Agata McCormac, Dragana Calic, and Meredith Lillie. Understanding susceptibility to phishing emails: Assessing the impact of individual differences and culture. In HAISA, pages 12–23, 2017.

[Bre20] Robert Lee Brewer. Semantics vs. Syntax vs. Pragmatics (Grammar Rules). "https://www.writersdigest.com/write-better-fiction/semantics-vs-syntax-vs-pragmatics-grammar-rules", 2020. Accessed: 2020-11-21.

[Bro18] Hannah S Brooks. Linguistic Persuasion Techniques in Phishing Emails: A Corpus and Critical Discourse Analysis. Hofstra University, 2018.

[BSMK16] Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. Frontiers in psychology, 7:1116, 2016.

[BSN19] Maria Bada, Angela M Sasse, and Jason RC Nurse. Cyber security awareness campaigns: Why do they fail to change behaviour? arXiv preprint arXiv:1901.02672, 2019.

[C+18] Gillian Cleary et al. 2018 symantec internet security threat report. `https://docs.broadcom.com/doc/istr-23-2018-en`, 2018. Accessed: 2020-06-21.

[Cal13] Tracey Caldwell. Spear-phishing: how to spot and mitigate the menace. Computer Fraud & Security, 2013(1):11–16, 2013.

[CBB14] Davide Canali, Leyla Bilge, and Davide Balzarotti. On the effectiveness of risk prediction based on users browsing behavior. In Proceedings of the 9th ACM symposium on Information, computer and communications security, pages 171–182, 2014.

[CBD+19] Tom Cuchta, Brian Blackwood, Thomas R. Devine, Robert J. Niichel, Kristina M. Daniels, Caleb H. Lutjens, Sydney Maibach, and Ryan J. Stephenson. Human risk factors in cybersecurity. In Proceedings of the 20th Annual SIG Conference on Information Technology Education, SIGITE '19, page 87–92, New York, NY, USA, 2019. Association for Computing Machinery.

[CFL13] Alexander Clark, Chris Fox, and Shalom Lappin. The handbook of computational linguistics and natural language processing. John Wiley & Sons, 2013.

[CG02] Robert B Cialdini and Noah J Goldstein. The science and practice of persuasion. Cornell Hotel and Restaurant Administration Quarterly, 43(2):40–50, 2002.

[CGR20] Rui Chen, Joana Gaia, and H Raghav Rao. An examination of the effect of recent phishing encounters on phishing susceptibility. Decision Support Systems, 133:113287, 2020.

[Cho55] Noam Chomsky. The Logical Structure of Linguistic Theory. PhD thesis, MIT, 1955.

[Cia16] Robert Cialdini. Pre-suasion: A revolutionary way to influence and persuade. Simon and Schuster, 2016.

[CJ17] Neelam Choudhary and Ankit Kumar Jain. Comparative analysis of mobile phishing detection and prevention approaches. In International Conference on Information and Communication Technology for Intelligent Systems, pages 349–356. Springer, 2017.

[CNU06] Madhusudhanan Chandrasekaran, Krishnan Narayanan, and Shambhu Upadhyaya. Phishing email detection based on structural properties. In NYS cyber security conference, volume 3. Albany, New York, 2006.

[DF18] Mohamed Dahmane and Samuel Foucher. Combating insider threats by user profiling from activity logging data. In 2018 1st International Conference on Data Intelligence and Security (ICDIS), pages 194–199. IEEE, 2018.

[dK19] JJJ Jorine de Koning. Countering persuasion in spear-phishing: The effect of forewarning about persuasion in spear-phishing in combination with providing if-then procedural knowledge on targets' susceptibility to spear-phishing. Master's thesis, Technical University Eindhoven, 2019.

[DL18] Guozhu Dong and Huan Liu. Feature Engineering for Machine Learning and Data Analytics. CRC Press, 2018.

[DLC17] Jesse Dunietz, Lori Levin, and Jaime Carbonell. Automatically tagging constructions of causation and their slot-fillers. Transactions of the Association for Computational Linguistics, 5:117–133, 2017.

[DMM08] Marie-Catherine De Marneffe and Christopher D Manning. The stanford typed dependencies representation. In Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation, pages 1–8, 2008.

