DETECTING AND ANALYZING CYBERCRIME IN TEXT-BASED COMMUNICATION OF CYBERCRIMINAL NETWORKS THROUGH COMPUTATIONAL LINGUISTIC AND PSYCHOLINGUISTIC FEATURE MODELING

by

Alex Vincent Mbaziira

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Information Technology

Committee:

_____	Dr. James H Jones, Dissertation Director
_____	Dr. Duminda Wijesekera, Committee Member
_____	Dr. Douglas J Wulf, Committee Member
_____	Dr. Hemant Purohit, Committee Member
_____	Dr. Stephen Nash, Senior Associate Dean
_____	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering
Date:_____	Spring Semester 2017 George Mason University Fairfax, VA

Detecting and Analyzing Cybercrime in Text-based Communication of Cybercriminal Networks Through Computational Linguistic and Psycholinguistic Feature Modeling

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Alex Vincent Mbaziira
Master of Science
George Mason University, 2014
Master of Science
Uganda Martyrs University, 2004
Bachelor of Science
Uganda Martyrs University, 2000

Director: James H Jones, Associate Professor
Department of Electrical & Computer Engineering

Spring Semester 2017
George Mason University
Fairfax, VA

DEDICATION

To my faithful God; my beautiful wife, Karen; and baby girls, Nissa and Arika.

ACKNOWLEDGEMENTS

I would like to express my appreciation and gratitude to my advisor and dissertation Chair, Dr. James H Jones for his patience, mentorship, expert guidance, and encouragement during this research. I also want to thank my committee members: Dr. Douglas Wulf, Dr. Duminda Wijesekera and Dr. Hemant Purohit for their comments, input and support that made this research possible.

I am grateful to the Fulbright program for my initial funding that enabled me to start and successfully complete the first stage of my doctoral work. Furthermore, I want to thank Dr. Diana Wang from the Department of Information Sciences and Technology Department in Volgenau School of Engineering for her support and additional funding that has enabled me to complete the last stage of my doctoral work. I cannot forget to express my gratitude to Dr. Joseph and Dr. Immaculate Kiiza from University of Tennessee in Chattanooga, who mentored, supported, and prepared me for this journey.

I extend my sincere gratitude to Dr. Bernice Griffith of Fairfax Radiology Consultants and her daughter, Dr. Candace Griffith for their prayers, support, generosity and encouragement towards me and my family during my doctoral work and also during that difficult time when I lost my mum and kid sister.

I also extend my appreciation and gratitude to my parents and relatives: Mr. Vincent Ssennyonjo and the late Maddy Lule Ssennyonjo, Dr. John and Mrs. Dorothy Kaboggoza, and Dr. Peter John and Ms Jennifer Opio for their prayers, encouragement, support and generosity towards my family.

Lastly, I thank all my siblings, extended family, friends, GMU colleagues, prayer families in New York and Toronto for your prayers and support.

May God bless and reward you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Facebook ..... FB
Enron ..... EN
Favorable Reviews ..... PR
Unfavorable Reviews ..... NR
Amazon Mechanical Turk ..... AMT
Part-of-Speech ..... POS
Naïve Bayes ..... NB
Support Vector Machines ..... SVM
kth Nearest Neighbors ..... kNN
Psycholinguistics ..... PL
Computational Linguistics CL
Receiver Operator Characteristic ..... ROC
True Positive Rate ..... TPR
False Positive Rate ..... FPR
True Positives ..... TP
True Negatives ..... TN
False Positives ..... FP
False Negatives ..... FN
Principal Component Analysis ..... PCA
Natural Language Processing ..... NLP
Linguistic Inquiry and Word Count ..... LIWC

ABSTRACT

DETECTING AND ANALYZING CYBERCRIME IN TEXT-BASED COMMUNICATION OF CYBERCRIMINAL NETWORKS THROUGH COMPUTATIONAL LINGUISTIC AND PSYCHOLINGUISTIC FEATURE MODELING

Alex Vincent Mbaziira, Ph.D

George Mason University, 2017

Dissertation Director: Dr. James H Jones

Cybercriminals are increasingly using Internet-based text messaging applications to exploit their victims. Incidents of deceptive cybercrime in text-based communication are increasing and include fraud, scams, as well as favorable and unfavorable fake reviews. In this work, we use a text-based deception detection approach to train models for detecting text-based deceptive cybercrime in native and non-native English-speaking cybercriminal networks. I use both computational linguistic (CL) and psycholinguistic (PL) features for my models to study four types of deceptive text-based cybercrime: fraud, scams, favorable and unfavorable fake reviews. The data is obtained from three web genres namely: email, websites and social media.

I build 1-dataset non-hybrid models as well as two types of hybrid models for native and non-native English speaking cybercriminal networks: 2-dataset and 3-dataset hybrid models. I use Naïve Bayes, Support Vector Machines and kth Nearest Neighbor to train and test all the models. All the 1-dataset non-hybrid models are trained on data from one web genre and then used to detect and analyze other types of cybercrime in other web genres that are not part of the training set. Furthermore, all the 2-dataset hybrid models are trained on data combined from two web genres and then used to detect cybercrime in other web genres that are not part of the training set. Further still, the 3-dataset models are trained on every triplet data in three web genres and used to detect and analyze cybercrime in the web genre which was not part of the training set.

Performance of the models on test datasets ranges from 60% to 80% accuracy with best performance on detection of fraud and unfavorable reviews. There were notable differences in models in detecting and analyzing scams in both native and non-native English speaking cybercriminal networks. This work can be applied as provider- or user-based filtering tools to identify cybercriminal actors and block or label messages before they reach their intended audience.

CHAPTER ONE

Text-based digital communication, especially email, has existed since the nascent years of the Internet. Over the years, the Internet has evolved with advances in computing and digital network technology giving rise to even more text-based communication channels like Internet Relay Chat (IRC), messaging boards, blogs, websites and social media. Text-based communication can be classified into two types: synchronous and asynchronous. In synchronous text-based communication, communicating actors get instant feedback whenever they transmit messages during a communication session. By comparison, in asynchronous text-based communication, feedback for transmitted messages is either delayed or ignored by the message recipient.

It is also important to note that the language styles used in each medium of text-based communication medium varies by web genre: web (Mehler, Sharoff, & Santini, 2011), email (Wollman-Bonilla, 2003), Facebook, Twitter (Westman & Freund, 2010) etc.

The emergence and growth of mobile technology has also enabled the ubiquity of Internet connectivity and text messaging, eventually rendering it more popular than voice communication (Shropshire, 2016). This shift to mobile communication and surge in text messaging is also encouraging cybercriminals to consider mobile and text messaging

applications as a potential vector for cybercrime (Vergelis, Shcherbakova, Demidova, & Gudkova, 2016).

Cybercrime has continued to surge causing organizations to invest vast amounts of money to secure their computer infrastructure against increasing cyber-attacks (Ponemon Institute, 2015). Cybercriminals continue to improve their capabilities in cyber exploitation to defeat cyber-defense infrastructures for economic gain. Many of these cyber-attacks are successful because the text-based communication infrastructure is designed and built and on top of the Internet's rapidly evolving but insecure and open architecture (McGrath, 2015). These vulnerabilities enable cybercriminals to exploit this infrastructure and also conceal their real identities by using dissociate anonymity to commit cybercrime (Jaishankar, 2011).

Cybercrime broadly falls into two categories: technology-based cybercrime and content-based cybercrime (ITU, 2009). Technology-based cybercrime involves abuse of computing and digital network infrastructures to commit crime. Some examples of technology-based cybercrime include online auction fraud, advance-free fraud, computer-related forgery, deceptive reviews, identity-theft, denial of service attacks, phishing and cyber-terrorism. Content-based cybercrime involves using text-based communication to generate and disseminate malicious content like spam, scams, fake reviews, child pornography, hate speech etc.

Motivation

This research is motivated by work in graph theory and the bag-of-words approach in machine learning which have attempted to detect cybercrime by identifying relationships and patterns between cybercriminal text messages in the form of email spam and social media scams (H. Chen et al., 2004; Sarvari, Abozinadah, Mbaziira, & McCoy, 2014; Yang, Harkreader, Zhang, Shin, & Gu, 2012). Research on detecting cybercrime using graph theory has mainly centered on centrality measures like degree centrality, betweenness centrality and eigenvector centrality. However, there are several drawbacks in using centrality measures to detect content-based cybercrime in cybercriminal networks as discussed below.

a) Firstly, degree centrality considers criminal nodes with numerous connections within a cybercriminal network as important nodes while eigenvector centrality identifies nodes with more important connections by considering neighboring nodes. The problem with degree and eigenvector centrality measures is that criminal nodes with low measures may seem uninteresting, however, from a cybercriminal network standpoint such nodes may be valuable because they contain incriminating evidence authored by hard core cybercriminals maintaining a low profile in a large criminal network. This can also be attributed to the fact that cybercriminals are discrete about their illegal activities especially when interacting in large social online communities, hence high degree centrality and eigenvector centrality does not imply

prominence of a node in a cybercriminal network (Aransiola & Asindemade, 2011).

b) Betweenness centrality measures how important nodes connect with other nodes in networks within giant components as shown in Figure 1. In a typical social network structure, nodes with high betweenness centrality may be information brokers, but in large criminal network such nodes may only be gregarious.

For the bag-of-words approach, I observed that the text messages in cybercriminal networks reveal that cybercriminals are not necessarily unilingual or English speaking. Cybercriminal networks have a global presence in countries that may or may not use English as a native language. I published two papers that explored linguistic variations in detecting content-based cybercrime in a unilingual Arabic speaking cyber-criminal network and a bilingual Nigerian cybercriminal network using bag of words as well as graph theory. The paper on the unilingual cybercriminal network uses the bag-of-words approach to detect patterns of cybercrime in tokens of Arabic abusive tweets used by cybercriminals to launch successful spam campaigns targeting the Middle East (Abozinadah, Mbaziira, & Jones, 2015). I also explored the problem of detecting cybercrime in the form of advance-fee fraud and online dating scams in a Nigerian bilingual cybercriminal that speaks both English as well as an English-based pidgin language widely spoken in West Africa (Mbaziira, Abozinadah, & Jones, 2015). This paper also applies machine learning and bag-of-words approach to detect patterns of cybercrime within such cybercriminal network.

Figure 1. A Cybercriminal Network (Sarvari et al., 2014)

The limitations above have motivated this research to explore a linguistic approach for detecting and analyzing cybercrime in text-based communication using CL and PL processes to detect deception and cybercrime. To achieve this, I use a text-based deception discourse to identify computational linguistic and psycholinguistic processes that can be mapped to deception and cybercrime.

Purpose Statement

The purpose of this research is to develop a model for detecting cybercrime in text-based communication channels that uses computational linguistics (CL) in Natural Language Processing (NLP), psycholinguistics (PL) and machine learning. I link fake reviews, scams and fraud to deception because cybercriminals use deception as a strategy for tricking and exploiting their victims. Cybercriminals are successful in exploiting their

victims because existing content filtering mechanisms have vulnerabilities, which exacerbate this problem.

There is still very limited work on detecting cybercrime using a text-based deception discourse. Also, existing research which have studied this problem use mock experiments with verifiable facts due to lack real world data (Hao, Chen, Cheng, Chandramouli, & Subbalakshmi, 2011; Torney, Vamplew, & Yearwood, 2012; Zhou, Burgoon, Twitchell, Qin, & Nunamaker, 2004). In this work, I study this problem using real world scams, fake reviews and fraud text-based messages from cybercriminal networks.

Research Problem

The cost of cybercrime continues to surge as cybercriminals exploit flaws in pervasive Internet technology (Morgan, 2016; Nakashima & Peterson, 2014). Increasing amounts of criminal content in form on scams, fake reviews and fraud as well as truthful text-based communication continues to be generated and disseminated on the web. Until recently unsolicited commercial or bulk email, called spam, was major problem in text-based communication media especially email. Existing content-based filtering mechanisms work well in containing email spam but fail in detecting scams, fake reviews and fraud in text-based communication. This because spam content filters use message headers and addresses to block spam but do not check the content of the text messages. The common factor that links fake reviews, fraud and scams but differentiates them from traditional email spam is deception. Gaps within existing research motivate this research to explore a natural language-based approach for detecting patterns fraud, fake online

views and scams. This will be crucial for developing a model for detecting deception and cybercrime in text-based digital communication.

Research Questions

The goal of this research is to create a model for detecting cybercrime in text-based communication media that addresses this general question: *how can we detect and analyze linguistic patterns of cybercrime in text-based communication of cyber-criminal networks?*

To address this question, we derive the following sub research questions (SRQ) to guide this work:

- **SRQ1**: Can we detect cybercrime in text-based communication of web genres?
- **SRQ2**: Can we generalize deception and cybercrime detection in text-based communication?
- **SRQ3**: Can we generalize cybercrime detection in text-based communication of native and non-native English speaking cybercriminal networks using hybrid models?
- **SRQ4**: What linguistic features are linked to deception and cybercrime in text-based communication?

Research Contributions

This research provides practical contributions to cybercrime research. Firstly, we develop cybercrime detection models using data from different web genres namely: email, social media and websites. A model trained on data from one web genre is used to

detect cybercrime in data of a different web genre. I build all the models using features from computational linguistic and psycholinguistic processes linked to deception and cybercrime in text-based messages.

Secondly, I develop hybrid models where I combine datasets in different web genres, which I use to train the cybercrime detection models. The models are then tested on a dataset of a completely different web genre. I use several real-world datasets, as shown in Table 6, to develop these models. For fraud, I study the Enron datasets which comprise of an email dump that was made public during the Enron scandal as well as email evidence used in prosecuting the two top executives of Enron for securities and wire fraud. The Facebook data is collected from leaked emails of Nigerian cybercriminals using an online data theft service. For reviews I use two publicly available datasets which comprise both favorable actual and fake reviews as well as unfavorable actual and reviews for hotels (Ott, Choi, Cardie, & Hancock, 2011). I developed several training hybrid models by combining 2-datasets and 3-datasets to detect for fraud, scams, unfavorable and favorable reviews. All these models are generalized.

Lastly, I develop deception and cybercrime detection models using dataset-specific features to detect cybercrime in non-native English-speaking cybercriminal networks. I accomplish this by building two sets of hybrid models, that is, 2-dataset and 3-dataset models for detecting scams in non-native English speaking cybercriminal network. I accomplish this by using dataset-specific features using a dataset obtained from non-native English speaking cybercriminals. The Facebook data was collected from Nigerian cybercriminals who are non-native English speakers. I build three 2-datasets

models and one 3-dataset model to detect scams in the non-native English speaking cybercriminal network.

Scope of work

Deception may be defined as the act of intentionally misleading another party or persons through falsified statements or fraudulent actions ("deception," 2006). Research on deception in form of fraud, scam and fake reviews in text-based communication is still very limited. This work focuses on fraud, scams and fake reviews because cybercriminals use deception as a strategy for exploiting their victims. I therefore use the term cybercrime to refer to a type of text-based online crime in form of fraud, scams and fake reviews where cybercriminals use deception for these types of cybercrime to exploit their victims.

This work investigates deception and cybercrime in text-based communication of criminal networks in three web genres: email, social media and e-commerce websites. The cybercrime detection models are trained on data from one web genre and evaluated on data from other web genres. Similarly, for hybrid models, the training sets are obtained by combining data from the two and three web genres and then the model is used to detect cybercrime in other web genres that not part of the training model.

I also need to point out that there are two types of cybercriminal networks: cybercriminal-to-cybercriminal network and cybercriminal-to-victim network. In cybercriminal-to-cybercriminal networks, cybercriminals use text-based communication media to plan and execute cybercrime and use coded messages that are specific such a network. Alternatively, in cybercriminal-to-victim networks, cybercriminals send

messages to victims to exploit them. Since messages in cybercriminal-to-victim networks have more research and linguistic value, my work focuses on this specific type of cybercriminal network. In this work, I shall use the term cybercriminal network to mean cybercriminal-to-victim networks.

CHAPTER TWO

LITERATURE REVIEW

Text-based digital communication continues to transform the way people exchange messages in online conversations. Despite its increasing popularity, text-based communication continues to be exploited by cybercriminals. There are various research investigating the context and impact of such cybercrime targeting text-based communication as well as approaches for detecting and analyzing such crime (Bohme & Moore, 2012; Nirkhi, Dharaskar, & Thakre, 2012). In this chapter, I discuss challenges in existing approaches for detecting cybercrime in text-based communication media. I also discuss the gaps of existing research with respect to cybercrime and deception, natural language processing, feature engineering and variable selection as well as machine learning.

Deception and Cybercrime Detection

Deception in text messages occurs when an actor generates and, or disseminates falsified information to manipulate and, or exploit his or her victims. In this research, I consider deception in text-based communication to occur when a criminal actor generates and disseminates false information in form of scam, fraud or online fake reviews with intent of making wrongful financial gain from targeted victims. Cybercriminals use deception as a strategy for committing cybercrime through social engineering, scams,

fraud, spam and distributing malware, etc. Even cyber terrorist organizations use deception to recruit unsuspecting victims into their networks (Engel, 2015; Hall, 2015). In the next section, I discuss theories on deception, which are linked to linguistic processes for cybercrime. I also review related work on detection of cybercrime and deception.

Deception Theories

Cybercrime prevalent in text-based communication is detrimental to both individuals and organizations because it causes financial loss to organizations. Some of the types of cybercrime in text-based communication include: fraud, scam and fake reviews. There are various types of fraud committed by cybercriminals when exploiting victims through text-based communication media. These include: fake contests and sweepstakes; advance-fee fraud where cybercriminals trick their victims to make upfront payments in exchange for something valuable. Scam is a form of a dishonest trick played on a victim to exploit them for wrongful financial gain. Some of the popular online scams are: work-at-home scams, online dating scams, investment scams, lottery scams etc. Fake reviews are also an emerging popular form of cybercrime targeting ecommerce websites. Online product reviews are great tools used by consumers in making informed purchase decisions for products and services in ecommerce. Despite the success of ecommerce and popularity of reviews, there are unscrupulous companies and individuals that are reaping financial rewards by either attacking competitors with unfavorable fake reviews or posting favorable fake reviews to promote inferior products which may be deemed dangerous to consumers' health (Streitfeld, 2011; Weise, 2015).

There is a growing body of research on deception especially for face-to-face, voice and video communication, however, research on deception in text-based communication is still limited (Hancock, Curry, Goorha, & Woodworth, 2007; Rowe, n.d.; Zhou et al., 2004). However, there is a limited body of work attempting to study some of the underlying theories on deception in text messages. Some of these theories on deception are : *Media Richness Theory, Channel Expansion Theory* and *Interpersonal Deception Theory* (Zhou et al., 2004; Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003).

The *Media Richness Theory* explains how deception happens in face-to-face conversations and voice communication through verbal and nonverbal cues. Some of the cues used in studying deception include: feedback, voice inflection, body language, emotion, feelings etc. The theory further asserts that there must be a continuous exchange of messages between communicating entities for the deceiver to initiate deceptive messages. From a deception and cybercrime standpoint, I consider psycholinguistic features which express positive and negative emotion and feelings in text messages. This theory is particularly useful in extracting some features for detecting deception and cybercrime since cybercriminals use also emotion to trick and exploit their victims.

The *Channel Expansion Theory* expands the *Media Richness Theory* by including experience and skill that two actors gain when exchanging messages. This theory asserts deception in messages improves as a malicious actor gains more experience in communicating with the victim. Gaining experience is important for such a malicious actor to enable him or her to craft better deceptive messages which are expressive and

emotional. We also consider this theory to study deception and cybercrime in text messages because there is increasing adoption web tools for expressivity and emotion in text messages.

The *Interpersonal Deception Theory,* on the other hand, asserts that during a communication session, malicious actors will continuously modify their messages to prevent their victims from detecting deception. This is possible if the malicious actor has established an interpersonal relationship with a victim to gain some form of trust which can be exploited to sustain a conversation for a reasonable time. We also use this theory to understand deception and cybercrime in text messages.

