CLASSIFICATION OF REAL AND FALSIFIED NARRATIVES IN A REPEATED-MEASURES TEXT CORPUS

by

Ran Wei

M.A., The University of New Brunswick, 2013

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE COLLEGE OF GRADUATE STUDIES

(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

January 2020

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, the dissertation entitled:

CLASSIFICATION OF REAL AND FALSIFIED NARRATIVES IN A REPEATED-MEASURES TEXT CORPUS

submitted by Ran Wei in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

Examining Committee:

Brian O'Connor, Irving K. Barber School of Arts and Sciences, University of British Columbia – Okanagan
Supervisor

Mike Woodworth, Irving K. Barber School of Arts and Sciences, University of British Columbia – Okanagan
Supervisory Committee Member

Jan Cioe, Irving K. Barber School of Arts and Sciences, University of British Columbia – Okanagan
Supervisory Committee Member

Fatemeh Fard, Irving K. Barber School of Arts and Sciences, University of British Columbia – Okanagan
University Examiner

Doug Twitchell, College of Business and Economics, Boise State University
External Examiner

Abstract

A great deal of research has been devoted to finding reliable ways of detecting deception. The current deception literature has recognized that human deceptive behavior is highly complex; behavioral differences due to deception (deception cues) are small and probably context dependent. In general, using a repeated-measure design to control variabilities at the individual level is an effective way of amplifying small behavioral effects; however, no study so far has explored the benefit of adding individual-level random-effects in deception detection models. This dissertation focused on a developing field of language-based deception detection research that utilizes natural language processing (NLP; e.g., Fitzpatrick, Bachenko, & Fornaciari, 2015; Heydari, Tavakolia, Salima, & Heydari, 2015). We tested a novel NLP-based deception detection scheme that utilizes multiple language samples from the same individual. A repeated-measures truthful-and-fabricated text corpus (4 truthful and 4 fabricated statements per individual) from 152 individuals was collected. Truth-telling and fabrication scenarios were created using video recordings of real-life negative events as stimuli. Various sets of cues including n-grams, POS tags, and psycholinguistic cues were extracted from the text sample using NLP techniques. Our results showed that mixed-effects variations of popular classification models including logistic regression, decision tree/random forest, and artificial neural network have better cross-context generalizability than their regular fixed-effects counterparts. This research should encourage further development of repeated-measure deception detection schemes and classification models that can fully utilize such a data structure.

Lay Summary

The current research studies deceptive language, focusing on changes in a person's language pattern during deception. We asked research participants to watch videos of four different real life events, and write one truthful and one deceptive statement about each event. We used computer software to extract a large number of linguistic features (e.g., word usage) from these writings, and explored a number of advanced modeling techniques so as to find an effective way of utilizing the uniqueness of individual language style in deception detection tasks. When we applied deception detection models learned from one set of language data to a new set of data, the models that factored in individuals' unique language patterns performed better than those that did not.

Preface

The experiment that generated data for the current research was designed and executed by Ran Wei. A student volunteer research assistant, Emma Smith, was involved in data screening and preparation. Data analysis was performed by Ran Wei. Ethics approval for this research was granted by the University of British Columbia's Behavioral Research Ethics Board on June 3rd, 2018. The ethics approval certificate number for the experiment is H18-00087. To date, the research included in this dissertation has not been published.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC: Akaike Information Criterion

ANN: Artificial Neural Network

CBCA: Criteria-Based Content Analysis

CMC: Computer-mediated communication

CQT: Control Question Test

DLT: Directed Lie Test

EEG: Electroencephalography

FACS: Facial Action Coding System

fMRI: Functional Magnetic Resonance Imaging

GATE: General Architecture for Text Engineering

GKT: Guilty Knowledge Test

GLM: Generalized Linear Models

GLMM: Generalized Linear Mixed-effects Models

LASSO: Least Absolute Shrinkage and Selection Operator

LIWC: Linguistic Inquiry and Word Count

MEM: Maximum Entropy Model

NB: Naïve Bayes

NLP: Natural Language Processing

PCFG: Probabilistic Context-Free Grammar

POS: Part-of-speech

RIT: Relevant-Irrelevant Test

RM: Reality Monitoring

ROC-AUC: Area under the Receiver Operating Characteristic Curve

SCAN: Scientific Content Analysis

SVA: Statement Validation Analysis

SVM: Support Vector Machine

Acknowledgements

I want to thank the faculty, staff and fellow students at the psychology department of UBC – Okanagan for their support of my work.

I would like to express my special gratitude to my supervisor Dr. Brian O'Connor, thank you for your guidance and continuous support. To other members my supervisory committee, Dr. Jan Cioe and Dr. Michael Woodworth, thank you for your valuable input and encouragement.

I would also like to express my appreciation to my volunteer, Ms. Emma Smith, for the time she spent helping me with critical tasks in this research.

To my dear family and friends, thank you for your unwavering support and love.

Dedication

To Veritas, the goddess of truth.

To Laozi, the great teacher of wisdom.

Chapter 1: Introduction

1.1 Overview

Deception is a common occurrence in everyday human social interactions (Cole, 2001; DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; Vrij, Floyd, & Ennis, 2003). Since many human endeavors rely on truthful communication, people have practiced the art of deception detection throughout history (Kerr, 1990; Larson, 1932/1969; Trovillo, 1939). Nevertheless, empirical evidence has shown that lie detection techniques based on human judgement and intuition are highly unreliable (Frank, Paolantonio, Feeley, & Servoss, 2004; Hartwig & Bond, 2011; Vrij, 2008). Generally, a wide range of deception cues and contextual information need to be considered in order to make relatively reliable distinction between truthful and deceptive communication. Meanwhile, the proliferation of information technology in the past couple of decades has made computer-mediated communication (CMC) an increasingly important mode of communication, where deception also frequently takes place (Hancock, Thom-Santelli, & Ritchie, 2004; Haythornthwaite & Wellman, 2002). CMCs (including, but not limited to text messaging, e-mail, and social media) post considerable challenges to deception detection by limiting potential cues of deception to mainly a single dimension – language.

Past research in language-based deception detection has demonstrated that psychological variables that are often associated with lying, such as emotion and memory, can affect language usage (e.g., Alonso-Quecuty, 1992; Pennebaker & Graybeal, 2001). A prominent approach to language-based lie detection known as reality monitoring suggests that statements based on true experience (i.e., memories of external origins) are quantifiably different from fabricated ones (i.e., based on internally generated memories; Johnson & Raye, 1998). On another note,

advancements in the study of computerized natural language processing (NLP) and data science have enabled the emergence of a new field of deception detection research based on automatic language classification. This research endeavors to build highly accurate automatic deception detection systems to cope with the abundance of digital deception in the online space, as well as perform traditional language-based lie detection tasks at beyond human levels. Deception detection studies that utilized NLP techniques only became available in the past 20 years, but have already shown promising results (Banerjee & Chua, 2014; Bond & lee, 2005; Fuller, Biros, & Wilson, 2009; Fusiliera, Montes-y-Gómezb, Rossoc, & Cabreraa, 2015; Hao, Chen, Cheng, Chandramouli, & Subbalakshmi, 2011; Long, Nghia, & Vuong, 2014; Rayson, Wilson, & Leech, 2001; Ott, Choi, Cardie, & Hancock, 2011; Ott, Cardie, & Hancock, 2013; Xu & Zhao, 2012).

Furthermore, given the complex nature of deceptive behavior, all deception detection methods are sensitive to moderator variables (e.g., Masip, Sporer, Garrido, & Herrero, 2005; Sporer & Schwandt, 2007). Research has shown that natural language usage, in general, is linked to individual differences such as social identity and cognitive styles (Pennebaker & Graybeal, 2001; Pennebaker, Mehl, & Niederhoffer, 2003). Language usage during deception can be heavily affected by the age and personality of the individual, as well as situational details of deception (e.g., the subject of the lie, and the liar's prior exposure to related information; Masip et al., 2005; Vrij, 2008). A repeated-measures or within-subject design is a common method for controlling random error when potentially uncontrolled covariates are numerous. In deception detection research, a within-subject design can be used to study behavioral changes during deception at the individual level, rather than general behavioral differences between lying and truth telling. For instance, polygraph lie detection tests (e.g., Larson, 1932/1969; Saxe, Dougherty, & Cross, 1985) monitor psycho-physiological activities of the individual in real time;

deception is signaled by elevated physiological activity from the individual's baseline. In traditional language-based deception studies, especially studies that use the reality monitoring approach, a within-subject design was often used to cope with small sample sizes (e.g., Alonso-Quecuty, 1992; DeCicco & Schafer, 2015; Masip et al., 2005).

Successful applications of NLP in automated authorship authentication (e.g., Stamatatos, 2009; Van Halteren, Baayen, Tweedie, Haverkort, & Neijt, 2005) demonstrated that each individual has a quantitatively distinctive language profile, which may provide critical information to language-based deception detection. Nevertheless, available NLP-based deception detection research has been limited to between-subject design studies, where deceptive and truthful statements were collected from different individuals. Existing studies generally control confounding variables by limiting the content and context of the statements. Results of these studies suggested that, when the content and context of the lie is strictly controlled, NLP-based techniques can detect deception with high rates of accuracy (80% - 90%); however, this strategy is restrictive in its applications. In the interest of exploring another direction of NLP-based deception detection, the current research aims to examine the usage of NLP techniques in a within-subject design lie detection experiment to obtain and analyze multiple true and false statements from each individual participant.

This dissertation will first review key findings of modern deception detection research; then, a novel NLP-based deception detection scheme will be examined. This research collected a repeated-measures true-and-false statement text corpus using experimental manipulations that are similar to traditional laboratory reality monitoring studies (e.g, Alonso-Quecuty, 1992; DeCicco & Schafer, 2015). From each research participant, we elicited multiple truthful statements describing video recordings of real-life events, and fabricated statements related to

these events. Various sets of linguistic cues that are commonly captured in NLP-based as well as traditional deception studies were extracted. The performance of multilevel/mixed-effects variations of popular classification models, including models of both traditional frequentist statistics and machine-learning, were evaluated.

1.2 Deceptive Behaviors

Deception can be broadly defined as a false communication that tends to benefit the communicator (Mitchell, 1986). Deceptive behaviors occur not only in humans. Many species of animals, even plants, have evolved mechanisms or behaviors that send false information in order to exploit the perception of their predator or prey (Bond & Robinson, 1988; Trivers, 1985). In humans, deception is usually associated with the intention to deceive, that is, to cause another person to believe in something the deceiver considers false (Rubin, 2010; Zuckerman, DePaulo, & Rosenthal, 1981). Consequentially, lying is generally considered a morally undesirable act, while at the same time being a part of our everyday social interaction (DePaulo et al., 1996; Schein, 2004).

Research has shown that children start to lie at a very early age (Newton, Reddy, & Bull, 2000; Reddy, 2007; Wilson, Smith, & Ross, 2003), possibility as soon as their cognitive development allows them to distinguish other people's minds from their own (Talwar & Lee, 2008). Based on a diary study conducted by DePaulo and colleagues (1996), on average, adults tell 1 to 2 lies a day, and lie in 20% of all social interactions; however, most of the lies that people tell tend to be inconsequential or could be considered white-lies. DePaulo and colleagues (1996) suggested that more lies were told for psychological reasons than for materialistic reasons. People typically lie to maintain a consistent image of themselves in front of others, or

to smooth social interactions; they also are mostly inclined to believe other people's lies (Ahlmeyer, Heil, McKee, & English, 2000; Stiff, Kim, & Ramesh, 1992; Vrij, Nunkoosing, Paterson, Oosterwegel, & Soukara, 2002). The permissibility of deception may also be judged differently according to the interest of the liar and the importance of accurate communication in the situation (Backbier, Hoogstraten, & Terwogt-Kouwenhoven, 1997; Pontari & Schlenker, 2006). So-called white-lies are deceptions with a prosocial intent when telling the truth is believed by the liar to cause harm to the listener (Sweetser, 1987). Nevertheless, malicious or self-interested lies can have serious personal, societal, and economic consequences.

How much people lie is influenced by many aspects of individual characteristics. For instance, teenagers lie a lot more than adults (Levine, Serota, Carey, & Messer, 2013). Men tend to lie more about themselves than women, while women lie more often about others than men (DePaulo & Bell, 1996; Feldman, Forrest, & Happ, 2002). Extraverted and socially self-conscious individuals tell more lies during social interactions than others (Kashy & DePaulo, 1996). Attachment styles are also related to the frequency of lying. People who have high attachment anxiety lie more frequently to both strangers and friends (Ennis, Vrij, & Chance, 2008), while avoidant attachment is related to more frequent lying in romantic relationships (Cole, 2001). Most importantly, individuals who have psychopathic, narcissistic, and Machiavellian personality traits are known to be especially disposed to manipulative and deceitful behaviors (Kashy & DePaulo, 1996; Porter & Woodworth, 2007); they are prolific liars who lie much more often and tell more serious lies than the average individual (Serota & Levine, 2015; Serota, Levine, & Boster, 2010). In a more recent prevalence study in America, Serota and colleagues (2010) found half of the reported lies were told by only 5.3% of the sample. They argued that, in light of such a drastic individual difference in the frequency of lying,

prevalence estimates based on sample average (i.e., adults lie 1 to 2 times per day; DePaulo et al., 1996; Serota et al., 2010) are misleading.

The mode of communication also has an effect on deceptive behavior. For instance, people appear to lie more often on telephone than during face-to-face communication, in e-mails, or in instant messages (DePaulo et al., 1996; Hancock et al., 2004). Hancock and colleagues (2004) argued that the influence of communication media on the frequency of deception is determined by multiple factors related to the design of the media. In the case of a telephone conversation, the communication allows for synchronous interaction, where the liar can monitor the reaction of the target; it is also recordless and distributed (i.e., the liar and the target are not co-present in the same location), thus making it the ideal medium for deception. In addition, the mode of communication is known to moderate the effect of motivation on how well a person lies. In face-to-face communications, lie catchers have observed a motivational impairment effect; that is, the more motivated is the liar to avoid detection, the easier it is for the lie catcher to detect the lie (DePaulo & Kirkendol, 1989). This effect is reversed in CMC environments, where motivated liars appear to be harder to detect (Hancock, Woodworth, & Goorha, 2010; Woodworth, Hancock, & Goorha, 2005).

1.3 Traditional Deception Detection Research

All forms of deception detection rely on the assumption that liars and truth tellers are distinguishable through either overt or covert behaviors. The task of lie detection involves perceiving and interpreting these behavioral signals, which are often referred to as deception cues. Prescientific methods of deception detection tended to rely on a few easily observable cues, such as dry mouth, face discoloration, avoiding eye contact, and finger or toe rubbing, and

simple associations of fear and guilt to deception (Trovillo, 1939). Popular culture also romanticizes the effectiveness of simple and easily executable methods of lie detection. In fact, empirical evidence suggests that, relying on human observation alone, both laypersons and professional lie catchers (e.g., police officers and judges) are generally no more accurate than chance in spotting lies (e.g., Bond & DePaulo, 2006; Levine, Kim, Park, & Hughes, 2006; Vrij, 2008). Furthermore, people often judge incorrectly their own ability to catch liars (Ekman & O'Sullivan, 1991); professional lie catchers, especially, have a tendency to over-estimate their own ability to detect lies. In fact, self-reported experience and level of confidence are not associated with higher accuracy (Bond & DePaulo, 2008; Wyman, Foster, Lavoie, Tong, & Talwar, 2017). These findings are unsurprising given decades of scientific research has found no reliable cue for deception akin to Pinocchio's nose (DePaulo et al., 2003). Today's scientific research in deception detection shows distinguishing truths from lies is a highly complex task that one may accomplish through utilizing evidence-based understanding of human behavior, specialized measurement tools, and systematic decision-making processes (Vrij, 2008).

Theories of deception have attributed the behavioral differences between liars and truth-tellers to a variety of causes. For instance, Davis (1961) suggested that the main source of deception cues is the liar's unconscious fear response to the possibility of being detected. Elaad (1990) argued that the traumatic nature of the event, the liar's internal conflict about telling the lie, and the threat of punishment will increase the intensity and frequency of occurrences of deception cues. Knapp, Hart, and Dennis (1974) highlighted a number of behavioral and emotional phenomena that may cause deception cues to appear, including nervousness, reserved speech, exaggerated behavior, insecurity, and conflict between external behavior and internal emotions. A multi-factor theory proposed by Zuckerman and colleagues (1981) argued for both

emotional and cognitive sources of deception cues. These sources included incriminating emotions, cognitive overload, and failed behavioral control. Furthermore, the interaction between liars and their targets may play an important role in shaping deceptive behaviors (Buller & Burgoon, 1996; Buller, Stiff, & Burgoon, 1996). An interpersonal deception theory proposed by Buller and Burgoon (1996) emphasizes the effect of motivations and goals on deceptive behavior. According to Buller and Burgoon, deceivers who are motivated by either instrumental or relational reasons will engage in various strategic behaviors in order to maintain credibility during the interaction. This may lead to overly managed behaviors, which will appear to be rigid, inexpressive, and non-spontaneous.

In practice, empirical studies of deception detection have taken place both in laboratory and in the field. Both sources of deception data have their advantages and limitations. Laboratory studies offer an abundance of data, usually with controlled conditions and clearly established ground truth. Researchers have used various experimental manipulations, such as games, material rewards, and simulated crimes to elicit deception (e.g., Blair, Levine, & Shaw, 2010; Levine, 2010; Porter & Yuille, 1996; Sip et al., 2010). However, experimentally simulated lying conditions usually involve sanctioned lies, which may not be representative of real-world high-stake deception scenarios (Frank & Ekman, 1997). On the other hand, real-world deception data, such as videotaped police interviews (e.g., Davis, Markus, Walters, Vorus, & Connors, 2005; Mann, Vrij, & Bull, 2002), provide opportunities to study deception when the liars had authentic motivations to lie. Nevertheless, there is always a degree of uncertainty regarding the ground truth when the deception is not experimentally controlled. Real-world deception situations for which ground truth can be identified with a high degree of certainty are rarely available.

Studies that evaluated the validity of individual deception cues have shown that, either in the laboratory or in the field, no proposed deception cue can be relied upon to consistently signal deception (Vrij, 2008). Deception detection is now understood as a complex decision-making process. To detect deception, lie catchers often need to examine a cluster of cues, that may or may not be directly observable, along with various contextual information (Fitzpatrick et al., 2015). The most successful schemes of deception detection usually involve structured procedures that systematically capture behavior data and use probabilistic decision-making (i.e., having a test score or measurement that is associated with the probability of deception). Investigative techniques and tools designed to induce or isolate the occurrence of deception cues may also be employed by these deception detection schemes. Given that clues of deception can manifest in various ways, deception detection approaches are often grouped by the types of cues on which they focus. In general, there are three types of deception cues: (a) non-verbal behaviors, (b) psycho-physiological activities, and (c) linguistic cues (Vrij, 2000; Yuille, 1989). Researchers who developed lie detection techniques based on a specific type of cue have also proposed deception theories that justify their choices of cues. The following review of deception detection methods and theories will be organized based on these three categories of cues.

1.3.1 Deception Detection Using Non-Verbal Cues

Traditionally, the majority of deception detection scenarios take place during face-to-face communication; therefore, a great deal of research effort has been dedicated toward finding deception cues that are readily observable to humans. Non-verbal and para-verbal (e.g., pitch and tone of voice) cues of deception were a major research focus in face-to-face lie detection. So far, research has revealed few clear differences in non-verbal and para-verbal behaviors between lying and truth-telling. Numerous deception cues related to facial expression, movements of

hands, feet, and the head have been studied, but none of these cues have had a consistent significant association with deception across all studies. In many cases, even the direction of the association was in question. For instance, while some studies found liars smile more than truth tellers (e.g., Gozna & Babooran, 2004), others reached the opposite conclusion (e.g., Vrij, 2008).

In a meta-analysis of deception cues, DePaulo and colleagues (2003) compiled a comprehensive list of proposed non-verbal and para-verbal deception cues that had been researched and calculated their average effect sizes across all available studies. This list showed only three non-verbal and para-verbal behaviors that had a significant average effect size in association with deception: A person who is lying tends to use a higher pitch of voice (r = .21), move his/her hand and finger less often (r = -.36), and use fewer hand gestures to illustrate his/her speech (r = -.14). Hypothesized behavioral cues of deception, such as gaze direction, speech hesitation, and leg and foot movements were shown to be unrelated to lying (r < .1). Moreover, Bond, Levine, and Hartwig (2015) observed a "decline effect" in the effect sizes of non-verbal deception cues from meta-analysis results over time; that is, the longer a deception cue was studied, the weaker was its effect. In a recent review of cue-based deception detection research, Levine (2018) noted that non-verbal deception cues were often found to be significantly related to deception in the opposite way that the literature expects. He argued that deception cues are highly situational. Their association with deception varies greatly across context and deceiver; therefore, the results of deception studies are often not replicable and have poor generalizability. These findings encourage a skeptical view regarding any claims of a single effective deception cue.

Furthermore, the effect of motivational mediators on the manifestation of deception cues has been demonstrated. Some studies have shown that the liars' incentive in getting away with

their lies does affect the occurrences of deception cues as predicted by theory. For instance, DePaulo and Morris (2004) found that the increase in voice pitch only occurred when the stake of the lie is high. The same was true for the decrease of foot and leg movement, and gaze avoidance, which were only observed when the liars were motivated to avoid embarrassment (DePaulo et al., 2003).

1.3.1.1 Leakage Theory and Micro-Expressions

One theory-driven approach to deception detection using non-verbal cues is the study of deception-related micro-expressions. Micro-expression is a series of subtle emotive behaviors conceptualized by Ekman and Friesen's (1969) in their theory of non-verbal deception leakage. Ekman and Friesen defined leakages as behaviors that give away concealed information, in contrast to deceptions cues, which only signal that deception is in progress. They argued that the main source of deception cues and leakages are evolved instinctual behaviors such as basic emotions; the less a behavior is controlled consciously, the more likely it will become a leakage during deception. Ekman's (1985) deception detection approach specifically focuses on emotive facial actions that express fear, guilt, and duping delight. He suggested that given the involuntary nature of emotive facial expressions, lying individuals who attempt to falsify or hide emotions will likely fail in subtle ways. These failures in hiding true emotions will manifest as micro-expressions, which are fleeting moments of emotion difficult to detect by the naked eye, but can be recognized by trained professionals or through frame-by-frame analyses of video recordings of the face based on a quantitative facial action coding system (FACS, Ekman & Friesen, 1978). A micro-expression may contain all or part of the muscle actions in a basic emotion; when identified, it leaks the true emotion a person is trying to mask (Ekman, 2009; Ekman & O'Sullivan, 1991).

The research on micro-expressions was driven mainly by Ekman and his research team. Both Ekman's early theoretical and empirical works seemed to suggest that micro-expression as a promising deception detection tool (Ekman & Rosenberg, 2005; Gladwell, 2005). Ekman claimed that individuals who are properly trained in reading micro-expressions can increase their lie detection accuracy in face-to-face interactions from chance level (50%) to 70% (Gladwell, 2005; Weinberger, 2010). Unfortunately, many details of Ekman's research cannot be found in published scientific literature because of his wish to keep certain information inaccessible to the general public and foreign governments (Henig, 2006; Weinberger, 2010). Nevertheless, in the recent decade, a number of researchers have independently investigated various aspects of Ekman's claims about micro-expression in relation to lie detection (e.g., Jack, Caldara, & Schyns, 2012; Porter & ten Brinke, 2008; Porter, ten Brinke, & Wallace, 2012).

Ekman (2009) claimed that micro-expressions were observed in at least half of the deceptions in his research (e.g., Frank & Ekman, 1997). Two laboratory studies of emotional deception leakage (Porter & ten Brinke, 2008; Porter et al., 2012) found that almost every individual who attempted to fabricate or mask an emotion displayed inconsistent facial expressions that betrayed their true emotion. However, most of these emotional leakages lasted for more than a second, and micro-expression, according to Ekman's (1985) definition (last for 1/25-1/5 second), rarely occurred. These studies provided empirical support for emotional leakages during deception, yet, their results challenged the concept of micro-expression. These findings seemed to suggest that emotional expressions are even more difficult to conceal than previously believed, in that most emotional leakages are full displays of expression rather than micro-expressions. Nevertheless, Ekman's studies were based on real-world high-stake cases, while the laboratory

studies used simulated scenarios. The difference in stakes and settings may be responsible for differences in the manifestations of emotional leakages.

1.3.1.2 Impressionistic Cues

It has been suggested that non-specific behavior cues such as abnormal and unexpected behaviors are what people intuitively use to make judgements of deception (Bond et al., 1992; Levine et al., 2000). Some studies have found that deception can be detected with a higher accuracy using some global impressions than using individual deception cues and reasoning (DePaulo et al., 2003). For instance, Anderson, DePaulo, and Ansfield (2002) found that people who were told a true story felt more comfortable, less suspicious, and thought the narratives were more informative compared to when they were told a fabricated story. Albrechtsen, Meissner, and Susa (2009) compared the accuracies of intuitive decision-making style and deliberative decision-making style in deception detection tasks. They found using intuition rather than a deliberative process can significantly improve an individual's accuracy in detecting lies. ten Brinke, Lee, and Carney (2019) found evidence of physiological reactions in observers of high-stakes lies. They suggested that by attending to these automatic reactions, human judges can improve their deception detection accuracy.