[Dom12] Pedro Domingos. A few useful things to know about machine learning. Communications of the ACM, 55(10):78–87, 2012.

[Dun17] Jonathan Dunn. Computational learning of construction grammars. Language and Cognition, 9(2):254–292, 2017.

[Fal16] Courtney Falk. Knowledge modeling of phishing emails. PhD thesis, Purdue University, 2016.

[Fal17] Courtney Falk. Hydra: Towards a parser for ontological semantics technology using genetic algorithms. 2017.

[FBoIF19] United States Federal Bureau of Investigation (FBI). 2019 internet crime report, 2019.

[FBoIF20] United States Federal Bureau of Investigation (FBI). 2020 internet crime report, 2020.

[FCK+03] Edward L Fink, Deborah A Cai, Stan A Kaplowitz, Sungeun Chung, Mark A Van Dyke, and Jeong-Nam Kim. The semantics of social influ-

ence: Threats vs. persuasion. Communication Monographs, 70(4):295–316, 2003.

[FF20] Edwin Donald Frauenstein and Stephen Flowerday. Susceptibility to phishing on social network sites: A personality information processing model. Computers & Security, page 101862, 2020.

[FT19] Ana Ferreira and Soraia Teles. Persuasion: How phishing emails can influence users and bypass security measures. International Journal of Human-Computer Studies, 125:19–31, 2019.

[GC08] ADELE E GOLDBERG and DEVIN CASENHISER. English constructions. The Handbook of English Linguistics, page 343, 2008.

[GDSV+20] Eder Souza Gualberto, Rafael Timoteo De Sousa, Thiago Pereira De Brito Vieira, João Paulo Carvalho Lustosa Da Costa, and Cláudio Gottschalg Duque. The answer is in the text: Multi-stage methods for phishing detection based on feature engineering. IEEE Access, 8:223529–223547, 2020.

[GM12] Günther Grewendorf and Georg Meggle. Speech acts, mind, and social reality: discussions with John R. Searle, volume 79. Springer Science & Business Media, 2012.

[Gol95] Adele E Goldberg. Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press, 1995.

[Gro20] Anti-Phishing Working Group. Phishing activity trends report, 4th quarter 2020. "https://docs.apwg.org/reports/apwg_trends_report_q4_2020.pdf", 2020. Accessed: 2021-03-04.

[GWD17] Sanjay Goel, Kevin Williams, and Ersin Dincelli. Got phished? internet security and human vulnerability. Journal of the Association for Information Systems, 18(1):2, 2017.

[HAK13] Isredza Rahmi A Hamid, Jemal Abawajy, and TH Kim. Using feature selection and classification scheme for automating phishing email detection. Studies in informatics and control, 22(1):61–70, 2013.

[Hei70] David R Heise. Potency dynamics in simple sentences. Journal of Personality and Social Psychology, 16(1):48, 1970.

[HH01] Roland Hausser and R Hausser. Foundations of computational linguistics. Springer, 2001.

[Hil14] Martin Hilpert. Construction grammar and its application to English. Edinburgh University Press, 2014.

[HK17] Felix Haeussinger and Johann Kranz. Antecedents of employees' information security awareness-review, synthesis, and directions for future research. ECIS Proceedings 2017, 2017.

[HMN15] Tzipora Halevi, Nasir Memon, and Oded Nov. Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks. Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks (January 2, 2015), 2015.

[IS19] Rahul Radhakrishnan Iyer and Katia Sycara. An unsupervised domain-independent framework for automated detection of persuasion tactics in text. arXiv preprint arXiv:1912.06745, 2019.

[Jak07] Markus Jakobsson. The human factor in phishing. Privacy & Security of Consumer Information, 7(1):1–19, 2007.

[JM12] Dan Jurafsky and Christopher Manning. Natural language processing. Instructor, 212(998):3482, 2012.

[JTRH19] Helen S Jones, John N Towse, Nicholas Race, and Timothy Harrison. Email fraud: The search for psychological predictors of susceptibility. PloS one, 14(1):e0209684, 2019.

[JZ07] Jing Jiang and ChengXiang Zhai. A systematic exploration of the feature space for relation extraction. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 113–120, 2007.