Linguistic Approaches

There several linguistic approaches for detecting cybercrime in text messages. The bag-of-words approach is a popular approach for detecting patterns of cybercrime in text messages (Abozinadah et al., 2015; Li, Huang, Yang, & Zhu, 2011; Mbaziira et al., 2015; Mukherjee, Liu, & Glance, 2012). This approach uses individual words or combined words as features, which are also called n-gram words. These features may vary from unigram, bigram, trigram to n-gram words where *n* is denoted as any counting number that is greater than zero. Other research, attempts to detect deception in online product reviews using n-grams and deeper syntax like weight, location, price etc., within text-based messages to discriminate between fake and truthful reviews (Feng & Hirst, 2013). However, in spite of popularity of the n-gram approach, some studies reveal that this approach is not robust enough in detecting patterns of cybercrime in text-based communication (Y. Chen, Zhou, Zhu, & Xu, 2012; Reynolds, Kontostathis, & Edwards,

2011). In this work, I do not adopt this approach because it is very difficult to generalize the cybercrime models trained specific types of cybercrime to detect other types of cybercrime. This is because models trained using approach tend to have specific characteristics and properties of the datasets, which makes it difficult for such models to generalize well.

Besides bag-of-words approach, other approaches use computational techniques in computer science to study relationships and patterns in lexical and syntactic properties of text messages. This approach is called computational linguistics (CL). Lexical features comprise of character-based and word-based features like total characters, unique characters, total words, characters per word, frequency of large words, and unique words while syntactic features include frequency of punctuation marks, occurrence of function words, parts-of-speech (POS) tagging (Afroz, Brennan, & Greenstadt, 2012; Shojaee, Murad, Azman, Sharef, & Nadali, 2013). Table 1 below shows a summary of lexical and syntactical features. Similarly, in POS tagging each word is marked and assigned a lexical category or word class that corresponds to its part-of-speech in a sentence (Bird, Klein, & Loper, 2009). The tags are classified in various broad lexical categories namely: nouns, pronouns, verbs, adjectives, adverbs, prepositions, interjection and conjunctions. Early research reveals that detection of content-based cybercrime like spam using POS tags depends on the relationship that exists between the genre and frequencies of POS tags in the corpora being investigated (Ott et al., 2011). Table 2 below shows the POS tags for various lexical categories (Crawford, Khoshgoftaar, Prusa, Richter, & Najada, 2015). For this work, I use lexical items and syntactical features to study

deception and cybercrime because the models I train will be more robust that can generalize in detecting cybercrime in test datasets that are not used in the training model.

Table 1 Description of Lexical and Syntactical Features

Type of Linguistic Features	**Linguistic Features**
Word-based Lexical Features	
	Frequency of lexical items
	Average sentence length
	Average lexical item length
	Ratios of characters in words
Character-based Lexical Features	
	Frequency of characters
	Frequency of special characters
	Number of white spaces
	Ratios of digits and letters to total characters
Syntactic Features	
	Frequency of punctuation marks
	Frequency of function words

Table 2 POS Tags and Description

Tag #	**POS Tag**	**POS Tag Description**
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural

14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Pre-determiner
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	To
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

Lastly, another linguistic approach worth considering is psycholinguistics (PL), which attempts to establish relationships between linguistic behavior and psychological processes. Since deception is a psychological process, I also use a deception discourse to study cybercrime in text messages. I combine word-based lexical features in Table 1 above, with psychological processes like positive and negative emotions, cognitive load, cognitive complexity, etc. (Tausczik & Pennebaker, 2010).

Research in Feature Modeling and Deception

There is very limited body of work on linguistic approaches for detecting deception and cybercrime in text-based communication. Some early work is an attempt to adapt deception research to text-based communication from well-studied areas of psychology as well as verbal and non-verbal communication. Some existing research on deception detection in text-based communication has been done on written statements and interviews to improve tools for interviewing bilingual and, or non-native English speaking criminals during criminal investigations (Sandoval, Matsumoto, Hwang, & Skinner, 2015). Some deception features identified are: quantity of words, non-immediacy, sentence complexity, distinction markers, emotion, expressivity, lexical diversity (Hancock et al., 2007; Zhou et al., 2004, 2003; Zhou & Zhang, 2008).

There is also research that observed that deceptive messages have fewer quantities of words (DePaulo et al., 2003). The quantities of words are measures of frequencies for lexical items, sentences as well as parts-of-speech in corpora like verbs, nouns adjectives, etc. However, later studies reveal that deceptive messages in asynchronous text messages are wordier than truthful messages because liars have time to plan and write their messages (Zhou et al., 2004). Another linguistic attribute for deception detection is lexical diversity. This is a ratio the frequency of unique words used in text messages to the total number of words. When this attribute is used in deception detection, researchers observe that untruthful messages will have lower lexical diversity (Zhou et al., 2004).

Other studies further reveal that deceptive messages have higher expressivity compared to truthful communication. In this case expressivity is measured by frequency of adverbs and adjectives. This work differs from that research because I use features

from CL and PL processes linked to deception and cybercrime and test our models for generalizability to detect new or emerging types for deceptive cybercrime.

Cybercrime in text-based communication can be also detected using PL features. This is because linguistic behavior in text-based communication can be mapped to criminal psychological processes. Tools like Linguistic Inquiry and Word Count (LIWC) have been widely used in studying various linguistic patterns like: suicides in poets and linguistic styles with self-references; deception detection and writing styles; social judgments (Crawford et al., 2015; Hancock et al., 2007; Tausczik & Pennebaker, 2010). LWIC supports several linguistic processes for various PL processes, however, I identify features which are relevant to deception and cybercrime in text-based messaging. For instance, linguistic features like word quantity, average sentence length, first-person singular and exclusive words have a relationship with psychological processes like talkativeness, cognitive complexity and truthfulness respectively (Tausczik & Pennebaker, 2010). Similarly features like frequency of lexical items, first person pronouns and exclusive words that can be linked to deception and PL processes (Keila & Skillicorn, 2005; Newman, Pennebaker, Berry, & Richards, 2003; Tausczik & Pennebaker, 2010). Furthermore, words of negative emotion are also linked to deception and cybercrime where deceptive messages have more negative emotion words than affect and positive emotion words (Newman et al., 2003).

An earlier paper attempted to automate deception detection using some CL features and classification methods like logistic regression, decision trees, discriminant analysis and neural networks (Zhou et al., 2004). The data for training models was

collected from subjects who were given a fictitious scenario and were required to lie about their experience. These subjects were paired so that they could exchange truthful and deceptive emails about this scenario. Two datasets of 180 and 204 instances were collected from which 70% was used for training and 30% for testing. Performance of the models ranged from 61% to 81% accuracy rate. My work differs from this work because I use real world data from several web genres. Also, all the instances used in testing my models are obtained from web genres that were not part of the training set to prevent the models from overfitting or underfitting.

Another early paper investigated the deception from a sample of 100 undergraduates who were tasked to give both their truthful and deceptive views on topics like abortion, feelings about friends and mock crime (Newman et al., 2003). The data from the experiment was collected from three main sources: video-tapes, hand written and typed manuscripts. The paper used psychological processes to deception to extract the features for the paper to create a multi-variate linguistic model. These features are: exclusive words like *but, except, without, exclude* as well as negations like *no, never*. This is because messages with more distinction markers require the author of a message to have good cognitive load hence liars avoid using them to prevent any contradiction in their communication lest they get caught (Hancock et al., 2007). The model performed better than human judges in detecting deception. My work differs from this paper because we use real world data from different web genres to classify deceptive messages from truthful ones. I also build generalizable models that can detect new forms of deceptive cybercrime.

Since my work is also investigates detection of cybercrime and deception in web genres, there is recent work on Twitter that uses n-gram words for feature modeling. Text messages in Twitter are limited to 140 character, therefore the tweets are preprocessed to remove punctuation marks and generate n-gram words (X. Chen, Chandramouli, & Subbalakshmi, 2014). That research then applies semi-supervised learning algorithms and principal components to generate a model that discriminates between deceptive and truthful tweets. A training set of 200 tweets is applied to Naïve Bayes and Suffix Tree classifiers to build models which have an accuracy rate ranging from 62% to 77%.

I also considered another study which explored models built from unigrams (1-GRAM), bi-grams (2-GRAM), n-grams (N-GRAM) and deep syntactical features (SYN) and principal components to discriminate between deceptive and truthful favorable reviews for travel websites (Feng & Hirst, 2013). The features for the SYN model are extracted from distinct aspects and descriptive aspects of a product. Distinct aspects mention known monuments, landmarks, etc., while descriptive aspects define general aspects of a product. Comparatively for the n-gram model, unigrams and bigrams are combined to generate the features. Five models were trained namely: SYN, N-GRAM, 1-GRAM+SYN, 2-GRAM+SYN, N-GRAM+SYN and the performance of these models ranged 87% to 90% accuracy rate. This research differs from my work because I use a different feature modeling approach comprising CL and PL features and supervised learning to build models for detecting deception and cybercrime in text messages. I also do not use the n-gram model because it is not robust enough and is very difficult to generalize. I also used an additional dataset of unfavorable truthful and fake reviews.

Another paper investigates fraud in emails using four PL features: first person pronouns, negative emotion, action verbs, and exclusive words to discriminate between deceptive and truthful emails (Keila & Skillicorn, 2005). Clustering is used to generate models that detect fraud in emails. My work differs from this work since my models use more PL features which are also complemented with CL features. I also use classification to build generalizable models for detecting deception and cybercrime in text messages. Since that paper uses an Enron dataset, I complement my research with an additional email dataset from DoJ.

There is recent research in stylometry that attempts to study deception and authorship attribution for disputed documents. One paper uses n-grams, n-gram words, syntactical features, function words, and specific keywords used by spammers in addition to nine CL features (Brennan, Afroz, & Greenstadt, 2012), while another paper investigates adversarial stylometry by applying n-gram words, character n-grams, vocabulary richness, function words, POS tags, personal pronouns and punctuation marks to detect imitation and obfuscation of documents through writing styles using data collected from 12 participants in a survey and 45 *Amazon Mechanical Turkers* (Brennan et al., 2012). Datasets from Turkers are analyzed against data from underground cybercriminal forums to investigate deception through imitation and obfuscation of documents through writing style. This work differs from stylometry research in that I use real world data comprising both truthful and deceptive text messages in form of scams, fraud and fake reviews. I do not use any tools to imitate or obfuscate to alter the text messages from cybercriminals to make them deceptive, an approach that was used by

Brennan et al., 2012. I also use more CL and PL features linked to deception and cybercrime and also normalize the datasets to enable the classifiers to utilize the entire feature-sets for training and test sets. I do not use n-grams words in my models because they are not robust.

I also identify CL and PL processes linked to deception and cybercrime from which linguistic features for the cybercrime detection models are extracted. In Table 3, I identify CL processes, while Table 4 summarizes PL processes linked to deception and cybercrime in text messages.

Table 3 Summary of linguistic features using Computational Linguistics (CL)

Linguistic Feature	**CL process linked to deception and cybercrime**
Quantity of words	Deceptive messages have more lexical items (i.e. lexical item rich)
Lexical diversity	Deceptive messages have fewer ratio of unique words (i.e. lexically poor)
Expressivity	Deceptive messages have higher frequencies of adjectives and adverbs (i.e. modifier rich)
Non-immediacy	Deceptive messages have fewer first-person pronoun (i.e. self-reference poor)
Sentence complexity	Deceptive messages have less complex sentence complexity i.e. lower average sentence length, average word length, pasuality and punctuation marks

Table 4 Summary of linguistic features using Psycholinguistics (PL)

Linguistic Feature	Psycholinguistic process linked to cybercrime
Quantity of words	Criminals are talkative to make their scams, online fake reviews and fraud forgettable hence low cognitive load
Average sentence length	Criminals are verbally fluent to make to scams, online fake reviews and fraud forgettable hence low cognitive complexity
First person pronoun singular (*i.e. I, me, mine)*	Criminals use less first pronouns to avoid accountability in their messages hence dissociate themselves from their messages
Exclusive words *(i.e. but, without, exclude)*	Criminals use more exclusive words to be more imprecise hence low cognitive complexity & deception
Emotion	Cybercriminals use more words of negative emotion but less words of positive emotion and affect. Negative emotion words are used to detect sublimated guilt.

Machine Learning and Cybercrime

Machine learning is branch of artificial intelligence that gives computers capabilities to learn and predict patterns in data using computational algorithms without being explicitly programmed. There are several categories of machine learning: supervised, unsupervised, and semi-supervised machine learning. In supervised machine learning, an algorithm will use labeled data to generate a model that maps inputs into desired outcomes output. On the other hand, unsupervised machine learning, uses an unlabeled data to create a function that models inputs into desired outcomes. Semi-supervised learning uses both labeled and unlabeled data to generate a learning model.

Table 5 Summary of Input and Output for Text Classification Model

Classifier Input	**Classifier Output**
A set of classes *C* for labeling instances in training set C=$\{c_1, c_2, \ldots c_n\}$	The trained text classification model that can maps each new document *t* of the test set to a specific class *c* $\gamma : t \rightarrow c$
A document *d* with a set of features *X* where X =$\{x_1,x_2,\ldots,x_n\}$	
A training set where each document *d* comprising of *n* records is manually assigned a class label *c* $\{(d_1,c_1), (d_2,c_2), \ldots (d_n,c_n)\}$	

Since I am using supervised machine learning to address the problem of detecting and analyzing deception and cybercrime, I regard this to be a classification problem because each of data objects will be assigned to one pre-defined category or class (Conway & White, 2012). In classification, the algorithms use the input data to generate a model which correctly predicts records it has never seen (Tan, Steinbach, & Kumar, 2006). A model with generalizable capabilities is constructed by creating a training set with records assigned to the class labels. This training set is input data which is mapped by a learning algorithm to generate a classification model. The classification model is then applied to test set to predict the class labels of unlabeled data. The text classification model will have input and output. The input constitutes a training set with a number of labeled documents while the output is a learned classifier that can map unlabeled documents by assigning each to a specific class. Table 5 is a summary of input and output for text classification model.

Figure 2 below illustrates the approach of building a text classification model with generalizable capabilities.

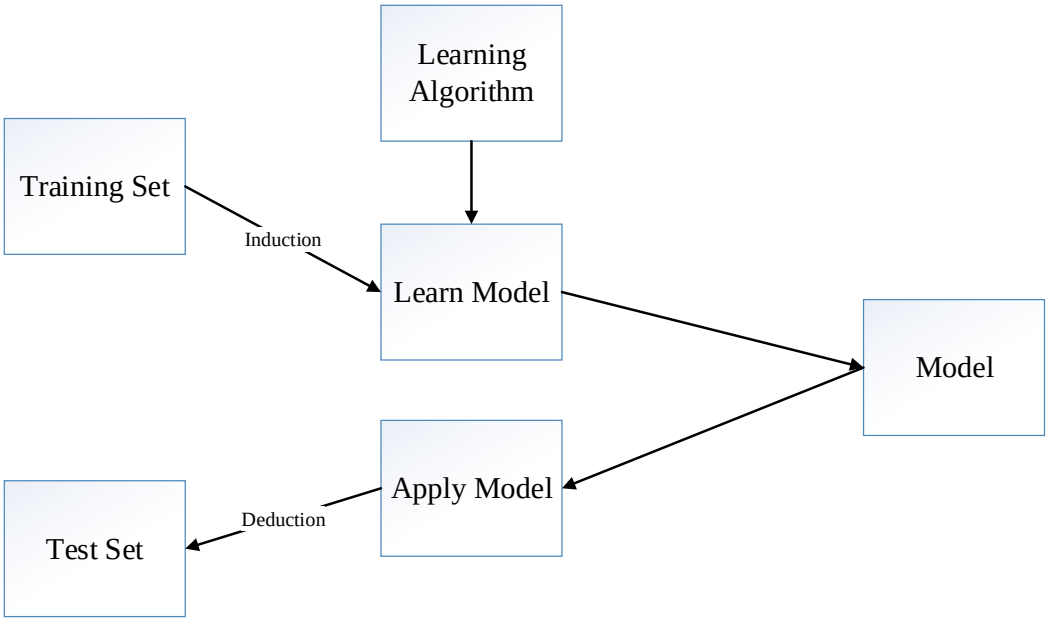

Figure 2. An Approach to Text Classification (Tan et al., 2006)

Text classification has been well studied for various types of cyber-crime like spam, identification of deceptive authors, detection of fake product reviews (Firte, Lemnaru, & Potolea, 2010; Pearl & Steyvers, 2012; Shojaee et al., 2013, 2013; Zheng, Qin, Huang, & Chen, 2003). Some of the popular algorithms used in text classification are: Naïve Bayes (NB), Support Vector Machines (SVM) and k-Nearest Neighbor (kNN).

NB is a popular classifier which has been applied to a number of text classification problems investigating different types of cybercrime like spam, phishing, intrusion detection (Abu-Nimeh, Nappa, Wang, & Nair, 2007; Fette, Sadeh, & Tomasic, 2007; Sommer & Paxson, 2010). When NB classifies records, it computes posterior probabilities for every class *C* instead of computing class conditional probabilities of each class *C* given a document *d* (Tan et al., 2006).

Equation 1 Naïve Bayes

$$\mathrm{P(c|d)} = \frac{P(c)\ \prod_{i=1}^{n} P(di\,|c)}{P(d)}$$

where d = {d_1, d_2, …, d_n} is a document with an attribute set of *n* attributes and *c* is a class label.

SVM is a popular classifier which is founded on statistical learning. The algorithm uses a concept of a maximal margin hyper-plane to linearly separate instances into two classes (Chang & Lin, 2001). For example given training examples, the class label *y* of a test example can be predicted using a linear function below (Tan et al., 2006):

Equation 2 Support Vector Machines

$$y = \begin{cases} 1, if\ w.z + b > 0; \\ -1, if\ w.z + b < 0 \end{cases}$$

where 1 and -1 are class labels, *w* and *b* are parameters of the decision boundary

There are also other classifiers, which are eager learners, like *kNN*, that delay to map the input data attributes to class labels until at that time when the training data is available. *kNN* uses a distance function to determine which instances are closest to the new example. When modeling the classifier, the class label of an instance is determined by a majority class of the nearest *k* neighbors, using a voting scheme as shown in the equation below (Aha, Kibler, & Albert, 1991; Tan et al., 2006).

Equation 3 k-Nearest Neighbor

$$y = \sum_{(xi,yi)\in D} I(v = yi)$$

where *v* is class label, y_i is the class label a nearest neighbors and I is a function that returns 1 if the function is true or zero.

CHAPTER THREE

METHODOLOGY

In this chapter, I discuss the methodology used to design my models for deception and cybercrime. I use four datasets from three web genres namely: Facebook, email and websites to study deception and cybercrime. Furthermore, I choose three well-studied classification algorithms to build the cybercrime detection models: NB, SVM and kNN. Figure 3 below is an illustration of the framework I will use to build my models for detecting and analyzing cybercrime in text messages. In this chapter, I discuss how data is collected and pre-processed and review techniques for selecting features as well as metrics for evaluating classifier performance for the models.

Data Collection

I use data from different sources of web genres namely: Facebook, email and a website. For Facebook data, I use a dataset for of 1036 publicly leaked email addresses of Nigerian cybercriminals who are using an online data theft service called *PrivateRecovery*, which was formerly called *BestRecovery* (Sarvari et al., 2014). This criminal network is notorious for scams like: advance-fee fraud, online dating scams and Nigerian chain letter scams. I conducted Facebook look-ups on each email address to identify corresponding public profiles for each account and its friends. I collected data

from timeline, groups and likes from each of 43,125 Facebook profiles. Table 6 below is a summary of data from 43,125 Facebook profiles.

I also use the Enron email dataset which was made public by the Federal Energy Regulatory Commission (FERC) during the Enron case (Cohen, 2015). This dataset contains about 500,000 emails from 150 employees in Enron's executive management team. I also obtain 89 emails that were part of the evidence used to prosecute Mr. Kenneth Lay, the former Chairman and Mr. Jeffrey Skilling, the Chief Executive Officer for securities and wire fraud in the Enron scandal. I use this court evidence from the Department of Justice (DoJ). The email dataset and court evidence were made public by the FERC and DoJ due to public interest in the Enron case.

Furthermore, I also use a publicly available dataset for online reviews which has both fake and truthful online reviews for hotels from travel websites (Ott et al., 2011). The dataset comprises 400 truthful reviews on hotels on a travel website and 400 fake reviews for the same hotels are collected from anonymous online workers within the United States using Amazon Mechanical Turk (AMT).