DePaulo and colleagues' (2003) meta-analysis showed that global impression cues, including vocal involvement, cooperation, and ambivalence, are more strongly associated with deception than specific behavioral cues. DePaulo and Morris (2004) suggested that a group of behaviors that occur simultaneously may contribute to a dishonest impression. For instance, liars tend to appear less certain in their use of words and tone of voice, but at the same time convey confidence through body language. Some researchers argued that impressions formed with less

conscious thought are more accurate in identifying deception than cue-based cognitive judgements (ten Brinke, Stimson, & Carney, 2014). If we consider general impressionistic cues as clusters of specific behaviors, these findings may suggest that deception is signaled by a complex yet recognizable behavioral pattern. Therefore, studying deception-related behavioral profiles using multivariate classification methods rather than isolated behavioral cues may be a more fruitful direction in deception detection research (Bond et al., 2015). This sentiment is supported by Levine's (2018) observation that deception detection models using multiple deception cues mostly reported significantly greater-than-chance detection accuracies, but the associations between deception and individual behavioral cues are inconsistent across different studies.

1.3.2 Psycho-Physiological Cues

Psycho-physiological deception cues are covert non-verbal behaviors that can only be observed through specialized instruments, such as the polygraph machine, thermographic cameras, electroencephalography (EEG), and functional magnetic resonance imaging (fMRI). Many widely adopted systematic lie detection tests by professional liar catchers are based on psycho-physiological cues.

1.3.2.1 Polygraph Tests

The polygraph is the most popular deception detection tool designed to capture psycho-physiological deception cues. Polygraph tests consist of a family of deception detection procedures based on three kinds of stress-related physiological activities measured by the polygraph machine, which are blood pressure, electro-dermal activity, and respiration (American Polygraph Association, 2010). These tests are popular practices among government agencies

and private organizations in assisting various deception detection-related tasks, such as crime investigation, insurance fraud inquiry, and employee screening (Raskin & Honts, 2002). The basic assumption of polygraph tests is that deception-related stress and anxiety will manifest as elevated levels of physiological activities. These tests generally involve a questioning procedure, during which the examinees' physiological activities are continuously monitored by the polygraph machine; changes in the levels of physiological activities from an established baseline may be interpreted as indicators of deception.

Currently, there are mainly two different approaches to polygraph testing, which are the concern approach and the orienting reflex approach. Concern-based polygraph tests are some of the oldest and most practiced modern lie detection tests. The fundamental principle of these tests is that the liar will experience greater anxiety and stress when answering crime-relevant questions (e.g., "did you steal that wallet?") about which they have to lie, than inconsequential control questions (e.g., "is your name John?"; Saxe, Dougherty, & Cross, 1985). The relevant-irrelevant test (RIT, Larson, 1932/1969) was a widely used early form of concern-based polygraph test that used precisely these two kinds of questions. Elevated levels of physiological activities when responding to crime-relevant questions were interpreted as detection anxiety from lying (Raskin & Honts, 2002). Today, the RIT is considered overly simplistic and ineffective by most experts because of its failure to account for intrapersonal differences in responses to considerably different questions (Raskin & Kircher, 2014; Vrij, 2008); that is, crime-related questions are inherently more stressful to answer than neutral questions regardless of whether the examinee was guilty or innocent. A significant development of concern-based polygraph testing was the control question test (CQT, Raskin, 1986; Reid, 1947), which improved upon RIT by adding generalized incriminating control questions, for example: "Have

you ever done something illegal," to control for the intrapersonal differences in response to questions of different nature.

Although CQT is currently the most common form of concern-based test practiced by professional lie catchers, its adequacy as lie a detection test is still controversial. A review of CQT research by the National Research Council (2003) concluded that the control questions used in CQT are insufficient for controlling intrapersonal differences; that is, the CQT cannot theoretically rule out the possibility that innocent examinees will react to the questions in the same manner as guilty examinees. In addition, the CQT is not a standardized test. The choice of questions and the scoring process are subjective to the individual decisions of the examiners, which can lead to disagreements between different examiners regarding the same case (Carroll, 1988). An attempt to standardize the CQT led to the development of the directed lie test (DLT, Raskin & Honts, 2002; Raskin, Honts, & Kircher, 1997), which added standardized control questions to the CQT, such as "In the first 27 years of your life, have you ever told one lie?" Examinees are instructed to always answer "No" to these questions. Nevertheless, the DLT still does not address the theoretical short comings of the control questions in CQT. Laboratory experiments showed that concern-based polygraph tests can reach 70% to 80% accuracy in distinguishing liars and truthtellers (e.g., Ben-Shakhar, 2002; Ben-Shakhar, Bar-Hillel, & Kremnitzer, M., 2002; Ben-Shakhar & Furedy, 1990; Saxe et al., 1985). Field studies have reported the CQT detecting 90% of the lies (Raskin & Honts, 2002), but they also produce a high rate of false positives (Vrij, 2008).

Compared to the weak theoretical foundation of the concern-based approach, the orienting reflex approach is widely considered by researchers to be grounded in solid behavioral science (e.g., Ben-Shakhar, 2002). Orienting reflex is a well understood psycho-physiological

phenomenon. It refers to a kind of adaptive response to immediate changes in the environment, in which stimuli that are novel or of significant value to the individual evoke greater responses (e.g., Bernstein, 1979; Sokolov, 1963). For example, a murderer will recognize the murder weapon and react more strongly to it than to a weapon that is unrelated to the crime. Guilty knowledge test (GKT), also known as concealed information test, is the polygraph testing procedure that was developed based on the principle of orienting reflex (Lykken, 1960). A GKT consists of a series of multiple-choice questions, each shows a relevant choice and several control choices separately to the examinee. Guilt knowledge is inferred by greater physiological reaction to the relevant choice compared to the control choices.

Both laboratory and field studies suggested that the effectiveness of GKT is at least on par with the CQT (70% - 80% accuracy; Vrij, 2008). In contrast to the CQT, GKT appears to produce a much lower rate of false positives, but a higher rate of false negatives, especially in field studies (Elaad, 1990; Elaad, Ginton, & Jungman, 1992; MacLaren, 2001). In a meta-analysis of laboratory studies of the GKT, Ben-Shakhar and Elaad (2003) reported a rather impressive average effect size across different types of experimental manipulations (d = 1.55). This study also highlighted a number of important moderators for the effectiveness of GKT. Most importantly, studies that most elaborately simulated real-world deception in forensic settings (i.e., using mocked-crime procedure) produced the largest effect size (d = 2.09, r = .65), whereas in studies that used the participants' personal items and memorized cards as stimulus, GKT was noticeability less effective (d = 1.58 and d = 1.35). The effect of motivational impairment was also apparent. Moreover, GKTs that required the examinees to respond to a larger number of questions, or provide verbal response (i.e., "no") to each choice, were more effective in detecting deception than those that did not.

Although, according to theoretical considerations and empirical evidence, the GKT appears to be a promising lie detection test, it has limited application in real-world lie detection (Vrij, 2008). In forensic settings, the examiners need to possess sufficient information regarding the crime in question in order to design the test. At the same time, they have to ensure that there is no other way the suspect may learn about crime-relevant information without committing the crime. These conditions are often too ideal for real-world criminal investigations.

1.3.2.2 Other Physiological Measures

Other physiological measures such as voice stress analysis, thermal imaging, and electroencephalograms (EEG) also can provide ways to detect deception. These alternative physiological measurements are often associated with GKT research. Although not as widely used as the polygraph, some of these technologies may potentially be more accurate than the polygraph. For instance, research seems to suggest that the P300 brain wave measured by the EEG is a slightly better indicator of guilty knowledge than the polygraph, and produces fewer false positives (Vrij, 2008). Functional magnetic resonance imaging (fMRI) is another promising technology that can be used to detect lies. The fMRI allows researchers to identify specific patterns of brain activities that are associated with different mental tasks; thus, lying should produce unique patterns of brain activity on the fMRI. A number of studies have demonstrated that fMRI can detect lies with reasonable accuracy in strictly controlled laboratory environments (e.g., Davatzikos et al., 2005; Kozel et al., 2005); however, the generalizability of these laboratory results has not been proven. In addition, the high cost of the equipment is a major barrier for field testing and application of fMRI-based lie detection.

1.3.3 Deception Detection Using Linguistic Cues

The hypothesis that words may betray deception can be dated back to Freud's (1959) notion of slips of the tongue, whereby liars would subconsciously reveal their deceit in their use of language. The interpersonal deception theory (Buller & Burgoon, 1996) predicts that liars will attempt to manipulate both nonverbal behaviors and verbal content. Buller and Burgoon (1996) argued that, to manage information, the liars' words will appear to be incomplete, non-veridical, indirect, vague, uncertain, hesitant, brief, and disassociated. Many individual verbal cues of deception have been studied in similar ways to non-verbal cues. DePaulo and colleagues' (2003) meta-analysis included nine individual verbal cues of deception, among which negative statements, generalizing terms, and self-references were found to have a significant effect in predicting deception.

Just like for polygraph testing, the fields of criminal justices and forensic psychology also gave rise to systematic approaches to language-based deception detection, some of which have gained wide recognition and are used to train professional lie catchers throughout the world (Vrij, 2008). The two most well-known examples of forensic statement analysis procedures for detecting deceptions are Sapir's Scientific Content Analysis (SCAN; Sappir, 1987/2000) and the Criteria Based Content Analysis (CBCA, e.g., Berliner & Conte, 1993; Trankell, 1972; Undeutsch, 1982). Here, we will briefly review SCAN and CBCA.

1.3.3.1 Sapir's Scientific Content Analysis

SCAN was developed and marketed by Laboratory for Scientific Interrogation (LSI, www.lsiscan.com). It was designed to analysis hand-written statements that are prepared without the influence of an interviewer. SCAN uses a wide range of criteria to differentiate

truthful statements from deceptive statements. Some examples of SCAN criteria include the use of words such as first-person singular pronouns and past tense, overall structure of the statement, spontaneous correction, spontaneous correction of mistakes, etc. Although SCAN has been adopted in training of law enforcement and military personnel in various countries, its scientific validity is still controversial (Porter & Yuille, 1996; Vrij, 2008). Only a few empirical studies of the SCAN exist, including three field studies (Adams & Jarvis, 2006; Driscoll, 1994; Smith, 2001) and two laboratory studies (Nahari, Vrij, & Fisher, 2012; Porter & Yuille, 1996), which differ greatly in both methodology and results. In general, field studies showed support for some SCAN criteria, such as denial of allegation and statement structure, whereas laboratory studies found no statistically significant difference in any of the SCAN criteria between truthful and deceptive statements. Aside from a lack of empirical support, SCAN was also criticized for its lack of theoretical framework (e.g., Smith, 2001; Vrij, 2008, 2015). The reason why SCAN criteria should differentiate truthful statements from deceptive ones has never been articulated, nor were the criteria quantified. As a result, SCAN lacks a standardized decision-making process or a cohesive total score (Vrij, 2008; Vrij, 2015). Different users of SCAN tend to use different criteria to justify their decision of whether a statement is truthful or deceptive (Smith, 2001). These characteristics of the SCAN make SCAN-based decisions largely subjective.

1.3.3.2 Criteria-Based Content Analysis

CBCA is one critical part of Statement Validation Analysis (SVA; Arntzen, 1970; Trankell, 1972; Undeutsch, 1982), an extensive procedure that was developed in forensic psychology to evaluate the credibility of witness testimonies. SVA combines a case-file analysis, a semi-structured interview, CBCA, and clinical judgement to answer the question

whether a statement given by a specific witness is likely fabricated. During CBCA, the transcript of the semi-structured interview is scored according to 19 criteria that are on a 3-point scale (0 = *criterion is absent*, 1 = *criterion is present*, 2 = *criterion is strongly present*). With each criterion that is present in a statement, the credibility of the statement increases. Although some CBCA criteria overlap with SCAN criteria, CBCA appears to be a more systematic way of analyzing the content of a statement than SCAN, which attempts to capture a wide range of unrelated verbal features. The 19 CBCA criteria are grouped into four categories, which are general characteristics, specific contents, motivations-related contents, and offence-specific elements. Published studies of the CBCA criteria showed that some of these criteria were more consistently present in truthful statements than others. For instance, Vrij (2008) summarized findings of 38 field and laboratory studies of the CBCA. He found four CBCA criteria (unstructured production, quantity of details, contextual embedding, and reproduction of conversation) occurred more frequently in truthful statements than in fabricated statements in at least 50% of the studies that examined these criteria, while two criteria (accurately reported misunderstood details and self-deprecation) were significant indicators of truthful statements in roughly 10% of the available studies. The majority (80%) of the studies that included the CBCA total score found truthful statements scored significantly higher than deceptive statements. Nevertheless, a CBCA score is not a standardized measurement of truthfulness; case details and other contextual information gathered in other steps of the SVA are crucial to its interpretation (Köhnken, 2004). For example, individuals' exposure to case-related details will directly influence the CBCA score; statements provided by children or mentally ill individuals are also likely to score much lower than normal adults. Therefore, CBCA scores should not be viewed as probabilistically linked to the fabrication of a statement.

Although both the SCAN and the CBCA originated from field experience of credibility assessment experts, the CBCA has generated significantly more empirical research than the SCAN in support of its validity as a lie detection tool (Vrij, 2008). Based on available evidence, it is safe to conclude that the CBCA criteria can, at least in laboratory settings, reflect real differences between truthful and fabricated narratives (Masip et al., 2005; Oberlader et al., 2016). These differences are often attributed to the psychological impacts of fabricating a story especially during face-to-face communication, where cognitive load and motivation are major factors (e.g., Undeutsch, 1982). As a result, false memory, as experienced by both children and adults in many credibility-related cases (e.g., Foley & Johnson, 1985; Parker, 1995), presents a substantial challenge to the CBCA/SVA approach. Indeed, laboratory experiments have showed that CBCA criteria have little discriminative power in differentiating narratives that are based on experimentally planted false memory from those based on true memory (Blandón-Gitlin, Pezdek, Lindsay, & Hagen, 2009; Kulkofsky, 2008; Porter, Yuille, & Lehman, 1999). Nevertheless, basic research in memory and cognition suggested that narratives based on true and false memories can be differentiated through their content, regardless of whether the narrators intended to deceive or not (Johnson & Raye, 1981). Another language-based deception-detection approach known as Reality Monitoring (RM, Johnson, 1988; Johnson & Raye, 1998) was created based on this line of research. The following section will discuss the RM approach.

1.3.3.3 Reality Monitoring

RM was developed as a scientific method of differentiating externally generated memories (i.e., memories generated by experience and perception) from internally generated memories (i.e., imagined or fabricated events). It posits that memory based on true experience

will be rich in emotional, sensory, and perceptual information; in contrast, fabricated memories tend to be less vivid, but contain more references to thinking and reasoning (Johnson & Raye, 1998). According to this framework, truthful narratives should contain larger quantities of sensory, contextual, and semantic information, whereas false narratives should contain a greater amount of information indicative of cognitive processes (Masip et al., 2005). Early research of RM included eight criteria for distinguishing true and false narratives. These are clarity/vividness, perceptual information, spatial information, temporal information, affect, re-constructability of the story, realism, and cognitive operations (Sporer, 1997). However, there is currently no standardized test based on the RM approach. Different deception studies may operationalize the same criteria in different ways, or use slightly different criteria.

The first study of RM criteria was conducted by Alonso-Quecuty (1992), where a video depicting a crime was shown to 22 individuals, who were then asked to write a truthful statement and a deceptive statement about the crime. Results of this study supported the notion that deceptive narratives contain more internally generated information, but were inconclusive regarding whether truthful narratives contain more sensory and contextual information, in that a 10-min delay between writing the two statements appeared to reverse the effect. A few subsequent studies conducted by Alonso-Quecuty and her research team using similar study designs (e.g., Alonso-Quecuty, 1992; Alonso-Quecuty, Hernandez-Fernaud, & Campos, 1997; Hemandez-Fernaud & Alonso-Quecuty, 1997) were largely supportive of the hypothesis that truthful narratives contain richer sensory and contextual information; however, they found the difference in internally generated information between truthful and deceptive narratives statistically non-significant. Alonso-Quecuty and colleagues (1997) also showed that the

difference between true and false narratives measured using RM criteria, may increase through repeated telling of the stories.

Studies conducted by other research teams have also found mixed results in regard to the utility of individual RM criteria. A few studies have produced results that indicated the relations between individual RM criteria and the truthfulness/falseness of the narratives to be non-significant or not in the hypothesized direction (e.g., Manzanero & Digest, 1996; Santtila, Roppola, & Niemi, 1999). However, when all RM criteria were utilized, the overall system does appear to discriminate with a similar accuracy as CBCA (Sporer, 1997; Vrij, Akehurst, Soukara, & Bull, 2004; Vrij, Edward, & Bull, 2001; Vrij, Edward, Roberts, & Bull, 2000). Furthermore, similar to all previously mentioned deception-detection cues and systems, the validity of the RM criteria and the effectiveness of the RM system are influenced by numerous factors. For instance, Hemandez-Fernaud and Alonso-Quecuty (1997) suggested that the effect of externally generated information is mediated by the style of the interview. They found truthful and deceptive statements obtained through cognitive interviews (CI, Fisher & Geiselman, 1992) had a greater difference in terms of contextual content than statements obtain through normal police interviews. RM scores were also found to be correlated with personality factors, gender (Vrij et al., 2001), and age (Vrij et al., 2004).

Compared to other language-based deception-detection systems that were mentioned in this section, researchers have noted that the support of a theoretical basis for selecting cues is a clear advantage of the RM (Masip et al., 2005; Vrij, 2008); however, empirical support for the validity of some of its criteria was weak. Furthermore, better discrimination can be achieved by combining RM and CBCA criteria (Sporer, 1997; Vrij et al., 2004), indicating that these two systems may be complementary to each other. Researchers have also attempted to use automatic

coding of RM criteria to solve issues of subjectivity associated with the implementation of the system (Bond & Lee, 2005), but due to the inability of the computerized coding system to take language context into account, manual coding appeared to be the superior approach for applying RM (Vrij, Mann, Kristen, & Fisher, 2007). Nevertheless, the use of automatic systems has become increasingly popular in language-based deception-detection research. The most successful of such systems generally employ machine-learning/data-mining algorithms to analyze massively multivariate data sets that included large numbers of machine-extracted language features. These studies may use the RM theory as a general justification for the possibility of discrimination between truthful and deceptive statements, but the selection of cues is treated mostly as a technical issue with the goal of maximizing predictive validity of the system (Fitzpatrick et al., 2015). This computational approach to language-based deception detection will be the focus of the following section.

1.4 Deception Detection Using Natural Language Processing

Natural language processing (NLP) is a developing field of computer science that is concerned with language-based human-machine interaction. Motivated by ideas such as automatic translator and speech-based human-computer interface, this field of research and technological developed aims to enable machines to understand human language (Nadkarni, Ohno-Machado, & Chapman, 2011; Turing, 1950/2007). A main goal of current NLP research is to create computer programs that can perform human language-based tasks at or above human capability. Notably, in clinical psychology, NLP has been used in a number of recent novel studies. For instance, Cook and colleagues (2016) used NLP-based prediction models to successfully identify suicidal ideations of recently discharged psychiatric patients based on their answer to a simple unstructured question: “How do you feel today?” Jarrold and colleagues

(2010) showed that NLP-based analyses can be used to identify cognitive impairment, depression, and pre-symptomatic Alzheimer's disease with up to 97% accuracy. He, Veldkamp, Glas, and de Vries (2015) developed an automatic posttraumatic stress disorder screening process based on patients' self-narration using NLP. Related to deception detection research, specifically the RM approach, Rayson and colleagues (2001) have found that imaginative and informative writings are computationally differentiable based on their word usages. Although automated deception detection is a relatively new area of NLP application, results from various studies have consistently demonstrated that even the simplest NLP techniques can outperform human judgement in differentiating true and false statements (Fitzpatrick et al., 2015).

1.4.1 Linguistic Features

Procedurally, NLP-based deception detection consists of two essential computerized tasks: The first is the automated extraction of linguistic features; the second is the classification of true and false statements based on the extracted features. Although speech recognition is a major part of NLP research (e.g., Cambria & White, 2014), so far, all deception studies were based on electronic text data. Language features can be coded at different linguistic levels by the computer, ranging from single characters to the structure of an entire conversation. Currently, the levels of features that have received the most empirical and theoretical supported in terms of their relevance to deception detection include n-gram, part-of-speech (POS), and psycholinguistic features (Fitzpatrick et al., 2015).

A lexicon n-gram is a contiguous sequence of *n* words. A one-word sequence is a unigram, a two-word sequence is a bigram, a three-word sequence is a trigram, et cetera. These features can provide information about content as well as context of a given text. For example, articles from the *Wall Street Journal* will contain a high frequency of bigrams such as "stock

price" and "debt increase," while a play of Shakespeare will have a high occurrence of "thou art" and "I pray." Language analysis based on n-gram frequencies has been used extensively in solving text classification problems, such as sorting online writings by interest, sentiment, and emotional state (e.g., Abbasi, Chen, & Salem, 2008; Lin, Yang, & Chen, 2008). It is also a frequently used method in authorship attribution research (e.g., Ishihara, 2014; Kešelj, Peng, Cercone, & Thomas 2003; Peng, Choo, & Ashman, 2016). Deception detection research has shown that n-gram features can be used alone to identify online deceptive writings with reasonable accuracies (above 75%), or combined with other types of features to achieve even greater accuracies (e.g., Mihalcea & Strapparava, 2009; Ott et al., 2011; Ott et al., 2013; Xu & Zhao, 2012).

Parts-of-speech (POS) are word categories according to their syntactic functions, for example, noun, pronoun, adjective, and verb. POS tagging is another common NLP technique used in linguistic analysis. Language typology research showed that POS distributions can be used to differentiate conversational speech and writing, as well as imaginative writings and informative writing (Rayson et al., 2001). According to Rayson and colleagues, informative writings tend to use more nouns, adjectives, prepositions, determiners, and coordinating conjunctions. Imaginative writings, on the other hand, tend to use more verbs, adverbs, pronouns, and pre-determiners. Some deception studies showed that POS tags are slightly less accurate predictors of deception than n-grams (e.g., Ott et al., 2011; Ott et al., 2013; Xu & Zhao, 2012), while being a much smaller set of features.

Psycholinguistic features are a selection of specific features, or groups of words, that are theoretically associated with deception, such as words related to positive/negative emotion and words that give sensory information. Text processing programs like the General Architecture for

Text Engineering (GATE, Cunningham, 2002) and the Linguistic Inquiry and Word Count (LIWC, Pennebaker, Boyd, Jordan, & Blackburn, 2015) provide means to extract this type of features from comprehensive lists of features. Many of these computer-coded psycholinguistic features were shown to appear in different frequencies between deceptive and truthful communications (Zhou, Burgoon, Nunamaker, & Twitchell, 2004a). Deception detection studies that were based on psycholinguistic profiles produced by the GATE and the LIWC generally reported about 70% accuracy (e.g., Fuller et al., 2009; Zhou, Burgoon, Twitchell, Qin, & Nunamaker, 2004b) in distinguishing deceptive statements from truthful statements.

Furthermore, grammatical features other than POS have occasionally been used for deception detection. Features that contain information about deeper grammatical structure of sentences are sometimes referred to as *deep features*. A study conducted by Xu and Zhao (2012) included POS n-gram and parse in addition to other commonly used feature types. POS n-grams are n-grams of POS tags, for example, a singular proper noun followed by a determiner (NNP_DT) is a POS bigram. Parse is a representation of the syntactic structure of a sentence. A number of different parsing approaches exist. Dependency parsing (de Marneffe, MacCartney, & Manning, 2006) and probabilistic context-free grammars (PCFGs, Jelinek, Lafferty, & Mercer, 1992) are two prominent examples. Dependency parsing, in particular, is based on dependency grammar, which describes the relation between two words in a sentence as the head and its dependent (Covington, 2001). For example, in the sentence "Tom loves cats," "love" is the head, "Tom" and 'cats" are the subject and object of "love" respectively. Xu and Zhao (2012) used various combinations of psycholinguistic cues, lexicon n-grams, POS n-grams, and dependency parses to analyze the dataset that was collected by Ott and colleagues in 2011 and

achieved marginally better classification results. Their highest accuracy (91.6%) was achieved while using a combined feature set of lexicon unigrams and dependency parses.

1.4.2 The Computational Approach to Text Classification

Given the large number of potential features that may be utilized to detect deception, traditional statistical methods of classification are usually inadequate. The use of machine-learning classifiers in the place of frequentist inferential statistics is often a major difference between deception studies conducted by the NLP research community and those based in psychological science. Machine-learning is a subfield of both computer science and statistics that studies algorithms capable of learning from data. The simplest form of machine-learning algorithms can be driven from traditional statistical models such as the logistic regression, but instead of using a statistical estimator (e.g., maximum likelihood), machine-learning algorithms procedurally try out a sequence of parameter values until the best solution for a given dataset is found. Advanced machine-learning algorithms allow the model itself to be highly dynamic; some algorithms such as artificial neural networks are even considered model agnostic (Dangti, 2017).

Furthermore, machine-learning includes cross-validation as part of the model building process by separating the data into training and testing data. The algorithm "learns" the model from the training data, then apples the learned model to the testing data for performance evaluation. An *n*-fold cross-validation means the original sample is divided into *n* sub-samples, and each sample takes a turn being the testing sample, while the model is estimated *n* times; the prediction results are averaged.