[KA19] R Kiruthiga and D Akila. Phishing websites detection using machine learning. Int. J. Recent Technol. Eng.(IJRTE), 8, 2019.

[KCA+09] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: a real-world evaluation of anti-phishing training. In Proceedings of the 5th Symposium on Usable Privacy and Security, pages 1–12, 2009.

[KHHW15] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. Advanced social engineering attacks. Journal of Information Security and applications, 22:113–122, 2015.

[KTTZ05] Koen Kerremans, Yan Tang, Rita Temmerman, and Gang Zhao. Towards ontology-based e-mail fraud detection. In 2005 portuguese conference on artificial intelligence, pages 106–111. IEEE, 2005.

[LCE+19] Tian Lin, Daniel E Capecci, Donovan M Ellis, Harold A Rocha, Sandeep Dommaraju, Daniela S Oliveira, and Natalie C Ebner. Susceptibility

to spear-phishing emails: Effects of internet user demographics and email content. ACM Transactions on Computer-Human Interaction (TOCHI), 26(5):1–28, 2019.

[LHWR10] Gastón L'Huillier, Alejandro Hevia, Richard Weber, and Sebastian Rios. Latent semantic analysis and keyword extraction for phishing classification. In 2010 IEEE international conference on intelligence and security informatics, pages 129–131. IEEE, 2010.

[LHY20] Xuecong Lu, Milena Head, and Junyi Yang. The impacts of individual emotional state and emotional framing of phishing attack on susceptibility to phishing: An emotional congruence perspective. In Proceedings of the Nineteenth Annual Pre-ICIS Workshop on HCI Research in MIS, Virtual Conference, December 12, 2020, 2020.

[Loc14] Scott Locklin. Neglected machine learning ideas. `https://scottlocklin.wordpress.com/2014/07/22/neglected-machine-learning-ideas/`, 2014. Accessed: 2020-11-07.

[LSY+16] Bin Liang, Miaoqiang Su, Wei You, Wenchang Shi, and Gang Yang. Cracking classifiers for evasion: a case study on the google's phishing pages filter. In Proceedings of the 25th International Conference on World Wide Web, pages 345–356, 2016.

[Luh58] Hans Peter Luhn. The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2):159–165, 1958.

[LZW20] Xue Li, Dongmei Zhang, and Bin Wu. Detection method of phishing email based on persuasion principle. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), volume 1, pages 571–574. IEEE, 2020.

[MA97] John P Meyer and Natalie J Allen. Commitment in the workplace: Theory, research, and application. Sage, 1997.

[MCB17] Naghmeh Moradpoor, Benjamin Clavie, and Bill Buchanan. Employing machine learning techniques for detection and classification of phishing emails. In 2017 Computing Conference, pages 149–156. IEEE, 2017.

[MF04] Rada Mihalcea and Ehsanul Faruque. Senselearner: Minimally spervised word sense disambiguation for all words in open text. In Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 155–158, 2004.

[MOWB09] Liping Ma, Bahadorrezda Ofoghi, Paul Watters, and Simon Brown. Detecting phishing emails using hybrid features. In 2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pages 493–497. IEEE, 2009.

[MRDM20] Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. Cxgbert: Bert meets construction grammar. arXiv preprint arXiv:2011.04134, 2020.

[MZP+17] Agata McCormac, Tara Zwaans, Kathryn Parsons, Dragana Calic, Marcus Butavicius, and Malcolm Pattinson. Individual differences and information security awareness. Computers in Human Behavior, 69:151–156, 2017.

[OST57] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. The measurement of meaning. University of Illinois press, 1957.

[Par18] Gilchan Park. Towards Ontology-Based Phishing Detection. PhD thesis, Purdue University, 2018.

[PBDL19] Kathryn Parsons, Marcus Butavicius, Paul Delfabbro, and Meredith Lillie. Predicting susceptibility to social influence in phishing emails. International Journal of Human-Computer Studies, 128:17–26, 2019.

[PCA16] Jian Peng, Kim-Kwang Raymond Choo, and Helen Ashman. User profiling in intrusion detection: A review. Journal of Network and Computer Applications, 72:14–27, 2016.

[PF20] Heather J Parker and Stephen V Flowerday. Contributing factors to increased susceptibility to social media phishing attacks. South African Journal of Information Management, 22(1):1–10, 2020.