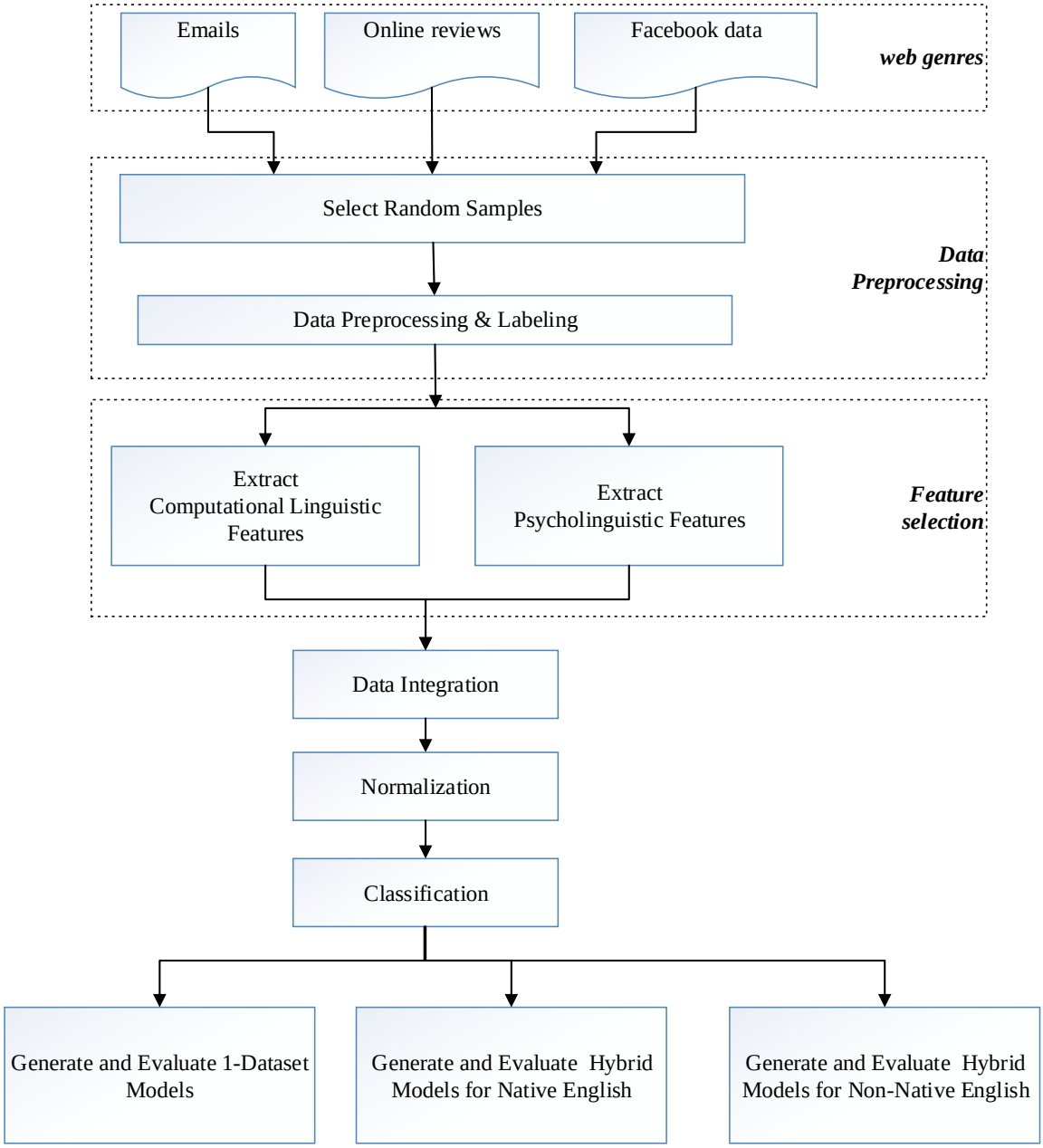

Figure 3. Approach for Detecting Cybercrime in Text-based Communication

Table 6 Summary of Datasets

Type of data	Type of Web Genre	Number of instances
Groups	Facebook	58,564
Likes	Facebook	56,755
Timeline posts	Facebook	1,897
Enron Email (FERC)	Email	500,000
Enron Court Evidence (DoJ)	Email	89
Favorable truthful reviews	Website	400
Unfavorable truthful reviews	Website	400
Favorable fake reviews	Website	400
Unfavorable fake reviews	Website	400

Ground Truth

The Facebook scams were obtained from a notorious cybercriminal network known for transmitting scams. I worked with team of three PhD students to manually verify and label scams and truthful for training and test sets. I considered the majority vote for both test sets

In the Enron datasets, all the 89 emails from DoJ which were used as evidence to prosecute the two Enron executives for securities and wire fraud were labelled as deceptive. Mr. Kenneth Lay was found guilty on all counts of securities and wire fraud while Mr. Jeffrey Skilling was found guilty on 19 counts of fraud and conspiracy. However, Mr. Skilling appealed only 1 count of fraud out of the 19 which was honest-services fraud, however, he was resentenced to 168 months instead of 292 months provided that he did not contest the original forfeiture and restitution order nor appeal or challenge his sentence or conviction (DoJ, 2013). The email dataset from DoJ was therefore labeled as deceptive because all the co-accused did not appeal all counts of

securities and wire fraud and may not do so. I labeled the emails released by FERC as truthful after excluding the evidence from the DoJ dataset.

For public e-commerce website dataset of favorable and unfavorable hotel reviews, all the fake reviews collected from the *Amazon Mechanical Turkers* were labeled as deceptive since the *Turkers* were tasked to generate fake favorable and unfavorable reviews for hotels. However, for truthful reviews, all hotel reviews with four-star and five-star rating were considered to be favorable truthful reviews while those with one-star and two-star rating were considered to be unfavorable truthful reviews. To verify the truthful reviews, I only considered reviews with transactional information like hotel booking deals as well as prices for valet services, meals, tips. This is because transactional information in reviews for e-commerce websites gives more assurance that a reviewer experienced a product or service from a provider (Fitzpatrick, Bachenko, & Fornaciari, 2015). I manually identified 84 out of 400 favorable truthful reviews and 78 out of 400 unfavorable truthful reviews with transactional information.

Data Preprocessing

I use real world data to create the model for detecting and analyzing cybercrime in text messages. The main problems with such real world data are: noise, missing values, inconsistencies and redundancy hence it has to first be cleaned and transformed to improve performance of machine learning algorithms (Larose, 2014). The Facebook data contained: non-ASCII symbols used in expressing emotion; emoticons; phrases in non-English languages local Nigerian dialects and pidgin; accented words in languages

like Spanish and French. To preprocess the email data, I removed the email headers; email addresses, prior email threads and text-based emoticons.

The next step in data preprocessing is data integration where the cleaned data from different repositories is merged for further processing. The data for this research is obtained from three repositories: Facebook, Enron email dump and US Department of Justice. After cleaning data from each repository, the Enron data is merged to create an Enron dataset. For the Facebook data, I randomly selected scams and non-scam messages, which I then merged to create a Facebook dataset.

I also considered using random sampling on Facebook and Enron data in Table 6 above to obtain subsets of data for analyzing cybercrime. This is because real world data is highly imbalanced. Considering the email data from Enron, we have a total of 500,000 emails compared to the 89 emails used as evidence in prosecuting the top two Enron executives. This implies that when selecting the random samples for the training data, the number of instances for both deceptive and truthful records should be the same to create representative and balanced datasets (He & Garcia, 2009; Tan et al., 2006).

All the instances for the datasets of the web genres were manually labelled as either truthful or deceptive since I am using supervised learning to build the cybercrime detection models. Deceptive instances are positive instances that have been manually identified as either scams, fake reviews or fraud while truthful instances are messages which are truthful.

All the instances in the datasets used in the experiments are then transformed using a technique called normalization. This is to ensure that values for all the features

are normalized between 0 and 1 since some values for some features were either extremely high or relatively low. I used normalization filter in WEKA to normalize all the instances for all the training and test sets. This is to ensure that the values are uniformly distributed to improve learning for the classifiers.

Feature Selection and Engineering

Feature selection is a technique in machine learning that is used to select a subset of features that are relevant for constructing a model that gives a good predictive accuracy. Feature selection is also useful for removing redundant features; noisy misleading data; and reducing the training time such that the algorithms run faster (Brownlee, 2014). There are there approaches to feature selection: embedded, filter and wrapper (Guyon & Elisseeff, 2003; Tan et al., 2006). Embedded approach is part of the learning algorithm, which when executed learns with the features to improve the predictive accuracy of the model. Filter methods use statistical models which are independent of the learning task to score and rank each feature. Ranks from filter methods determine whether a feature will be either removed or retained. Alternatively, wrapper methods are like black boxes in the learning algorithm that consider the selection task as search problem to determine the best subset of features. With a wrapper method, sets of attributes with high or low pair-wise correlation can be selected.

I identified features for the cybercrime detection models from both CL and PL processes which are linked to deception and cybercrime. The CL processes linked to deception and cybercrime are: quantity of words, lexical diversity, expressivity, non-immediacy and sentence complexity as summarized in Table 3 (Zhou et al., 2004). The

first CL process linked to deception and cybercrime is quantity of words. Criminals use messages with more quantity of lexical items especially verbs, nouns, modifiers and characters in deceptive messages compared to truthful messages. The second CL process is lexical diversity which measures the ratio of unique words in messages. Deceptive messages always have low lexical diversity compared to truthful messages. The third CL process is expressivity which measures the number of adjectives and adverbs. Deceptive messages have more adjectives and adverbs than truthful messages. The fourth CL process is non-immediacy which refers to the use self-references in messages to determine accountability. In text communication, deceptive messages will also have few self-references because cybercriminals do not want to be held accountable in their messages hence use more other refers like *she, he.* Lastly, I also considered sentence complexity as a CL process linked to deception and cybercrime. There are several attributes for measuring sentence complexity: average sentence length, redundancy, average word length, punctuation marks, pausality. I extracted all the CL features from the text messages in our datasets using python's natural language processing toolkit.

I also included PL features in the feature-set of the learning models to link linguistic behavior in text-based communication to deception and cybercriminal processes. I extracted these features using a tool called *Linguistic Inquiry and Word Count* (LIWC) because it has been widely used to study psychological relationships in text messages (Newman et al., 2003; Sandoval et al., 2015; Tausczik & Pennebaker, 2010). The PL processes relevant to deception and cybercrime are: quantity of words, emotion (i.e. negative emotion, positive emotion), first-person pronoun singular,

exclusive words as summarized in Table 4. I use these PL features to determine psychological behavior like verbosity, accountability, sublimated guilt, emotion, cognitive load and cognitive complexity can be linguistically linked to deception and cybercrime. Cybercriminals are verbose in cybercrime like scams and online fake reviews forgettable so that the messages can be recycled to exploit more victims. In deceptive text messages, cybercriminals will limit use of the first-person pronouns singular like *I, me and mine* to avoid being held accountable or even liable for their communication. To reduce cognitive complexity in their deceptive messages, cybercriminals will limit use of analytical words or phrases that express insight or certainty in communication. Furthermore, such deceptive messages will have more words of empathy which these cybercriminals use to share or express emotion to that they can manipulate and exploit their victims.

Evaluating Classifier Performance

There are various techniques that we adopt to evaluate performance of the classifiers: confusion matrix, precision, recall, f-measure and ROC area under curve. I am addressing a binary classification problem where the classes are either *deceptive* for text messages with scams, fraud and fake reviews or *truthful* messages. The positive class is the *deceptive* class since this is the class I am interested in predicting, while the negative class is the *truthful* class.

A confusion matrix is visual representation of performance of a classification model. As shown in the Table 7 below, $\mathbf{X_{11}}$ and $\mathbf{X_{00}}$ are the *true positive* and *true negative* examples respectively, that are correctly predicted. $\mathbf{X_{01}}$ are negative examples

which are predicted as positive (i.e. *false negative*) while $\mathbf{X_{10}}$ are positive examples which are predicted as negative (i.e. *false positive*). Also, using the confusion matrix, we can determine the accuracy and error rates of the classification model. Accuracy rate is the ratio of correct predictions (i.e. X_{11} and $\mathbf{X_{00}}$) to the total number of predictions while error rate is the ratio wrong prediction (i.e $\mathbf{X_{01}}$ and $\mathbf{X_{10}}$) to the total number of predictions. The accuracy and error rates are formally defined as below:

Equation 4 Accuracy Rate

$$Accuracy\ Rate = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 5 Error Rate

$$\boldsymbol{Error\ Rate = \frac{FN + FP}{TP + TN + FP + FN}}$$

Table 7 Confusion Matrix for Content-based Cybercrime Prediction Model

		Predicted Class	
		Class = 1	*Class = 0*
Actual Class	*Class = 1*	$\mathbf{X_{11}}$	$\mathbf{X_{10}}$
	Class = 0	$\mathbf{X_{01}}$	$\mathbf{X_{00}}$

Precision measures the proportion of records that are actually positive in the group which the classifier has declared as positive class. This is therefore a measure of the proportion of selected examples that are correct which is *Class 1* of the predicted class in Table 7 above.

Equation 6 Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall measure the proportion of positive examples that are predicted correctly hence true positive rate. This therefore is the proportion of correct examples that are selected which is class 1 of the actual class in Table 7 above.

Equation 7 Recall

$$Recall = \frac{TP}{TP + FN}$$

The F measure is a combined measure evaluates the trade-off between precision and recall. This measure is a harmonic mean between precision and recall as show in the equation below.

Equation 8 F1 Measure

$$F1\ Measure = \frac{2PR}{P + R}$$

A ROC curve is a graph that illustrates the trade-off between the benefits of true positives (TPR) and costs (FPR) of a binary classifier. The TPR is also called sensitivity while the FPR is called specificity. Each curve corresponds to the performance of a single

binary classifier. One classifier may perform better than another if its curve is much higher on the upper left most part of the curve.

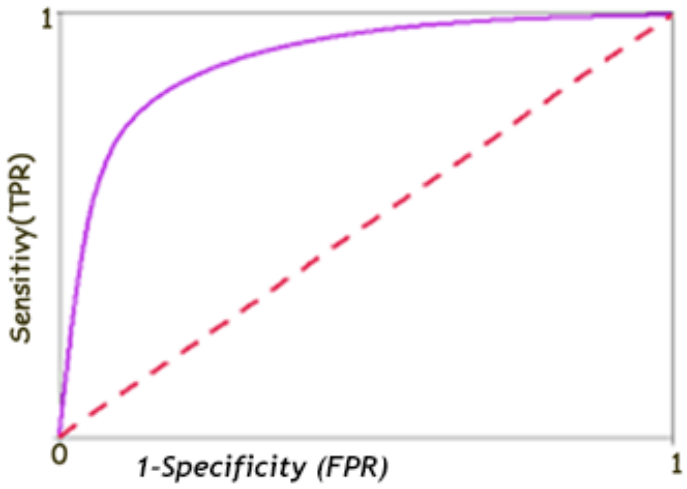

Figure 4 ROC Curves

Popularity of text messaging and ubiquitous adoption of text-based communication media as modern communication tools is attracting cybercriminals to exploit these tools for cybercrime. Filtering scams, online fake reviews and fraud using existing content filtering mechanism is becoming more challenging hence the need to explore an automated linguistic approach that use CL in NLP, psycholinguistics and machine learning. Using such an approach and text-based deception detection discourse, we explore how to use CL and PL can be implemented in constructing generalizable models for detecting scams, fraud and deceptive reviews in text-based communication.

CHAPTER FOUR

FINDINGS AND ANALYSIS

Dataset Description

I prepare three types of datasets for our cybercrime detection models namely: 1-dataset, 2-dataset and 3-dataset training sets. I use 1-dataset training sets to build individual models for each of the four types of cybercrimes for this research i.e. fraud, scams, favorable fake reviews and unfavorable fake reviews. Each 1-dataset model trained on a specific type of cybercrime and then used to detect and analyze cybercrime which was not part of the training model. For instance, a model trained on a dataset of fraud is used to detect scams, fraud, favorable fake reviews and unfavorable fake reviews. Table 8 describes the 1-dataset training sets for models. Similarly, for 2-dataset training sets, which are used in building the hybrid models, are generated by combining any two of the four single datasets, which are then used to detect cybercrime in datasets that are not part of the training model. For example, a 2-dataset training set trained on scam and fraud is used to detect and analyze unfavorable and favorable fake reviews. Table 9 describes the 2-dataset training sets for our hybrid models. Lastly I also generated 3-dataset training sets from any three of the four datasets representing each type of cybercrime. The training model was then used to detect and analyze cybercrime from a dataset which was not part of the training model. For instance, a model trained on fraud, scam and favorable fake reviews was used to detect and analyze unfavorable fake

reviews. Table 10 describes 3-dataset training train sets for the second type of hybrid models.

Table 8 Description of Sample for Single Datasets (1-Datasets)

Dataset	Model	Type of Cybercrime Within the Dataset	Web genre	# Instances in Train Set	# Instances in Test Set
Enron	EN	Fraud	Email	100	20
Facebook	FB	Scam	Facebook	100	20
Unfavorable Reviews	NR	Unfavorable Fake Reviews	Website	100	20
Favorable Reviews	PR	Favorable Fake Reviews	Website	100	20

Table 9 Description of Sample for 2-Datasets for Hybrid Models

Hybrid Model	Description for Hybrid Datasets	#Instances in Train Set
EN + FB	Enron & Facebook	200
EN + NR	Enron & Unfavorable Reviews	200
EN + PR	Enron & Favorable Reviews	200
FB + NR	Facebook & Unfavorable Reviews	200
FB + PR	Facebook & Favorable Reviews	200
NR + PR	Favorable & Unfavorable Reviews	200

Table 10 Description of Sample for 3-Datasets for Hybrid Models

Hybrid Model	Description for Hybrid Datasets	# Instances in Train Set
EN+FB+NR	Enron, Facebook & Unfavorable Online Review datasets	300
EN+FB+PR	Enron, Facebook & Favorable Reviews	300
FB+NR+PR	Facebook, Unfavorable & Favorable Reviews	300
EN+PR+NR	Enron, Favorable & Unfavorable reviews	300

Feature Selection and Engineering

For each of the training and testing sets I extract 29 features from CL and PL processes linked to deception and cybercrime. We identify 16 features for the CL process and 13 features for the PL process as described in Table 11. Using python's NLP and POS tagging, we derive the CL features for the learning model while for the PL features we use Linguistic Inquiry Word Count (LIWC) tool. LIWC is a text analysis tool for analyzing words in respect to behavior and psychological processes (Tausczik & Pennebaker, 2010).

The relationship of these features to deception and cybercrime is explained below:

- verbs - this feature measures the frequency of the verbs in the datasets. Deceptive messages with scams and fraud will contain more verbs because the criminals sending such messages want to be non-committal. Some examples of such non-committal verbs expressing deception include think, sort of, guess, believe (Clikeman, 2012).
- modifiers - these are words used to add sense to head nouns. Deceptive messages will use many modifiers like *kind of, somewhat*, etc., because non-truth tellers are non-committal to avoid being held responsible for their messages.
- average sentence length - the average sentence length of deceptive messages will be greater than that of truthful messages (Zhou et al., 2003). This because non-truth tellers are verbose to dominate the conversation.
- average word length – non-truth tellers will use shorter phrases in attempt to reduce cognitive load in their communication.

- punctuation marks - measures the frequency of punctuation marks used in sentences. Deceptive messages will have more punctuation marks than truthful messages since cybercriminals tend to be verbose in their scams and fraud communication.
- Pausality – this measure that ratio of function words per sentence. Deceptive messages will have a higher rate of pausality than truthful messages.
- modal verbs - this is a frequency of modal verbs. Modal verbs indicate obligation or ability like can, shall, might, will etc. Deceptive messages will have less modal verbs because cybercriminals are non-committal in their communication.
- emotiveness - measures emotion as a ratio of number of adjectives and adverbs to number of verbs and nouns (Zhou et al., 2004).
- lexical diversity - measures the ratio of unique words per sentence. Deceptive messages will have lower lexical diversity.
- Number of function words - frequency of function words in messages. Function words are used in constructing grammatical relationships within sentences for example: prepositions (e.g. of, at without, etc.), pronouns (e.g. he, they, it, etc.), determiners (e.g. the, a, that, etc.), conjunctions (e.g. and, when, while etc.), auxiliary verbs (e.g. is, am, be, have, got, do etc.) and particles (e.g. no, not nor, etc.). Deceptive messages will have more function words than truthful messages since cybercriminal are

verbose and attempt to dominate conversations. Hence the ratio of function words in messages will be higher in deceptive messages than in truthful messages.