Compared with traditional statistical methods, many machine-learning algorithms can utilize a much larger number of predictors and estimate complex interactions that are not feasible

in traditional statistics. They can also handle nonlinearity and noise in the data automatically (Yarkoni & Westfall, 2017). These features make machine-learning approaches ideal tools for analyzing large real-world data sets. Although basic assumptions still should be made for the underlining data-generation process in machine-learning (different assumptions will result in different machine-learning approaches), they are much more relaxed than in traditional statistics. Furthermore, machine-learning algorithms in general are very robust to assumption violations in that they can make good predictions even if the prediction model is somewhat different from the true model (Dangeti, 2017). A variety of machine-learning classifiers have been applied to text-based deception detection. Some successful examples include decision tree (Quinlan, 1986), artificial neural networks (ANNs; see Hassoun, 1995), and support vector machine (SVM; Cortes & Vapnik, 1995)

Decision tree as a decision-making tool uses tree-like flowcharts in order to identify the path through a number of nodes to reach a final goal. The same process was adapted in machine-learning, where predictor variables are selected as "nodes" at different levels in a tree model. Different observed values at a node will lead to different "branches" that connect to other nodes, and eventually lead to a "leaf" representing the final class label (e.g., truth or lie). One advantage of the decision tree over many other machine-learning approaches is its simplicity. It is a so-called white-box model, in that the relation between input variables and the outcome can be easily understood and interpreted following Boolean logic. When used for deception detection, tree classifiers may identify important deception cues from a large selection of cues, and produce combinations of cues and decision pathways that make the most accurate distinction between truthful and deceptive statements. Furthermore, when dealing with massively multivariate problems, such as text classification, an ensemble of trees can be computed to create

a so-called “random forest” (Breiman, 2001). The final prediction made by a random forest model summarizes the output of a large number of trees. This technique sacrifices some computational efficiency for much better accuracy and stability.

The ANNs are a family of biologically motivated algorithms that simulate neural networks in the brain. The fundamentals of ANNs were laid by psychologists and computer scientists in the 1940s and 50s (Hebb, 1949; Rosenblatt, 1958). Similar to neurons in a biological neural system, the input neurons in an ANN have the potential to be activated by stimuli (values of input variables). Activated input neurons will transmit a signal that will propagate a hierarchical neuronal system with multiple layers. Signals from input neurons will be summarized by neurons in a higher hierarchical layer (often referred to as a hidden layer in ANN terminology), which, upon activation, will pass the signal to the next layer. This process will repeat through a number of hidden layers in the network until the signal reaches a final layer of output neurons. The classification is given according to the type of output neuron that was activated. Compared with decision trees, ANNs are more proficient in handling nonlinearity in the data; they are able to approximate any hypothetical function between the input and the outcome values. As a trade-off, the ANN cannot provide insight on the form of the function that it was trained to approximate. In other words, ANNs can be used to make good predictions, but cannot show the user any simple link between the predictors and the outcome, or even which predictor is important; hence, they are sometimes referred to a black-box model. Because of their wide adaptation in industrial and medical settings, ANNs often are considered an important performance benchmark for evaluating other machine-learning models.

SVM is another powerful machine-learning classifier that may offer performance on par, if not better, than ANNs while being less complex. SVM algorithms combines a mathematical

technique called kernelling and the simple idea of finding a straight decision line that separates two categories of data (e.g., truthful versus deceptive) with the best margin between data points nearest to the line. When the data appear to be not linearly separable, these algorithms expand the existing feature space (the space with the number of dimensions equal to the number of observed predictor variables) to a higher dimension using a kernel function, then they find the optimal separation hyperplane (i.e., a multidimensional generalization of a straight line; Cortes & Vapnik, 1995). Due to its efficiency in handling long data sets that are typical to NLP research, the SVM is a frequently used analytic strategy in NLP-based deception studies (e.g., Ott et al., 2011; Ott et al., 2013; Xu & Zhao, 2012).

Due to their flexibility and superior performance when handling massively multivariate data, machine-learning classifiers are also the tool-of-choice for deception researchers who attempted to build a multi-dimensional deception detection model that utilizes several categories of behavioral data. For instance, Krishnamurthy, Majumder, Poria, and Cambria (2018) analyzed a set of recordings of courtroom testimonies (60 truthful, 61 deceptive) using a neural network that combined four types of inputs, including video and audio signals, transcribed texts, and FACS-coded micro-expressions. Their multimodal model achieved an impressive 96.14% classification accuracy. Aside from providing the opportunity to build highly complex deception detection models, machine-learning also promotes the use of cross-validation, which could be the solution for the issue of poor generalizability in traditional deception detection research (Levine, 2018).

1.4.3 Examples of NLP-Based Lie Detection Research

In this section we will summarize a few examples of text-based deception detection studies that employed automatic feature extraction and machine-learning classifiers. These

studies provide an overview of different methodologies and diverse classification results found in NLP-based lie detection research. The average accuracies of the best performing classifier in each of these studies are reported in Table 1. All reported classification results in Table 1 are measured by out-sample accuracy (classification accuracy of the testing sample).

An early laboratory study conducted by Zhou and colleagues (2004b) [Table 1: Zhou2004] focused on theoretically motivated features exclusively. They compared the performances of four different classifiers, which included logistic regression, discriminant analysis, decision tree, and neural network. To generate the text corpus, Zhou and colleagues recruited students from a management information systems course to participate in a decision-making task based on the desert survival problem designed by Lafferty and Eady (1974). During the task, participants were put into pairs to discuss the importance of 12 items after a hypothetical jeep crash in the Kuwaiti desert with the goal of reaching a consensus ranking of these items. Two experimental conditions were randomly assigned, one of which instructed the participants to lie to their partners during the task. All discussions took place via email messages; linguistic features were extracted using GATE. Zhou and colleagues tested the classification models' ability to discriminate truthful and deceptive messages as well as truthful and deceptive individuals. Individual discrimination had better accuracies than message discrimination. In terms of message discrimination, different classification models gave greatly varied performances, some had an accuracy rate no different from chance, and some achieved accuracy rates as high as 80%. The best performing classifier was the ANN, which had an 80% accuracy for discriminating truthful and deceptive messages and an 89% accuracy for discriminating truthful and deceptive individuals.

Fuller and colleagues (2009) [Table 1: Fuller2009] applied a similar classification strategy to a real-world forensic text sample of 366 "person of interest statements" obtained from law enforcement agencies. These statements were originally handwritten by the "people of interest." They were transcribed to electronic texts by the researchers. Among these statements, 79 were contradicted by credible evidence and thus were labeled deceptive. Fuller and colleagues tested various combinations of GATE and LIWC coded features, and compared the performances of three classification methods, which were logistic regression, decision tree, and ANN. The classification models were trained and tested using sub-samples that contained equal numbers of truthful and deceptive statements, so that chance-level performance was kept at 50%. Overall, different combinations of classification methods and feature sets tested by Fuller and colleagues had very similar performances. The classification accuracies ranged between 67% to 74%; ANN gave the best performances. The highest classification accuracy was achieved using the ANN and a constructed feature set that contained only 12 theoretically motivated features among the 31 GATE and LIWC features that were included in this study.

Mihalcea and Strapparava (2009) [Table 1: M&S2009] employed a lexicon-based classification strategy and reported a similar level of performance as the two previously mentioned studies. They conducted an online study that instructed participants to provide truthful and deceptive opinion statements about three topics, which were abortion, death penalty, and best friend. One hundred truthful statements and 100 deceptive statements for each topic were obtained. Using Naïve Bayes and SVM classifiers with lexicon unigram as input, this study reported slightly above 70% average classification accuracy when statements about different topics were analyzed separately. Classification accuracies were much better on the personal topic - best friend, than on the two political topics (76% vs. 68% on average). Mihalcea

and Strapparava also reported, on average, about 59% cross-topic classification accuracy when two of the three topics were included in the training samples, and the third topic was used as the testing sample. Notably, when the best friend topic was used as the testing sample and the training sample consisted of only political topics, the classification accuracies were the worst (56% on average). These results appear to highlight the contextual dependency of lexicon n-gram-based deception detection models.

Some studies have attempted to improve lie detection accuracy by including a more comprehensive list of features. Ott and colleagues conducted two studies (2011; 2013) that had almost identical designs. In both studies, 400 genuine online reviews of 20 hotels and 400 elicited fake reviews of the same selection of hotels were collected and analyzed. The 2011 study collected only positive reviews, while the 2013 study collected only negative reviews. Ott and colleague employed all three levels of linguistic cues - psycholinguistic, lexicon n-grams, and POS, as well as their various combinations. Naïve Bayes and SVM classifiers were used in these studies. SVM had the better performance between the two types of classifiers, but the differences in accuracies was only about 1% in each case. When used alone, the LIWC-coded psycholinguistic feature set and the POS-tags set achieved classification accuracies in the 70% to 80% range, which was much lower than the close to 90% accuracy reached by lexicon unigrams. Adding psycholinguistic features to the already highly accurate lexicon n-gram models further increased the accuracy rates by roughly 1-2%. Overall, classification results for positive and negative reviews were very similar. The 2013 study [Table 1: Ott2013] reported classification results of the combined sample; the average accuracy of both positive and negative reviews was 87%. The same study also reported classification results when one type of review was used as the training sample and the other type was used as the testing sample. The classification

accuracy of positive reviews was 81% when negative reviews were used to train the model; in the opposite case, the accuracy rate of negative reviews was 75%. It is worth mentioning that human judges were able to correctly classify as much as 65% of the reviews in Ott and colleagues' corpora, indicating that it was a relatively easy classification task.

Xu and Zhao (2012) [Table 1: X&Z2012] reanalyzed Ott and colleagues' 2011 positive review corpus. They used a maximum entropy model (MEM) for the classification task, and added a more comprehensive selection of grammatical features, including POS bigrams, POS trigrams, and dependency parses. Three of their models achieved above 90% classification accuracies. The inputs of these models were lexicon unigrams plus dependency parses (accuracy = 91.6%), lexicon unigrams plus LIWC features (accuracy = 91.4%), and POS unigrams plus POS bigrams (accuracy = 90.5%). Aside from the marginally better classification accuracies compared with the results reported by Ott and colleagues, Xu and Zhao showed that POS unigrams and POS bigrams performed much better as a combined features set than individually: The classification accuracies of the POS unigram model and the POS bigram model were only 74% and 77% respectively. Furthermore, Xu and Zhao also applied their classification models to a Chinese corpus that was collected from a photography forum. This corpus contained 900 genuine reviews of photographs and 900 spam posts. The average classification accuracy of Xu and Zhao's various models on this Chinese corpus was 79%.

Unlike previously mentioned studies, which all used entire statements as the units of analysis, a study conducted by Fornaciari and Poesio (2013) [Table 1: F&P2013] analyzed a real-world corpus of Spanish court hearings using utterances as the analysis units. Utterances are defined as strings of text separated by punctuation marks. The corpus was comprised of 35 transcribed court hearings, which consisted 3015 utterances. Nine hundred and forty-five of

these utterances were labeled as false, 1202 were labeled as true, and the remaining 868 as uncertain. Fornaciari and Poesio employed the SVM classifier and two different sets of features as predictors. One of these feature sets was the LIWC psycholinguistic feature set, the other was a "surface feature" set, which consisted of a selection of lexicon n-grams and POS n-grams (up to 7-gram) based on the frequencies of their appearances in the data. Although there were three classes of utterances, the classifications were binary. The uncertain utterances were either removed or grouped with the truthful ones in different analyses. The cross-validation was based on different hearings, that is the training and testing sample always consisted of different hearings. Given the unbalanced data, depending on whether uncertain utterances were included or excluded, chance-level performances were estimated to be 60% and 55% respectively. The reported classification accuracies ranged between 68% to 72%. Fornaciari and Poesio also found that the number of hearings used to train the model only made small differences in classification accuracies and the weights of individual features. Models trained using data from just one hearing was able to achieve 66% accuracy. Based on this observation, Fornaciari and Poesio argued that false testimonies given at court hearings contain stereotyped deceptive language, probably due to high cognitive load.

The sample of studies discussed above showed that the choice of classifiers and input features have a relatively small impact on lie detection accuracy compared to the type of the language sample. Different categories of language features seemed to perform differently in different situations. We will discuss these points in greater detail in the following sections. Moreover, complex classification models that utilize a wider range of features appeared to have an advantage over simpler models, but adding more features to a classification model as a means of improving performance has diminishing returns.

Table 1. Results from Selected NLP-Based Lie Detection Studies

Study	Best Classifier	Feature type	Average Performances		Sample Type	Language
			Random	Context		
F&P2013	SVM	Various Combinations	/	70%*	Transcribed Speech Forensic	Italian
Fuller2009	ANN	Psycholinguistic	73%	/	Handwritten Forensic	English
M&S2009	NB	N-grams	71%	60%	CMC Opinions	English
Ott2013	SVM	Various Combinations	87%	78%	CMC Hotel review	English
X&Z2012[1]	MEM	Various Combinations	86%	/	CMC Hotel review	English
X&Z2012[2]	MEM	Various Combinations	79%	/	CMC Forum posts	Chinese
Zhou2004	ANN	Psycholinguistic	75%	/	CMC Interpersonal	English

Note: Studies were arranged alphabetically.
Average performances were calculated based on the best classifier within each study. Results from randomized cross-validation and context-based cross-validation were separated.
X&Z2012 reported results from two sets of data: [1.] Ott et al.'s (2011) positive hotel reviews; [2.] Chinese photography reviews.
Due to unbalanced sample, F&P2013 had a chance-level performance benchmark that was not 50%. The result of this study was marked with *.

1.4.4 Reasoning for Repeated-Measures in NLP-Based Lie Detection

Deception detection studies conducted by the NLP research community have demonstrated that the combination of automatic feature extraction and machine-learning classification models can be used to effectively separate deceptive statements from truthful ones (e.g., Ott et al., 2011; Ott et al., 2013; Xu & Zhao, 2012). Currently, most NLP-based deception

detection studies were conducted to solve specific online deception problems, such as filtering scam emails and opinion spams (e.g., fake hotel reviews). These studies have generated several text corpora containing truthful and deceptive statements written by a large number of individuals over the Internet. In order to control confounding variables, the content and context of the statements in these data sets were highly restrictive. For instance, the hotel review corpus collected by Ott and colleagues (2011) only included reviews of 20 selected hotels in Chicago. Although deception detection models built based on these data sets can reach close to 90% in accuracy, this accuracy can be partially attributed to the systems' ability to utilize situational cues (e.g., the phrase "fish tank" will have a higher probability of appearing in truthful statements about a hotel that has a fish tank in the lobby). Therefore, different deception detection scenarios will require different strategies for controlling variability in the data. In the case of forensic-oriented lie detection research, obtaining large samples of individuals is difficult; in field studies, situational cues cannot be relied upon, since each instance of deception will likely be about a different event. Nevertheless, it is sometimes possible to collect multiple language samples of truth-telling and lying from the same individual. In such a case, the ideal model should utilize consistencies in an individual's language style, and recognize changes in language pattern between truth-telling and lying that are unique to the individual. So far, no deception detection study has explored this strategy.

There are compelling reasons why individual language style may be an effective control of variability in language data. It is a common belief among linguists that humans acquire language based on examples of its usage, and due to the uniqueness of the set of examples that each individual was exposed to while learning a language, his/her language style is also unique (e.g., Chomsky, 1999; Pinker, 2002). Research that utilize computerized language analyses has

provided ample evidence for individual differences in language use (Pennebaker & Graybeal, 2001; Pennebaker & King, 1999; Pennebaker et al., 2003). The uniqueness of individual language style can manifest in various analytic dimensions, such as the distributions of n-grams and POSs (e.g., Stamatatos, 2009; Van Halteren et al., 2005). Using a large collection of linguistic features, computerized language analyses can accurately attribute writing samples to their correct author. For instance, an authorship-attribution study conducted by Van Halteren and colleagues (2005) was able to correctly attribute 97.8% of the sampled texts from eight different authors.

Moreover, language use can also reflect an individual's personality and cognitive abilities (Dewaele & Furnham, 1999; Pennebaker & King, 1999). Research has demonstrated that n-gram and psycholinguistic features extracted from language samples can be used to identify an individual's personality traits (e.g., extraversion and psychopathy, Gill & Oberlander, 2002; Hancock, Woodworth, & Porter, 2013; Mairesse & Walker, 2006), or detect cognitive impairment caused by developmental disorders and neurodegenerative diseases (Jarrold et al., 2010; Orimaye, Wong, Golden, Wong, & Soyiri,., 2017). These individual difference variables not only affect how individuals normally use language, but may also influence how their language patterns change during deception (e.g., Hancock et al., 2013; Vrij, 2008; Vrij et al., 2004; Vrij et al., 2001). By collecting multiple truthful and deceptive statements from the same individual, a mixed-effects model can be built to account for effects of the individual on deceptive language.

1.4.5 Mixed-Effects Modeling

Mixed-effects modeling (a.k.a. mixed modeling or multilevel modeling) is the most common method of modeling repeated-measure data in regression-type models. It is

distinguished from mixed-design in analysis of variance (ANOVA), which means the mixing of between-subject and within-subject independent variables. In the context of repeated-measure linear regression, mixed-effects models divide the effect of each predictor into a fixed-effect and a random-effect (Snijders & Bosker, 2012). A fixed-effect is a consistent association between the predictor and the outcome that does not depend on the individual. It is represented in a regression model by a fixed-effect parameter that does not vary across the sample. For instance, in a simple linear regression, both the intercept (β_0) and the slope (β_1) parameters are fixed-effect parameters. The model written in mathematical form as the following:

$$Y = \beta_0 + \beta_1 X$$

In contrast, a random-effect is a variable association between the predictor and the outcome that depends on the individual. It manifests in a repeated-measure regression model as a series of parameters, each of which is associated with a specific individual. The following is an example of mixed-effect linear regression:

$$Y = \beta_0 + \beta_{0j} + (\beta_1 + \beta_{1j})X_j$$

Comparing it with the simple linear regression, here we added random-effects β_{0j} and β_{1j} to the intercept and the slope respectively. The subscript j is used to index different individuals in the dataset. Consider a repeated-measure dataset that consists 20 different individuals, the resulted mixed-effects model will use 21 parameter values to represent the relationship between the predictor and the outcome variable. One of these is the fixed-slope parameter (β_1), which represents an overall effect that is consistent across different individuals. The rest are 20 random-effect parameters ($\beta_{1.1}$ to $\beta_{1.20}$), representing 20 unique effects of the 20 individuals in the sample.

Mixed-effects modeling is available for both traditional inferential statistic and machine-learning; however, in the case of machine-learning classification, because most its application is concerned with out-of-sample validity, repeated-measures models have received little attention. The next chapter describes an NLP-based deception detection experiment that has a repeated-measures design. The data generated by this experiment were used to explore the utility of mixed-effects modeling in deception detection tasks using classical regression models as well as machine-learning classifiers.

Chapter 2: The Current Research

2.1 Overview

The current research is motivated by both the psychological science of deception detection and the computational literature that confronts the same problem. Analytic techniques developed by NLP research and data science have shown an enormous potential for text-based deception detection (e.g., Fitzpatrick et al., 2015; Heydari et al., 2015). A key task for advancing this line of research is to collect suitable deceptive language data upon which models of deception detection can be built (Gokhman, Hancock, Prabhu, Ott, & Cardie, 2012). Existing text corpora for training deception detection models consist of mostly between-subjects data sets, where true and false statements were provided by different individuals. Psychological research has shown that deceptive behaviors can be highly individualized, specifically in the usage of language (e.g., Chomsky, 1999; Hancock et al., 2013; Vrij, 2008). Individuals' personality, experience, and cognitive abilities affect how well they construct both true and false statements (Santtila et al., 1999; Vrij, 2005; Vrij et al., 2004). Therefore, it is important to examine the impact of modelling individual effect on the accuracy of deception detection systems. To this end, the current research collected a repeated-measures text corpus.

Deception detection studies conducted by the NLP research community tend to be atheoretical and focus only on the performance of the systems. Nevertheless, the linguistic differences between truth telling and deception in these studies can be explained under the reality-monitoring framework (Fitzpatrick et al., 2015). The current research elicited true and false statements using a similar design as some classical RM studies (e.g., Alonso-Quecuty, 1992; DeCicco & Schafer, 2015; Hemandez-Fernaud & Alonso-Quecuty, 1997), where video

recordings were used to create real memory. In this research, participants were asked to watch short videos depicting memorable real-life events. Immediately after viewing each video, they provided a truthful and a false statement about the video. To improve the generalizability of the current research to real-life applications, which mostly involve detecting lies about negative events such as crime, all four selected videos have a negative emotional tone and depict minor violence.

Similar to most NLP-based deception studies, the current research collected text data over the Internet. The Internet provides an opportunity to collect large amounts of data quickly and is more convenient for the participants. But online data collect makes controlling confounding variables, such as motivation, a challenge. Nevertheless, deception theories that motivated the RM approach suggest that some linguistic deception cues are not related to motivation (Johnson & Raye, 1998). The motivational enhancement effect of deception in CMC environments also predicts that lies told by motivated liars over the Internet will be less distinguishable from truth (Woodworth et al., 2005). In fact, studies that have produced the most accurate classifications of true versus false statements using online data generally provided little incentive for lying (Ott et al., 2011; Ott et al., 2013; Xu & Zhao, 2012). The current research also did not attempt to manipulate the participants' incentive for lying in a significant way. All false statements were produced in a low motivation setting, where the participants were told that they can win a prize if their stories are judge by the researcher to be the most convincing. In addition, to screen for over preparation, the time participants took to complete each statement was recorded.

The current research used three categories of linguistic cues – lexical cues (n-grams), POS, and theoretically motivated psycholinguistic cues – as input of our deception detection

models. Lexicon n-grams and POS tags are two of the most commonly used feature types in NLP research. Psycholinguistics cues are typically used in psychological studies of deception. These three categories of cues encode different domains of information: Lexical cues provide highly detailed and concrete information about language usage; POS tags provide basic grammatic information; psycholinguistic cues provide psychologically meaningful abstract information. By comparing the effectiveness of these three categories of cues as the inputs of our deception models, we were able to discern the most important domain of information for language-based deception detection tasks.

The main focus of the current research was to evaluate performance gains of various classification models from utilizing a repeated-measure data structure. This was accomplished mainly through mixed-effects modeling. Regular classification models do not allow the effects of the input variables to vary depending on the individual, in other words they are fixed-effects models. Even though we have a repeated-measure dataset, fixed-effects models would treat all input variables as between-subject variables, and thus provide the same classification performance as if the dataset was a between-subject dataset. Using mixed-effects modeling we were able to add individual-level random-effects to our classification models, so that the repeated-measure data structure would be correctly represented in the models. By comparing the performances of the original fixed-effects classifier with the performances of their mixed-effects variations, we were able to show the usefulness of individual language pattern in deception detection tasks and argue for the use of the more accurate modeling approach whenever repeated-measure data is available.

2.2 Method

The following sections detail of the experimental procedure and data analysis strategy of the current research.

2.2.1 Participant Recruitment

The current research used a convenient student sample. All participants were recruited through the SONA system at the University of British Columbia Okanagan campus. All participants were enrolled in a psychology course at the time of their participation and received two bonus credits toward their course as compensation. In addition, to motivate the participants to provide higher quality responses to the writing tasks, a $100 gift card was promised as a reward for the participant who produced the best responses (see Appendix B.1: General Instruction for detail). This was not an incentive for only deception, but a measure to increase participant engagement in all writing tasks. The judgement was made subjectively based on diligence and the appearance of truthfulness (apply to both truthful and deceptive statements).

In order to participate in the current experiment, the individual must have been at least 19 years old, a native English speaker, and had normal vision and hearing. To enforce these exclusion criteria, a brief qualification questionnaire was given to the participants before they started the experiment (Appendix A.1). Disqualified individuals were directed to exit the study.

2.2.2 Data Collection

Data collection was accomplished through an online survey created using the UBC survey tool. The survey asked each participant to provide four pairs of truthful and fabricated written statements (eight statements in total). Four short videos (2-3 min in length) depicting real-life events were selected from the website Liveleak.com. Liveleak.com is a video-sharing site that promote citizen journalism, where footages of numerous real-life events, such as violent

crime, can be found. Participants were directed to watch the selected videos in a random order; immediately after watching each video, they were instructed to write a truthful and a false statement about the video also in random order. Each statement had a suggested length of 150 to 300 words. It was expected that the experiment could take participants roughly 2 hr to complete. The survey tool allowed research participants to save their progress in the middle of a survey and resume at a later time. Due to the length of the experiment, participants were encouraged to complete the writing tasks in separate sessions to reduce the influence of fatigue on the writings (See Appendix B.1 for the general study instructions). In accordance with the recommendations given by Gokhman and colleagues (2012) for collecting deception data over the Internet, the researcher and a research assistance reviewed each collected statements to complete two tasks: (a) Clean the texts for computer analyses and (b) screen for low quality responses. Statements that are under 100 words were considered too short. We removed cases that contained four or more statements that are too short. In order to obtain a balanced dataset, all retained cases must consist all eight statements. Cases that contained any statements that are off-topic or written without following the study instructions (described in the next section) were also removed.

2.2.3 Materials

2.2.3.1 Writing Tasks

The four publicly available videos each depicted a memorable real event and were used to create four pairs of truthful and four pairs of deceptive writing tasks. The videos were downloaded and re-uploaded to a Youtube channel as private videos, which meant they would not appear in Internet search results and were only accessible through a link distributed by the

uploader of the video. Original titles of the videos were replaced by generic names (e.g., experiment video A, B, C, and D). All four videos were 2 to 3 min in length. Each writing task was accompanied by a specific instruction that was designed to control potentially unwanted variabilities in the statements. Based on the deception literature one would expect, given the choice, participants would take a first-person perspective more often in the truthful writing tasks than in the deceptive writing tasks (Vrij et al., 2007). This effect was treated as a confounding factor in the current research. Our instructions forced a specific perspective for each video: For video A and D, participants were instructed to take a third-person perspective; for video C and B, they were instructed to take a first-person perspective. To reduce noise variance caused by varied interpretations of the videos (for instance, who is at fault in video D) and to obtain a more uniform set of responses, sometimes additional details about the events and suggestions about the scope of the lie were also given. The contents of the four videos and specific instructions to the participants are briefly described below. To view the full instructions that appeared on the data collection survey please see Appendix B.2.