[PHG04] David L Pepyne, Jinghua Hu, and Weibo Gong. User profiling for computer security. In Proceedings of the 2004 American Control Conference, volume 2, pages 982–987. IEEE, 2004.

[PHS18] Tianrui Peng, Ian Harris, and Yuki Sawa. Detecting phishing attacks using natural language processing and machine learning. In 2018 ieee 12th international conference on semantic computing (icsc), pages 300–301. IEEE, 2018.

[Pia14] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic bulletin & review, 21(5):1112–1130, 2014.

[PJBC09] James L Parrish Jr, Janet L Bailey, and James F Courtney. A personality based model for determining susceptibility to phishing attacks. Little Rock: University of Arkansas, pages 285–296, 2009.

[PM15] Sean Palka and Damon McCoy. Dynamic phishing content using generative grammars. In 2015 IEEE Eighth International Conference on

Software Testing, Verification and Validation Workshops (ICSTW), pages 1–8. IEEE, 2015.

[PR12] Mayank Pandey and Vadlamani Ravi. Detecting phishing e-mails using text and data mining. In 2012 IEEE International Conference on Computational Intelligence and Computing Research, pages 1–6. IEEE, 2012.

[PSTR14] Gilchan Park, Lauren M Stuart, Julia M Taylor, and Victor Raskin. Comparing machine and human ability to detect phishing emails. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 2322–2327. IEEE, 2014.

[PT15a] Gilchan Park and Julia M Taylor. Poster: Syntactic element similarity for phishing detection. 2015.

[PT15b] Gilchan Park and Julia M Taylor. Using syntactic features for phishing detection. arXiv preprint arXiv:1506.00037, 2015.

[PT17] Alan Partington and Charlotte Taylor. The language of persuasion in politics: An introduction. Routledge, 2017.

[RBDG20] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. IEEE Access, 8:42200–42216, 2020.

[SARG20] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. What makes phishing emails hard for humans to detect? In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 64, pages 431–435. SAGE Publications Sage CA: Los Angeles, CA, 2020.

[SK19] Teodor Sommestad and Henrik Karlzén. A meta-analysis of field experiments on phishing susceptibility. In 2019 APWG Symposium on Electronic Crime Research (eCrime), pages 1–14. IEEE, 2019.

[SL87] Lynn Smith-Lovin. Impressions from events. Journal of Mathematical Sociology, 13(1-2):35–70, 1987.

[Son20] Gunikhan Sonowal. Phishing email detection based on binary search feature selection. SN Computer Science, 1(4):1–14, 2020.

[SS05] Shigeaki Sakurai and Akihiro Suyama. An e-mail analysis method based on text mining techniques. Applied Soft Computing, 6(1):62–71, 2005.

[ST+14] S Sarju, Riju Thomas, et al. Spam email detection using structural features. International Journal of Computer Applications, 89(3), 2014.

[SW19] Harminder Singh and Jocelyn Williams. How contextualisation affects the vulnerability of individuals to phishing attempts. Twenty-Third Pacific Asia Conference on Information Systems, 2019.

[SZQ17] Wenqing Sun, Bin Zheng, and Wei Qian. Automatic feature learning using multichannel roi based on deep structured algorithms for computerized lung cancer diagnosis. Computers in Biology and Medicine, 89:530–539, 2017.

[Tay10] Julia Taylor. Computational semantic detection of information overlap in text. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 32, 2010.

[TP91] Matthew Turk and Alex Pentland. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1):71–86, 1991.

[UQ14] Sven Uebelacker and Susanne Quiel. The social engineering personality framework. In 2014 Workshop on Socio-Technical Aspects in Security and Trust, pages 24–30. IEEE, 2014.

[Vem20] Vijay K Vemuri. The Hundred-Page Machine Learning Book. Taylor & Francis, 2020.

[Ver19] Verizon. 2019 data breach investigations report. "https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf", 2019.

[VH13] Rakesh Verma and Nabil Hossain. Semantic feature selection for text with application to phishing email detection. In International Conference on Information Security and Cryptology, pages 455–468. Springer, 2013.