- redundancy - measures the ratio of function of words used per sentence. Deceptive messages will have higher redundancy than truthful messages.
- Number of characters - Frequency of characters in sentences. Deceptive messages will have higher character frequency because liars are more verbose than truth-tellers.
- Number of Sentences - Frequency of sentences in messages. Deceptive messages will have more sentences.
- Number of adjectives - Number of adjectives in messages. Adjectives are words that describe other words. Certain adjectives will be used by deceptive messages by cybercriminals to make their conversations or messages vague and, or ambiguous.
- Number of adverbs - these are phrases that modify verbs and adjectives with respect to a place, time and circumstance e.g. quite, suddenly etc. Cybercriminals can also abuse adverbs to craft deceptive messages that are vague and ambiguous.
- Number of nouns - frequency of nouns in messages. Since adverbs modify nouns the frequency of nouns in deceptive messages should also be more compared to truthful messages.

- Analytical - this psycholinguistic feature measures analytical phrases within messages. Cybercriminals use less analytical words in their scams since their messages have low cognitive complexity.
- Words per sentence - this is a measure of the number of words per sentence. Deceptive messages will have less words per sentence since deceptive messages have short average length but more sentences in message compared to truthful messages.
- Six letter words - Words with more than six characters measure longer words hence high cognitive complexity. Deceptive messages will have fewer words greater than six characters to keep messages less cognitively complex.
- *I* - Cybercriminals will use fewer first personal pronouns singular, *I,* in their deceptive messages to avoid being held accountable hence dissociate themselves from their messages.
- *we* - Cybercriminals will use fewer first personal pronouns plural, *we,* in their deceptive messages to avoid being held accountable hence dissociate themselves from their messages.
- *you* - Cybercriminals will use more second personal pronouns, *you,* in their deceptive messages to hold others accountable hence dissociate themselves from their messages.

- *she/he* - Cybercriminals will use more third personal pronouns *she/he* in their deceptive messages to hold others accountable hence dissociate themselves from their messages.
- affect - cybercriminals use more affect language in their messages to express emotions like pity, sympathy etc.
- positive emotion - Cybercriminals use more positive emotion in deceptive messages to express enthusiasm to trick their victims
- negative emotion - cybercriminals use more negative emotion in deceptive messages to express to pain, fear, sympathy, loneliness to attract and exploit their victims. Deceptive messages will have more negative emotion words which can be used to determine sublimated guilt.
- insight - cybercriminals use less words of insight in their text messages to reduce cognitive complexity in their communication.
- cause - cybercriminals use less words on causation like because, effect to reduce cognitive complexity in their communication.
- certain - Cybercriminals use less words on certainty in their text messages since their messages have low cognitive complexity and are also vague and ambiguous

Table 11 Summary of Feature-set for cybercrime models

Feature	Linguistic Process	Description
Verbs	CL	Measures frequency of verbs in the datasets where liars use non-committal verbs
Modifiers	CL	Frequency of modifiers will be greater in deceptive messages compared to truthful messages
Averages sentence length	CL	Average sentence length of deceptive messages will be greater than in truth messages.
Average word length	CL	Deceptive messages will usually be shorter than truthful messages.
Pasuality	CL	Ratio of punctuation marks in the messages. Deceptive messages will have more punctuation marks.
Modal verbs	CL	Frequency of verbs that indicate obligation. Since cybercriminals are non-committal, fewer modal verbs will be used in deceptive messages.
Emotiveness	CL	Measure of emotion as a ratio of number of adjectives and adverbs to number of nouns and verbs. Deceptive messages will have higher ratio of emotiveness
lexical diversity	CL	Measure ratio of unique words per sentence. Deceptive messages will have lower lexical diversity.
redundancy	CL	Measure ratio of function words in sentences. deceptive messages will have more redundancy.
Number of characters	CL	Frequency of characters in sentences. Deceptive messages will have higher character frequency.
Number punctuation marks	CL	Frequency of punctuation marks. Deceptive messages will have more punctuation marks than truthful messages.
Number of sentences	CL	Frequency of sentences in messages. Deceptive messages will have more sentences
Number of adjectives	CL	Deceptive messages can be vague and ambiguous hence adjectives will be used for this purpose
Number of adverbs	CL	Deceptive messages can be vague and ambiguous hence adverbs will be used for this purpose. Adverbs are words that modify nouns
Number of nouns	CL	Frequency of nouns. Deceptive messages will have more nouns compared to truthful messages
Number of function words	CL	Frequency of function words. Deceptive messages will have more function words compared to truthful messages
Analytic	PL	Cybercriminals use less Analytical words in their scams since their messages have low cognitive complexity
Word per Sentence	PL	Words per sentence
Six letter words	PL	Words with more than six characters. Measures longer words hence cognitive complexity. Deceptive messages will have fewer words greater than six characters to keep messages less cognitively complex
I	PL	Cybercriminals will use fewer first personal pronouns

		singular *I* in their deceptive messages to avoid being held accountable.
we	PL	Cybercriminals will use fewer first personal pronouns *we* plural in their deceptive messages to avoid being held accountable
you	PL	Cybercriminals will use more second personal pronouns *you* in their deceptive messages to hold others accountable
She/he	PL	Cybercriminals will use more third personal pronouns singular *she/he* in their deceptive messages to hold others accountable
affect	PL	Cybercriminals use more affect language in their messages to express emotions like pity, sympathy etc.
positive emotion	PL	Cybercriminals use more positive emotion cybercrime to express enthusiasm to trick their victims
negative emotion	PL	Cybercriminals use more negative emotion in cybercrime (like romance scams) to express to pain, fear, loneliness
insight	PL	Cybercriminals use less words on insight in their scams since their messages have low cognitive complexity
cause	PL	cybercriminals use less words on causation like because, effect, hence in their scams since their messages have low cognitive complexity
certain	PL	Cybercriminals use less words on certainty in their scams since their messages have low cognitive complexity and are also vague and ambiguous
class {deceptive, truthful}		These are the class labels for the binary classifiers

Feature Analysis

In this section, I show that it is possible discriminate cybercrime text messages from truthful messages using binary classifier with truthful and deceptive labels. Figure 5 and Figure 6 represent stacked line graphs for CL and PL features respectively, and reveal that it is possible to discriminate between cybercrime and truthful text messages even within cross section of CL and PL features of the training sets.

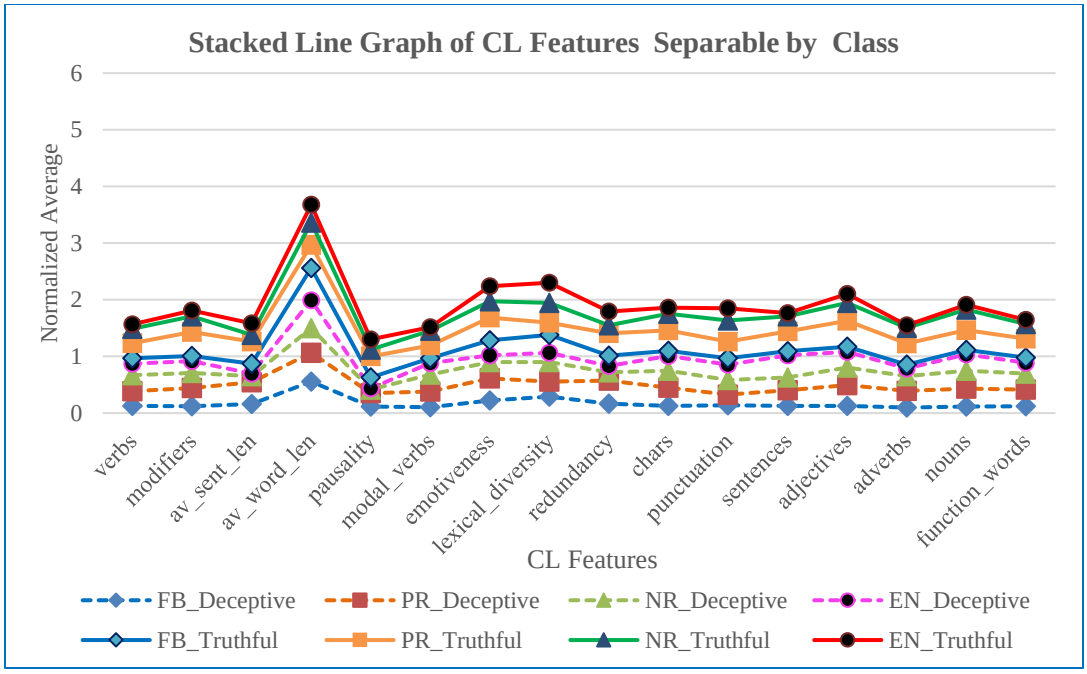

Figure 5 Stacked Line Graph for CL Features Separable by Binary Class

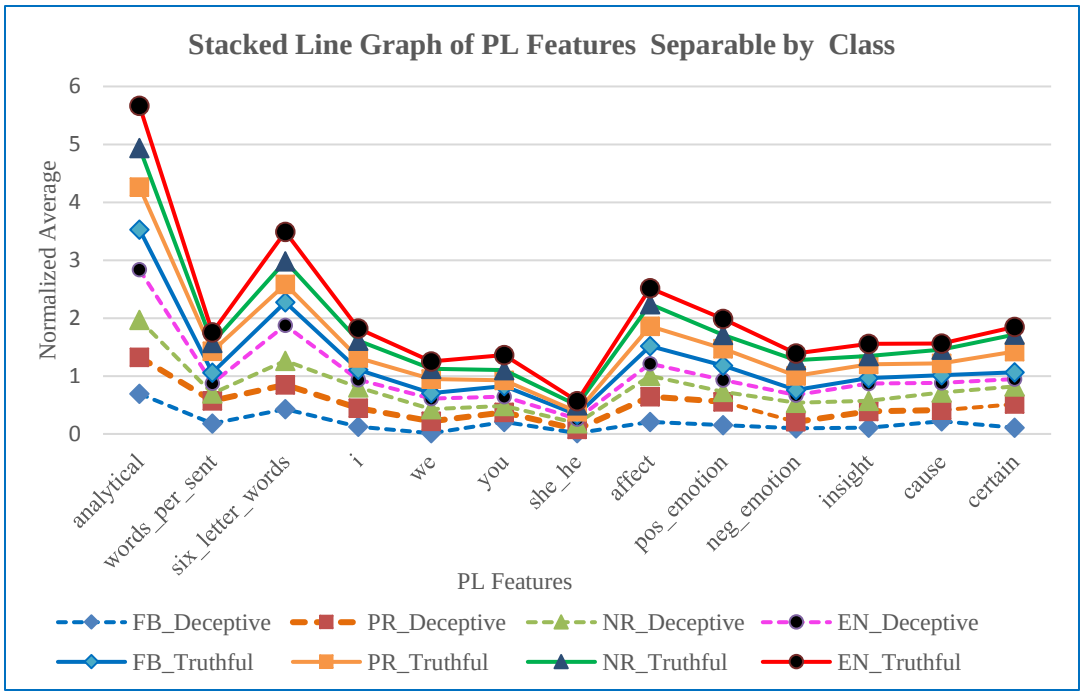

Figure 6 Stacked Line Graph for PL Features Separable by Binary Class

I build hybrid 2-dataset and 3-dataset models for detecting and analyzing cybercrime in native and non-native English speaking cybercriminal networks. In Figures 5 and 6, I observe that the Facebook dataset, which represents the non-native English speaking cybercriminal networks, has low values in both graphs that illustrate the deception and truthful messages can be discriminated in both CL and PL feature-set learning sets. This motivates my work to identify and generate dataset-specific features from the Facebook dataset for non-native English speaking cybercriminals. Using class-based CL and PL Feature graphs, shown in Figures 7-10, I eliminate all features in both classes for CL and PL features whose normalized average is below 0.1 so that we retain Facebook dataset-specific features for hybrid models for detecting and analyzing cybercrime in non-native English speaking cybercriminal networks.

The CL features I identify for detecting cybercrime in non-native English speaking cybercriminal networks are average sentence length, average word length, pausality, emotiveness, lexical diversity, redundancy and punctuation. For PL features we identify I, we, you, affect, positive emotion, cause and certain.

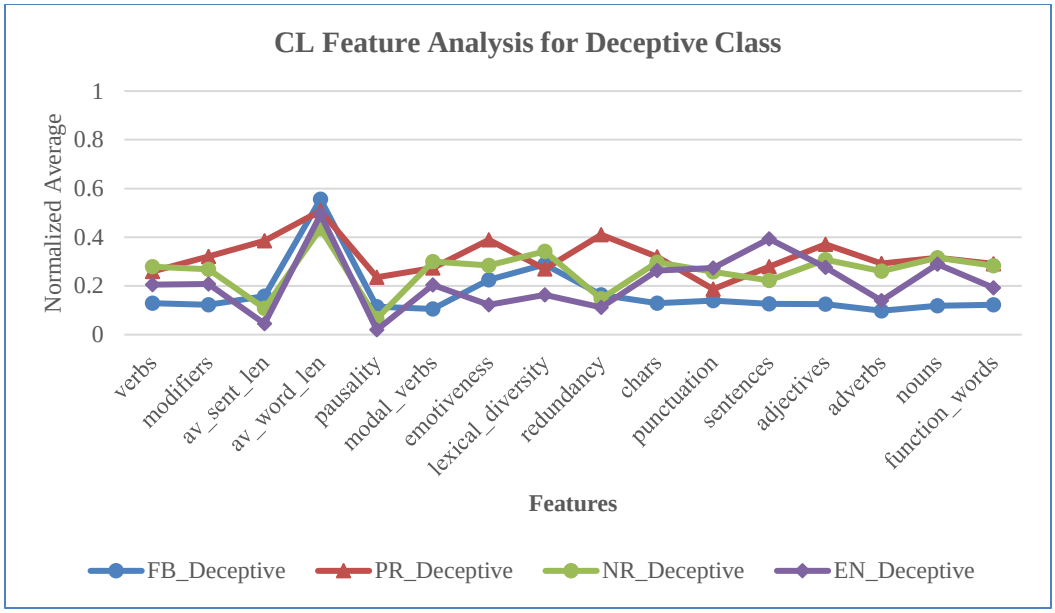

Figure 7 CL Feature Analysis Graph for the Deceptive Class

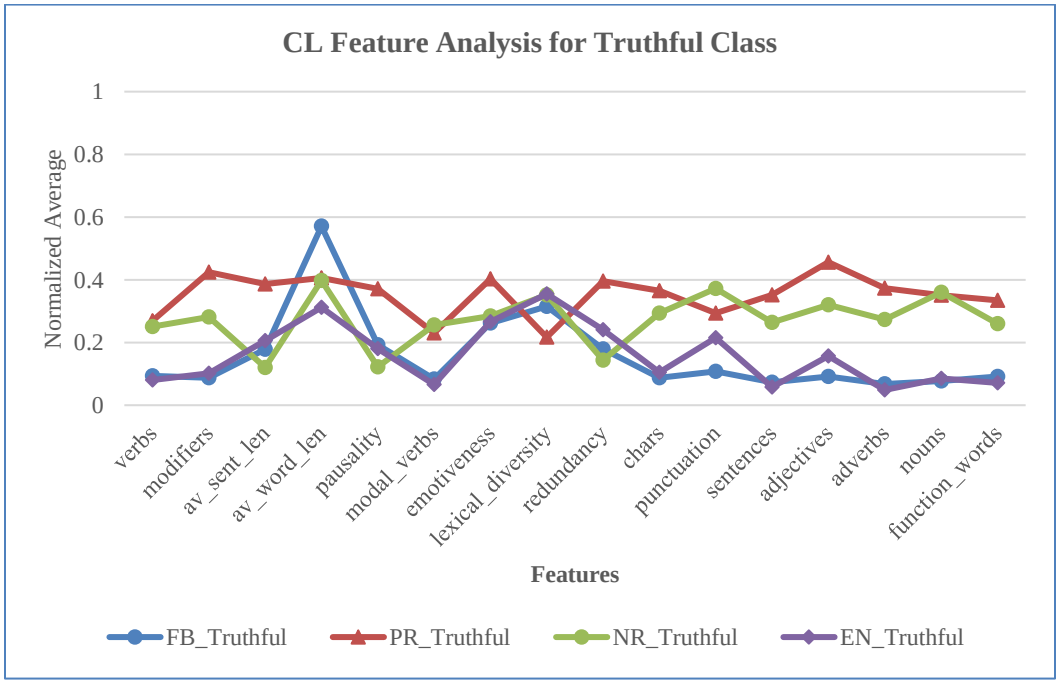

Figure 8 CL Feature Analysis Graph for the Truthful Class

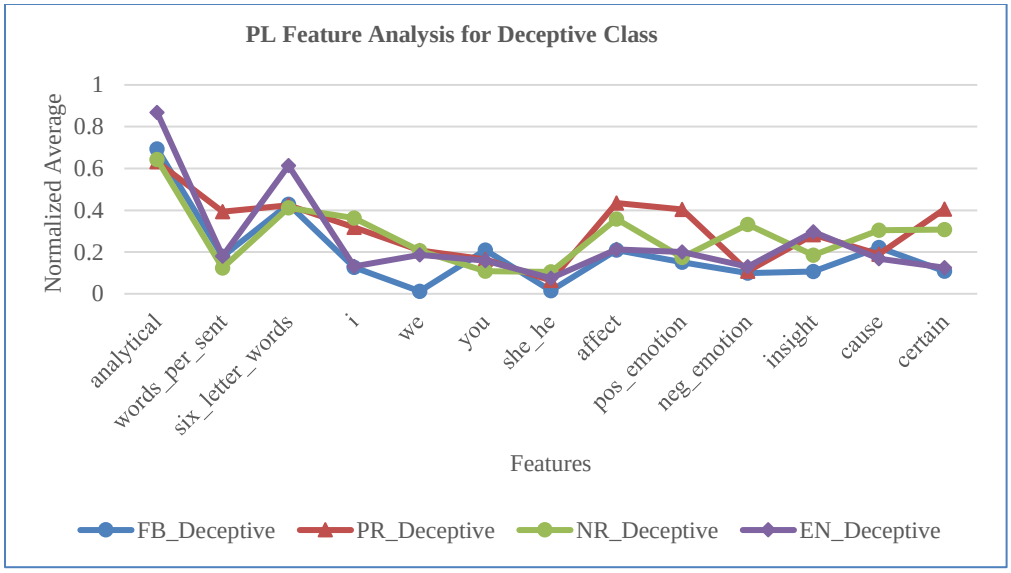

Figure 9 PL Feature Analysis Graph for the Deceptive Class

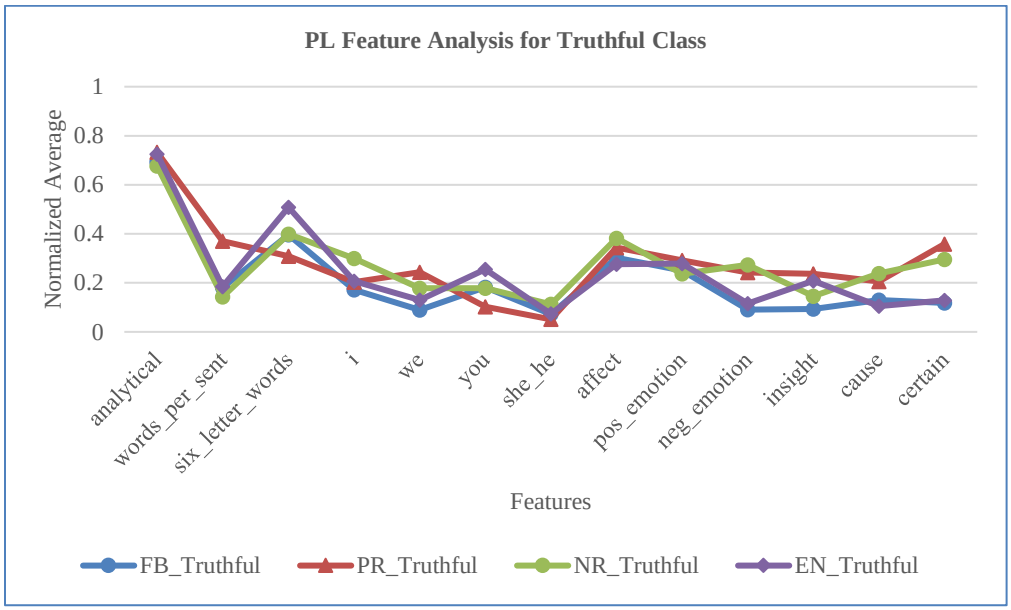

Figure 10 PL Feature Analysis Graph for the Truthful Class

1-Dataset Non Hybrid Models for Cybercrime

Table 12 Evaluation for EN and FB Models using PCA Components

Classifier	Model	P	R	F	ROC AREA
NB_pca	EN	0.927	0.927	0.927	0.966
SVM_pca	EN	0.967	0.966	0.966	0.966
kNN_pca	EN	0.881	0.876	0.876	0.866
NB_pca	FB	0.918	0.91	0.91	0.927
SVM_pca	FB	0.941	0.94	0.94	0.94
kNN_pca	FB	0.845	0.84	0.839	0.84

I build classifiers using NB, SVM and kNN, I first explored PCA for two models, that is, EN and FB models. Table 11 is a summary of the evaluation of the classifiers for the models. As shown in Figure 10 the NB_{PCA} classifier for EN model predicts scams with 50% accuracy, while the NB_{PCA} classifier for FB predicts fraud with 60% accuracy. On the hand, the SVM_{PCA} classifier for the EN model predicts scams with 40% accuracy while in the FB model, the classifier predicts with 30% accuracy. Lastly for the kNN_{PCA} classifier, the EN model predicts scams with 50% accuracy while the FB model predicts fraud with 40%.