Video A is a security camera footage of a robbery. It shows a male robber attacking a lone female clerk at a cellphone store and forcing her to open the cash register. Participants were asked to assume the perspective of a witness. They provided a truthful statement that accurately described the details of the incident, and a deceptive account of the robber's action that minimized his crime.

Video B is a footage shot from a head-mounted camera that captured the first-person view of a motorcyclist during a road rage incident. The video shows the motorcyclist walking onto a bus and confronting the bus driver aggressively, after the bus had cut into his lane at an intersection. Participants were asked to assume the perspective of the motorcyclist when

completing the writing tasks. In the truthful statement, they described the incident accurately based on the video; in the deceptive statement, they were instructed lie to justify the motorcyclist's action, and minimize any wrongdoing.

Video C is a home security camera footage that shows a woman trying to steal a parcel from the front of a house, but she was caught by the homeowner. The homeowner tackled and restrained the thief. To complete the writing tasks, participants were asked to assume the perspective of the thief. In the truthful statement, they told a story that accurately reflected the camera footage. In the deceptive statement, they were instructed to lie to conceal the crime and exaggerate the homeowner's aggression.

Video D is a cell phone camera footage that shows an altercation between a security guard and a customer at a grocery store. Suspecting the customer had stolen something, the security guard asked the customer to show him the receipt for his purchases, but the customer refused to comply. Participants were asked to assume the perspective of the person who recorded the footage. For the truthful statement, they were instructed to describe what they saw in the video as accurately as they could; in the false statement, they were instructed to lie about the situation in favor of the customer and paint the security guard in a bad light.

2.2.3.2 Demographic and Study Qualification Questionnaire

The qualification questionnaire (see Appendix A.1) appeared on the first page of the online survey. It included the aforementioned qualification criteria. After being qualified, participants were instructed to carefully read and sign the informed consent form (Appendix C), which was followed by a brief demographic questionnaire that asked about the participants' gender and ethnicity (Appendix A.2). Disqualified participants were directed to an exit page

explaining the qualification criteria of the study and thanking them for taking an interest in the experiment. After completing the writing tasks, participants were given the option to provide their email address in order to receive follow-up communications.

2.2.4 Language Feature Extraction

Our psycholinguistic feature set was extracted using the most current version of LIWC (LIWC2015). The LIWC is representative of a traditional manually-designed way of coding abstract information, while the NLP literature offer a plethora of automatic approaches for condensing information at a sentence or document level. For instance, rather than using word counts within predetermined categories to represent the emotional tone of a statement, machine-learned sentence or document vectors can be used to represent such information. Research showed that automatically induced dense vectors are useful in sentiment analysis models, and potentially better than LIWC features for opinion-based deception detection tasks (e.g., Ren & Ji, 2017). Although the deception detection task presented by the current research is fact-based, it may still benefit from a greater level of data-driven automation. Nevertheless, using such an approach would greatly complicate our model building process and distract from the thesis of the current dissertation.

LIWC2015 offers a list of 93 features, among which 38 were included. Seven of the selected features are summary features (e.g., word count, analytical thinking, and emotional tone); 31 have specific psychological references (e.g., positive emotion, anxiety, causation, and achievement). The selection was made based on relevance to our corpus, and to avoid redundancy with other feature sets used in the current research. For instance, LIWC2015 captures grammatical features and punctuations, which are essentially POS features. In addition, LIWC2015's psychological process features are organized hierarchically. Higher-order

categories, such as negative emotion, encapsulate lower-order features such as anger and anxiety. For categories that are more relevant to the current corpus, including affect, cognition, perception, drive, time orientation, and personal concern, the lower-order features were included to provide more information; for categories that are less relevant, including social/relationship, biological processes, and informal words, the higher-order categories were used instead to reduce the number of model parameters that needed to be estimated. For a complete feature list, please see Appendix D.

Five lexicon-n-gram feature sets were created using the Python package NLTK (Loper & Bird, 2002) and custom code. These sets were unigram, bigram, trigram, bigram+, and trigram+ (+ sign indicates the set includes preceding n-gram set/s). N-gram sets retained the top 300 most frequent n-grams in the sample or all n-grams that appeared at least twice, whichever was less; n-gram+ sets retained the top 500 most frequent n-grams. It is a common practice in NLP research to remove "stop words" from n-gram features. "Stop words" are the most common words in a language that have little meaning by themselves, for example "is," "the," and "that" to name a few (Luhn, 1966). In the current research, no "stop words" were removed from n-gram feature sets because we were concerned about individuals' language habits in relation to deception. Any word, regardless of meaning, can be a potential deception cue.

Three POS feature sets (POS tag, POS bigram, and POS bigram+) were generated also using Python and NLTK. The POS tags set included frequencies of all POS labels in the NLTK universal tags set, such as singular noun (NN), past tense verb (VBD), and pre-determiner (PDT). For the full list of POS tags in the NLTK universal tags set please see Appendix E. POS bigrams are simply bigrams of POS tags. Excluding punctuations, the NLTK universal tags set includes 35 POS tags; that means that there are potentially 1190 POS bigrams (Permutations:

$_{35}P_2$). POS bigrams+ combined both POS unigrams and bigrams and was the largest single feature set used by the current research.

2.2.5 Classification Methods

The current research employed three classification methods, including one traditional statistical method and two types of machine-learning methods; these were generalized regression, classification tree/forest, and artificial neural network. These methods were chosen because they have well documented mixed-effects implementations for repeated-measure or clustered data. Details of the use of these methods in the current research are discussed below.

2.2.5.1 Logistic Regression

The current research presented a classification problem with binary outcomes (truth vs. fabrication). A commonly used traditional statistical classifier is the logistic regression. The data for the current research were generated by a repeated-measures experiment; multilevel/mixed-effects modelling is required to accurately represent this data generation process. The logistic mixed-effects regression is a special case of generalized linear mixed-effects model (GLMM) as defined by Goldstein (1991). Estimation for GLMMs based on least square and maximum likelihood are highly sensitive to violations of the normality assumption (Fotouhi, 2003). Currently, this type of model is most commonly estimated using the expectation maximization method.

Furthermore, regular GLMMs cannot include more model parameters than the number of data points in a dataset. The number of random slope parameters in a mixed-effects model increases by a factor of the number of groups/clusters (in our case, the number of individuals). In the current dataset, each individual contributes to only eight data points, which means fewer than eight predictors can be allowed to have random slope parameters. In order to allow the

GLMM to account for all possible random-effects in even our smallest feature sets (i.e., the psycholinguistic set and the POS unigram set), a parameter elimination strategy needed to be employed. LASSO regression (Least Absolute Shrinkage and Selection Operator; Efron, Hatie, Johnstone, & Tibshirani, 2004; Tibshirani, 1996) was used to fill this role in our analysis. LASSO is a well-documented popular automated variable selection strategy for regression. It reduces the number of parameters in a regression model by shrinking small parameter values to zero, effectively eliminating unimportant parameters. This method not only can be used as an empirical feature selection process for the regular logistic regression, but it also enables us to include random-effects for all the features in the mixed-effects model. Our regular regression models (without variable selection and mixed-effects modeling) were fitted using the R package "lme4" (Bates, Maechler, Bolker, & Walker, 2015). LASSO logistic regression and mixed-effects logistic regression were fitted using the R package "glmmLasso" (Groll, 2017).

In the current research, a regular fixed-effects logistic regression model that uses only the psycholinguistic feature set as input was identified as our performance baseline, with which all other models were compared. This is because this model is one of the most parsimonious (least complex) models among all of our models, and it is representative of the typical deception detection studies found in the psychological literature, which tend to focus on theoretically motivated deception cues.

It is important to note that in other context logistic regression was also adopted as a machine-learning model. The difference is that, as a machine-learning model, the parameters of the model are estimated using cost functions instead of frequentist estimation methods such as maximum likelihood and least squares. Various optimization routines and variable selection schemes are often integrated within the model. In the current dissertation, the use of the term

"logistic regression" always refers to the frequentist statistical model instead of its machine-learning counterpart.

2.2.5.2 Decision Tree and Random Forest

Although some machine-learning classifiers, such as the decision tree and random forest, can handle repeated-measures data by simply treating the individual as an input variable, modifications need to be made for the models to fully take advantage of the hierarchical data structure. However, adapting classification algorithm for longitudinal/repeated-measures data analysis is currently a developing area of research, where few packaged software solutions are available.

The Random-Effect Expectation Maximization (RE-EM) Trees (Sela & Simonoff, 2012) and the Mixed-effects Random Forest (MERF, Hajjem, Bellavance, & Larocque, 2014) are two variations of the decision-tree/forest algorithm that incorporated mixed-effects modeling. Nevertheless, these two algorithms were designed for continuous outcome variables. The Binary Mixed Model (BiMM) tree and its random forest counterpart (Speiser et al., 2019) are a more recent development in this area of research. They were specifically designed to solve binary classification problems. The R package for these algorithms is still under development, but an R implementation is available via request from the author. Two other versions of mixed-effects classification tree/forest also exist (Glmertree, Fokkema, Smits, Zeileis, Hothorn, & Kelderman, 2018; GMERF, Hajjem, Larocque, and Bellavance, 2017), but their current implementations only allow a small number of random-effects to be included in the model; therefore they are not suited for the current research. The current research compared the performance of the regular random forest algorithm with the BiMM tree and forest algorithms. The performance of all

tree/forest models were also compared with the baseline model. The R package "randomForest" (Liaw & Wiener, 2002) was used to fit the regular random forest model.

2.2.5.3 ANN

For ANNs, no specific modification is available or necessary to handle repeated-measures data; however, different architecture perform differently for different data structures. It is sensible to simply include participant ID as a categorical predictor. Nevertheless, available research suggested that feedforward ANN with multi-layered perceptron (MLP) using a back-propagation algorithm, where individual identifiers and predictor variables are connected to different hidden layers, is probably most suited for repeated-measures data (Hallner & Hasenbring, 2004; Maity & Pal, 2013).

In the current research, an ANN with four hidden layers and no individual identifier input was used to represent a fixed-effects model. A second ANN with the architecture described above was created to represent a mixed-effects model. Individual identifiers in the second model was a $N \times N$ matrix of binary cells, N being the number of individuals in the sample; each column represents a single individual. These inputs were connected to the second hidden layer of the four-layer ANN. Hidden layers in both ANNs used a rectified linear unit activation function; their output layers used a sigmoid activation function. The loss function that we used to estimate the model was the binary cross-entropy/log loss function (see Murphy, 2012). The Python package "keras" (Chollet, 2015) was used to create both ANNs. We compared the performance of the "fixed-effects" ANN and the performance of the "mixed-effects" ANN to test the main thesis of the current research. Both ANNs were compared with the baseline fixed-effects logistic regression model.

2.2.5.4 Cross-Validations

Performance of all classification models used in the current research were evaluated through cross-validation of their predictions. Two cross-validation schemes were used: A conventional 10-fold cross-validation, and a cross-validation scheme based on the content of the texts. In a conventional *n*-fold cross-validation scheme, the sample is divided into *n* subsamples randomly. Each of these subsamples takes a turn to be the testing sample while the model is estimated using the rest of data. Using this randomized cross-validation scheme for our data meant that statements about each of the four videos had an equal chance to be represented in the training set; thus, the testing set would not provide cross-context validation. A major point of training a classification model using repeated-measure data is that individual-level information can help generalize the model to a wider context; therefore, a cross-context validation scheme was needed.

In the second cross-validation scheme, every time a classifier was estimated, one set of truthful and deceptive statements about a same video were left out as the testing sample. Each of the four sets of statements rotated to be the testing sample. This is a four-fold cross-context validation scheme, where the classifier was learned through statements about three videos, then tested on statements about a video to which it had no exposure.

Both cross-validation schemes were used for fixed-effects only classifiers. It was expected that these classifiers would perform better under the 10-fold random sampling scheme than under the cross-context scheme. Only the cross-context scheme was used for the mixed-effects models, because random sampling would cause each individual to be unevenly represented in the training and testing samples, making estimations of the random-effects unstable. Since we designed the data collection to generate a balanced dataset (containing the

same number of truthful and deceptive statements) and had no preference for false positives or false negatives, performances of the models were measured using accuracy (instead of precision, recall, or F1-score) and area under the receiver operating characteristic curve (*ROC-AUC*; DeLong, DeLong, & Clarke-Pearson, 1988).

2.3 Results

2.3.1 The Sample

A total of 228 individuals responded to the survey, out of which, 198 qualified to participate and provided text data. After reviewing the text data, the researcher removed 46 participants' responses according to standards described in section 2.2.2. Among the 152 participants whose data were retained, 110 (72.3%) were female, 42 (27.7%) were male; 103 (67.8%) were native English speakers, 49 (33.2%) were multilingual but speak English as a primary language. The average age of the sample was 20.89 years. All participants in the retained sample completed all eight writing tasks; thus, the final text sample is consisted of 608 truthful and 608 deceptive statements. The average word count of the statements was 180 (*SD* = 57.89).

2.3.2 Logistic Regressions

2.3.2.1 Regular Logistic Regression

Regular logistic regressions without random-effects (fixed-effects only) were performed on the psycholinguistic feature set, POS tags set, and the combination of these two sets. Fixed-effects logistic regression trained using the psycholinguistic set and the 10-fold randomized cross-validation had an in-sample prediction accuracy of .706 (*ROC-AUC* = .768). The same model trained using the four-fold cross-context validation had an in-sample accuracy of .720 (*ROC-AUC* = .783). Various iterations of the model had an average Akaike information criterion

(*AIC*) of 1076.9. Out-sample prediction results are reported in Table 2. Statistically significant predictors included analytic ($z\beta$ = -4.60, $p < .001$), words per sentence ($z\beta$ = 3.45, $p < .001$), anxiety ($z\beta$ = 2.147, $p = .03$), anger ($z\beta$ = -2.97, $p < .01$), cause ($z\beta$ =2.67, $p < .01$), risk ($z\beta$ = 2.40, $p = 0.02$), future focus ($z\beta$ = -2.29, $p = .02$), motion ($z\beta$ = -2.97, $p < .01$), space ($z\beta$ = -2.27, p = .02), home ($z\beta$ = 3.06, $p < .01$), and money ($z\beta$ = 6.29, $p < .01$).

Regular logistic regression trained using the POS tags set and the 10-fold randomized cross-validation had an in-sample prediction accuracy of .666 (*ROC-AUC* .720). The same model trained using the cross-context validation had an in-sample accuracy of .675 (*ROC-AUC* = .734). Various iterations of the model had an average *AIC* of 1223. Out-sample prediction results of these models are reported in Table 2. Statistically significant predictors included coordinating conjunction ($z\beta$ = -3.04, $p < .01$), preposition/subordinating conjunction ($z\beta$ = -2.58, $p < .01$), plural noun ($z\beta$ = 2.34, $p = .02$), personal pronoun ($z\beta$ = 4.31, $p < .01$), adverb ($z\beta$ = 3.08, $p < .01$), particle ($z\beta$ = -3.05, p < .01), past participle verb ($z\beta$ = 4.12, $p < .01$), and present non-third person verb ($z\beta$ = -2.41, $p = .02$).

Regular logistic regression trained using the combined psycholinguistic and POS feature set and the10-fold randomized cross-validation had an in-sample accuracy of .728 (*ROC-AUC* = .802). The same model trained using the cross-context validation had an in-sample accuracy of .721 (*ROC-AUC* =.809). Various iterations of the model had an average *AIC* of 1164. Significant predictors include analytic ($z\beta$ = -4.02, $p < .01$), six letter words ($z\beta$ = -2.21, $p = .03$), anxiety ($z\beta$ = 2.80, $p < .01$), anger ($z\beta$ = -2.33, $p = .02$), cause ($z\beta$ = 3.33, p < .01), future ($z\beta$ = -2.11, $p = .03$), motion ($z\beta$ = -2.69, $p < .01$), home ($z\beta$ = 2.58, $p < .01$), money ($z\beta$ = 4.90, $p < .01$), personal pronoun ($z\beta$ = 2.25, $p = .02$), present non-third person verb ($z\beta$ = -2.79, p < .01),

present third person verb ($z\beta$ = -2.02, p = .04), and wh-determiner (e.g., which, whatever; $z\beta$ = -2.14, p = .03).

Out-sample prediction results of these models are reported in Table 2. Overall, predictions under randomized cross-validation were evidently better than under cross-context validation. The baseline model (regular regression using the psycholinguistic feature set as input) had out-sample prediction accuracies of .661 and .547 under randomized cross-validation and cross-context validation respectively. The prediction was slightly less accurate when POS tags were used as input. Combining psycholinguistic features and POS tags did not improve the prediction results. For a visual representation of the classification results, see figure 1.

Table 2. Classification Results of the Fixed-Effect Logistic Regression

	Test Samples									
Feature Set	A		B		C		D		Random	
	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
PSYC	.526	.537	.510	.493	.586	**.645**	.566	**.586**	.661	**.702**
POS	.592	**.605**	.599	**.630**	.579	**.634**	.553	**.585**	.632	**.693**
PSYC+POS	.474	.492	.529	.510	.552	**.594**	.474	.492	.654	**.712**

Note: AUC at .5 means accuracy of the prediction is at chance level. 95% CI of the AUCs were calculated using Hanley and McNeil's (1982) formula. Bolded AUCs had a 95% CI that did not include .5.
PSYC: The Psycholinguistic feature set. POS: The POS tags set.

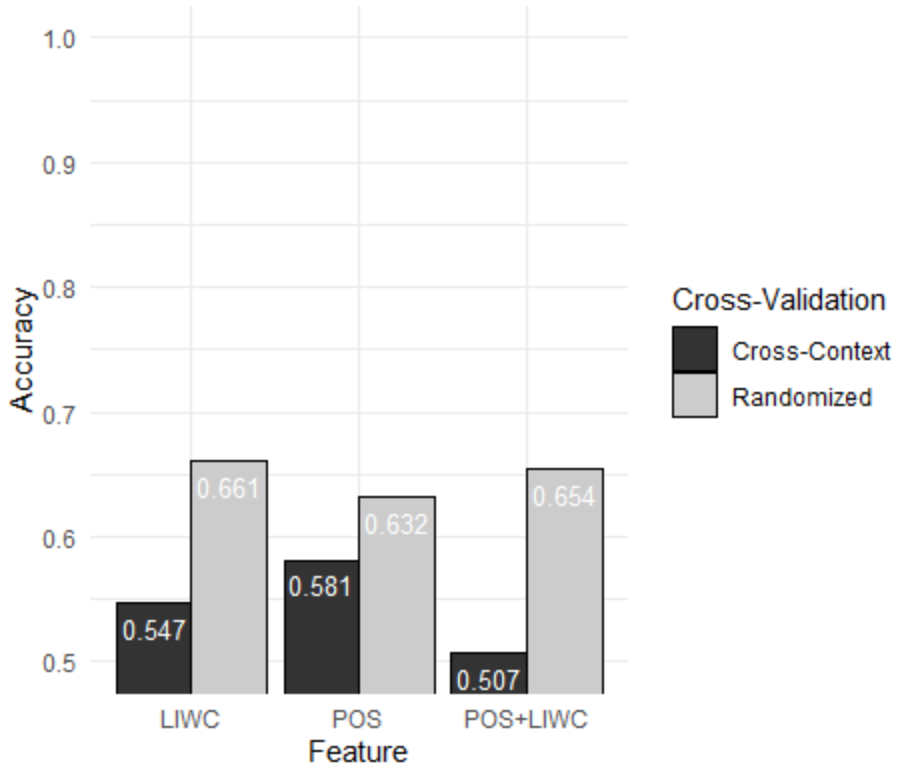

Figure 1. Classification Results of the Fixed-Effect Logistic Regression

2.3.2.2 LASSO Regressions

Two types of LASSO logistic regressions were estimated, one with random-effect (GLMM) and one without (GLM). LASSO models require a penalization parameter λ (Lambda) to be defined a priori. λ controls the amount of shrinkage. The number of regression coefficients being set to zero increases as the value of λ increases. The optimal λ is typically chosen based on a bias-variance tradeoff. In our analysis, we chose the λ values that minimized cross-validation residuals using a 10-fold cross-validation.

The fixed-effects LASSO logistic regression model trained using 10-fold randomized cross-validation and the psycholinguistic set had an in-sample prediction accuracy of .691 (*ROC-AUC* =.750). The same model trained using cross-context-validation had an in-sample prediction

accuracy of .727 (*ROC-AUC* = .781). Various iterations of the model had an average *AIC* of 1170. Out-sample prediction results of these models are reported in Table 3. Predictors that were frequently rejected (set to zero) by the LASSO procedure included discrepancy, tentative, differentiation, affiliation, and achievement.

The fixed-effects LASSO logistic regression trained using 10-fold randomized cross-validation and the POS tags set had an in-sample prediction accuracy of .661 (*ROC-AUC* = .710). The same model trained using cross-context-validation had an in-sample prediction accuracy of .649 (*ROC-AUC* = .704). Various iterations of the model had an average *AIC* of 1225. Out-sample prediction results of these models are reported in Table 3. None of the predictors were consistently being rejected by the LASSO procedure.

The fixed-effects LASSO logistic regression trained using 10-fold randomized cross-validation and the combined psycholinguistic and POS tags set had an in-sample prediction accuracy of .718 (*ROC-AUC* = .787). The same model trained using cross-context validation had an in-sample prediction accuracy of .720 (*ROC-AUC* = .782). Various iterations of the model had an average *AIC* of 1103. Out-sample prediction results of these models are reported in Table 3. Feel and affiliation were the only two predictors that were frequently rejected by the LASSO procedure.

For reasons discussed in section 2.2.5.4, the training of mixed-effects LASSO logistic regressions only used cross-context validation. Using the psycholinguistic feature set as input, the LASSO model had an in-sample prediction accuracy of .693 (*ROC-AUC* = .776). The average *AIC* was 1276. The mixed-effects LASSO logistic regression using POS tags as input had an in-sample prediction accuracy of .681 (*ROC-AUC* = .740). The average *AIC* was 1457. The LASSO model trained using the combined set of psycholinguistic features and POS tags had

an in-sample prediction accuracy of .705 (*ROC-AUC* = .762), and an average *AIC* of 1798. Out-sample prediction results for these models are reported in Table 3. None of these models had any fixed-effects coefficients that were consistently being set to zero by the LASSO procedure.

Overall, the fixed-effects LASSO logistic regression repeated the behavior of regular logistic regression in out-sample prediction. The predictions were much better under randomized cross-validation than under cross-context validation, except in one case. When statements about video-A were used as the testing data, and the combined feature set was the input, the fixed-effects LASSO logistic regression achieved a remarkably high prediction accuracy. In nearly all cases, the mixed-effects LASSO regression outperformed the fixed-effects LASSO regression as well as the baseline model under cross-context validation. The fixed-effects LASSO regression outperformed the baseline model under randomized cross-validation using the same feature set. Figure 2 shows a visual comparison of performances between fixed-effect logistic regression and mixed-effect logistic regression under cross-context validation.

The lexicon n-gram sets and the POS bigram set were not used for either regular logistic regressions or LASSO logistic regressions, because these feature sets each contain hundreds of features. Not only would they lead to large standard error of estimation, mixed-effects logistic regressions would also require an extremely long time to calculate and often fail to converge.

Table 3. Classification Results of the LASSO Logistic Regression Models

		Test Samples									
Model	Feature Set	A		B		C		D		Random	
		Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
Fixed	PSYC	.533	.537	.520	.494	.595	**.649**	.543	**.595**	.677	**.720**
	POS	.513	.537	.559	.595	.582	**.630**	.553	**.581**	.602	**.654**
	PSYC+POS	.780	**.841**	.532	.521	.602	**.644**	.532	**.595**	.635	**.678**
Mixed	PSYC	.737	**.751**	.624	**.654**	.576	**.611**	.612	**.664**	NA	NA
	POS	.648	**.657**	.561	**.582**	.625	**.675**	.634	**.684**	NA	NA
	PSYC+POS	.623	**.658**	.579	**.600**	.635	**.680**	.597	**.655**	NA	NA

Note: AUC at .5 means accuracy of the prediction is at chance level. 95% CI of the AUCs were calculated using Hanley and McNeil's (1982) formula. Bolded AUCs had a 95% CI that did not include .5.
PSYC: The Psycholinguistic feature set. POS: The POS tags set.