[VHC+11] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H Raghav Rao. Why do people get phished? testing individual differences in phishing vulnerability within an integrated, information processing model. Decision Support Systems, 51(3):576–586, 2011.

[VOW+13] Timothy Vidas, Emmanuel Owusu, Shuai Wang, Cheng Zeng, Lorrie Faith Cranor, and Nicolas Christin. Qrishing: The susceptibility of smartphone users to qr code phishing attacks. In International Conference on Financial Cryptography and Data Security, pages 52–69. Springer, 2013.

[vRMB+19] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, et al. Informed machine

learning–a taxonomy and survey of integrating knowledge into learning systems. arXiv preprint arXiv:1903.12394, 2019.

[VSH12] Rakesh Verma, Narasimha Shashidhar, and Nabil Hossain. Detecting phishing emails the natural language way. In European Symposium on Research in Computer Security, pages 824–841. Springer, 2012.

[Wes15] June West. Strategic Communication to Inform or Persuade. "https://ideas.darden.virginia.edu/2015/02/strategic-communication-toinform-or-persuade/", 2015. Accessed: 2021-01-21.

[WHJ18] Emma J Williams, Joanne Hinds, and Adam N Joinson. Exploring susceptibility to phishing in the workplace. International Journal of Human-Computer Studies, 120:1–13, 2018.

[WHZ+15] Allaire K Welk, Kyung Wha Hong, Olga A Zielinska, Rucha Tembe, Emerson Murphy-Hill, and Christopher B Mayhorn. Will the "phishermen" reel you in?: Assessing individual differences in a phishing detection task. International Journal of Cyber Behavior, Psychology and Learning (IJCBPL), 5(4):1–17, 2015.

[WJT+14] Ryan T Wright, Matthew L Jensen, Jason Bennett Thatcher, Michael Dinger, and Kent Marett. Research note—influence techniques in phishing attacks: an examination of vulnerability and resistance. Information systems research, 25(2):385–400, 2014.

[WLR16] Jingguo Wang, Yuan Li, and H Raghav Rao. Overconfidence in phishing email detection. Journal of the Association for Information Systems, 17(11):1, 2016.

[Wor08] Michael Workman. Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. Journal of the American Society for Information Science and Technology, 59(4):662–674, 2008.

[YA16] Adwan Yasin and Abdelmunem Abuhasan. An intelligent classification model for phishing email detection, 2016. arXiv:1608.02196.

[YG12] Zheng Yan and Hamide Y Gozu. Online decision-making in receiving spam emails among college students. International Journal of Cyber Behavior, Psychology and Learning (IJCBPL), 2(1):1–12, 2012.

[ZZJ+17] Xi Zhang, Yu Zeng, Xiao-Bo Jin, Zhi-Wei Yan, and Guang-Gang Geng. Boosting the phishing detection performance by semantic analysis. In 2017 ieee international conference on big data (big data), pages 1063–1070. IEEE, 2017.

[ZZMIV17] Azar Zafari, Raúl Zurita-Milla, and Emma Izquierdo-Verdiguier. Integrating support vector machines and random forests to classify crops in time series of worldview-2 images. In Image and Signal Processing for Remote Sensing XXIII, volume 10427, page 104270W. International Society for Optics and Photonics, 2017.

Appendix A

Corpora

A.1 Public Domain Corpora

- The Podesta Emails (overwhelmingly non-phishing), downloaded from `https://wikileaks.org/podesta-emails/` (n=57,000) (Date: 2015)
- 2005-2007 Phishing Email Corpora (phishing), downloaded from `http://monkey.org/~jose/wiki/doku.php?id=phishingcorpus(2015-02-01)` (n=2,700) (Date: 2005-2007)
- The "Mbox" Corpus (phishing), downloaded from `https://github.com/diegoocampoh/MachineLearningPhishing` (n=2,500) (Date: 2007)

A.2 Non-Public Corpora

- Phishing Emails Curated by the Author (phishing) (n=1,795) (Date: 2018-2020)

A.3 Curated Corpora (for testing purposes)

- Sample Corpus - A combined set of emails comprised of a random sample from all phishing corpora (n=3,757) and a random sample from all non-phishing corpora (n=3,859)

A.4 Email Corpus Processing Issues

The public domain corpora were available as ZIP format compressed files with the exception of the MBOX corpus which was in the MBOX format. Initially, the MBOX format posed some challenges to accessing it properly. Eventually the best approach to obtaining a proper set of EML format compliant email instances was to simply import the MBOX file into a folder within the MacOS Mail application and then selecting all emails and copying them to a folder outside the application.