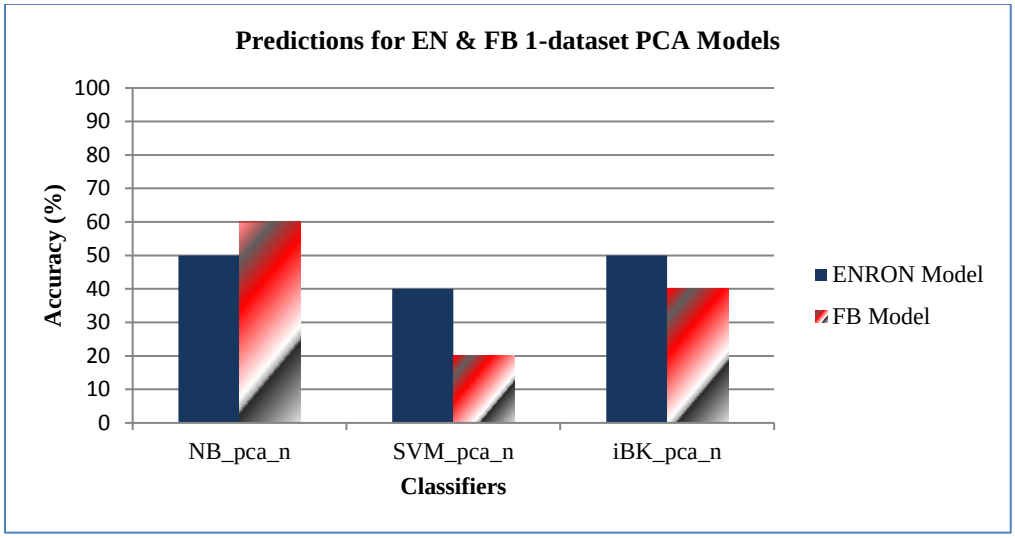

Figure 11 Accuracy for EN and FB using PCA

The training models for 1-dataset using PCA analysis in Table 12 were promising especially the NB_{PCA} which is trained on scams and detects fraud with 60% accuracy. These results motivated me to explore ways to train 1-Dataset generalizable cybercrime models that could detect cybercrime. I considered adding more datasets to the experiment as well as identifying better techniques in feature selection and engineering to improve our 1-Dataset learning models. I use four normalized 1-dataset training sets namely: EN, PR, NR and PR and this time we do not use PCA components. The overall results for the four models from 1-dataset training sets in Figures 12-16 are indicate an improvement compared to the results in Figure 11 where we use PCA components.

Table 13 Evaluation for 1-Dataset Training Models

MODEL	CLASSIFIER	P	R	F	ROC
EN	NB	0.918	0.91	0.91	0.972
EN	SVM	0.97	0.97	0.97	0.97
EN	KNN	0.891	0.86	0.857	0.96
FB	NB	0.622	0.62	0.619	0.625
FB	SVM	0.667	0.66	0.657	0.66
FB	KNN	0.603	0.6	0.597	0.586
NR	NB	0.622	0.62	0.619	0.662
NR	SVM	0.7	0.7	0.7	0.7
NR	KNN	0.688	0.68	0.677	0.715
PR	NB	0.735	0.73	0.729	0.835
PR	SVM	0.802	0.8	0.8	0.8
PR	KNN	0.78	0.78	0.78	0.839

I build four 1-dataset models where each has three classifiers. Table 13 is a summary of the performance evaluation for each of the three classifiers used in the training these models. I use precision, recall, F, and ROC area under curve as the performance measurement metrics for the models. Our results indicate that classifiers for 1-dataset training models perform well in all the four metrics that we use.

The first 1-dataset model is the EN model which is trained on fraud and non-fraudulent text messages in the email web genre to detect cybercrime in the Facebook and website web genres. This model detects favorable fake reviews with 60% accuracy both the NB and kNN classifiers. The summary of this model's performance in detecting cybercrime is summarized in Figure 12.

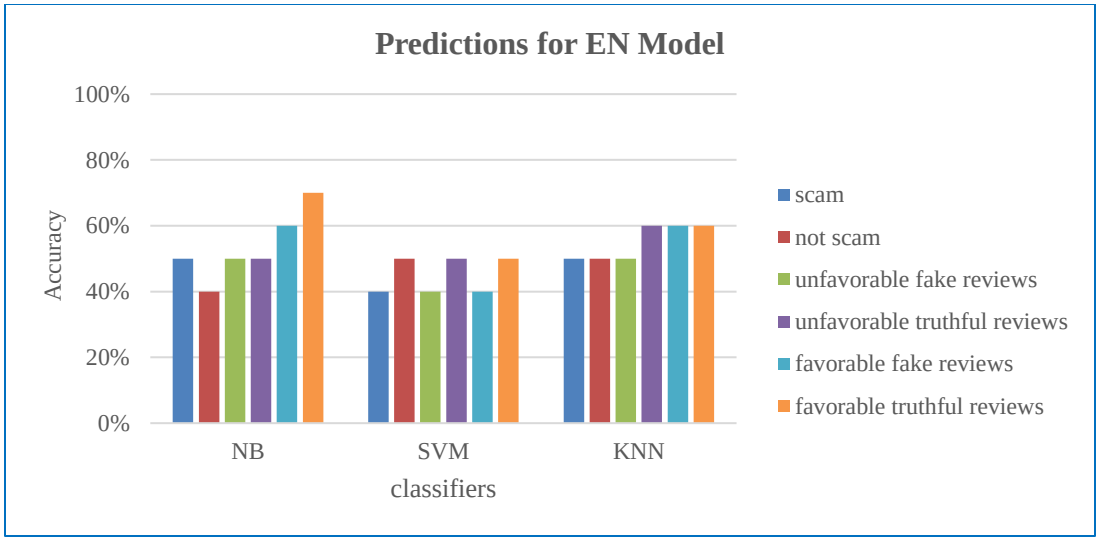

Figure 12 Predictions for 1-Dataset EN Model

The second 1-dataset model is the FB model, which is trained on both scams and messages which are not scam in the Facebook web genre to detect cybercrime in email and website web genres. This model detects fraud with 60% accuracy with the NB classifier. The summary of this model's performance in detecting cybercrime is shown in Figure 13.

The third 1-dataset model is the NR model, which is trained on both unfavorable truthful and fake reviews in the website web genre to detect cybercrime in the email, Facebook and website web genres. This model detects fraud with 80% accuracy using the SVM classifier and 70% accuracy with kNN classifier. For scams detection, the NB classifier predicts with 70% accuracy while SVM classifier detects scams with 60% accuracy. Furthermore, the same also detects unfavorable fake reviews are detected with 70% accuracy with SVM and 80% accuracy with kNN classifier. Summary of the model's performance in detecting cybercrime is summarized in Figure 14.

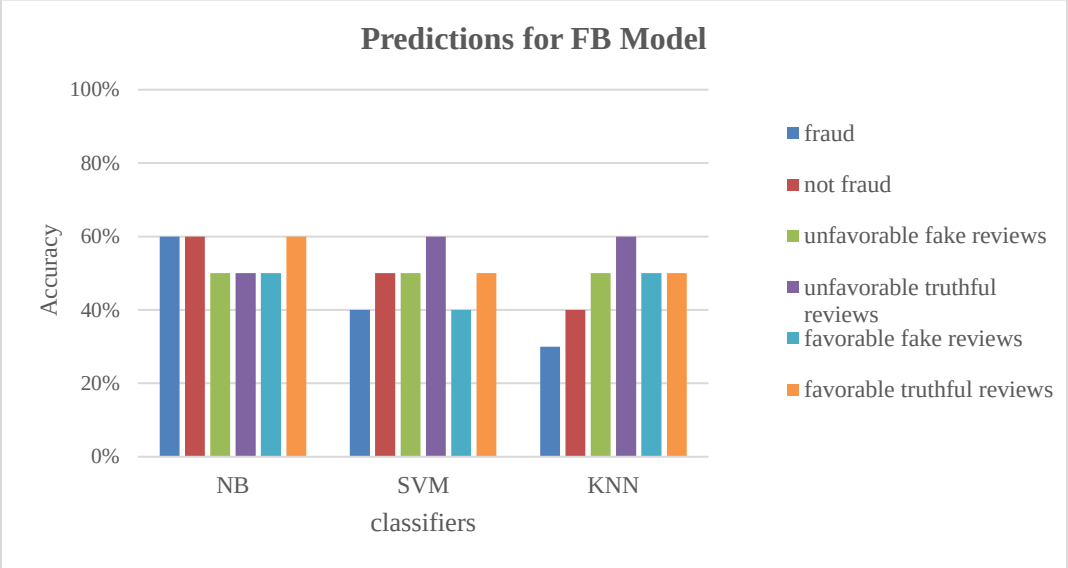

Figure 13 Predictions for 1-Dataset FB Model

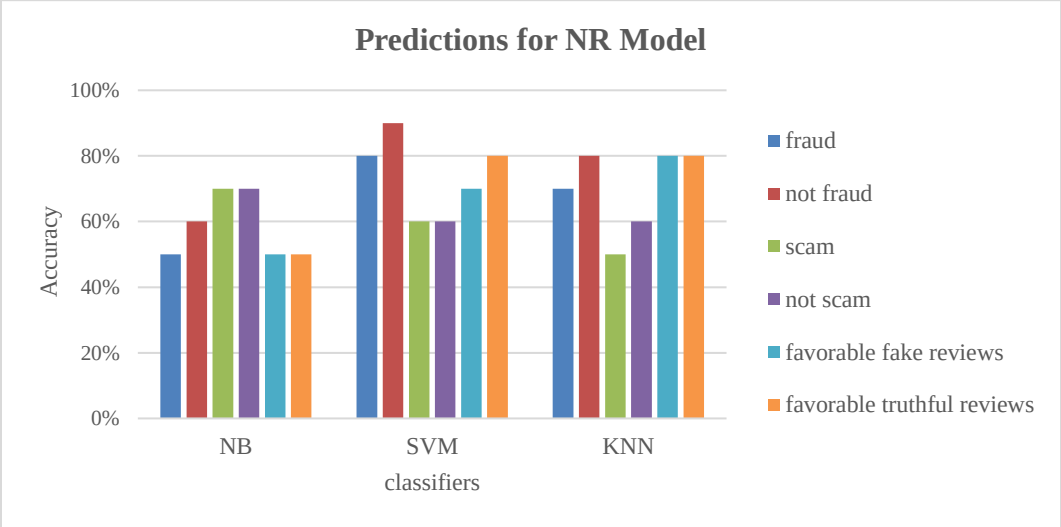

Figure 14 Predictions for 1-Dataset NR Model

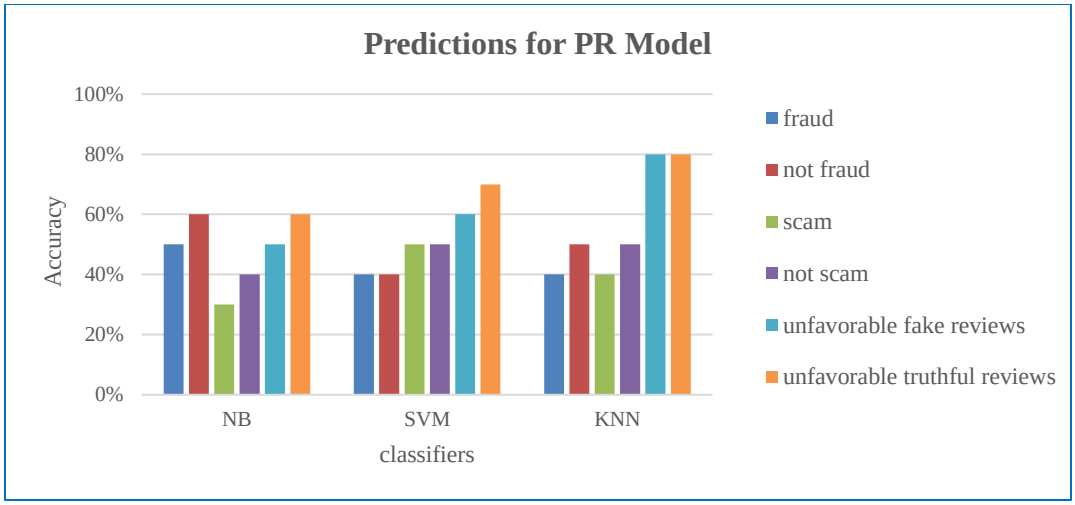

Figure 15 Predictions for 1-Dataset PR Model

Lastly, the fourth 1-dataset model, which is the PR model, is trained on favorable reviews in the website web genre to detect cybercrime in the email, Facebook and website web genres. This model detects unfavorable fake reviews with 70% accuracy with the SVM classifier and 80% with the kNN classifier as shown in Figure 15.

2-Dataset Hybrid Models for Native English Cybercriminal Networks

In this section, I review performance of 2-dataset hybrid models which are trained on two datasets combined from different web genres. I build six 2-dataset hybrid models each of which has three classifiers. I use these 2-dataset hybrid models to detect cybercrime in native English cybercriminal networks. Table 9 summarizes how I combine two datasets to generate training sets for the 2-datasets hybrid models. Table 14 summarizes results for performance of the classifiers for these hybrid models. I use precision, recall, F, and Receiver Operator Curve area as the performance measurement

metrics for the models. The results indicate that classifiers for 2-dataset training models perform well in all the four metrics that we use.

Table 14 Evaluation for 2-Dataset Models for Native English Cybercriminal Networks

MODEL	CLASSIFIER	P	R	F	ROC
EN + FB	NB	0.936	0.935	0.935	0.935
EN + FB	SVM	0.751	0.75	0.75	0.75
EN + FB	KNN	0.777	0.765	0.762	0.754
EN + NR	NB	0.668	0.665	0.664	0.731
EN + NR	SVM	0.777	0.77	0.769	0.77
EN + NR	KNN	0.753	0.75	0.749	0.753
EN + PR	NB	0.731	0.72	0.717	0.819
EN + PR	SVM	0.805	0.805	0.805	0.805
EN + PR	KNN	0.786	0.785	0.785	0.788
FB + NR	NB	0.599	0.595	0.59	0.593
FB + NR	SVM	0.652	0.65	0.649	0.65
FB + NR	KNN	0.612	0.61	0.608	0.603
FB + PR	NB	0.626	0.61	0.597	0.654
FB + PR	SVM	0.727	0.715	0.711	0.715
FB + PR	KNN	0.656	0.655	0.655	0.646
NR + PR	NB	0.667	0.66	0.657	0.709
NR + PR	SVM	0.762	0.76	0.76	0.76
NR + PR	KNN	0.677	0.675	0.674	0.736

The first 2-dataset hybrid model is EN+FB, I train using two combined datasets from email and Facebook web genres respectively to detect favorable and unfavorable reviews in the website web genre. This model is trained on both fraudulent and non-fraudulent emails as well as Facebook posts with scams and without scams. The NB classifier for this model detects unfavorable fake reviews with 60% accuracy. Results on performance of this model in detecting cybercrime are shown in Figure 16.

The second 2-dataset hybrid model is EN+NR and is trained on combined two datasets from the email and website web genres respectively, to detect cybercrime in the Facebook and website web genres respectively. This model is trained on both fraudulent and non-fraudulent emails as well as unfavorable truthful and fake reviews. NB and SVM classifiers for this model only detect favorable fake reviews. The model detects favorable fake reviews 60% accuracy with SVM classifier as shown in Figure 17.

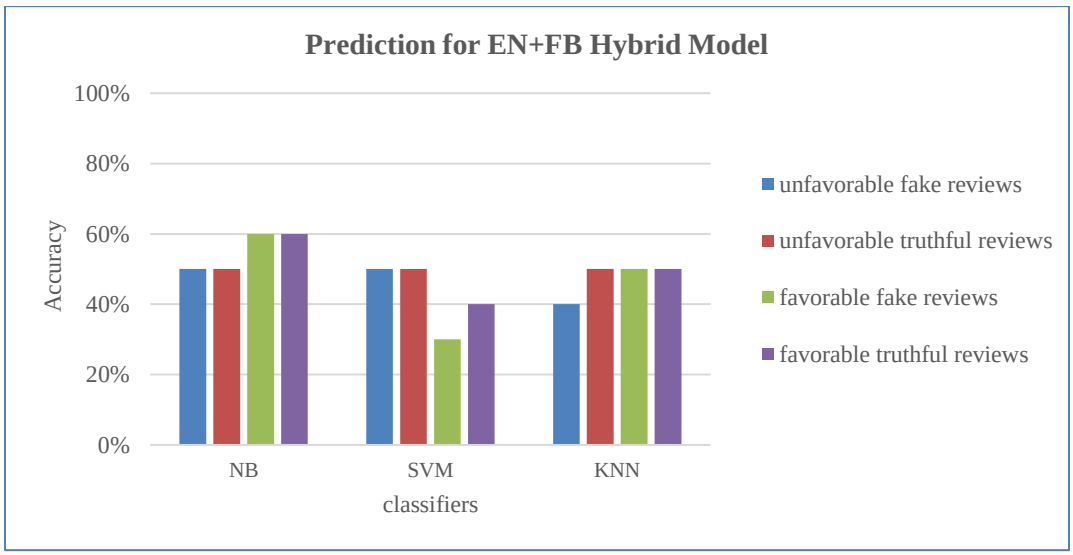

Figure 16 Predictions for 2-Dataset EN+FB Model

The third 2-dataset model is the EN+PR model, which is trained on combined datasets of Enron and favorable reviews from the email and website web genres respectively to detect cybercrime in the Facebook and website web genres. The SVM and kNN classifiers detect only unfavorable fake reviews with 70% accuracy respectively. Figure 18 summarizes the predictive performance of this model.

The fourth 2-dataset hybrid model is FB+NR which is trained on combined datasets from the Facebook and website web genres respectively to detect cybercrime from email and website web genres. This model detects cybercrime with all three classifiers for detecting detect fraud as shown in Figure 19. The model detects only detect fraud with 70% and 80% accuracy with the NB and SVM classifiers respectively.