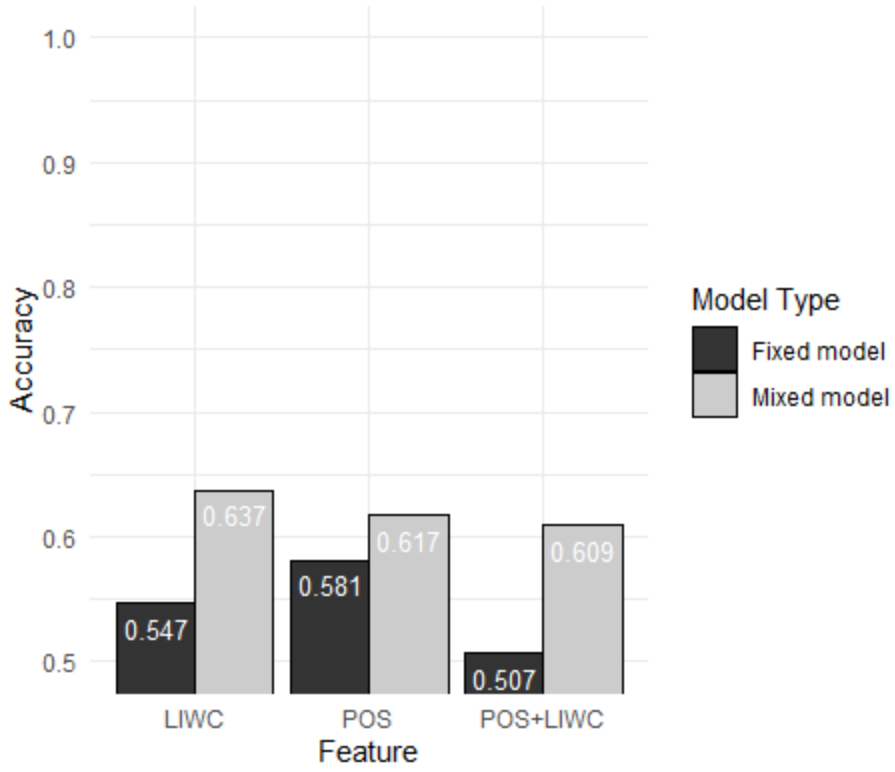

Figure 2. Mixed- vs. Fixed-Effect Logistic Regression under Cross-Context Validation

2.3.3 Tree and Forest Classifiers

Three tree/forest classifiers were used in the current research. They were the regular random forest, the BiMM tree, and the BiMM forest. The regular random forest model is not a mixed-effects model; the BiMM tree and forest are both mixed-effects models. In the same way previous logistic regression models were treated, the regular random forest was trained using both the 10-fold randomized cross-validation scheme and the four-fold cross-context validation scheme, while the two mixed-effects models were trained using only cross-context-validation. All feature sets described in section 2.2.4 were used as input of the tree and forest models. Combinations of any two feature sets were also explored. Due to the large number of possible

combinations, only the best examples are reported in Table 4. Figure 3 and Figure 4 provide visual comparisons between selected models.

In terms of in-sample accuracy, both forest algorithms consistently produced near perfect fit for training data, where only one or two cases were misclassified, while the average in-sample accuracy of the BiMM tree was .625 (95% CI [.563 .687]). The control parameter (a.k.a. hyperparameter; i.e. the number of variables randomly sampled at each split) of the regular random forest model was tuned based on a 10-fold cross-validation to maximize out-sample accuracy. One thousand trees were grown for each estimated forest. The same settings were used for the BiMM forest. The random forest model highlights important features with a measure called "mean decrease accuracy," which is the estimated accuracy loss from removing a specific feature. The top 10 most important features for our regular random forest models are listed in Table 4.

Out-sample prediction results are reported in Table 5. Overall, the tree and forest models mirrored the behavior of the regression models; their prediction accuracies were also comparable to the regression models. The tree/forest models appeared to favor the psycholinguistic feature set and the lexicon unigram set over other feature sets. The best prediction results were achieved when both sets were included as input; but adding the psycholinguistic set to the unigram set had very little impact on an already good performance. When lexicon bigrams and trigrams were used as input, these models struggled to make above-chance predictions. In the case of the trigram set, the BiMM tree classified all cases as truth. Since there were equal numbers of truth and lies in the testing sample, both accuracy and *ROC-AUC* were .5.

When compared with the baseline model, which had a .661 accuracy under randomized cross-validation and a .547 accuracy under cross-context validation, the regular random forest

generally performed better under randomized cross-validation, but offered comparable performance under cross-context validation. Both mixed-effects models – the BiMM tree and forest – performed better than the baseline model under cross-context validation in majority of the cases.

Table 4. Most Important Features Weighted by the Random Forest Models

Feature Set	Top Ten Features	Mean Decrease Accuracy [range]
PSYC	Analytic, money, motion, cause, risk, anger, space, Clout, insight, hear	[1.63% - 0.27%]
POS	PRP$, DT, VBN, NN, VBZ, RP, RB, VBD, PRP, VBG	[0.57% - 0.19%]
POS-bi	DT_NN, WRB_TO, VB_VBN, JJ_NN, VBZ_JJ, VBD_RB, VBG_PRP, VBD_TO, PRP_RB, RB_VB	[0.16% - 0.06%]
POS-bi^{+}	PRP$, WRB_TO, DT_NN, JJ_NN, JJ, NN_VBZ, VBZ, RP, CC, IN_DT	[0.14% - 0.06%]
Unigram	the, fell, arm, was, car, just, face, so, help, money	[0.39% - 0.17%]
Bigram	at_him, the_face, his_arm, the_car, fell_to, my_friend, so_he, open_the, to_explain, to_open	[0.59% - 0.14%]
Bigram^{+}	the, arm, face, would, was, me, my, friend, wrong, car	[0.43% - 0.18%]
Trigram	in_the_face, not_to_move, him_to_the, to_get_in, to_come_out, how_to_open, and_the_customer, and_the_man, me_to_the, ran_back_to	[1.11% - 0.18%]
Trigram^{+}	the, was, me, would, my, out, and, so, it, to	[0.35% - 0.01%]

Note: Top ten most important features of the models are listed from the most important to the least important. “Mean Decrease Accuracy” is the estimated loss of accuracy if the feature is removed from the model. The reported “Mean Decrease Accuracy” range is the range of the top ten most important features.
PSYC: The Psycholinguistic feature set. POS: The POS tags set. POS-bi: The POS bigram set. POS-bi+: The POS bigram+ set. Unigram: The lexicon unigram set. Bigram: The lexicon bigram set. Trigram: The lexicon trigram set. Bigram+: The lexicon bigram+ set. Trigram+: The lexicon trigram+ set.

Table 5. Classification Results of the Tree and Forest Models

	Random Forest				BiMM Tree		BiMM Forest	
Feature Set	Context		Random		Context		Context	
	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
PSYC	.533	.537	.668	**.725**	.615	**.639**	.602	**.637**
POS	.563	**.571**	.605	**.656**	.543	.558	.556	**.579**
POS-bi	.546	**.591**	.611	**.643**	.542	**.583**	.569	**.574**
POS-bi^{+}	.566	**.580**	.592	**.610**	.543	.543	.588	**.592**
Unigram	.536	**.574**	.734	**.782**	.610	**.655**	.639	**.696**
Bigram	.457	**.591**	.658	**.711**	.522	.540	.523	.550
Bigram^{+}	.504	.511	.688	**.776**	.525	.536	.584	**.625**
Trigram	.480	.554	.661	**.722**	.500	.500	.538	.551
Trigram^{+}	.500	.526	.707	**.770**	.497	.536	.525	**.576**
PSYC+POS	.549	.556	.678	**.735**	.631	**.651**	.629	**.667**
PSYC+Unigram	.526	**.568**	.697	**.771**	.634	**.696**	.612	**.662**
POS+Unigram	.559	**.601**	.737	**.806**	.643	**.681**	.636	**.683**
Mean	.527	.563	.670	.726	.567	.593	.583	.616
SD	.035	.028	.048	.061	.055	.067	.042	.052

Note: AUC at .5 means accuracy of the prediction is at chance level. 95% CI of the AUCs were calculated using Hanley and McNeil's (1982) formula. Bolded AUCs had a 95% CI that did not include .5. Numbers under the column title "Context" are based on cross-context validation. PSYC: The Psycholinguistic feature set. POS: The POS tags set. POS-bi: The POS bigram set.
POS-bi+: The POS bigram+ set. Unigram: The lexicon unigram set. Bigram: The lexicon bigram set. Trigram: The lexicon trigram set. Bigram+: The lexicon bigram+ set. Trigram+: The lexicon trigram+ set.

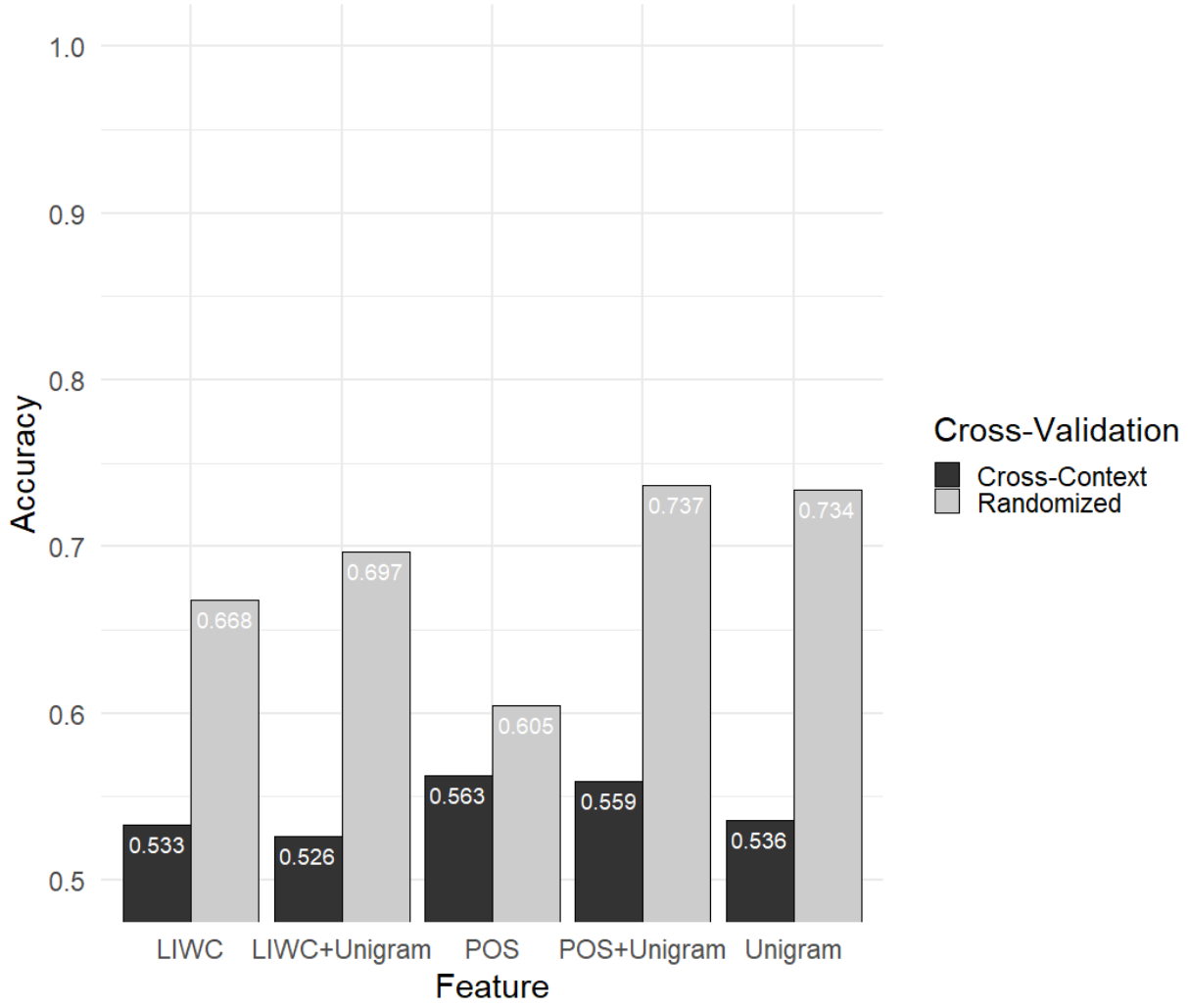

Figure 3. Performances of the Random Forest Under Different Cross-Validation Schemes

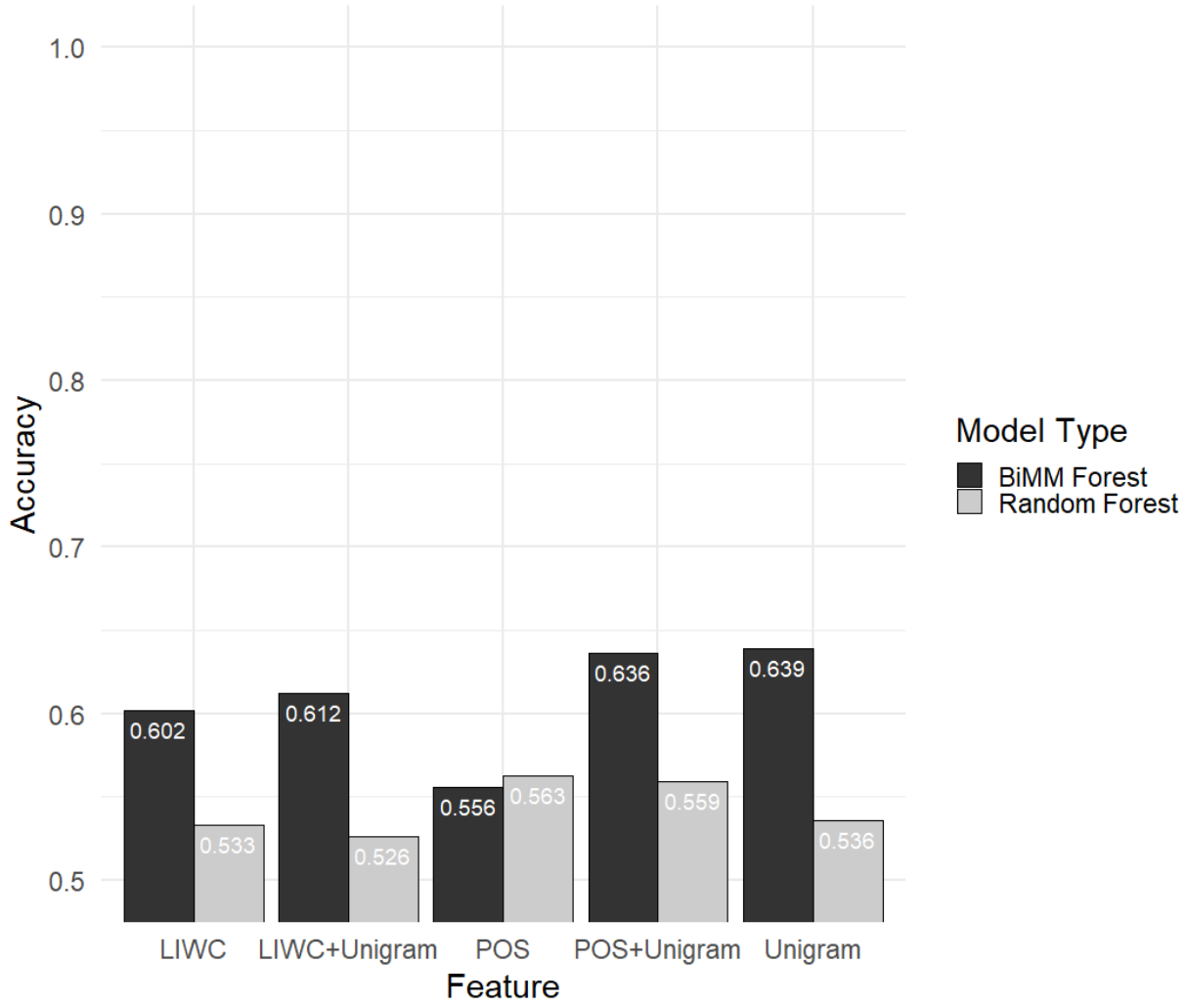

Figure 4. Random Forest vs. BiMM Forest under Cross-Context Validation

2.3.4 ANNs

Both the regular ANN and the "mixed-effects" ANN contained four hidden layers. The only difference between the two models was that the "mixed-effects" ANN added individual identifiers as input and connected them to the second hidden layer. Similar to the random forest, ANNs easily produced near perfect fit for the training data. Early stop was used to reduce overfitting, that is, each ANN was estimated using only 10-20 epochs (iterations).

Out-sample prediction results are reported in Table 6. Similar to all previously tested models, the "fixed-effects" ANN performed much better under randomized cross-validation than under cross-context validation. The "mixed-effects" ANN outperformed the "fixed-effects" ANN under cross-context validation; however, in most cases the difference was small. The ANN models favored POS bigram+, lexicon bigram+, and trigram+ as input. Predictions that included the POS bigram+ feature set as input were not only more accurate than predictions made using any other feature sets, but were also more generalizable across contexts. This effect was not observed for models that included either one of the two components of the POS bigram+ set (POS unigrams and bigrams) as separate feature sets.

When compared with the baseline model (accuracies = .661 and .547 for randomized cross-validation and cross-context validation respectively), the "fixed-effects" ANN generally performed better under randomized cross-validation. Under cross-context validation, it only performed better than the baseline model in some cases, noticeably when either the psycholinguistic set or the POS bigram+ set were included in the input. The "mixed-effects" ANN outperformed the baseline model under cross-context validation in all cases; however, sometimes the difference was negligible. Figure 5 and Figure 6 show visual comparisons between selected models.

Table 6. Classification Results of the ANNs

Feature Set	ANN (Fixed) Context		ANN (Fixed) Random		ANN (Mixed) Context	
	Acc.	AUC	Acc.	AUC	Acc.	AUC
PSYC	.569	**.669**	.658	**.726**	.595	**.649**
POS	.536	.562	.616	**.659**	.586	**.629**
POS-bi	.529	**.575**	.595	**.630**	.602	**.622**
POS-bi^{+}	.642	**.730**	.751	**.817**	.675	**.730**
Unigram	.543	.543	.688	**.783**	.625	**.652**
Bigram	.510	.470	.655	**.719**	.533	**.588**
Bigram^{+}	.559	**.575**	.730	**.805**	.593	**.611**
Trigram	.523	.484	.672	**.735**	.553	**.551**
Trigram^{+}	.487	.506	.720	**.783**	.552	**.588**
PSYC+POS	.572	**.633**	.636	**.698**	.592	**.639**
PSYC+POS-bi^{+}	.623	**.691**	.740	**.758**	.745	**.788**
PSYC+Bigram^{+}	.539	.550	.737	**.802**	.605	**.636**
POS+Unigram	.536	**.580**	.702	**.795**	.596	**.622**
POS-bi^{+}+Bigram^{+}	.655	**.714**	.719	**.791**	.692	**.754**
Mean	.566	.599	.687	.747	.610	.647
SD	.065	.096	.049	.058	.058	.066

Note: AUC at .5 means accuracy of the prediction is at chance level. 95% CI of the AUCs were calculated using Hanley and McNeil's (1982) formula. Bolded AUCs had a 95% CI that did not include .5. Numbers under the column title "Context" are based on cross-context validation. POS-bi+: The POS bigram+ set. Unigram: The lexicon unigram set. Bigram: The lexicon bigram set. Trigram: The lexicon trigram set. Bigram+: The lexicon bigram+ set. Trigram+: The lexicon trigram+ set.

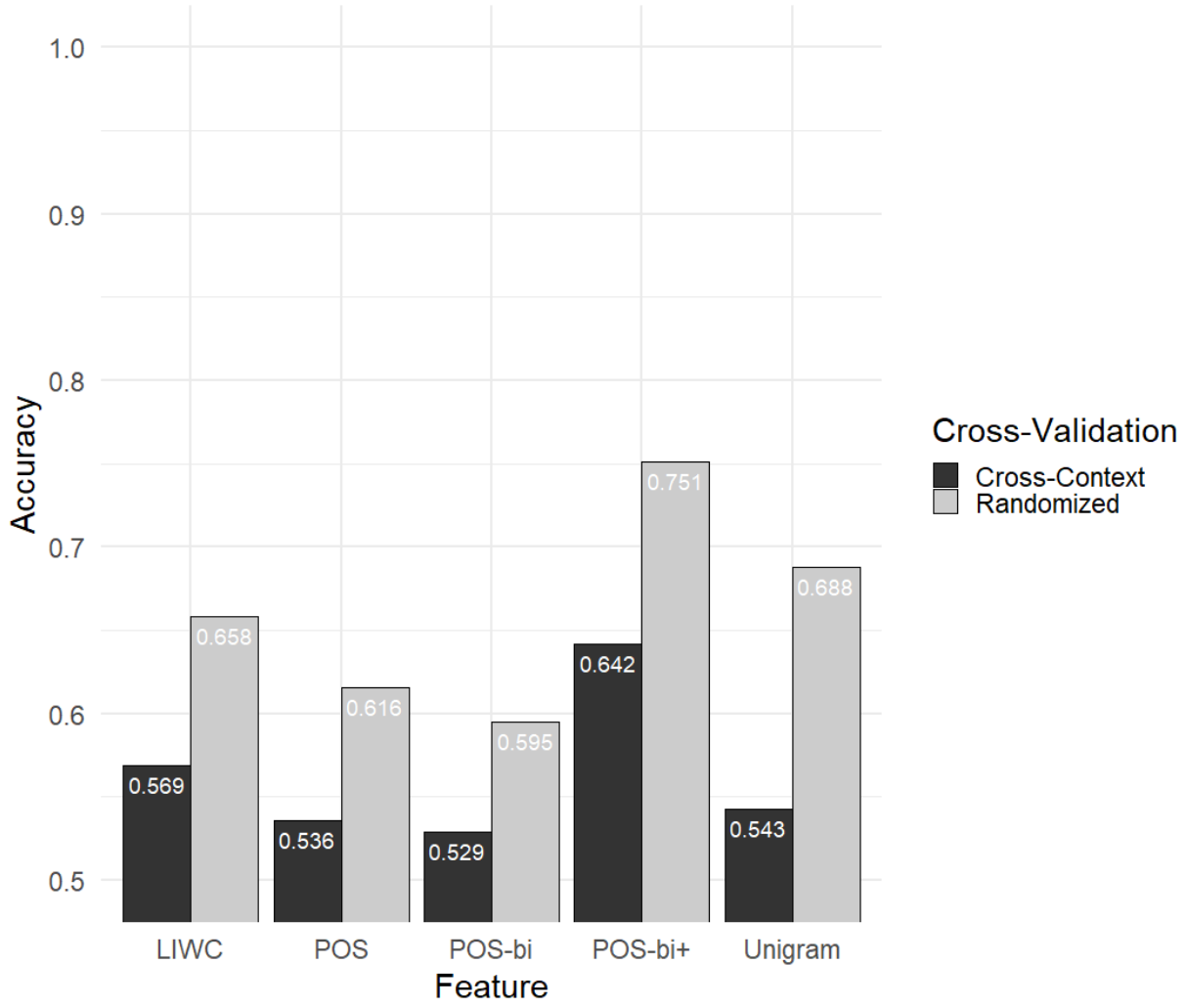

Figure 5. Performances of the Regular ANN Under Different Cross-Validation Schemes

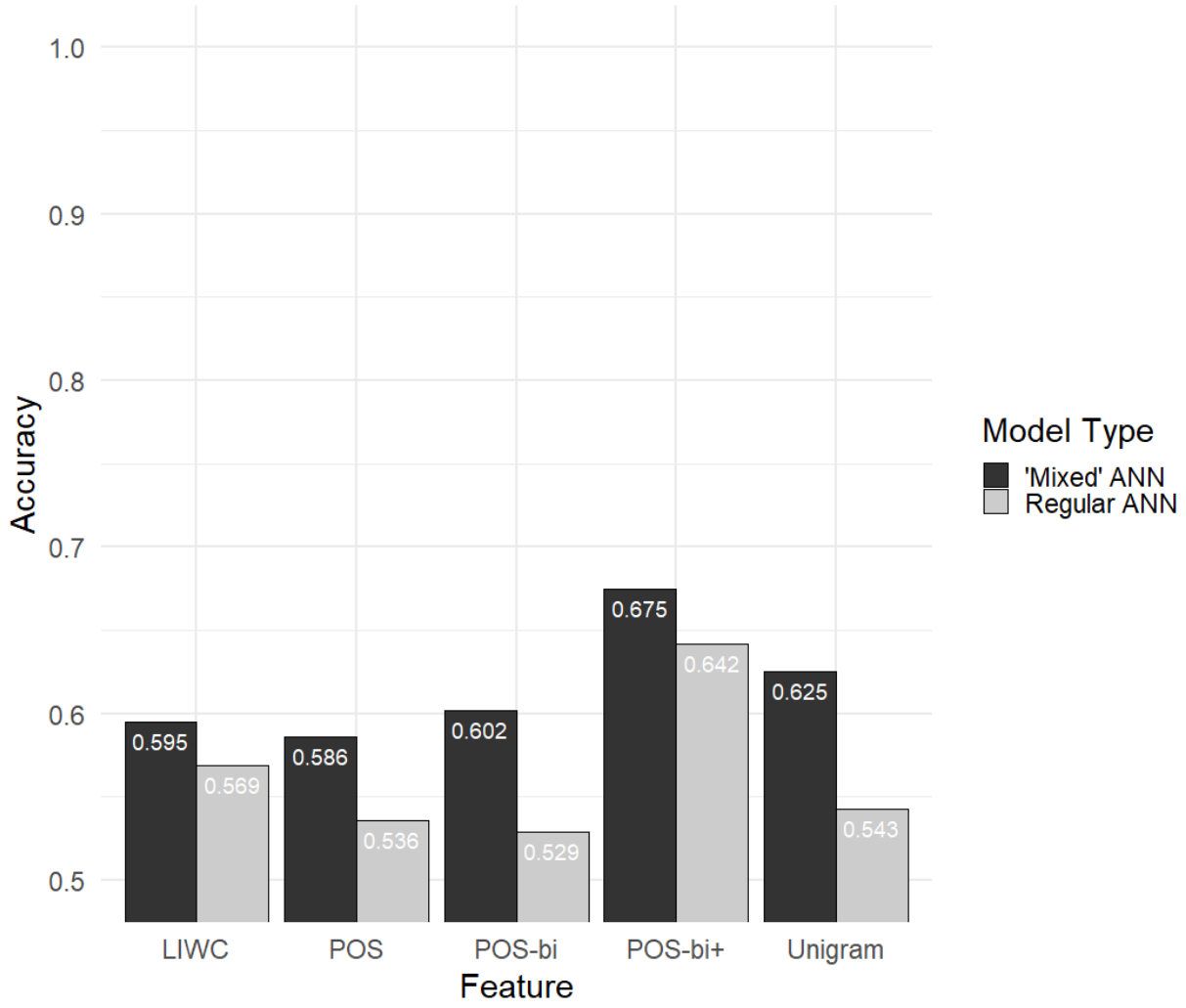

Figure 6. Regular ANN vs. “Mixed” ANN under Cross-Context Validation

2.3.5 Cross-Perspective Validation

Given our text corpus included statements written in either a first-person (writing task B and C) or a third-person (writing task A and D) perspective, we also tested the cross-perspective generalizability of our classification models. We estimated regular and mixed-effects versions of LASSO regression, random forest, and ANN using statements written in one perspective and tested them on statements written in the other perspective. This constituted a two-fold cross-validation. Only single feature sets were used for this cross-validation scheme. The average classification accuracies and ROC-AUCs for the tested models are reported in Table 7. The overall average accuracies for the regular random forest, the BiMM forest, the regular ANN, and the "mixed-effects" were .534 (*SD* = .026), .545 (*SD* = .032), .538 (*SD* = .043), and .582 (*SD* = .046) respectively; their overall average ROC-AUCs were .554 (*SD* = .035), .568 (*SD* = .031), .543 (*SD* = .052), and .604 (*SD* = .051) respectively. In general, the tested classification models performed worse than they did under cross-context validation, but most of the classifications were still significantly better than chance. The mixed-effects models outperformed their regular counterparts in all but a few cases: When using the psycholinguistic set, the lexicon bigram set, or the lexicon trigram+ set as input, the regular random forest slightly outperformed the BiMM forest in terms of classification accuracy.