The various corpora contained emails in both the MSG as well as the EML format. MSG is a proprietary file format used by Microsoft to store individual email messages. EML (short for Electronic MaiL), is an email file format for storing individual email messages according to the MIME RFC 822 standard. Easily usable API's for parsing both these formats were readily available for the Java environment.

Appendix B

Phishalyzer Implementation

B.1 Hardware Used

All development and testing was completed on a 2018 2.2 GHz 6-Core Intel Core i7 Apple MacBook Pro with 16 GB RAM and a 256GB SSD.

B.2 Phase 1: From ZIP to Serialization

Phishalyzer, a configurable email processing pipeline, was developed early on using the Java language. Phishalyzer has the following features:

- The application runs from the command line and has a number of configurable variables to tailor the pipeline to the output requirements for a particular run.
- N-gram lists and corresponding feature descriptors are imported from an easily editable configuration file.
- It reads and processes individual emails in both the MSG and EML formats from a specified file directory.
- Application run time dialog boxes permit selection of the configuration file and corpus for processing.

- Available API's are used to extract various components from an email file instance, including the subject line text and body text.
- During processing, individual email instance data for every email is stored in a global Java Class EmailInstance object in memory, and attributes are modified/added to this object as processing proceeds.
- Depending on the processing task, post processing results are written to file on a per email basis or for global procedures at the end of processing, in either a Comma Separated Value (CSV) format or Resource Description Framework (RDF) format.

The application includes classes and methods which perform the following functionality:

- Return an "in sequence" list of feature n-grams found in a passed String.
- Return a list of n-grams and their cardinality from a passed EmailInstance Object.
- Generate a list of the frequency of n-grams found within a corpus.
- Generate a list of features and their cardinality in each email within a corpus.
- Return embedded text from a multipart MIME email (Multipart MIME messages are often created as a result of forwarding an email or where attachments are sent with an email.)
- Return a tokenized version of a passed String.
- Remove "excluded strings" from a passed String. Excluded strings are loaded from a configuration file upon application execution and are limited to artifacts left by a phishing email forwarding service and some email signature and footer

based Strings specific to certain corpora, which contained terms also present in the feature n-gram lists.

- Remove all tags from a passed string (using the Jsoup library[1]
- Return a list of strings matching a regular expression, found in a passed String.
- Replace abbreviated word snippets with their expanded version (for example, turning "isn't" into "is not" etc.).
- Generate feature co-occurrence datasets.
- Count all the words found in a passed String.

Phishalyzer does not incorporate any stop word, stemming or lemmatization processing of the various email text data.

B.3 Phase 2: From Serialized Data to Insights

The processing for this phase was all completed within various Jupyter Notebooks[2] using the Python3[3] programming language and its extensive available libraries, in particular SciKitLearn[4].

In this phase, testing followed these iterative steps, repeatedly:

1. The import of a data set into a Python dataframe object.
2. Visual inspection of the head and the tail of the data set, to confirm the import was generally successful.
3. Determine if there are any duplicate entries, and if so remove them.

[1] `https://jsoup.org/`.
[2] `https://jupyter.org/`
[3] `https://www.python.org/`
[4] `https://scikit-learn.org/stable/index.html`

4. Assess data for any null values and if present, remediate.

5. Explore the feature correlation and visualize this with a Python *seaborn* library[5] plot. Assess any obviously correlated columns as candidates for feature reduction.

6. Verify that the phishing and non-phishing distribution within the data set was still roughly equal.

7. Split the data between a training and test data set at a ratio of usually 70 percent training set, 30 percent testing test, and verify.

8. Train the chosen algorithm.

9. Test the performance of the generated model against both the training and test data and compare results.

10. If acceptable, generate Confusion Matrix and F1-Score metrics.

11. If not acceptable, seek to determine the cause and adjust any required inputs.

[5] `https://seaborn.pydata.org/api.html`