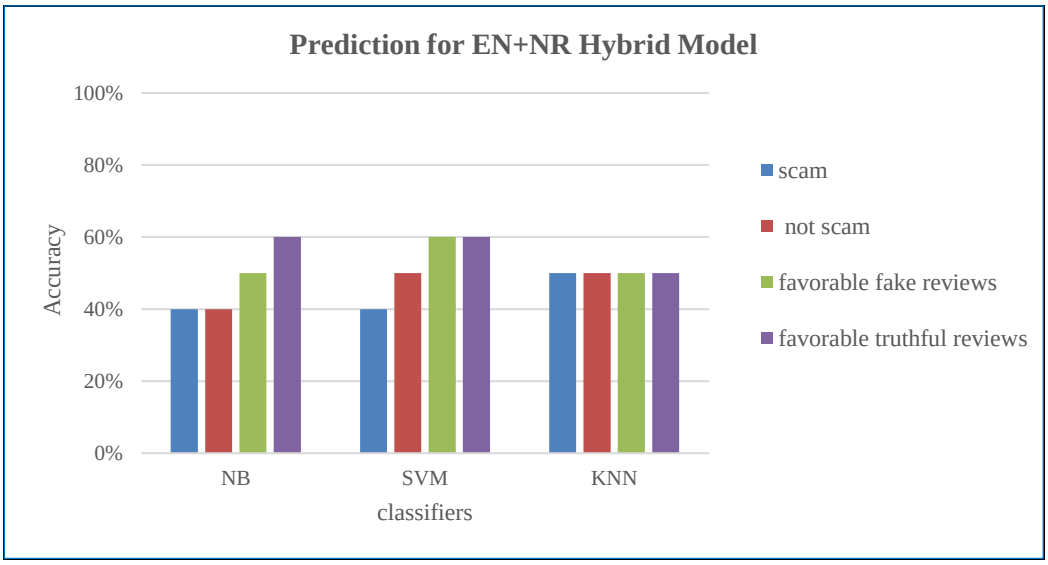

Figure 17 Predictions for 2-Dataset EN+NR Model

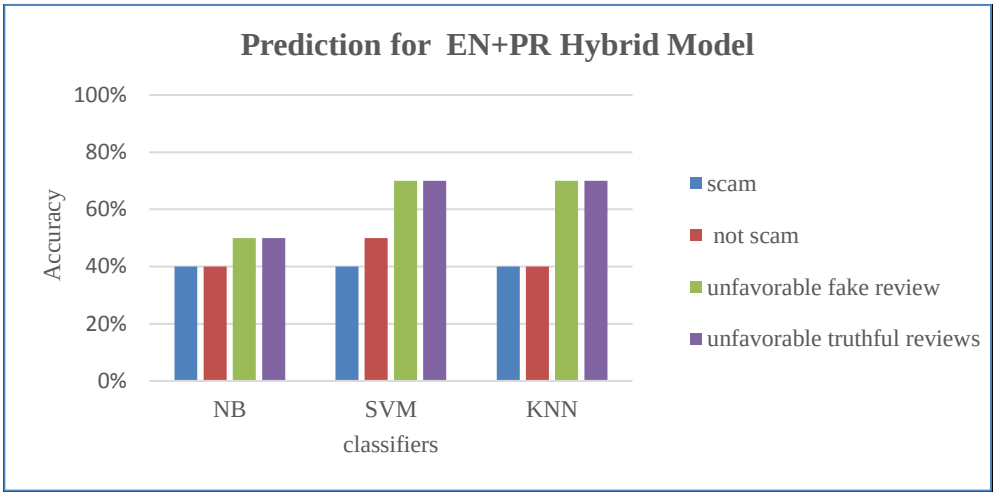

Figure 18 Predictions for 2-Dataset EN+PR Model

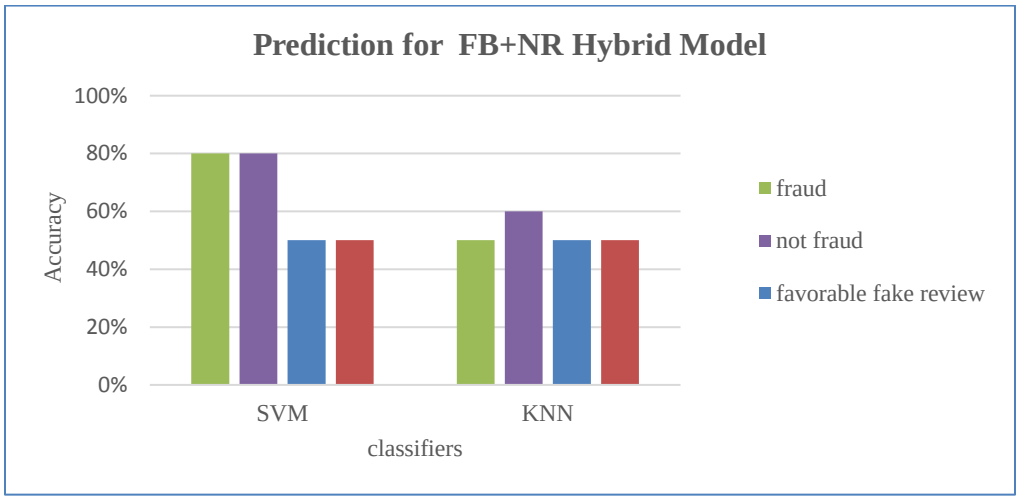

Figure 19 Predictions for 2-Dataset FB+NR Model

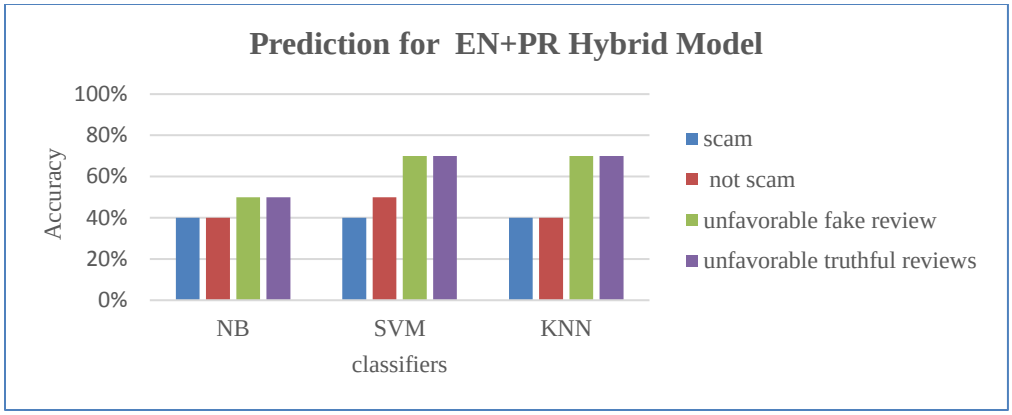

Figure 20 Predictions for 2-Dataset FB+PR Model

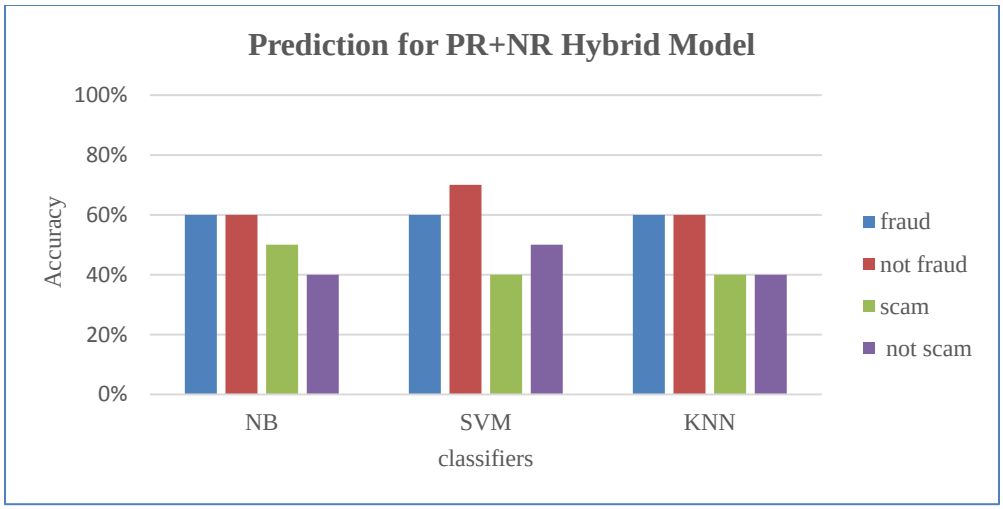

Figure 21 Predictions for 2-Dataset PR+NR Model

The fifth 2-dataset hybrid model is FB+PR, which is trained on two combined datasets from Facebook and website web genres to detect cybercrime in the Facebook and email web genres. The model detects fraud and unfavorable fake reviews with three classifiers as shown in Figure 20. The model detects fraud with 70% accuracy using NB classifier, 60% accuracy with SVM classifier and 60% accuracy with KNN classifier.

Futhermore the detects unfavorable fake reviews with 60% accuracy with NB classifier, 80% accuracy with SVM classifier and 70% accuracy with KNN classifier.

Lastly, the sixth 2-dataset model is PR+NR, is trained on combined datasets favorable and unfavorable reviews to detect fraud and scams in the email and Facebook web genre respectively. Results for performance of this model which are shown in Figure 21 reveals that it only detects fraud with all the three classifiers with 60% accuracy.

3-Dataset Hybrid Models for Native English Cybercriminal Networks

In this section, I evaluate 3-dataset hybrid models and the predictive accuracy of the generalized models in detecting deception and cybercrime in native English cybercriminal networks. I generated three 3-dataset hybrid models for detecting cybercrime in native English cybercriminal networks each of which has three classifiers. Table 14 summarizes the results on performance of the classifiers for the 3-dataset models. I use precision, recall, F, and Receiver Operating Curve area under curve as the performance measurement metrics for the models. The results reveal that classifiers for 3-dataset training models perform well in all the four metrics on evaluation of classifier performance.

The first 3-dataset hybrid model, which is EN+FB+NR, is trained on triplet datasets comprising Enron, Facebook and favorable online review datasets from the email, Facebook and website web genres to detect favorable fake reviews in the website web genre. This model fails to detect favorable fake reviews as shown in Figure 22.

Table 15 Evaluation for 3-Dataset Models for Native English Cybercriminal Networks

DATASET	CLASSIFIER	P	R	F	ROC
EN+FB+NR	NB	0.664	0.657	0.653	0.68
EN+FB+NR	SVM	0.705	0.703	0.703	0.703
EN+FB+NR	KNN	0.706	0.703	0.702	0.699
EN+FB+PR	NB	0.667	0.657	0.651	0.705
EN+FB+PR	SVM	0.751	0.747	0.746	0.747
EN+FB+PR	KNN	0.733	0.72	0.716	0.807
FB+NR+PR	NB	0.624	0.617	0.611	0.632
FB+NR+PR	SVM	0.64	0.637	0.635	0.637
FB+NR+PR	KNN	0.628	0.623	0.62	0.664

The second 3-dataset hybrid model, which is EN+FB+PR, is trained in Enron, Facebook and favorable reviews from the email, Facebook and website web genres respectively. All the three classifiers detect unfavorable reviews, as shown in Figure 23, with 60% accuracy for NB, 70% accuracy for SVM and 70% accuracy for kNN classifiers.

The third hybrid model, which is FB+PR+NR is trained on Facebook, favorable and unfavorable reviews, is used to detect fraud. All the three classifiers detect fraud with 70% accuracy for NB, 70% accuracy for SVM and 60% accuracy for kNN classifiers as shown in Figure 24.

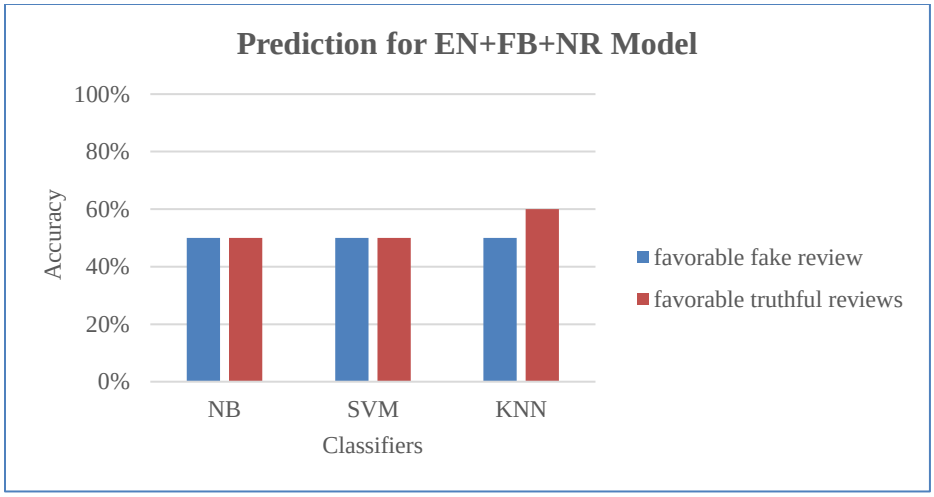

Figure 22 Predictions for 3-Dataset EN+FB+NR Model

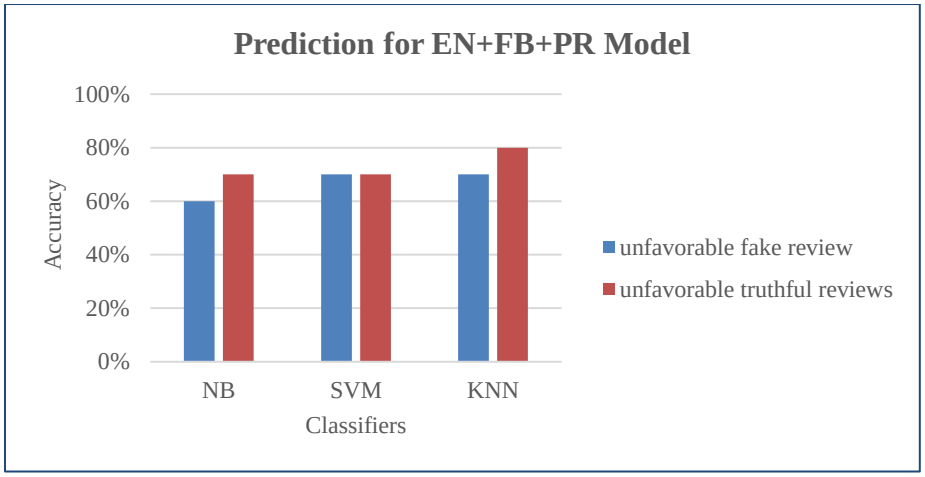

Figure 23 Predictions for 3-Dataset EN+FB+PR Model

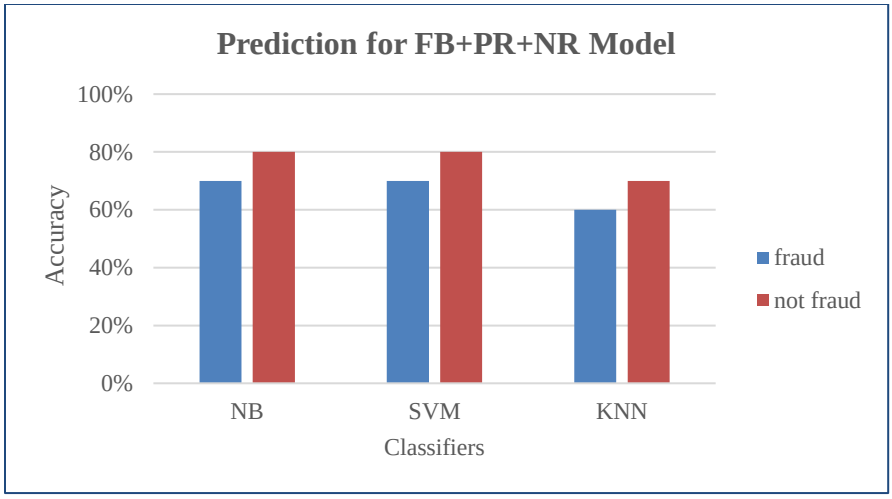

Figure 24 Predictions for 3-Dataset FB+PR+NR Model

Models for Cybercrime in Bilingual Cybercriminal Networks

This work on detecting cybercrime in non-native English speaking cybercriminals networks extends a paper on evaluating classifiers in detecting cybercrime in bilingual cybercriminal networks (Mbaziira et al., 2015). In that paper, we evaluated three classifiers NB, SVM and kNN in detecting scams with native English and Nigerian Pidgin, which is a broken form English, widely spoken in West Africa. I used unigrams and bigrams to build models that discriminate between scams and truthful messages. To evaluate the classifiers, I use four sub-datasets from comprising both deceptive and truthful data as show in Table 16 to train four models from these datasets: sub-dataset A, sub-dataset B, sub-dataset C and sub-dataset D.

Table 16 Classifier Evaluation for Bilingual Cybercrime Models

Sub-Dataset #	Language	N-Gram Words	# Words
A	English	Unigram	2081
B	English	Bigram	12070
C	English & Nigeria Pidgin	Unigram	1875
D	English & Nigeria Pidgin	Bigram	3057

Table 17 Classifier Evaluation for Bilingual Cybercrime Models

Classifier	Sub Dataset #	Precision	Recall	F-Measure	ROC Area
NB	A	0.915	0.911	0.911	0.964
SVM	A	0.886	0.885	0.885	0.947
kNN	A	0.833	0.78	0.771	0.822
NB	B	0.72	0.565	0.473	0.895
SVM	B	0.673	0.656	0.648	0.742
kNN	B	0.695	0.515	0.371	0.644
NB	C	0.964	0.964	0.964	0.994
SVM	C	0.962	0.962	0.962	0.993
kNN	C	0.851	0.79	0.781	0.915
NB	D	0.887	0.861	0.859	0.981
SVM	D	0.898	0.895	0.895	0.94
kNN	D	0.844	0.796	0.789	0.901

Table 17 shows the results for performance of the three classifiers of models for *sub-dataset A* is trained on English unigram words. The results in this table reveal has precision of 0.915, recall of 0.911, f-measure of 0.911, and ROC Area of 0.964. SVM has a precision of 0.866, recall of 0.885, f-measure of 0.885, and ROC Area of 0.947 while kNN has a precision of 0.833, recall of 0.78, f-measure of 0.771, and ROC-curve of 0.822.

The second model, *sub-dataset B* is trained on English bigram words. Results results on performance of the classifiers reveal NB has a precision of 0.72, recall of 0.565, f-measure of 0.473, and ROC Area of 0.895. Comparatively, the SVM classifier

has precision of 0.673, recall of 0.656, f-measure of 0.648, and ROC Area of 0.742. While kNN has precision of 0.695, recall of 0.515, f-measure of 0.371, and ROC of 0.644.

Furthermore, performance of classifiers for model *sub-dataset C* which is trained on unigram words in both English and Nigerian Pidgin reveals that NB has a precision of 0.964, recall of 0.964, f-measure of 0.964 and ROC area of 0.994. SVM classifier has a precision of 0.962, recall of 0.962, f-measure of 0.962, and ROC area of 0.963, while kNN classifier has a precision of 0.851, recall of 0.79, f-measure of 0.781, and ROC area of 0.915.

Lastly, results for performance of the classifiers for the model *sub-dataset D* which is trained on bigrams words in both English and Nigerian Pidgin. The results in this table indicate that SVM has a precision of 0.898, recall of 0.895, f-measure of 0.895, and ROC Area of 0.94. NB has a precision of 0.887, recall of 0.861, f-measure of 0.859, and ROC area of 0.981 while kNN has precision of 0.844, recall of 0.796, f-measure of 0.789, and ROC area of 0.901.

I use hypotheses to evaluate performance of classifiers using ROC Area and F-measure. In Table 18, I evaluate classifier performance using ROC and test the hypotheses as shown below:

- H_0: SVM's ROC area is greater than kNN's ROC Area for English unigrams while for H_1: SVM's ROC area is not greater that kNN's ROC Area for English unigrams. I reject the null hypothesis H_0 because SVM's ROC area is significantly worse at 0.8 with a standard deviation of 0.02.

- H_0: SVM's ROC area is greater than NB's ROC Area for both English and Nigerian Pidgin unigrams while for $H_{1:}$ SVM's ROC area is not greater that NB ROC Area for English and Nigerian Pidgin unigrams. I accept the null hypothesis H_0 because SVM's ROC area is significantly better at 1.00.
- H_0: SVM's ROC area is greater than kNN's ROC Area for English and Nigerian Pidgin unigrams while for $H_{1:}$ SVM's ROC area is not greater that kNN's ROC Area for English and Nigerian Pidgin unigrams. I reject the null hypothesis H_0 because SVM's ROC area for both English and Nigerian unigrams is significantly worse at 0.92 and standard deviation of 0.02.
- H_0: SVM's ROC area is greater than NB's ROC area for English and Nigerian Pidgin bigrams while for H_1 : SVM's ROC area is not greater that NB's ROC area for English and Nigerian Pidgin bigrams. I accept the null hypothesis H_0 because SVM's ROC area for both English and Nigerian bigrams is significantly better at 0.99.