Table 7. Classification Results under Cross-Perspective Validation

Feature Set	LASSO Regression Accuracy (AUC)		Random Forest Accuracy (AUC)		ANN Accuracy (AUC)	
	Regular	Mixed	Regular	BiMM	Regular	"Mixed"
PSYC	.540 (**.571**)	.590 (**.596**)	.563 (**.579**)	.546 (**.581**)	.556 (**.573**)	.569 (**.601**)
POS	.527 (.542)	.582 (**.591**)	.538 (**.573**)	.554 (**.567**)	.523 (.538)	.579 (**.572**)
POS-bi	\	\	.532 (.537)	.559 (**.569**)	.553 (**.562**)	.602 (**.654**)
POS-bi^{+}	\	\	.549 (**.580**)	.577 (**.600**)	.636 (**.651**)	.660 (**.688**)
Unigram	\	\	.572 (**.609**)	.600 (**.622**)	.532 (**.549**)	.637 (**.649**)
Bigram	\	\	.515 (.495)	.502 (.521)	.498 (.486)	.548 (**.556**)
Bigram^{+}	\	\	.510 (.535)	.531 (**.558**)	.538 (**.546**)	.591 (**.619**)
Trigram	\	\	.492 (.522)	.506 (.536)	.518 (.506)	.532 (.539)
Trigram^{+}	\	\	.535 (**.557**)	.528 (**.562**)	.488 (.477)	.521 (**.563**)

Note: AUC at .5 means accuracy of the prediction is at chance level. 95% CI of the AUCs were calculated using Hanley and McNeil's (1982) formula. Bolded AUCs had a 95% CI that did not include .5.

PSYC: The Psycholinguistic feature set. POS: The POS tags set. POS-bi+: The POS bigram+ set. Unigram: The lexicon unigram set. Bigram: The lexicon bigram set. Trigram: The lexicon trigram set. Bigram+: The lexicon bigram+ set. Trigram+: The lexicon trigram+ set.

.

Chapter 3: Discussion and Conclusion

3.1 Overview

The main goal of the current research was to explore the utility of repeated-measure and mixed-effects modeling in text-based deception detection. It was expected that adding individual-based random-effects to an existing model would improve cross-context generalizability of the model. This premise was supported by the behavior of all three types of classification algorithms used in the current research. In almost all cases, the mixed-effects versions of the tested classifiers performed better than their fixed-effects counterparts in cross-context validation, although sometimes, the differences were very small.

Overall, it appeared that the text corpus collected in the current research presented a fairly difficult lie detection task. In general, we observed relatively low classification accuracies (out-sample) compared with similar studies (e.g., Ott et al., 2011; Ott et al., 2013; Xu & Zhao, 2012), which regularly reported above 75% classification accuracy. Many of our classification models had an accuracy between 60%-70% under 10-fold randomized cross-validation. This level of performance was perhaps reasonable given that our text corpus is more diverse than those used by studies that reported high prediction accuracy. Although the four videos that we used to generate text responses all had a similar negative emotional tone and all depicted minor violence, the situations are quite different from video to video. They should lead to more varied statements than narrow topics such as hotel reviews. In fact, studies that reported similar levels of classification accuracy as the current research also had diverse text corpora (e.g., Fuller et al., 2009; Mihalcea & Strapparava, 2009).

In this chapter, we will summarize the performances of each implemented classification models, compare these results with other relevant research, and discuss the factors that may have influenced these results. The effectiveness of each of the three categories of linguistic features as deception cues will also be discussed; important individual cues will be highlighted.

3.2 The Classifiers

3.2.1 General Observations

Considering only fixed-effects models, random forest and ANN showed small improvements over regular logistic regression in terms of average prediction accuracies. Under randomized cross-validation, average classification accuracies for fixed-effects logistic regression, random forest, and ANN were 64% (*SD* = 3.5%), 67% (*SD* = 4.8%), and 69% (*SD* = 4.9%) respectively. These results suggested that when the training samples covered the same stories as the testing samples, the fixed-effects models were somewhat effective at separating truthful statements from deceptive ones. We can also see this from the fact that the *ROC-AUC*s of these models were all significantly higher than .5 regardless of what predictors were used as input. The ANN on average had the best performance, followed by the random forest; logistic regressions had the worst performance. Nevertheless, if we compare their performances when they used the same feature sets, the ANN and the random forest did not appear to outperform logistic regressions. Considering the three sets of features that were tested across all three types of classifiers (i.e., the psycholinguistic set, the POS tags set, and the combined set of psycholinguistic and POS unigram features), different classifiers had very similar performances. For instance, when using the psycholinguistic set as input, the LASSO logistic regression provided the best performance, which was an accuracy of 67.7%; yet, it only correctly classified 2% more statements from the testing samples than the worst performing model in this case,

which was the ANN. The random forest and the ANN gave better predictions when larger feature sets were used, thus, surpassing logistic regressions on average. The ANN even achieved above 80% accuracy when the input included the POS-bigram+ set, which was the largest feature set used in the current research.

On the other hand, under cross-context validation, average classification accuracies for the fixed-effects logistic regressions, random forest, and ANN were 56% (*SD* = 6.2%), 53% (*SD* = 3.5%), and 57% (*SD* = 6.5%) respectively. These numbers were not significantly higher than .5, which meant that all three classifiers, on average, failed to separate truthful and deceptive statements when their training sample did not include the same story as the testing sample. This result suggested that the fixed-effects classification models were not generalizable to contexts that they were not exposed to during training. In the case of the random forest and the ANN, both models were able to produce near perfectly fitting models for the training data. The failure to fit the testing data was a clear sign of overfitting, which meant that these models are susceptible to noise due to content differences between the training and testing sample. Nevertheless, the exceptional success of the ANN when using the POS bigram+ feature set as input showed that it is possible to find a cross-context generalizable language pattern of deception within our text corpus.

The average prediction accuracies of the mixed-effects LASSO logistic regression, the BiMM tree, the BiMM forest, and the "mixed-effects" ANN were 62% (*SD* = 3.1%), 57% (*SD* = 5.5%), 58% (*SD* = 4.2%), and 61% (*SD* = 5.8%) respectively. In general, the mixed-effects classifiers outperformed their fixed-effects counterparts under cross-context validation, but performances of specific combinations of classifiers and feature sets varied greatly. Although the average improvement in cross-context generalizability appeared to be small, some features

and classifier combinations clearly demonstrated the advantage of the mixed-effects models over the fixed-effects models. Detailed discussions of each type of the classifiers are provided below.

3.2.2 Logistic Regression Models

In psychological research logistic regression is a standard statistical tool for binary classification. A regular logistic regression model with the LIWC coded psycholinguistic features as input was used in the current research as the baseline model. This was because the use of both the model and the feature set was typical of earlier studies in this field, which were often surpassed in terms of classification accuracy by later studies that employed machine-learning models and more elaborated linguistic feature sets. In the current research, this baseline model was able to, on average, correctly classify 66% of the text in randomly selected testing samples; but with cross-context testing samples, its average classification accuracy was reduced to 55%. The *ROC-AUC*s under the two cross-validation schemes were .702 and .565 respectively. Only the *ROC-AUC* under randomized cross-validation was significantly above .5, meaning the model was correctly classifying truths and lies at a level above chance. A very similar model with comparable performance can be found in a study conducted by Newman, Pennebaker, Berry, and Richards (2003). With a mixed-topics opinion corpus, Newman and colleagues reported a 61% classification accuracy under randomized cross-validation, and chance levels of cross-topic classification accuracies.

In addition to the baseline psycholinguistic model, the current research also tested the performance of the logistic regression using POS tags as input. The POS model performed slightly worse than the baseline model under randomized cross-validation but was slightly more accurate under cross-context validation (63% and 58% respectively). Unlike the baseline model,

the *ROC-AUC* of the POS model under cross-context validation (.614) was significantly higher than chance, although it was only just above the critical value (i.e., .610).

Furthermore, we also estimated a logistic regression model using both the psycholinguistic set and the POS tags set. Despite of the increase in the number of predictors, this model did not outperform the baseline model. Under randomized cross-validation, the accuracy was slightly lower (65%) than the baseline model, but the *ROC-AUC* was slightly higher (.712). Under cross-context validation, the performance of the combined model (accuracy = 50%, *ROC-AUC* = .522) became noticeably worse than both the baseline model and the POS model. Although the difference was too small to be statistically significant, such a result could still be a hint that the logistic regression tends to produce less generalizable models when its input consisted of a larger number of predictors.

LASSO regression adds an automated variable selection and model regulation procedure to regular regression models. In the current research, the fixed-effects logistic regressions with LASSO offered very similar performances to regular logistic regressions. Under randomized cross-validation, prediction accuracies of the LASSO model when using the psycholinguistic set, POS tags set, or the combined set were 68%, 60%, and 64% respectively. In each iteration of the model, relatively few predictors were rejected compared with the total number of predictors, and there was no obvious impact on model fit and parsimony as indicated by the differences in *AIC* scores. Under cross-context validation, performances of the fixed-effects LASSO logistic regressions were also roughly equal to the regular logistic regressions when the psycholinguistic set and the POS tags set were used individually (accuracy in both cases were 55%). However, the LASSO logistic regression performed much better than the regular logistic regression when the combine feature set was used as input: The average accuracy over the four-fold cross-context

validation was 61% compared with the 50% accuracy of the model without LASSO. In fact, this level of performance was largely due to one exceptional case. When video A was used as the training sample, the classification accuracy of the model reached 78% (*ROC-AUC* = .841). There is no apparent explanation for such an abnormally high level of performance, but given the large number of tests conducted, such an anomaly could be purely due to chance.

The main reason why LASSO models were used in the current research in addition to regular logistic models was to enable the estimation of mixed-effects logistic regression models. As expected, adding mixed-effects to the regression models had a clear impact on classification accuracy under cross-context validation. The average accuracies of the mixed-effects models were 64%, 62%, and 61% respectively when using the psycholinguist feature set, the POS tags set, and the combined set of the two sets as input. These levels of performance were only slightly worse than that of the fixed-effects models under randomized cross-validation. The model that used psycholinguistic features as input benefited most from the added random-effects. On the other hand, adding random-effects to the regression models did worsen the *AIC* scores. Given the similar in-sample prediction accuracies between the two types of models, such a loss in *AIC* was indicative of worsened parsimony as a result of the increase in model complexity.

Overall, the frequentist logistic regression models appeared to be a viable tool for deception detection tasks with or without mixed-effects modeling in certain situations. When exploring a small number of higher-order or abstract deception cues, these models provided comparable performances to more advanced machine-learning classifiers. However, they are limited in model complexity, thus ill-suited for deception detection research that focuses on improving classification performance, which often require the input of hundreds of features. In addition, relatively complex mixed-effects LASSO models are time consuming to calculate. Our

various mixed-effects LASSO models that included around 40 input variables took 30 mins to an hour to compute on a mid-range PC. In comparison, the computations of BiMM forest models with hundreds of input features only took 1 to 2 mins on the same computer.

3.2.3 Decision Tree and Random Forest Models

Mixed-effects modification of machine-learning models is currently an underdeveloped area of research. There are few attempts at building machine-learning models specifically to take advantages of the repeated-measure data structure. Nevertheless, decision tree and random forest methods appear to have the best documented mixed-effects modifications among all other machine-learning methods. There are two published R packages that implemented different versions of mixed-effects decision tree and random forest models (Fokkema et al., 2018; Sela & Simonoff, 2012). Unfortunately, as discussed in section 2.2.5.2, both of these models are not suited for the current research. Speiser and colleagues' (2019) BiMM tree and forest are the only variations of mixed-effects tree and forest algorithms that matched with the proposed classification method of the current research. According to Speiser and colleagues, the BiMM random forest offers a state-of-the-art level of performance.

In the current research, the regular random forest models offered similar levels of performance as the baseline fixed-effects logistic regressions when using the psycholinguistic set and/or the POS tags set as input. Like the fixed-effects regression models, the classification accuracies of the random forest models generally ranged between 60% to 70% under randomized cross-validation and were not much different than chance under cross-context validation. The best classification accuracy under randomized cross-validation was 74% (*ROC-AUC* = .806). It was achieved using both the psycholinguistic set and the lexicon unigram set as input; however, the model using only unigram as input achieved a nearly identical level of performance

(accuracy = 73%, *ROC-AUC* = .782) suggesting that the addition of psycholinguistic cues was unnecessary. Moreover, under randomized cross-validation, the random forest models appeared to offer better performances when any n-gram feature set were used; but under cross-context validation, the random forest models using n-gram features had the worst performances: In most of these cases, the classification accuracies were nearly identical to 50% or below 50%. It is worth restating that the random forest models, regardless of input, classified the truthful and deceptive statements in the training samples with nearly perfect accuracies. These results suggested that the random forest models have poor cross-context generalizability when used in deception detection tasks.

As its mixed-effects counterparts, both the BiMM tree and the BiMM forest outperformed the regular random forest under cross-context validation; however, their average classification accuracies were still below 60%. The BiMM forest models outperformed the tree models in most of the cases, except whenever the psycholinguistic features were included in the input. Lexicon unigram was again the best feature set for both the BiMM tree and forest models. Classification accuracies in these two cases were 61% (ROC-AUC = .655) for the tree model, and 64% (ROC-AUC = .696) for the forest model. The performances of other n-gram models remained at a level barely different from chance. From these results, we can clearly see the benefit of adding mixed-effects modeling to the decision tree and random forest methods, but this benefit appeared to be dependent upon the model input. For instance, the performance differences between regular random forest and the BiMM forest were much larger when n-gram features were used as input than when POS features were used as input. It is also worth noting that the current implementations of the BiMM tree and forest lack any customization option that

affects the random-effect portion of the model. It is conceivable that better tuning on all parts of the model could lead to better performances.

3.2.4 ANNs

Neural network is widely considered the state-of-art machine-learning method. It is highly versatile and can potentially be customized to accommodate any data structure. In deception detection research, we frequently see ANNs outperform other classification models (e.g., Fuller et al., 2009; Zhou et al., 2004b). The only disadvantage is that they do not directly offer information regarding the relative importance of the predictors. It is possible to systematically remove features and re-estimate an ANN model to assess the importance of each feature, but such a measure is not very helpful when the feature set is large and contains correlated features.

Few studies so far have specifically discussed the issue of implementing ANN models for repeated-measure data. Examples of longitudinal prediction models can be sparsely found in medical and biological research (e.g., Tandon, Adak, & Kaye, 2006). Maity and Pal's (2013) article is the only document that made general recommendations for building repeated-measure ANN models. The current research employed a widely used simple architecture of ANN for solving classification problems and adopted it according to Maity and Pal's recommendation for repeated-measure data. Both the "fixed-effects" and the "random-effects" ANN models offered the best performance within their respective categories among all models tested in the current research.

Most notably the ANN models achieved the best performances in the current research when the POS bigram+ feature set was included in its input. The classification accuracies of the "fixed-effects" ANN under randomized cross-validation was the highest among all models

(accuracy = 75% *ROC-AUC* = .817); it even had relatively good generalizability across contexts (accuracy = 64%, *ROC-AUC* = .730). Its "mixed-effects" counterpart only showed marginal performance improvement under cross-context validation (accuracy = 68%, *ROC-AUC* = .730). Adding psycholinguistic features to the POS bigram+ model did not change the performance of the "fixed-effects" ANN but had a significant impact on the performance of the "mixed-effects" ANN, making it the best performing mixed-effects model in the current research (accuracy = 74%, *ROC-AUC* =.788). On the other hand, the behaviors of the lexicon n-gram models largely mirrored their decision tree and random forest counterparts, in that they gave good performances under randomized cross-validation but generalized poorly across contexts. Adding individual identifiers as input to the ANN did not appear to improve cross-context generalizability of the lexicon n-gram models. These results show that a generalizable language pattern of deception can be detected using the correct combination of features and classification method. The benefit of "mixed-effects modeling" for the ANN appeared to be significant, but only when certain types of features are used as input.

3.2.5 Cross-Perspective Generalization

In addition to testing cross-context generalizability we also tested the cross-perspective generalizability of our classification models. The results showed that classification models trained using statements written in one perspective have difficulty generalizing to statements written in a different perspective. Both the regular models and their mixed-effects counterparts gave worse classification results under cross-perspective validation than under cross-context validation. This reduction in accuracy was expected, because under our cross-perspective validation scheme the training samples were different from the testing sample not only in context but also in perspectives. These results confirmed that perspectives matter to language production

during deception. Patterns that can be used to detect deception in statements written in the first-person may not apply to statements that were written from an observer perspective, and vice versa. On the other hand, most of our classification models still provided above-chance level performance and the benefit of modelling individual-level random-effects was consistent in cross-perspective validation. These findings suggested that we can still apply what was learned from an individual lying in one perspective to their lies in another perspective. Nevertheless, the overall poor cross-perspective testing results suggested that it is better to train the classifiers using statements written in both perspectives whenever they are available.

3.3 Language Features

3.3.1 The LIWC Psycholinguistic Set

Theory-driven psycholinguistic cues were the focus of deception detection studies in psychology. In the current research, the average accuracy of all the classification models using the LIWC coded psycholinguistic feature set as input was 67% (*SD* = 0.84%) under randomized cross-validation, and 58% (*SD* = 3.72%) under cross-context validation; the *ROC-AUC*s were .718 (*SD* = .011) and .617 (*SD* = .052) respectively. This level of performance is comparable with many similar classification models found in past research (e.g., Fuller et al., 2009; Newman et al., 2003). However, studies that were conducted on restrictive corpora, such as Ott and colleagues' (2013) hotel review corpus, had clearly better classification results when using psycholinguist features exclusively. The general observation that linguistic and contextual diversity within a corpus lead to lower classification accuracies appeared to hold true for psycholinguistic features used in the current research. Almela, Valencia-Garcia, and Cantos (2012) also found that deception detection accuracies using LIWC cues vary greatly across

different topics. They argued that better alignment between the topic and LIWC word categories as well as greater emotional involvement would lead to better classification results. In the current research, our classification models that involved the psycholinguistic set had more varied performances during cross-context validation than other feature sets. This could attest to the context-dependency of the psycholinguistic features as deception cues.

Different classification models that used the psycholinguistic set exclusively as input varied very little in performance. Combining the psycholinguistic set with other feature sets largely made no significant differences. These results seemed to suggest that the abstract psychological constructs in the psycholinguistic set are effective deception cues by themselves, but do not add to the detection of deception above and beyond the more concrete linguistic features captured by lexicon n-grams and POS tags. Nevertheless, in one case, adding psycholinguistic cues to the already well-performing POS bigram+ ANN models had a clear positive impact on the cross-context generalizability of the ANN models especially when individual identifiers were included in the input of the ANN (see Table 6). This may suggest that the psycholinguistic features offer information related to personal deceptive language style in addition to the grammatical information contained in POS unigrams and bigrams.

3.3.2 POS Feature Sets

In psychological studies of deception, some POS features were often included as a dimension of the LIWC set. It was widely accepted that the use of pronouns is related to deception, in that first-person pronouns appear more frequently in truthful statements, while third-person pronouns are more abundant in deceptive statements (e.g., DePaulo et al., 2003; Fuller et al., 2009; McQuaid, Woodworth, Hutton, Porter, & ten Brinke, 2015; Newman et al., 2003). In the current research, POS features are extracted according to categories commonly

used by the NLP research community, which does not give first- and third-person pronouns different tags. Nevertheless, first- and third- person pronouns were captured by the n-gram feature sets as individual words. We will discuss them in section 3.3.4.

Previous deception detection studies that have used the same POS tags as the current research showed that they were by themselves less effective than n-gram and psycholinguistic features (e.g., Ott et al., 2011; Xu & Zhao, 2012). The findings of the current research agreed with this conclusion. The POS tags set achieved an average accuracy of 61% (*SD* = 1.36%) across all models under randomized cross-validation, and 57% (*SD* = 2.66%) under cross-context validation; the *ROC-AUC*s were .666 (*SD* = .018) and .593 (*SD* = .033) respectively. These numbers were lower than that of the psycholinguistic set and the lexicon unigram set. The POS bigram set performed slightly worse than the POS unigram set. However, when the both POS unigrams and bigrams (POS bigram+) are included as the input of the ANN models, the classification accuracy exceeded all other models that did not use this feature set by a noticeable margin. This finding echoed a conclusion of Xu and Zhao's study, that the POS features alone cannot fully represent the differences between truthful statements and deceptive statements, but richer grammatical information makes the classification much easier.

3.3.3 Lexicon N-gram Sets

In NLP-based deception detection studies, lexicon n-grams were shown to be potentially the most effective features. Ott and colleagues (2011; 2013) and Xu and Zhao's (2012) research using hotel reviews achieved their best classification accuracy using n-gram features. Adding other types of features to n-gram models only improved the performance very slightly. In Xu and Zhao's study, a unigram model was able to reach 90.6% accuracy. On the other hand, studies that were conducted on more diverse corpora reported around 70% accuracies when

using n-gram features to detect deception (Mihalcea & Strapparava, 2009; Sánchez-Junquera, Villaseñor-Pineda, Montes-y-Gómez, & Rosso, 2018). These results are much more comparable with ours. Our unigram models achieved average accuracies of 71% (*SD* = 3.25%) and 59% (*SD* = 6.32%) under randomized cross-validation and cross-context validation respectively. Other n-gram models performed similarly to the unigram model under randomized cross-validation but gave noticeably worse predictions under cross-context validation.

By observing the highest rated n-gram features in Table 4, we found that many of these highest ranked features are related to specific contexts in the videos (e.g., "car," "face," "his_arm," "in_the_face," "and_the_customer"). The prevalence of context-specific words among features that were given higher importance by the models could explain poor cross-context generalizability, but we also see generic words such as "the," "so," "was," "me," and "my" being equally prevalent. Most curiously, the top ranked features in the trigram+ model consisted of only generic words, yet the model was one of the least cross-contextually generalizable models. This is likely due to the low impacts of individual n-gram features on the overall model performance (see Table 4: *Mean Decrease Accuracy*).

The lack of cross-contextual generalizability of our n-gram models was contrary to our expectation based on authorship attribution research. N-grams were found to be an effective means of matching writings with their authors (e.g., Ishihara, 2014; Peng et al., 2016); therefore, n-gram features should be able to represent individuals' personal language patterns. However, both the fixed-effects and mixed-effects n-gram models in the current research, with the expectation of the unigram models, performed poorly in cross-context validations compared to models using other types of linguistic features. It is likely that the general length of the statements was the issue. In authorship attribution studies, the data often consisted of entire

essays or books. Perhaps short statements with a couple hundred words are not enough for establishing individual-level language patterns based on n-gram frequencies.

3.3.4 Individual Cues

Although the current research focused on the effectiveness of groups of deception cues, the regression models and the random forest models were able to highlight the most important individual cues. With regression models, the importance of individual predictors is assessed using the value and statistical significance of the regression weight coefficients. Our baseline logistic regression model suggested that deceptive statements had higher word count per sentence, and contained more words related to anxiety, risk, causation, home, and money. Truthful statements were ranked higher in analytical thinking and contained more words associated with anger, the future, motion, and space. In comparison, Newman et al. (2003) reported that negative emotion, exclusive words (replaced by differentiation in the current version of LIWC), motion, and sense were the most important psycholinguistic predictors of deception among the LIWC features. The negative emotion category of the LIWC consisted of three smaller sub-categories, which are anxiety, anger, and sadness. Unlike in Newman and colleagues' study, we used the smaller sub-categories of negative emotion in our classification models. Our results showed that anxiety and anger had opposite associations with deception. We also replaced the LIWC sense category with its sub-categories, which included see, hear, and feel. None of these features contributed significantly to the logistic regression.

Total word count is another potentially significant predictor highlighted by past studies (e.g., Van Swol & Braun, 2014; Zhou et al., 2004a). In the current research, despite the relatively large variations in the length of the statements (M = 180, SD = 57.89), this feature did not contribute significantly to the logistic regression model. Instead, word count per sentence

was estimated to have a significant weight in the model, suggesting that the participants used longer sentences while writing the deceptive statements.