Furthermore, conclude the evaluation of classifier performance and hypothesis testing with f-measure as below:

- H_0: SVM's f-measure is greater than kNN's f-measure for English unigrams while for $H_{1:}$ SVM's f-measure is not greater than kNN's f-measure for English unigrams. I reject the null hypothesis H_0 because SVM's f-measure for English unigrams is significantly worse at 0.76 and standard deviation of 0.01.
- H_0: SVM's f-measure is greater than NB's F-measure for English bigrams while for $H_{1:}$ SVM's f-measure is not greater than NB's f-measure for English

bigrams. I reject the null hypothesis H_0 because SVM's f-measure for English bigrams is significantly worse at 0.48 and standard deviation of 0.04.

- H_0: SVM's f-measure is greater than kNN's f-measure for English bigrams while for H_1: SVM's f-measure is not greater than kNN's f-measure for English bigrams. I reject the null hypothesis H_0 because SVM's f-measure for English bigrams is significantly worse at 0.37 and standard deviation of 0.02.
- H_0: SVM's f-measure is greater than NB's f-measure for English and Nigerian Pidgin unigrams while for $H_{1:}$ SVM's f-measure is not greater than NB's f-measure for English and Nigerian Pidgin unigrams. I accept the null hypothesis H_0 because SVM's f-measure for English and Nigerian Pidgin unigrams is significantly better at 0.97 and standard deviation of 0.01.
- H_0: SVM's f-measure is greater than kNN's f-measure for English and Nigerian Pidgin unigrams while for $H_{1:}$ SVM's f-measure is not greater than kNN's f-measure for English and Nigerian Pidgin unigrams. I reject the null hypothesis H_0 because SVM's f-measure for English and Nigerian Pidgin unigrams is significantly worse at 0.79 and standard deviation of 0.03.
- H_0: SVM's f-measure is greater than NB's f-measure for English and Nigerian Pidgin bigrams while for $H_{1:}$ SVM's f-measure is not greater than NB's f-measure for English and Nigerian Pidgin bigrams. I accept the null hypothesis H_0 because SVM's f-measure for English and Nigerian Pidgin bigrams is significantly better at 0.85 and standard deviation of 0.03.

Table 18 Hypotheses for Testing Bilingual Cybercrime Models at 95% Confidence

Sub-Dataset Model #	Classifier Metric	SVM	SVM vs NB	Hypothesis	SVM vs kNN	Hypothesis (α=0.05)
A	ROC Area	0.93±0.02	0.95 ±0.01	Not Reject	0.80 ±0.02	Reject
B	ROC Area	0.94±0.03	0.88 ±0.03	Not Reject	0.84 ±0.12	Not Reject
C	ROC Area	0.99±0.00	1.00 ±0.00	Accept	0.92 ±0.02	Reject
D	ROC Area	0.89±0.02	0.99 ±0.00	Accept	0.94 ±0.02	Accept
A	F-Measure	0.86±0.03	0.89 ±0.02	Not Reject	0.76 ±0.01	Reject
B	F-Measure	0.80±0.05	0.48 ±0.04	Reject	0.37 ±0.02	Reject
C	F-Measure	0.94±0.01	0.97 ±0.01	Accept	0.79 ±0.03	Reject
D	F-Measure	0.77±0.04	0.85 ±0.03	Accept	0.79 ±0.04	Not Reject

The SVM classifiers for English models perform better than NB and kNN classifiers of Nigerian Pidgin because Nigerian Pidgin has limited English vocabulary of words compared to native English. The models reveal that even with unigram and bigram models, the classifiers can discriminate deception in native English and Nigerian Pidgin text-based communication. This finding motivates us to explore linguistic models that use CL and PL processes to detect deception and cybercrime in non-native English speaking cybercriminal networks.

Hybrid Models for Non-Native English Cybercriminal Networks

I evaluate performance and predictive accuracy for 2-dataset and 3-dataset hybrid models for non-native English speaking cybercriminals. All the hybrid models for detecting cybercrime for native English speaking cybercriminal networks fail to detect cybercrime in non-native English cybercriminal networks. I have three 2-dataset hybrid models and one 3-dataset hybrid models each of which has three classifiers for detecting

cybercrime in non-native English cybercriminal networks. I use dataset-specific features for these hybrid models using techniques explained earlier in feature selection and engineering.

I build three 2-dataset hybrid models: EN+NR, EN+PR and NR+PR. Table 19 summarizes the results on performance of the classifiers for the 2-dataset models. I use precision, recall, F, and Receiver Operator Curve area as the performance measurement metrics for the models. The results indicate that classifiers for 2-dataset training models perform well in all the four metrics.

Table 19 Evaluation for 2-Dataset Models for Non-Native English Cybercriminals

MODEL	CLASSIFIER	P	R	F	ROC AREA
EN + NR	NB	0.767	0.74	0.733	0.766
EN + NR	SVM	0.736	0.73	0.728	0.73
EN + NR	KNN	0.721	0.1	0.709	0.71
EN + PR	NB	0.755	0.71	0.697	0.78
EN + PR	SVM	0.851	0.85	0.85	0.85
EN + PR	KNN	0.75	0.74	0.737	0.8
NR + PR	NB	0.569	0.545	0.502	0.612
NR + PR	SVM	0.637	0.635	0.634	0.635
NR + PR	KNN	0.619	0.6	0.583	0.62

The first hybrid 2-dataset model for detecting scams in non-native English cybercriminal networks is EN+NR, which is trained on Enron and unfavorable online review datasets from the email and website web genres respectively. The model detects scams with 60 % accuracy with the NB classifiers.

The second 2-dataset model, which is EN+PR, is trained on Enron and favorable reviews from email and website web genres. This model detects scams with 60% accuracy with the NB classifier.

The third model, which is NR+PR, is trained on both favorable and unfavorable reviews from the website web genre. This model detects scams with 60% accuracy with both the NB and kNN classifiers. Results of predictive accuracy of 2-dataset hybrid models for non-native English speaking cybercriminals are in Figure 25.

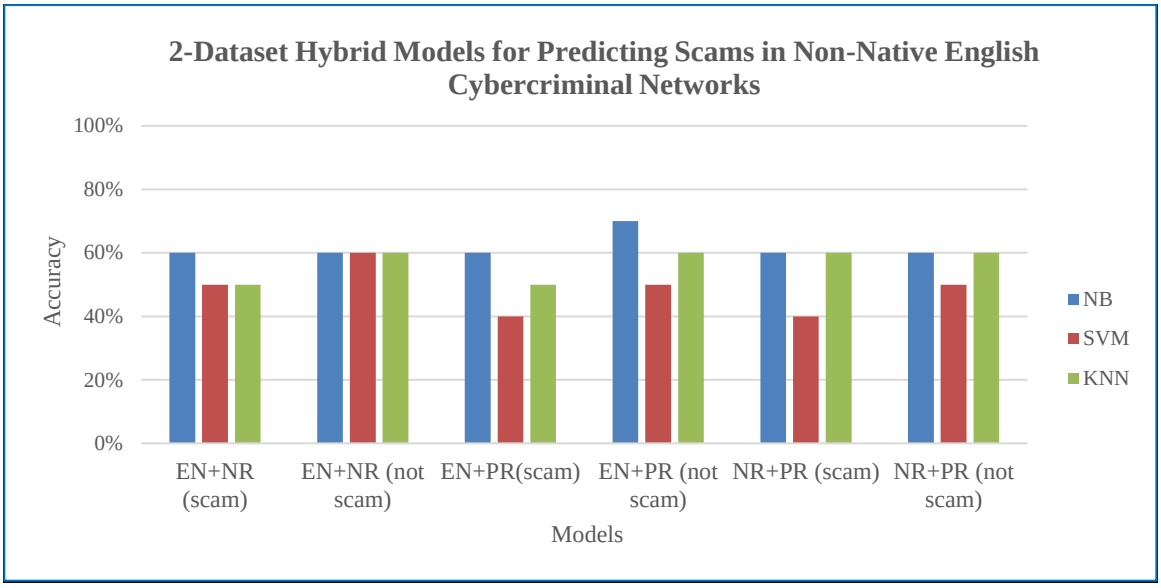

Figure 25 Predictions for 2-Dataset Models for Non-Native English Cybercrime

Table 20 Evaluation for 3-Dataset Models for Non-Native English Cybercriminals

MODEL	CLASSIFIER	P	R	F	ROC AREA
EN+PR+NR	NB	0.656	0.65	0.647	0.723
EN+PR+NR	SVM	0.997	0.997	0.997	0.997
EN+PR+NR	KNN	0.732	0.727	0.725	0.833

I also build one 3-dataset hybrid model for detecting scams in non-native English cybercriminal network. Table 20 summarizes the results on performance of the classifiers for the 3-dataset models for non-native English speaking cybercriminal networks. I use precision, recall, F, and Receiver Operator Curve area as the performance measurement metrics for the models. The results indicate that classifiers for 3-dataset training models perform well in all the four metrics.

The 3-dataset hybrid model, EN+PR+NR, is trained on Enron, favorable and unfavorable reviews from the email and website web genres. The model detects scams in the Facebook web genre with 60% accuracy with NB classifier and 70% accuracy with while the SVM classifier as shown in Figure 26.

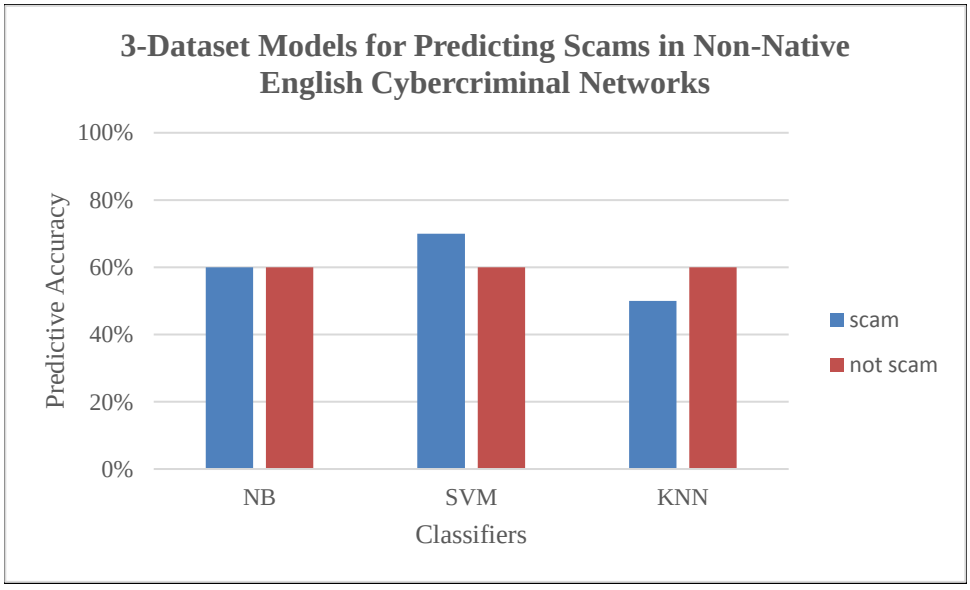

Figure 26 Predictions for 3-Dataset Models for Non-Native English Cybercrime

The results for the 2-dataaset and 3-dataset hybrid models for detecting detection and cybercrime are promising. These results reveal that it is possible to build models for that discriminate between deceptive cybercrime and truthful messages in native and non-native English speaking cybercriminal networks. The results also reveal that the models generalize well in detecting new forms deceptive cybercrime in non-native English speaking cybercriminal networks.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

Incidents of cybercrime where cybercriminals use deception to exploit their victims through text-based communication continue to surge. In this dissertation, fraud, scams and online fake reviews as types of cybercrime where cybercriminals use deception to trick and exploit their victims. Fraud is a very serious criminal offense because criminals deliberately alter facts and information about financial health of their companies to deceive by investors, shareholders, and stockholders. Similarly, cybercriminals use deceptive language in scams and fake reviews to either trick and exploit their victims. The consequences of deceptive language in scams, fraud and fake reviews are serious. These include: lost livelihoods, destroyed brands, and vast financial losses among others. In this section, we evaluate the models linguistic model in detecting cybercrime, highlight contributions of this work as well as limitations and future work.

Discussion

This research mainly focus on building models that generalize well in detecting deception and cybercrime in cybercriminal networks. The sizes of the samples used in training and testing all the models in our experiments are small but this does not affect the conclusions drawn from generalizability of the models (Domingos, 2012).

Accuracy of automated or text classification models in detecting deception in text measured against human ability to determine whether information in a message is truthful or deceptive. According to early research, human beings detect deception with an accuracy rate of 50%, which means a person can only detect whether the information in a communication session is deceptive or truthful (Bond & DePaulo, 2006; Newman et al., 2003). I disregard all results for classifiers of all 1-dataset, 2-dataset and 3-dataset models that failed to generalize detection of deception and cybercrime above the 50% accuracy rate. This is because such models either performed at human deception detection rate or worse.

Figures 5-8 reveal that deception can be discriminated within CL and PL features for all the four training sets we build models for detecting deception and cybercrime. In the first CL process, which is quantity of words, reveals that FB and EN training sets compared to PR and NR training sets. I observe a similar consistent pattern of more verbs, nouns, characters, punctuation marks, and sentences. The second CL process is lexical diversity, where deceptive messages have lower lexical diversity compared to truthful messages, which is consistent in FB, EN and NR training sets. The third CL process is expressivity, where by deceptive messages have more adjectives and adverbs than truthful messages, which is consistent in EN and FB training sets. The fourth CL process is non-immediacy, where deceptive messages use less self-references to be avoid being held accountable of their communication. I observe this in FB and EN training sets. Lastly, the fifth CL process is sentence complexity, where deceptive messages have less

sentence structure that is average sentence length, average word length and punctuation marks. We observe this in EN, FB, PR and NR training sets.

Similarly, for PL processes linked to deception and cybercrime, I first consider quantity of words or verbosity, a technique that cybercriminal use to make text-based messages forgettable to commit cybercrime using the same content. Figures 5-8, reveal that FB and EN training sets have more quantity of words which is consistent with deception. The second PL process is average sentence length because cybercriminals make their fake reviews, fraud and scams fluent to make them forgettable and this is consistent with EN, FB, PR and NR training sets. Thirdly, cybercriminals use less first-person pronouns to avoid being held accountable for their messages. I observe this discrimination in EN and FB training sets, however, cybercriminals tend to use more other pronouns like he, she etc., in deceptive messages, which we observe in PR and NR training sets. Another PL process linked to cybercrime and deception is emotion. Deceptive messages have more negative emotion words and we observe this in FB, NR and EN training sets, however, deceptive messages have less positive emotion and words of affect. Similarly, FB, NR and EN has less positive emotion and words of affect.

In the first research question, I inquired whether it is possible to detect deception and cybercrime by web genre and train models can that detect cybercrime in messages of web genres that were not part of the training model. I built four models each representing a web genre that is: website, email and Facebook. All the three classifiers for the 1-dataset models detected and analyzed cybercrime in text messages that where part of the training with over 50% predictive accuracy. The results revealed that it is possible to

build cybercrime detection models that generalize well in detecting cybercrime in text messages from web genres that were not part of training model hence addressing that research question.

In the second research question, I investigated whether it is possible to generalize deception and cybercrime in text-based communication. All the 1-dataset models generalize well in detecting and analyzing cybercrime. This was achieved first using 10-fold cross validation and then a held-out method detect cybercrime from test-sets which were not used in training the models. All the 1-dataset models as well as hybrid models generalize well in detecting cybercrime hence addressing that research question.

Since, this research mainly focus on building models that generalize well in detecting deception and cybercrime in cybercriminal networks. The sizes of the samples used in training and testing all the models in our experiments are small but it does not affect the conclusions drawn from generalizability of the models (Domingos, 2012).

A total of twelve classifiers were trained for all the four 1-dataset models of which eight classifiers for all the four models generalized well in detecting cybercrime in other web genres which not part of the training model with accuracies ranging from 60% to 80%. There were three NB and kNN classifiers respectively and two SVM classifiers for the 1-dataset models that generalized well in detecting cybercrime in native English speaking cybercriminal networks. These classifiers generalized well in detecting favorable fake reviews and fraud as shown in Table 21. I also observed the NR 1-dataset model generalizes better than the PR, FB and EN models because it has better deception

patterns for detecting cybercrime. On the other hand, the FB model has the fewest classifiers which only generalized well in detecting in detecting fraud.

Model	fraud	scam	favorable fake reviews	unfavorable fake reviews
EN		0	2	0
FB	1		0	0
PR	0	0		2
NR	2	2	2	

Table 21 Number of classifiers in 1-dataset models for native English that generalize well

Generally, I observed that 1-dataset models did not generalize well in detecting scams which were obtained from non-native English speaking cybercriminal networks. I conducted further experiments using unigram and bigram analyses to analyze this dataset from non-native English speaking cybercriminals. Since this dataset was obtained from bilingual non-native English speakers we observed that the classifiers for detecting scams in pidgin English performed significantly better compared to those trained on non-native English scams. This explains why 1-dataset models trained on fraud, favorable and unfavorable fake reviews from native English speaking cybercriminal networks did not generalize well in detecting cybercrime from non-native English speaking cybercriminal networks.

I also investigated whether it is possible to detect deception and cybercrime using hybrid models in native English and non-native English speaking cybercriminal networks. I built 2-dataset and 3-dataset models for detecting cybercrime in native and non-native English speaking cybercriminal networks. All the three classifiers that is NB, SVM and kNN for the 2-dataset models detect and analyze cybercrime in native English

speaking cybercriminal networks. I generated dataset-specific models for non-native English speaking cybercriminal networks and generally, all the three classifiers detect and analyzing cybercrime. I also generated 3-dataset models for detecting and analyzing deception and cybercrime in native English and non-native English speaking cybercriminal networks. In native English speaking cybercriminal networks, NB and kNN classifiers detect and analyze deception and cybercrime. However, in non-native English speaking cybercriminal networks, the three classifiers of the 3-dataset models detect and analyze cybercrime.

Eighteen 2-dataset models were trained to detect deception and cybercrime in native English cybercriminal networks. Twelve of these models generalized well in detecting cybercrime with predictive accuracies ranging from 60% to 80%. The twelve models comprised four NB classifiers, five SVM classifiers and three kNN classifiers as shown in Table 22. The 2-dataset models for native English cybercriminal networks detect fraud better than scams, favorable and unfavorable fake reviews. I observe that messages have more patterns of deception compared to other types of cybercrime since this dataset was obtained from skilled, eloquent and elite native English speakers. All the classifiers in the FB+PR model generalize well in detecting cybercrime compared to all other 2-dataset hybrid models for native English speaking cybercriminal networks because they contain the most patterns of deception. All the other models except the EN+FB model have relatively a good number of patterns with deception features and generalize well with specific one of type of cybercrime unlike FB+PR model which

generalizes well with two types of cybercrime. I also observed these models fail in detecting scams from non-native English cybercriminal networks.

Models	fraud	scam	favorable fake reviews	unfavorable fake reviews
FB+NR	2		0	
FB+PR	3			3
NR+PR	3	0		
EN+NR		0	1	
EN+FB			1	1
EN+PR		0		2

Table 22 Number of classifiers in 2-dataset hybrid models for native English that generalize well

Nine classifiers for 3-dataset hybrid models were trained to detect deception and hybrid in native English cybercriminal networks. A total of six out of nine classifiers detect cybercrime in the 3-dataset models as shown in Table 23. I also observe that all the classifiers for the FB+NR+PR and EN+FB+PR models generalize well in detecting cybercrime with predictive accuracies ranging from 60% to 70%. These models detect fraud and unfavorable fake reviews better than other types of cybercrime. However, for EN+FB+NR model only kNN classifier generalizes well in detecting cybercrime. There are more patterns of deception in FB+NR+PR and EN+FB+PR models that enable these models to generalize well in detecting cybercrime.