A more recent study conducted by Levitan, Maredia, and Hirschberg (2018) using a corpus of transcribed face-to-face interviews reported a larger number of significant predictors of deception among the LIWC features. However, contrary to the findings of the current research, Levitan and colleagues found higher ranking in analytical thinking and higher counts of words related to the future and motion were associated with deception rather than truth-telling. These conflicting findings suggest that the associations between some LIWC features and deception may be sensitive to context. Interestingly, the LIWC summary variable – analytical thinking – was considered the most important feature among the LIWC coded predictors by multiple regression models as well as the random forest model in our research. These convergent results suggested that this particular finding of our research was not due to type II error. Based on the current deception literature one may expect this conflict to occur due to the effect of the medium of communication on deceptive language usage. That is in face-to-face deception, cognitive overload may take effect, thus reducing analytical thinking; meanwhile, in CMC deception, deceivers are not affected by cognitive load. Motivated deceivers may think even harder, thus showing an increase in analytical thinking. However, the directions of the effects of analytical thinking in Levitan and colleagues' study and our research were the opposite of this expectation. This is likely due to the low-stake settings of these studies.

Our POS logistic regression model found deceptive statements had higher counts of plural nouns, personal pronouns, adverbs, and past tense verbs; whereas truthful statements had higher counts of coordinating conjunctions, prepositional conjunctions, particles, and present non-third person verbs. This list of discriminating features has obvious similarities with the POS

profiles of informative and imaginary writings reported by Rayson and colleagues (2001), who found that imaginative writings contain more verbs, adverbs, pronouns, and pre-determiners, while informative writings contain more nouns, adjectives, prepositional conjunctions, coordinating conjunctions, and determiners. Among statistically significant predictors in our logistic regression model, personal pronouns obtained the highest regression weight. Xu and Zhao (2012) reported the same finding while using a qualitatively much different corpus as well as different classification models. These convergent finding may indicate a cross-contextually generalizable pattern of deceptive language.

Prior research and psychological theories suggested that truthful statement should contain higher frequencies of the first-person pronouns; and higher frequencies of third-person pronouns are associated with deception (e.g., DePaulo et al., 2003; McQuaid et al., 2015; Newman et al., 2003). We can see a manifestation of this pattern in the current research from the most important n-gram features in the random forest models (Table 4). The words "me," "my," and "him" and phrases that contained these words were frequently treated as the most important features by the random forest algorithm. Although each individual feature contributed very little toward the overall performances of these complex models, personal pronouns as a group made up 20% of the top 10 features of the five n-grams models reported in Table 4. In combination with the previously mentioned finding regarding the significance of overall personal pronoun usage, these findings further highlighted the importance of personal pronouns in language-based deception detection.

3.4 Strengths and Limitations

Among the multitude of deception detection studies, some had within-subject design, but few paid attentions to the within-subject effect. The current research is the first one that focused on exploring the benefit of the repeated-measure data structure. In theory, describing a story and fabricating a story should involve different psychological processes and thus influence language usage (Johnson, 1988; Johnson & Raye, 1998). However, with language-based deception detection, it can be difficult to discern whether the classification was made based on content differences between truthful statements and deceptive statements or real effects of deception on language production. The fact that many language-based deception detection studies lack generalizability across different contexts or topics (Fitzpatrick et al., 2015; Mihalcea & Strapparava, 2009) is perhaps indicative of a relatively small effect of deception compared to the effect of content differences. Modeling mixed-effects at the individual level is a way to amplify any linguistic differences that is truly due to deception. Nevertheless, in the current research we relied on out-of-the-box solutions for modeling mixed-effects in machine-learning. Given the complexity of the task, it is likely additional optimization and/or custom model regulation routines are needed to improve the performance of these models for deception detection specifically.

Furthermore, our solutions to mixed-effects modeling for the random forest classifier and the ANN were highly experimental. The BiMM tree and forest (Speiser et al., 2019) are newly developed methods with incomplete software implementations. So far, there has been no published article that documented the used of these models aside from the original paper published by their creator. Maity and Pal (2013) provided an intuitive and simple treatment of repeated-measure data for ANN; however, examples of its implementation are also sparse.

Google Scholar search returned only four studies in the field of medicine and biology that implemented this method. More information is needed in order for us to fully understand the properties of these classification models and how to optimize them for analyzing language data.

In addition to potential shortcomings of our classification models, our experimental design may also be the cause of some noise and/or biases in our data. Given the study was conducted online, there was no control of our participants' levels of engagement. Even though this limitation was partially addressed by our manual data screening process, disinterested participants may still introduce noise in subtle ways. There is also a possibility that the instructions of our writing tasks could artificially inflate the difference between truthful and deceptive statements by prompting the participants to use different words in different types of statement.

3.5 Implications and Future Research

As a highly complex human behavior, the effect of deception on language production can be subtle and easily drowned out by noise. In order to improve the accuracy of deception detection models, researchers may implement control procedures that reduce noise variance from certain sources. For instance, previous research showed large gain in deception detection accuracy as the result of limiting the context of the deception. However, many applications of lie detection, such as forensic interviews, require the lie detection model to be able to cope with diverse language data. The current research explored a potential avenue for enhancing the performances of deception detection models especially in the aspect of cross-context generalizability. We showed that modeling individual-level random-effects can be an effective way of improving cross-context generalizability of deception detection models. This approach

should enhance our ability to detect deception in cases where multiple language samples can be obtained from an individual.

Nevertheless, the deception detection scheme proposed by this dissertation is currently limited by the state of research in mixed-effects/multilevel classification models. Although the mixed-effects GLM is well documented (Fotouhi, 2003; Goldstein, 1991) in statistics literature, such models can only handle a small number of individual-level random-effects, and are limited by parametric assumptions. The choices of suitable machine-learning methods for modeling high-dimensional repeated-measure data are few. The few choices that exist all have an origin in medical and biological research. The current dissertation highlighted a need for this type of models beyond these fields.

The results of the current research also demonstrated the importance of grammatical information for the generalizability of deception detection models. In comparison, the notion that lexicon features and psycholinguistic features are context-dependent deception cues was reinforced. This suggest that future research that aims to improve the generalizability of text-based deception detection models should focus on grammatical cues over other types of cues. In the future, we shall extend the input of the mixed-effects models to include more types of grammar-related features such as dependency parse (de Marneffe et al., 2006) and probabilistic context-free grammars (Jelinek et al., 1992) discussed in section 1.4.1.

For real-world applications of the mixed-effects deception detection models, there is still an important issue that needs to be addressed: How should new individuals be added to an already trained model? Adding new individuals to a trained model will require new random-effect parameters to be added and estimated; the values of fixed-effects parameters will probably also need to be updated. A solution is conceivable for true mixed-effects models such as LASSO

GLMM and the BiMM forest; but in the current version of "mixed-effects" ANN there is no separate fixed- and random-effect parameters within the model. Therefore, an entirely different ANN architecture that clearly differentiate fixed- and random-effects will probably be needed.

For the purpose of real-world application, the lie detection accuracies achieved by the models tested in the current research were likely too low. However, the main goal of this research was reached, which is to provide evidence for the usefulness of explicitly modeling repeated-measure data in lie detection tasks. Although the current research is limited to language-based lie detection, we expect that our finding will generalize to deception detection using other types of cues. Furthermore, recent research has showed that a deep ANN (i.e. an ANN with a very large number of layers) can utilize multiple types of deception cues to achieve high levels of deception detection accuracy. For instance, Krishnamurthy and colleagues' (2018) multidimensional deep learning model was able to achieve a 96.12% classification accuracy using four different types of deception cues. Although the generalizability of Krishnamurthy and colleagues' model still needs to be investigated, their high accuracy is indicative of the potential of highly sophisticated automatic lie detection system. The current research showed that we may improve the performance of lie detection system through not only expanding the feature space, but also by taking advantage of certain data structures.

Given the fact that humans, in general, cannot distinguish lies from truth at a level better than chance (e.g., Vrij, 2008), it will be desirable to replace human judgements with highly accurate automatic lie detection systems in all real-world applications of deception detection. So far, the generalizability issue of automated lie detection systems might be a major obstacle of wide adoption. There is a trade-off between generalizability and accuracy. Currently, highly accurate systems tend to be contextually specific, thus, are limited in application. High-stake

situations such as criminal investigation require the lie detection system to be adaptive to varied contexts and has high accuracy at the same time. The current research showed that training the system to recognize different individuals is a way to improve accuracy without sacrificing generalizability. In the future, we may develop information gathering procedures to accommodate the training needs of a repeated-measure lie detection system.

3.6 Conclusion

In the current dissertation, we reviewed literature on deception detection emphasizing language-based deception detection methods that employed automated linguistic feature extraction and machine-learning classifiers. We designed a repeated-measure experiment to test the benefit of mixed-effects modeling at the level of individual in lie detection tasks using three types of classifiers, which were frequentist logistic regression, random forest/decision tree, and ANN. We also examined the effectiveness of three types of linguistic features as deception cues; these were psycholinguistic features (LIWC), grammatic features (POS), and lexical features (n-grams).

The conclusions of the current research are summarized as the following: (a) By comparing mixed-effects and fixed-effects variations of the same types of classifiers, we showed that adding individual-level random-effects improved the cross-context generalizability of all three types of classifiers used by the current study. The magnitude of the improvement is dependent upon the specific combination of the classification model and the type/s of input features. (b) The psycholinguistic features extracted by LIWC can be used to distinguish truthful statements from deceptive ones, but their cross-context generalizability is questionable. (c) POS tags by themselves are slightly less effective deception cues than psycholinguistic features; but

when combined with POS bigrams, the classification models can become highly accurate and have good cross-context generalizability. It is possible that cross-context deception detection will rely on grammatical cues like POS n-grams. (d) Lexicon n-gram features appeared to be superior deception cues than both psycholinguistic cues and POS cues, but the cross-context generalizability of n-gram based deception detection models are poor. N-gram features are probably better suited for detecting deception within a narrow context. (e) Finally, despite of the large differences in text corpora and model performances, the current research was able to reaffirm a widely held conclusion from previous deception research that the usage of personal pronouns is important for differentiating truth from lies (e.g., DePaulo et al., 2003; Fuller et al., 2009; Newman, 2003; Xu & Zhao, 2012).

The current deception detection literature highlighted a lack of generalizability among lie detection studies (Levine, 2018). The fact that individuals may have unique behavior patterns during deception is potentially a major contributor to this problem. In this dissertation, we proposed a possible modelling solution to address this particular issue. The results of our research suggest this is a direction that is worthy of further exploration. Many real-world applications of lie detection, such as forensic interview, can generate repeated-measure behavioral data and thus benefit from our data analysis approach. The current dissertation encourages deception researchers to consider using advancements in statistics and data science to augment their efforts.

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems, 26*(3), 1-34. doi:10.1145/1361684.1361685

Adams, S., & Jarvis, J. (2006). Indicators of veracity and deception: An analysis of written statements made to the police. *International Journal of Speech, Language and the Law, 13,* 1-22. doi:10.1558/sll.2006.13.1.1

Ahlmeyer, S., Heil, P., McKee, B., & English, K. (2000). The impact of polygraphy on admissions of victims and offenses in adult sexual offenders. *Sexual Abuse: Journal of Research and Treatment, 12,* 123-138. doi:10.1177/107906320001200204

Albrechtsen, J. S., Meissner, C. A., & Susa, K. J. (2009). Can intuition improve deception detection performance? *Journal of Experimental Social Psychology, 45,* 1052-1055. doi:10.1016/j.jesp.2009.05.017

Almela, A., Valencia-Garcia, R., & Cantos, P. (2012). Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the EACL2012 Workshop on Computational Approaches to Deception Detection* (pp. 15 - 22). Stroudsburg, PA: Association for Computational Linguistics.

Alonso-Quecuty, M. (1992). Deception detection and reality monitoring: A new answer to an old question? In F. Lösel, D. Bender, T. Bliesener, F. Lösel, D. Bender, & T. Bliesener (Eds.), *Psychology and law: International perspectives* (pp. 328-332). Oxford, England: Walter De Gruyter.

Alonso-Quecuty, M. L., Hernandez-Fernaud, E., & Campos, L. (1997). Child witnesses: Lying about something heard. In S. Redondo, V. Garrido, J. Perez, & R. Barberet (Eds.), *Advances in psychology and law: International contributions* (pp. 129-135). Berlin, Germany: Walter de Gruyter.

American Polygraph Association. (2010). American polygraph association standards and principles of practice. Linthicum Heights, MD: American Polygraph Association.

Anderson, D. E., DePaulo, B. M., & Ansfield, M. E. (2002). The development of deception detection skill: A longitudinal study of same-sex friends. *Personality and Social Psychology Bulletin, 28,* 536-545. doi:10.1177/0146167202287010

Arntzen, F. (1970). *Psychologie der zeugenaussage*. Goettingen, Germany: Hogrefe.

Backbier, E., Hoogstraten, J., & Terwogt-Kouwenhoven, K. M. (1997). Situational determinants of the acceptability of telling lies. *Journal of Applied Social Psychology, 27,* 1048-1062. doi:10.1111/j.1559-1816.1997.tb00286.x

Banerjee, S., & Chua, A. Y. (2014, August). *Applauses in hotel reviews: Genuine or deceptive?* Paper presented at Science and Information Conference (SAI), London, United Kingdom. London, United Kingdom: IEEE. doi:10.1109/SAI.2014.6918299

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1-48. doi:10.18637/jss.v067.i01.

Ben-Shakhar, G. (2002). A critical review of the Control Questions Test (CQT). In M.Kleiner (Ed.), *Handbook of polygraph testing* (pp. 103–126). London, UK: Academic Press.

Ben-Shakhar, G., Bar-Hillel, M., & Kremnitzer, M. (2002). Trial by polygraph: Reconsidering the use of the guilty knowledge technique in court. *Law and Human Behavior, 26,* 527-541. doi:10.1023/A:1020204005730

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology, 88,* 131-151. doi:10.1037/0021-9010.88.1.131

Ben-Shakhar, G., & Furedy, J. J. (1990). *Theories and applications in the detection of deception.* New York, NY: Springer-Verlag.

Bernstein, A. S. (1979). The orienting response as novelty and significance detector: Reply to O'Gorman. *Psychophysiology, 16,* 263-273. doi:10.1111/j.1469-8986.1979.tb02989.

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research, 36,* 423-442. doi:10.1111/j.1468-2958.2010.01382.x

Blandón-Gitlin, I., Pezdek, K., Lindsay, D. S., & Hagen, L. (2009). Criteria-based content analysis of true and suggested accounts of events. *Applied Cognitive Psychology, 23,* 901-917. doi:10.1002/acp.1504

Berliner, L., & Conte, J. R. (1993). Sexual abuse evaluations: Conceptual and empirical obstacles. *Child Abuse & Neglect, 17,* 111-125. doi:10.1016/0145-2134(93)90012-T

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Report, 10,* 214-234. doi:10.1207/s15327957pspr1003_2

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134,* 477-492. doi:10.1037/0033-2909.134.4.477

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19,* 313-329. doi:10.1002/acp.1087

Bond, C. F., Levine, T. R., & Hartwig, M. (2015). New findings in non-verbal lie detection. In P. A. Granhag, A. Vrij, B. Verschuere, P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting deception: Current challenges and cognitive approaches* (pp. 37-58). Hoboken, NJ: Wiley-Blackwell.

Bond, C. F., Omar, A., Pitre, U., Lashley, B. R., Skaggs, L. M., & Kirk, C. T. (1992). Fishy-looking liars: Deception judgment from expectancy violation. *Journal of Personality and Social Psychology, 63,* 969-977. doi:10.1037/0022-3514.63.6.969

Bond, C. F., & Robinson, M. (1988). The evolution of deception. *Journal of Nonverbal Behavior, 12,* 295-307. doi:10.1007/BF0098759

Breiman, L. (2001). Random forest. *Journal of machine-learning, 45,* 5-32. doi:10.1023/A:1010933404324

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication theory, 6,* 203-242. doi:10.1111/j.1468-2885.1996.tb00127.x

Buller, D. B., Stiff, J. B., & Burgoon, J. K. (1996). Behavioral adaptation in deceptive transactions: Fact or fiction: Reply to Levine and McCornack. *Human Communication Research, 22,* 589-603. doi:10.1111/j.1468-2958.1996.tb00381.x

Chomsky, N. (1999). On the nature, use, and acquisition of language. In W. Ritchie & T. Bhatia (Eds.). *Handbook of child language acquisition* (pp. 33-54). San Diego, CA: Academic Press.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine, 9,* 48-57. doi:10.1109/MCI.2014.2307227

Carroll, D. (1988). How accurate is polygraph lie detection? In A. Gale (Ed.), *The polygraph test. Lies, truth and science* (pp. 19-28). Oxford, England: British Psychological Society.

Chollet, F. (2015) *keras*. GitHub. Retrieved from https://github.com/fchollet/keras

Cole, T. (2001). Lying to the one you love: The use of deception in romantic relationships. *Journal of Social and Personal Relationships, 18,* 107-129. doi:10.1177/0265407501181005

Cook, B. L., Progovac, A. M., Chen, P., Mullin, B., Hou, S., & Baca-Garcia, E. (2016). Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Computational & Mathematical Methods in Medicine,* 1-8. doi:10.1155/2016/8708434

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20,* 273-297. doi:10.1023/A:1022627411411

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference* (pp. 95-102). New York, NY: ACM.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities, 36*(2), 223-254. doi:10.1023/A:1014348124664

Dangti, P. (2017). *Statistics for machine learning*. Birmingham, UK: Packt.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., ... & Langleben, D. D. (2005). Classifying spatial patterns of brain activity with machine

learning methods: application to lie detection. *Neuroimage, 28,* 663-668. doi:10.1016/j.neuroimage.2005.08.009

Davis, R. C. (1961). Physiological responses as a means of evaluating information. In A. Biderman & H. Zimmer (Eds.), *Manipulation of human behavior* (pp. 142-168). New Yok: John Wiley & Sons, Inc.

Davis, M., Markus, K. A., Walters, S. B., Vorus, N., & Connors, B. (2005). Behavioral cues to deception vs topic incriminating potential in criminal confessions. *Law and human Behavior, 29,* 682-704. doi:10.1007/s10979-005-7370-z

DeCicco, A. J., & Schafer, J. R. (2015). Grammatical differences between truthful and deceptive narratives. *Applied Psychology in Criminal Justice, 11,* 75-92.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44,* 837-845. doi:10.2307/2531595

de Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). *Generating typed dependency parses from phrase structure trees*. Paper presented at the 5th edition of the International Conference on Language Resources and Evaluation, Genoa, Italy.

DePaulo, B. M., & Bell, K. L. (1996). Truth and investment: Lies are told to those who care. *Journal of Personality and Social Psychology, 71,* 703-716. doi:10.1037/0022-3514.71.4.703

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70,* 979-995. doi:10.1037/0022-3514.70.5.979

DePaulo, B. M., & Kirkendol, S. E. (1989). The motivational impairment effect in the communication of deception. In J. C. Yuille (Eds.), *Credibility Assessment* (pp. 51-70). New York, NY: Kluwer Academic/Plenum Publishers.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129,* 74-118. doi:10.1037/0033-2909.129.1.74

DePaulo, B. M., & Morris, W. L. (2004). Discerning lies from truths: Behavioral cues to deception and the indirect pathway of intuition. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 15–40). Cambridge, UK: Cambridge University Press.

Dewaele, J. M., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning, 49,* 509-544. doi:10.1111/0023-8333.00098

Driscoll, L. N. (1994). A validity assessment of written statements from suspects in criminal investigations using the scan technique. *Police Studies, 17,* 77-88.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32,* 407-499. doi:10.1214/009053604000000067

Ekman, P. (1985). *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York, NY: W. W. Norton & Company, Inc.

Ekman, P. (2009). Lie catching and microexpression. In C. Martin (Ed.), *The philosophy of deception* (pp. 118-133). Oxford, UK: Oxford University Press.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry, 32,* 88–106. doi:10.1521/00332747.1969.11023575

Ekman, P., & Friesen, W. V. (1978). *The facial action coding system (FACS)*. Palo Alto, California: Consulting Psychologists Press.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46,* 913-920. doi:10.1037/0003-066X.46.9.913

Ekman, P., & Rosenberg, E. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system* (FACS, 2nd ed.). New York, NY: Oxford University Press.

Elaad, E. (1990). Detection of guilty knowledge in real-life criminal investigations. *Journal of Applied Psychology, 75,* 521-529. doi:10.1037/0021-9010.75.5.521

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology, 75,* 757-767. doi:10.1037/0021-9010.77.5.757

Ennis, E., Vrij, A., & Chance, C. (2008). Individual differences and lying in everyday life. *Journal of Social and Personal Relationships, 25,* 105-118.

Feldman, R. S., Forrest, J. A., & Happ, B. R. (2002). Self-presentation and verbal deception: Do self-presenters lie more? *Basic and applied social psychology, 24,* 163-170. doi:10.1207/S15324834BASP2402_8

Fisher, R. P., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview.* Springfield, IL: Charles C Thomas Publisher.

Fitzpatrick, E., Bachenko, J., & Fornaciari, T. (2015). Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies, 8*(3), 1-119. doi:10.2200/S00656ED1V01Y201507HLT029

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods, 50,* 2016-2034. doi:10.3758/s13428-017-0971-x

Foley, M. A., & Johnson, M. K. (1985). Confusions between memories for performed and imagined actions: A developmental comparison. *Child development, 56,* 1145-1155. doi:10.2307/1130229

Fornaciari, T., & Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artificial intelligence and law, 21,* 303-340. doi:10.1177/0265407507086808

Fotouhi, A. R. (2003). Comparisons of estimation procedures for nonlinear multilevel models. *Journal of Statistical Software, 8*(9), 1-39. doi:10.18637/jss.v008.i09

Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology, 72,* 1429-1439. doi:10.1037/0022-3514.72.6.1429

Frank, M. G., Paolantonio, N., Feeley, T. H., & Servoss, T. J. (2004). Individual and small group accuracy in judging truthful and deceptive communication. *Group Decision and Negotiation, 13,* 45-59. doi:10.1023/B:GRUP.0000011945.85141.af

Freud, S. (1959). *Collected papers*. New York, NY: Basic Books.

Fuller, C. M., Biros, D. P., & Wilson, R. L. (2009). Decision support for determining veracity via linguistic-based cues. *Decision Support System, 46,* 695-703. doi:10.1016/j.dss.2008.11.001

Fusiliera, H. D., Montes-y-Gómezb, M., Rossoc, P., & Cabreraa, R. G. (2015). Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management, 51,* 433-443. doi:10.1016/j.ipm.2014.11.001

Gill, A. J., & Oberlander, J. (2002, January). Taking care of the linguistic features of extraversion. In *Proceedings of the Cognitive Science Society Volume 24*. (n.p.): Cognitive Science Society, Inc.

Gladwell, M. (2005). *Blink: The power of thinking without thinking.* New York City, NY: Back Bay Books.

Gokhman, S., Hancock, J., Prabhu, P., Ott, M., & Cardie, C. (2012). In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 23-30). Stroudsburg, PA: Association for Computational Linguistics.

Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika, 78,* 45-51. doi:10.2307/2336894

Gozna, L., & Babooran, N. (2004). Non-traditional interviews: Deception in a simulated customs baggage search. In A. Czerederecka, T. Jaskiewicz-Obydzinska, R. Roesch, & J. Wojcikiewicz (Eds.), *Forensic psychology and law* (pp. 153-161). Krakow, Poland: Institute of Forensic Research Publishers.

Groll, A. (2017). *glmmLasso: Variable selection for generalized linear mixed models by L1-penalized estimation*. Retrieved from https://cran.r-project.org/web/packages/glmmLasso/glmmLasso.pdf

Hajjem, A., Larocque, D., & Bellavance, F. (2017). Generalized mixed effects regression trees. *Statistics & Probability Letters, 126,* 114-118. doi:10.1016/j.spl.2017.02.033

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation, 84,* 1313-1328. doi:10.1080/00949655.2012.741599

Hallner, D., & Hasenbring, M. (2004). Classification of psychosocial risk factors (yellow flags) for the development of chronic low back and leg pain using artificial neural network. *Neuroscience Letter, 361*(1-3), 151-154. doi:10.1016/j.neulet.2003.12.107

Hancock, J. T., Thom-Santelli, J., & Ritchie, T. (2004). Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 129-134). New York, NY: ACM. doi:10.1145/985692.985709

Hancock, J. T., Woodworth, M. T., & Goorha, S. (2010). See no evil: The effect of communication medium and motivation on deception detection. *Group Decision & Negotiation, 19,* 327-343. doi:10.1007/s10726-009-9169-7

Hancock, J. T., Woodworth, M. T., & Porter, S. (2013). Hungry like the wolf: A word-pattern analysis of the language of psychopaths. *Legal and Criminological Psychology, 18,* 102-114. doi:10.1111/j.2044-8333.2011.02025.x

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29-36. doi:10.1148/radiology.143.1.7063747

Hao, P., Chen, X., Cheng, N., Chandramouli, R., & Subbalakshmi, K. P. (2011, August). Adaptive context modeling for deception detection in emails. In *International Workshop on Machine-learning and Data Mining in Pattern Recognition* (pp. 458-468). Berlin, Germnay: Springer.

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgements. *Psychological Bulletin, 137,* 643-659. doi:10.1037/a0023589

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks*. Cambridge, MA: The MIT Press.

Haythornthwaite, C., & Wellman, B. (2002). The Internet in everyday life: An introduction. In B. Wellman & C. Haythornthwaite (Eds.), *The Internet in everyday life* (pp. 3-41). Oxford, UK: Blackwell.