Models	fraud	scam	favorable fake reviews	unfavorable fake reviews
EN+FB+NR			0	
EN+FB+PR				3
FB+NR+PR		3		

Table 23 Number of classifiers in 2-dataset hybrid models for native English that generalize well

As more datasets with different types of cybercrime are added to the learning model, the classifiers for the hybrid learning models continue to generalize well in detecting new types of cybercrime. For the 2-dataset models for native English speaking cybercriminal networks, all the classifier in the FB+PR model generalized well in detecting cybercrime. In the 3-dataset models we observe that there are two models (i.e. EN+FB+PR and EN+NR+PR models) where all the classifiers generalize well in detecting new cybercrime.

All hybrid models for native English speaking cybercriminals failed in detecting scams from non-native English speaking cybercriminal networks. Compared to fraud, favorable and unfavorable fake reviews, deception patterns for scams were characterized by short sentences, low pausality and lexical diversity as well as limited use of punctuation marks. The messages also had low positive emotion as limited use of causality and certainty words. After re-engineering the feature-set for both 2-dataset and 3-dataset hybrid models, scam detection tremendously improves with predictive accuracies to 60% for 2-dataset hybrid models and while for the 3-dataset hybrid model the performance ranges from 60% to 70% accuracy.

Generally, the models generalize well in detecting cybercrime within datasets that are not part of the training model. In the hybrid models, all the three classifiers detect

fraud and unfavorable reviews because these types of cybercrime had more patterns of deception and cybercrime. The patterns summarized in Table 24 are:

- Cybercriminals are less committal in their text-based communication such that they use less verbs and modal verbs. This pattern was consistent in fraud, scams and unfavorable fake reviews.
- Cybercriminals are verbose in their text-based communication whereby they use more punctuation marks and function words. This pattern was consistent in scams, favorable and unfavorable fake reviews.
- Cybercriminals are vague and ambiguous in their text-based communication whereby they use more function words, adverbs and adjectives. I observed this pattern in scams, fraud and unfavorable reviews.
- Cybercriminals avoid being held accountable for their text-based communication whereby they use less self-pronouns or self-references. I observed this in scam and fraud.
- Cybercriminals use messages with low cognitive complexity. This text-based communication is characterized by less analytical, shorter, insight and causation words. I observed this pattern in scams, favorable and unfavorable fake reviews.
- Cybercriminals are emotional such that they use more negative and positive emotion words. I observed this pattern in fraud, favorable and unfavorable fake reviews.

Feature Description	Feature Type	fraud	scam	favorable fake reviews	unfavorable fake reviews
less committal	CL	x	x		x
verbose	CL		x	x	x
avoid accountability	CL	x	x		
vague & ambiguous	PL	x	x		x
low cognitive complexity	PL	x		x	x
emotion	PL	x		x	x

Table 24 Patterns on deception and cybercrime from the hybrid models

Contributions

This dissertation makes three main contributions to scientific and cyber-forensic research. Firstly, I demonstrate that it is possible to detect and analyze deception and cybercrime as well as generalize the models using web genres namely like: email, social media and websites. The features of the training models are derived from computational linguistic and psycholinguistic processes linked to deception and cybercrime in text-based communication. With increasing incidents of cybercrime, the generalizable models demonstrate that deception and cybercrime models trained in specific types of cybercrime can use be used to detect other types of cybercrime not used in training the models. This is particularly useful when detecting and analyzing certain types of cybercrime where data for training models may be limited. This contribution also demonstrates that there are similar patterns of deception and cybercrime in web genres of text-based communication which can be linked to CL and PL processes.

Secondly, I develop hybrid models where I combined two and three datasets from different web genres to generate 2-dataset and 3-dataset hybrid models respectively to

detect and analyze deception and cybercrime. I use linguistic features from CL and PL processes linked to deception and cybercrime to generate these hybrid 2-dataset and 3-dataset models.

Lastly, I also develop deception and cybercrime detection models using dataset-specific features to detect cybercrime in non-native English-speaking cybercriminal networks. I accomplish this by using dataset-specific features from a dataset obtained from non-native English speaking cybercriminal network. This contribution reveals that in non-native English speaking cybercriminal networks, there are linguistic variations in CL and PL processes linked to specific types of cybercrime. This work reveals that is it possible to use dataset-specific features to build linguistic models that detect and analyze deception and cybercrime.

Implications of the Research

This research can be applied to systems or applications that use content filters in discriminate between deceptive and truthful messages. In such applications, it is crucial to have systems that can block and label messages from cybercriminals which contain deceptive messages for exploiting targeted victims. Some of such applications include but are not limited to: online dating websites, customs and immigration interview processes, e-commerce, social media etc., (Papenfuss, 2016).

Since this research mainly focused on building models that generalize well in detecting deception and cybercrime in text messages. The results on performance of the models reveal that it is possible to detect new forms of cybercrime in text-based communication of web genres by training models on existing messages with both

deceptive and truthful messages. I achieved this objective by training hybrid models on that detect text-based communication of native and non-native English speaking cybercriminal networks whereby the models detected cybercrime with accuracy rates ranging from 60% to 80% accuracy. This work demonstrates that it is possible to predict cybercrime using models with PL and CL features and can be implemented in user- and provider-based content filtering systems.

Limitations and Future Directions

There is still very limited work in detecting cybercrime in text-based communication using machine learning and Natural Language Processing (NLP). However, there is increasing interest in using NLP and machine learning to solve real world problems like cybercrime. Limitations of this research are inherent on existing challenges in NLP application to addressing real world problems. Some of the limitations of this work in respect to detecting and analyzing deception and cybercrime are: ambiguity of messages expressed in English language; use of phrases in English that have more than one meaning, etc.

Conclusion

With increasing incidents of cybercrime in Internet-based text communication, I demonstrate that it is possible to build cybercrime detection models with features from both CL and PL processes linked to deception and cybercrime. The models perform well in detecting unseen types of cybercrime, which means that it is possible to detect new forms of deceptive cybercrime.

This work also demonstrates that hybrid models can generated by combining web genres to discriminate cybercrime from truthful text-based communication in both native and non-native English speaking cybercriminal networks. My approach for generalizing the hybrid cybercrime detection models improved performance of the training models. This implies that it is possible to detect new forms of deception and cybercrime in text-based communication.

REFERENCES

Abozinadah, E., Mbaziira, A., & Jones, J. (2015). Detection of Abusive Accounts with Arabic Tweets. *International Journal of Knowledge Engineering*, *1*(2), 6. https://doi.org/10.7763/IJKE.2015.V1.19

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A Comparison of Machine Learning Techniques for Phishing Detection. In *Proceedings of the Anti-phishing Working Groups 2Nd Annual eCrime Researchers Summit* (pp. 60–69). New York, NY, USA: ACM. https://doi.org/10.1145/1299015.1299021

Afroz, S., Brennan, M., & Greenstadt, R. (2012). Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *2012 IEEE Symposium on Security and Privacy (SP)* (pp. 461–475). https://doi.org/10.1109/SP.2012.34

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66. https://doi.org/10.1007/BF00153759

Aransiola, J. O., & Asindemade, S. O. (2011). Understanding cybercrime perpetrators and the strategies they employ in Nigeria. *Cyberpsychology, Behavior and Social Networking*, *14*(12), 759–763. https://doi.org/10.1089/cyber.2010.0307

Bird, S., Klein, E., & Loper, E. (2009). Categorizing and Tagging Words. In *Natural Language Processing with Python*. O'Reilly Media, Inc. Retrieved from http://proquest.safaribooksonline.com/book/programming/python/9780596803346

/1dot-language-processing-and-python/sec-computing-with-language-texts-and-words

Bohme, R., & Moore, T. (2012). How do consumers react to cybercrime? In *eCrime Researchers Summit (eCrime), 2012* (pp. 1–12). https://doi.org/10.1109/eCrime.2012.6489519

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234.

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Trans. Inf. Syst. Secur.*, *15*(3), 12:1–12:22. https://doi.org/10.1145/2382448.2382450

Brownlee, J. (2014, March 12). Feature Selection to Improve Accuracy and Decrease Training Time. Retrieved from http://machinelearningmastery.com/feature-selection-to-improve-accuracy-and-decrease-training-time/

Chang, C., & Lin, C.-J. (2001). *LIBSVM: a Library for Support Vector Machines*.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: a general framework and some examples. *Computer*, *37*(4), 50–56.

Chen, X., Chandramouli, R., & Subbalakshmi, K. P. (2014). Scam detection in Twitter. In *Data Mining for Service* (pp. 133–150). Springer.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece*

on Social Computing (SocialCom) (pp. 71–80). https://doi.org/10.1109/SocialCom-PASSAT.2012.55

Clikeman, P. (2012, February). The 10 Tell-Tale Signs of Deception - The Words Reveal. Retrieved April 12, 2016, from http://www.fraud-magazine.com/article.aspx?id=4294971184

Cohen, W. (2015, May 8). Enron Email Dataset. Retrieved March 29, 2016, from http://www.cs.cmu.edu/~enron/

Conway, D., & White, J. M. (2012). This or That: Binary Classification. In *Machine Learning for Hackers*. O'Reilly Media, Inc. Retrieved from http://proquest.safaribooksonline.com.mutex.gmu.edu/book/programming/machine-learning/9781449330514/machine-learning-for-hackers/id3077792

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Najada, H. A. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, *2*(1), 1–24. https://doi.org/10.1186/s40537-015-0029-9

deception. (2006). *Collins Dictionary of Law*. Retrieved from http://legal-dictionary.thefreedictionary.com/deception

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118.

DoJ. (2013, June 21). Former Enron CEO Jeffrey Skilling Resentenced to 168 Months for Fraud, Conspiracy Charges. Retrieved April 2, 2017, from https://www.justice.gov/opa/pr/former-enron-ceo-jeffrey-skilling-resentenced-168-months-fraud-conspiracy-charges

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Commun. ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Engel, P. (2015, May 9). ISIS has mastered a crucial recruiting tactic no terrorist group has ever conquered. Retrieved March 16, 2016, from http://www.businessinsider.com/isis-is-revolutionizing-international-terrorism-2015-5

Feng, V. W., & Hirst, G. (2013). Detecting Deceptive Opinions with Profile Compatibility. In *International Joint Conference on Natural Language Processing* (pp. 338–346). Nagoya, Japan.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to Detect Phishing Emails. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 649–656). New York, NY, USA: ACM. https://doi.org/10.1145/1242572.1242660

Firte, L., Lemnaru, C., & Potolea, R. (2010). Spam detection filter using KNN algorithm and resampling. In *2010 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 27–33). https://doi.org/10.1109/ICCP.2010.5606466

Fitzpatrick, E., Bachenko, J., & Fornaciari, T. (2015). Automatic Detection of Verbal Deception. *Synthesis Lectures on Human Language Technologies*, *8*(3), 1–119. https://doi.org/10.2200/S00656ED1V01Y201507HLT029

Fornaciari, T., & Poesio, M. (2014). Identifying fake Amazon reviews as learning from crowds. In *Proceedings of 14th conference of the European Chapter of the Association for Computational Linguistics* (pp. 279–287). Gothenburg, Sweden.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, *3*, 1157–1182.

Hall, E. (2015, March 11). How ISIS Uses Twitter To Recruit Women [BuzzFeed News]. Retrieved March 16, 2016, from http://www.buzzfeed.com/ellievhall/how-isis-uses-twitter-to-recruit-women

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, *45*(1), 1–23. https://doi.org/10.1080/01638530701739181

Hao, P., Chen, X., Cheng, N., Chandramouli, R., & Subbalakshmi, K. P. (2011). Adaptive Context Modeling for Deception Detection in Emails. In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition* (pp. 458–468). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23199-5_34

He, H., & Garcia, E. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

ITU. (2009). *Understanding Cybercrime:A Guide For Developing Countries* (p. 225). Switzerland: International Telecommunications Union. Retrieved from http://www.itu.int/ITU-D/cyb/cybersecurity/docs/itu-understanding-cybercrime-guide.pdf

Jaishankar, K. (2011). *Cyber Criminology: Exploring Internet Crimes and Criminal behavior*. Boca Raton, FL, USA: CRC Press, Taylor and Francis Group.

Keila, P. S., & Skillicorn, D. B. (2005). Detecting Unusual Email Communication. In *Proceedings of the 2005 Conference of the Centre for Advanced Studies on Collaborative Research* (pp. 117–125). Toranto, Ontario, Canada: IBM Press. Retrieved from http://dl.acm.org/citation.cfm?id=1105634.1105643

Larose, D. T. (2014). Why do We Need to Preprocess the Data? In *Discovering Knowledge in Data: An Introduction to Data Mining, 2nd Edition* (2nd ed.). John Wiley & Sons. Retrieved from http://proquest.safaribooksonline.com/book/databases/business-intelligence/9781118873571/chapter-2-data-preprocessing/c02_xhtml

Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). Learning to identify review spam. *Proceedings of International Joint Conference on Artificial Intelligence*, *22*(3).

Mbaziira, A., Abozinadah, E., & Jones, J. (2015). Evaluating Classifiers in Detecting 419 Scams in Bilingual Cybercriminal Communities. *International Journal of Computer Science and Information Security*, *13*(7), 7.

McGrath, S. (2015, June 11). Winning The War: The Evolution Of Cybercrime [Technology]. Retrieved from http://www.informationweek.com/interop/winning-the-war-the-evolution-of-cybercrime/a/d-id/1320817

Mehler, A., Sharoff, S., & Santini, M. (Eds.). (2011). *Genres on the Web* (Vol. 42). Dordrecht: Springer Netherlands. Retrieved from http://link.springer.com/10.1007/978-90-481-9178-9

Morgan, S. (2016, January 17). Cyber Crime Costs Projected To Reach $2 Trillion by 2019. Retrieved February 29, 2016, from

http://www.forbes.com/sites/stevemorgan/2016/01/17/cyber-crime-costs-projected-to-reach-2-trillion-by-2019/

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting Fake Reviewer Groups in Consumer Reviews. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 191–200). New York, NY, USA: ACM. https://doi.org/10.1145/2187836.2187863

Nakashima, E., & Peterson, A. (2014, June 8). Report: Cybercrime and espionage costs $445 billion annually. *The Washington Post*. Retrieved from https://www.washingtonpost.com/world/national-security/report-cybercrime-and-espionage-costs-445-billion-annually/2014/06/08/8995291c-ecce-11e3-9f5c-9075d5508f0a_story.html

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: predicting deception from linguistic styles. *Personality & Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

Nirkhi, S. M., Dharaskar, R. V., & Thakre, V. M. (2012). Analysis of online messages for identity tracing in cybercrime investigation. In *2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec)* (pp. 300–305). https://doi.org/10.1109/CyberSec.2012.6246131

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies -*

Volume 1 (pp. 309–319). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2002472.2002512

Papenfuss, M. (2016, December 19). Germany Weighs Stiff Fines for Social Media Sites That Carry Fake News. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/entry/german-fake-news-fines_us_585843d5e4b03904470a1dfb

Pearl, L., & Steyvers, M. (2012). Detecting authorship deception: a supervised machine learning approach using author writeprints. *LLC*, *27*, 183–196.

Ponemon Institute. (2015). *2015 Cost of Cyber Crime Study: Global* (p. 29). Retrieved from https://ssl.www8.hp.com/ww/en/secure/pdf/4aa5-5207enw.pdf

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using Machine Learning to Detect Cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops (ICMLA)* (Vol. 2, pp. 241–244). https://doi.org/10.1109/ICMLA.2011.152

Rowe, N. (n.d.). Detecting Online Deception and Responding to It. Retrieved March 16, 2016, from http://www.au.af.mil/au/awc/awcgate/nps/decep_detec.htm

Sandoval, T., Matsumoto, D., Hwang, H., & Skinner, L. (2015). Exploiting Verbal Markers of Deception Across Ethnic Lines: An Investigative Tool for Cross-Cultural Interviewing. Retrieved November 27, 2016, from https://leb.fbi.gov/2015/july/exploiting-verbal-markers-of-deception-across-ethnic-lines-an-investigative-tool-for-cross-cultural-interviewing

Sarvari, H., Abozinadah, E., Mbaziira, A., & McCoy, D. (2014). Constructing and Analyzing Criminal Networks. *IEEE Security and Privacy Workshops*, 8. https://doi.org/DOI 10.1109/SPW.2014.22

Shojaee, S., Murad, M. A. A., Azman, A. B., Sharef, N. M., & Nadali, S. (2013). Detecting deceptive reviews using lexical and syntactic features. In *2013 13th International Conference on Intelligent Systems Design and Applications (ISDA)* (pp. 53–58). https://doi.org/10.1109/ISDA.2013.6920707

Shropshire, C. (2016, March 26). Americans prefer texting to talking, report says [News]. Retrieved April 1, 2016, from http://www.chicagotribune.com/business/ct-americans-texting-00327-biz-20150326-story.html

Sommer, R., & Paxson, V. (2010). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *2010 IEEE Symposium on Security and Privacy (SP)* (pp. 305–316). https://doi.org/10.1109/SP.2010.25

Streitfeld, D. (2011, August 19). Ferreting Out Fake Reviews Online. *The New York Times*. Retrieved from http://www.nytimes.com/2011/08/20/technology/finding-fake-reviews-online.html

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining* (Vol. 1). Pearson Addison Wesley Boston.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Torney, R., Vamplew, P., & Yearwood, J. (2012). Using Psycholinguistic Features for Profiling First Language of Authors. *J. Am. Soc. Inf. Sci. Technol.*, *63*(6), 1256–1269. https://doi.org/10.1002/asi.22627

Vergelis, M., Shcherbakova, T., Demidova, N., & Gudkova, D. (2016, February 5). Kaspersky Security Bulletin. Spam And Phishing In 2015 - Securelist. Retrieved March 31, 2016, from https://securelist.com/analysis/kaspersky-security-bulletin/73591/kaspersky-security-bulletin-spam-and-phishing-in-2015/

Weise, E. (2015, October 19). Amazon cracks down on fake reviews. Retrieved March 15, 2016, from http://www.usatoday.com/story/tech/2015/10/19/amazon-cracks-down-fake-reviews/74213892/

Westman, S., & Freund, L. (2010). Information Interaction in 140 Characters or Less: Genres on Twitter. In *Proceedings of the Third Symposium on Information Interaction in Context* (pp. 323–328). New York, NY, USA: ACM. https://doi.org/10.1145/1840784.1840833

Wollman-Bonilla, J. E. (2003). E-mail as genre: A beginning writer learns the conventions. *Language Arts*, *81*(2), 126–134.

Yang, C., Harkreader, R., Zhang, J., Shin, S., & Gu, G. (2012). Analyzing Spammers' Social Networks for Fun and Profit: A Case Study of Cyber Criminal Ecosystem on Twitter. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 71–80). New York, NY, USA: ACM. https://doi.org/10.1145/2187836.2187847

Zheng, R., Qin, Y., Huang, Z., & Chen, H. (2003). Authorship Analysis in Cybercrime Investigation. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics* (pp. 59–73). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=1792094.1792100

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker, J. F. (2004). A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication. *Journal of Management Information Systems*, *20*(4), 139–165.

Zhou, L., Twitchell, D. P., Qin, T., Burgoon, J. K., & Nunamaker, J., J. F. (2003). An exploratory study into deception detection in text-based computer-mediated communication. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003* (p. 10 pp.-). https://doi.org/10.1109/HICSS.2003.1173793

Zhou, L., & Zhang, D. (2008). Following Linguistic Footprints: Automatic Deception Detection in Online Communication. *Commun. ACM*, *51*(9), 119–122. https://doi.org/10.1145/1378727.1389972

BIOGRAPHY

Alex Vincent Mbaziira graduated from Uganda Martyrs University, with a Bachelor of Science in Economics and Computer Science. He was first employed as Business Analyst and Executive Assistant to the Chief Executive Officer in MOGAS/Castrol Oil Head Office for two years. He received his Master of Science in Information Systems from Uganda Martyrs University in 2004. He worked for Kampala International University for three years as Director for Information Technology and then moved to St Lawrence University, where he was Dean for Information Technology for another three years. He received another Master of Science in Information Security and Assurance from George Mason University in Fairfax, VA in 2014. He is a Fulbright Fellow and a recipient of several service, academic and excellence awards.