He, Q., Veldkamp, B. P., Glas, C. A., & de Vries, T. (2015). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment, 24,* 157-172. doi:10.1177/107319111560255

Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley & Sons.

Henig, R. M. (2006, February 5). Looking for the lie. *New York Times*. Retrieved April 10, 2016, from http://www.nytimes.com.

Hernandez-Fernaud, E., & Alonso-Quecuty, M. L. (1997). The cognitive interview and lie detection: A new magnifying glass for Sherlock Holmes? *Applied Cognitive Psychology, 11,* 55-68. doi:10.1002/(SICI)1099-0720(199702)11:1<55::AID-ACP423>3.0.CO;2-G

Heydari, A., Tavakolia, M., Salima, N., & Heydari, Z. (2015). Detection of review spam: A survey. *Expert Systems with Applications, 42,* 3634-3642. doi:10.1016/j.eswa.2014.12.029

Ishihara, S. (2014). A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *International Journal of Speech, Language & the Law, 21,* 23-49. doi:10.1558/ijsll.v21i1.23

Jack, R. E., Caldara, R., & Schyns, P. G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General, 141,* 19-25. doi: 10.1037/a0023463

Jarrold, W. L., Peintner, B., Yeh, E., Krasnow, R., Javitz, H. S., & Swan, G. E. (2010, August). Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer's disease. In Y. Y. Yao et al. (Eds.), *International Conference on Brain Informatics* (pp. 299-307). Berlin, Germany: Springer.

Jelinek, F., Lafferty, J. D., & Mercer, R. L. (1992). Basic methods of probabilistic context free grammars. In P. Laface & R. De Mori (Eds.), *Speech recognition and understanding* (pp. 345-360). Berlin, Germany: Springer.

Johnson, M. K. (1988). Reality monitoring: An experimental phenomenological approach. *Journal of Experimental Psychology: General, 117,* 390-394. doi:10.1037/0096-3445.117.4.390

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88,* 67-85. doi:10.1037/0033-295X.88.1.67

Johnson, M. K., & Raye, C. L. (1998). False memories and confabulation. *Trends in Cognitive Sciences, 2*(4), 137-146. doi:10.1016/S1364-6613(98)01152-8

Kashy, D. A., & DePaulo, B. M. (1996). *Who lies? Journal of Personality and Social Psychology, 70,* 1037-1051. doi:10.1037/0022-3514.70.5.1037

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003, August*). N-gram-based author profiles for authorship attribution.* Paper presented at the Conference of Pacific Association for Computational Linguistics, Halifax, NS, Canada. Halifax, NS, Canada: Computer Science Dept. at Dalhousie University.

Kerr, P. (1990). *The Penguin book of lies*. New York, NY: Penguin.

Knapp, M. L., Hart, R. P., & Dennis, H. S. (1974). An exploration of deception as a communication construct. *Human Communication Research, 1,* 15-29. doi:10.1111/j.1468-2958.1974.tb00250.x

Köhnken, G. (2004). Statement validity analysis and the 'detection of the truth'. In P. A. Granhag & L. A. Stromwall (Eds.), *Detection of deception in forensic contexts* (pp. 41-63). Cambridge, England: Cambridge University Press.

Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biological psychiatry, 58,* 605-613. doi:10.1016/j.biopsych.2005.07.040

Krishnamurthy, G., Majumder, N., Poria, S., & Cambria, E. (2018). A deep learning approach for multimodal deception detection. *arXiv preprint arXiv:1803.00344*.

Kulkofsky, S. (2008). Credible but inaccurate: Can Criterion-Based Content Analysis (CBCA) distinguish true and false memories? In M. J. Smith (Ed.), *Child sexual abuse: Issues and challenges* (pp. 21-42). Hauppauge, NY: Nova Science Publishers.

Lafferty, J., & Eady, P. (1974). *The Desert Survival Problem*. Plymouth, MI: Experimental Learning Methods.

Larson, J. A. (1932/1969). *Lying and its detection. A study of deception and deception tests*. Montclair, NJ: Patterson Smith.

Levine, T. R. (2010). A few transparent liars. In C. Salmon (Ed.), *Communication yearbook, 34* (pp. 40–61). Thousand Oaks, CA: Sage.

Levine, T. R. (2018). Scientific evidence and cue theories in deception research: reconciling findings from meta-analyses and primary experiments. *International Journal of Communication, 12,* 2461-2479.

Levine, T. R., Anders, L. N., Banas, J., Baum, K. L., Endo, K., Hu, A. S., & Wong, N. H. (2000). Norms, expectations, and deception: A norm violation model of veracity judgments. *Communication Monographs, 67,* 123-137. doi:10.1080/03637750009376500

Levine, T. R., Kim, R. K., Park, H. S., & Hughes, M. (2006). Deception detection accuracy is a predictable linear function of message veracity base-rate: A formal test of Park and Levine's probability model. *Communication Monographs, 73,* 243-260. doi:10.1080/03637750600873736

Levine, T. R., Serota, K. B., Carey, F., & Messer, D. (2013). Teenagers lie a lot: A further investigation into the prevalence of lying. *Communication Research Reports, 30,* 211-220. doi:10.1080/08824096.2013.806254

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

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News 2*(3), 18-22.

Lin, K. H. Y., Yang, C., & Chen, H. H. (2008, December). Emotion classification of online news articles from the reader's perspective. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 220-226). Washington, DC: IEEE Computer Society. doi:10.1109/WIIAT.2008.197

Long, N. H., Nghia, P. H. T., & Vuong, N. M. (2014). Opinion spam recognition method for online reviews using ontological features. *Tạp chí Khoa học, 61*, 44-59.

Loper, E., & Bird, S. (2002). NLTK: The Naturel Language Toolkit. In *Proceedings of the ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics - Vol 1* (pp. 163-70). Stroudsburg, PA: Association for Computational Linguistics.

Luhn, H. P. (1966). 11 Keyword-in-Context Index for Technical Literature (KWIC Index). *Readings in automatic language processing, 1*, 159-167. doi:10.1002/asi.5090110403

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology, 44,* 258-262. doi:10.1037/h0044413

MacLaren, V. V. (2001). A quantitative review of the Guilty Knowledge Test. *Journal of applied psychology, 86,* 674-683. doi:10.1037/0021-9010.86.4.674

Mairesse, F., & Walker, M. (2006, January). Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the Cognitive Science Society Volume 28* (pp. 543-548). (n.p.): Cognitive Science Society, Inc.

Mann, S., Vrij, A., & Bull, R. (2002). Suspects, lies, and videotape: An analysis of authentic high-stake liars. *Law and Human Behavior, 26,* 365-376. doi:10.1023/A:1015332606792

Manzanero, A. L., & Digest, M. (1996). Effects of preparation on internal and external memories. In G. Davies, S. Lloyd-Bostock, M. McMurran & C. Wilson (Eds.), *Psychology, Law, and Criminal Justice: International developments in research and practice* (pp. 56-63). Berlin, Germany: de Gruyter.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law, 11,* 99-122. doi:10.1080/10683160410001726356

Maity, T. K., & Pal, A. K. (2013). Subject specific treatment to neural networks for repeated measures analysis. In *Proceedings of the International Multi-Conference of Engineers and Computer Scientists* (Vol. 1). Hong Kong, China: International Association of Engineers.

McQuaid, S. M., Woodworth, M., Hutton, E. L., Porter, S., & ten Brinke, L. (2015). Automated insights: Verbal cues to deception in real-life high-stakes lies. Psychology, Crime & Law, 21, 617-631. doi:10.1080/1068316X.2015.1008477

Mihalcea, R., & Strapparava, C. (2009, August). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 309-312). Stroudsburg, PA: Association for Computational Linguistics.

Mitchell, R.W. (1986). A framework for discussing deception. In R. W. Mitchell & N. S. Thompson (Eds.), *Deception: Perspectives on human and nonhuman deceit* (pp. 3-40). Albany, NY: State University of New York Press.

Murphy, K. (2012). *Machine-learning: A probabilistic perspective.* Cambridge, MA: The MIT Press.

National Research Council. 2003. *The Polygraph and Lie Detection.* Washington, DC: The National Academies Press. doi:10.17226/10420

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

Newton, P., Reddy, V., & Bull, R. (2000). Children's everyday deception and performance on false-belief tasks. *British Journal of Developmental Psychology, 18,* 297–317. doi:10.1348/026151000165706

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association, 18,* 544–551. doi:10.1136/amiajnl-2011-000464

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Does the truth come out in the writing? Scan as a lie detection tool. *Law and Human Behavior, 36,* 68-76. doi:10.1037/h0093965

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and Human Behavior, 40,* 440-457. doi:10.1037/lhb0000193

Orimaye, S. O., Wong, J. S.-M., Golden, K. J., Wong, C. P., & Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics, 18,* 34. doi:10.1186/s12859-016-1456-0

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, June). *Finding deceptive opinion spam by stretch of the imagination.* Paper presented at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR.

Ott, M., Cardie, C., & Hancock, J. T. (2013, June). *Negative deceptive opinion spam*. Paper presented at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA.

Parker, J. F. (1995). Age differences in source monitoring of performed and imagined actions on immediate and delayed tests. *Journal of Experimental Child Psychology, 60*, 84-101. doi:10.1006/jecp.1995.1032

Peng, J., Choo, K. R., & Ashman, H. (2016). Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Network & Computer Applications, 70*, 171-182. doi:10.1016/j.jnca.2016.04.001

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10*(3), 90-93. doi:10.1111/1467-8721.00123

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547-577. doi:10.1146/annurev.psych.54.101601.145041

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77,* 1296-1312. doi:10.1037/0022-3514.77.6.1296

Pinker, S. (2002). *The Blank slate: The modern denial of human nature.* London, UK: Allen Lane.

Pontari, B. A., & Schlenker, B. R. (2006). Helping friends manage impressions: We like helpful liars but respect nonhelpful truth tellers. *Basic and Applied Social Psychology, 28,* 177-183. doi:10.1207/s15324834basp2802_7

Porter, S., & ten Brinke, L. (2008). Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science, 19,* 508-514. doi:10.1111/j.1467-9280.2008.02116.x

Porter, S., ten Brinke, L, & Wallace, B. (2012). Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity. *Journal of Nonverbal Behavior, 36,* 23-37. doi:10.1007/s10919-011-0120-7

Porter, S., & Woodworth, M. (2007). "I'm sorry I did it… but he started it": A comparison of the official and self-reported homicide descriptions of psychopaths and non-psychopaths. *Law and human behavior, 31,* 91-107. doi:10.1007/s10979-006-9033-0

Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior, 20,* 443-458. doi:10.1007/BF01498980

Porter, S., Yuille, J. C., & Lehman, D. R. (1999). The nature of real, implanted, and fabricated memories for emotional childhood events: Implications for the recovered memory debate. *Law and Human Behavior, 23,* 517-537. doi:10.1023/A:1022344128649

Quinlan, J. R. (1986). Induction of decision tree. *Journal of machine-learning, 1,* 81-106. doi:10.1023/A:1022643204877

Raskin, D. C. (1986). The polygraph in 1986: Scientific, professional, and legal issues surrounding acceptance of polygraph evidence. *Utah Law Review, 29,* 29-74.

Raskin, D. C., & Honts, C. R. (2002). The comparison question test. In M. Kleiner & M. Kleiner (Eds.), *Handbook of polygraph testing* (pp. 1-47). San Diego, CA: Academic Press.

Raskin, D. C., & Honts, C. R., & Kircher, J. C. (1997). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (pp. 565-582). St Paul, MN: West Law.

Raskin, D. C., & Kircher, J. C. (2014). Validity of polygraph techniques and decision methods. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility assessment: Scientific research and applications* (pp. 63-129). San Diego, CA: Elsevier Academic Press. doi:10.1016/B978-0-12-394433-7.00003-8

Rayson, P., Wilson, A., & Leech, G. (2001). Grammatical word class variation within the British National Corpus sampler. *Language and Computers, 36,* 295-306.

Reddy, V. (2007). Getting back to the rough ground: Deception and 'social living'. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362,* 621-637. doi:10.1098/rstb.2006.1999

Reid, J. E. (1947). A revised questioning technique in lie-detection tests. *Journal of Criminal Law & Criminology, 37,* 542-547. doi:10.2307/1138979

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences, 385*, 213-224. doi:10.1016/j.ins.2017.01.015

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information Storage And organization in the brain. *Psychological Review, 65,* 386–408. doi:10.1037/h0042519

Rubin, V. L. (2010). On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the Association for Information Science and Technology, 47, 1-10*. doi:10.1002/meet.14504701124

Sánchez-Junquera, J., Villaseñor-Pineda, L., Montes-y-Gómez, M., & Rosso, P. (2018, September). Character N-Grams for Detecting Deceptive Controversial Opinions. In: Bellot P. et al. (eds) *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2018. Lecture Notes in Computer Science, vol 11018*. Springer, Cham.

Santtila, P., Roppola, H, & Niemi, P. (1999). Assessing the truthfulness of witness statements made by children (age 7-8, 10-11, and 13-14) employing scales derived from Johnson and Raye's model of Reality Monitoring. *Expert Evidence, 6,* 273-289. doi:10.1023/A:1008930821076

Sapir, A. (1987/2000). *The LSI course on scientific content analysis (SCAN).* Phoenix, AZ: Laboratory for Scientific Interrogation.

Saxe, L., Dougherty, D., & Cross, T. (1985). The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist, 40,* 355-366. doi:10.1037/0003-066X.40.3.355

Schein, E. H. (2004). Learning when and how to lie: A neglected aspect of organizational and occupational socialization. *Human Relations, 57,* 260-273. doi:10.1177/0018726704043270

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine-learning, 86,* 169-207. doi:10.1007/s10994-011-5258-3

Serota, K. B., & Levine, T. R. (2015). A few prolific liars: Variation in the prevalence of lying. *Journal of Language and Social Psychology, 34,* 138-157. doi:10.1177/0261927X14528804

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The prevalence of lying in America: Three studies of self-reported lies. *Human Communication Research, 36,* 2-25. doi:10.1111/j.1468-2958.2009.01366.x

Sip, K. E., Lynge, M., Wallentin, M., McGregor, W. B., Frith, C. D., & Roepstorff, A. (2010). The production and detection of deception in an interactive game. *Neuropsychologia, 48*(12), 3619-3626. doi:10.1016/j.neuropsychologia.2010.08.013

Smith, N. (2001). *Reading between the lines: An evaluation of the scientific content analysis technique (SCAN).* London, UK: Home Office, Policing and Reducing Crime Unit, Research, Development and Statistics Directorate.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis, 2nd ed*. Sage Publications.

Sokolov, E. N. (1963). *Perception and the conditioned reflex.* New York, NY: Macmillan.

Speiser, J. L., Wolf, B. J., Chung, D., Karvellas, C. J., Koch, D. G., & Durkalski, V. L. (2019). BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. *Chemometrics and Intelligent Laboratory Systems, 185,* 122-134. doi:10.1016/j.chemolab.2019.01.002

Sporer, S. L. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experiences events. *Applied Cognitive Psychology, 11,* 373-397. doi:10.1002/(SICI)1099-0720(199710)11:5<373::AID-ACP461>3.0.CO;2-0

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, And Law, 13,* 1-34. doi:10.1037/1076-8971.13.1.1

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology, 60,* 538-556. doi:10.1002/asi.21001

Stiff, J. B., Kim, H. J., & Ramesh, C. N. (1992). Truth biases and aroused suspicion in relational deception. *Communication Research, 19,* 326-345. doi:10.1177/009365092019003002

Sweetser, E. (1987). The definition of "lie": An examination of the folk models underlying a semantic prototype. In D. Hollard & N. Quinn (Eds.), *Cultural models in language and thought.* New York, NY: Cambridge University Press.

Talwar, V., & Lee, K. (2008). Social and cognitive correlates of children's lying behavior. *Child Development, 79,* 866-881. doi:10.1111/j.1467-8624.2008.01164.x

Tandon, R., Adak, S., & Kaye, J. A. (2006). Neural networks for longitudinal studies in Alzheimer's disease. *Artificial intelligence in medicine, 36,* 245-255. doi:10.1016/j.artmed.2005.10.007

ten Brinke, L., Lee, J. J., & Carney, D. R. (2019). Different physiological reactions when observing lies versus truths: Initial evidence and an intervention to enhance accuracy. *Journal of Personality and Social Psychology, 117,* 560-578. doi:10.1037/pspi0000175

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science, 25,* 1098-1105. doi:10.1177/0956797614524421

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58,* 267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

Trankell, A. (1972). *Reliability of evidence*. Stockholm, Sweden: Beckmans.

Trivers, R. (1985). *Social evolution*. Menlo Park, CA: Benjamin/Cummings Publishers.

Trovillo, P. V. (1939). A history of lie detection. *Journal of Criminal Law and Criminology, 29,* 848-881. doi:10.2307/1136489

Turing, A. M. (1950/2007). Computing machinery and intelligence. In B. Gertler, L. Shapiro, B. Gertler, & L. Shapiro (Eds.), Arguing about the mind (pp. 471-494). New York, NY: Routledge/Taylor & Francis Group.

Undeutsch, U. (1982). Statement reality analysis. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 27-56). Deventer, The Netherlands: Kluwer.

Wilson, A. E., Smith, M. D., & Ross, H. S. (2003). The nature and effects of young children's lies. *Social Development, 12,* 21-45. doi:10.1111/1467-9507.00220

Wyman, J., Foster, I., Lavoie, J., Tong, D., & Talwar, V. (2017). Detecting children's false allegations and recantations of a crime. *Psychology, Crime & Law.* Advance online publication. doi:10.1037/0033-2909.134.4.477

Van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M., & Neijt, A. (2005). New machine-learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics, 12,* 65-77. doi:10.1080/09296170500055350

Van Swol, L. M., & Braun, M. T. (2014). Communicating deception: Differences in language use, justifications, and questions for lies, omissions, and truths. *Group Decision and Negotiation, 23,* 1343-1367. doi:10.1007/s10726-013-9373-3

Vrij, A. (2000). *Detecting lies and deceit. The psychology of lying and the implications for professional practice*. Chichester, England: Wiley.

Vrij, A. (2005). Criteria-Based Content Analysis: A Qualitative Review of the First 37 Studies. *Psychology, Public Policy, and Law, 11,* 3-41. doi:10.1037/1076-8971.11.1.3

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Hoboken, NJ: John Wiley & Sons.

Vrij, A. (2015). Verbal lie detection tools: Statement validity analysis, reality monitoring and scientific content analysis. In P. A. Granhag, A. Vrij, B. Verschuere, P. A. Granhag, A. Vrij, & B. Verschuere (Eds.*), Detecting deception: Current challenges and cognitive approaches* (pp. 3-35). Hoboken, NJ: Wiley-Blackwell.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. *Human Communication Research, 30,* 8-41. doi:10.1111/j.1468-2958.2004.tb00723.x

Vrij, A., Edward, K., & Bull, R. (2001). Stereotypical verbal and nonverbal responses while deceiving others. *Personality and Social Psychology Bulletin, 27,* 899-909. doi:10.1177/0146167201277012

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal behavior, 24,* 239-263. doi:10.1023/A:1006610329284

Vrij, A., Floyd, M., & Ennis, E. (2003). Telling lies to strangers or close friends: its relationship with attachment style. In S. Shohov (Ed.), *Advances in psychology research V. 20* (pp. 61-74). New York, NY: Nova Science Publishers.

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior, 31,* 499-518. doi:10.1007/s10979-006-9066-4

Vrij, A., Nunkoosing, K., Paterson, B., Oosterwegel, A., & Soukara, A. (2002). Characteristics of secrets and the frequency, reasons and effects of secrets keeping and disclosure. *Journal of Community & Applied Social Psychology, 12,* 56-70. doi:10.1002/casp.652

Weinberger, S. (2010). Airport security: Intent to deceive? *Nature, 465,* 412-415. doi:10.1038/465412a

Woodworth, M., Hancock, J. T., & Goorha, S. (2005, January). *The motivational enhancement effect: Implications for our chosen modes of communication in the 21st century*. Paper presented at the 38th Annual Hawaii International Conference on System Sciences. Big Island, HI. doi:10.1109/HICSS.2005.607

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12,* 1100-1122. doi:10.1177/1745691617693393

Yuille, J. C. (1989). Preface. In J. C. Yuille (Ed.), *Credibility assessment.* Dordrecht, The Netherlands: Kluwer Academic.

Xu, Q., & Zhao, H. (2012, December). Using deep linguistic features for finding deceptive opinion spam. In *Proceedings of COLING 2012: Posters* (pp. 1341-1350). Mumbai, India: The COLING 2012 Organizing Committee.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004a). Automating linguistics-based cues for detecting deception in text-based asynchronous computer mediated communications. *Group Decision and Negotiation, 13,* 81–106. doi:10.1023/B:GRUP.0000011944.62889.6f

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker, J. F. (2004b). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20,* 139–166. doi:10.1080/07421222.2004.11045779

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In U. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 1-59). New York, NY: Academic Press.

Appendices

Appendix A Qualification Questions and Demographics

A.1 Study Qualification Questions

(appear at the beginning of the study)

The following questions will help us determine if you are a member of our target demographics. After you submit your answers to these questions, if you meet our inclusion criteria, you will be directed to read our informed consent form. If you do not meet our inclusion criteria, we thank you for taking an interest in our research, and your Internet browser will exit this study automatically.

1. What is your age, in years? _____

2. Choose the option that best describes you.

□ I am a native English speaker.

□ I am bilingual/multilingual and was raised speaking English as a primary language.

□ I learned to speak English as a second language.

3. In this experiment, you will be asked to describe details of what you see and hear from a series of videos, some of which have low video and audio quality (e.g., cell phone camera videos and security camera videos). Do you think you are able to see and hear well enough to complete these tasks?

□ Yes

□ No

A.2 Demographic Questions

(appear after consent)

1. What is your Gender?

- Male
- Female
- Others

2. To which ethnic group do you feel you most belong?

- Caucasian
- Asian
- African
- Arabian/Middle Eastern
- Aboriginal/First Nations
- Latino/Hispanic
- Other (please specify) _____

Appendix B Study Instructions

B.1 General Instructions

In this study, you will watch four video recordings of real-life events (A, B, C, and D) in a random order. After watching each video, you will be instructed to write a truthful statement and a deceptive statement (eight statements in total). Please read the instructions after each video carefully.

Please note, your SONA credit award is conditioned upon a sincere attempt at completing the study. Providing statements that are too short (far below 150 words), not following study instruction, or providing nonsensical responses will result in deduction of the credits.

To encourage you to provide higher quality responses, we offer a prize of a $100 gift card (from an online vendor of the winner's choice), based on diligence and the appearance of truthfulness of your statements. Once our participant quota (200 including dropouts) is reached, a researcher will review all the statements, and decide on a winner. The decision will be made subjectively, considering all eight statements that you submit; therefore, in order to win, you need to be consistent with the quality of your responses.

We recommend that you use a word processing program, such as Microsoft Word, on your computer to write these statements, then copy and paste them into their respective fields. Please correct typos.

Given the length of this study, we do not recommend you complete it in one session. During the study, every time you click on the arrow to continue, your progress will be saved. If you close the study and come back later, you will return to where you left. Please note: In order to receive SONA credits for a course, you will need to complete the study before the course ends.

B.2 Instructions for Each Tasks

Video A
(Truthful)
Please assume you are the only witness of the incident, in 150 to 300 words, provide a truthful and accurate description of what had happened. **[This statement should be written in third person.]**

(Deceptive)
Please assume you are the only witness of the incident, and a friend of the robber. In 150 to 300 words, provide a false description of the incident that minimizes the robber's crime as much as possible. (The only established truth is that the robber took money from the store.) **[This statement should be written in third person.]**

Video B
(Truthful)
Please assume the perspective of the motorcyclist, in 150 to 300 words, describe the incident in this video as truthfully and accurately as possible. **[This statement should be written in first person.]**

(Deceptive)
Please assume the perspective of the motorcyclist, in 150 to 300 words, describe the incident in a way that justifies and/or minimizes and/or deny any inappropriate behavior of your own. (No other witness is available to contradict your statement.) **[This statement should be written in first person.]**

Video C
(Truthful)
Please assume the perspective of the thief, in 150 to 300 words, describe what happened in this video as truthfully and accurately as possible. **[This statement should be written in first person.]**

(Deceptive)
Please assume the perspective of the thief, write a false statement about what happened. For example, you may try to minimize or deny your own crime, and/or exaggerate the owner's use of physical force. (150 to 300 words) **[This statement should be written in first person.]**

Video D
(Truthful)
Please assume the perspective of the person who was holding the camera, in 150 to 300 words, describe what happened as truthfully and accurately as you can. (The person who was holding the camera appears to be a friend of the customer in this video. You may choose to either ignore or acknowledge this detail.) **[This statement should be written in third person.]**

(Deceptive)
Please assume the perspective of the person who was holding the camera, in 150 to 300 words, lie about the situation in favor of the customer as much as possible, for example, exaggerate/make up wrong doings on the part of the security officer. (No other witness is available to contradict your statement.) **[This statement should be written in third person.]**

Appendix C Informed Consent Form

You are invited to participate in an online study being conducted by psychology researchers at the University of British Columbia Okanagan Campus. This study, titled "Classification of real and falsified narratives in a repeated-measure text corpus," is intended to help us investigate changes in people's language pattern during deception.

This research is being conducted by Ran Wei (MA, Experimental Psychology), and Dr. Brian O'Connor (Professor, psychology) at UBC Okanagan, Department of Psychology.

Contact Information:
Principal Investigator: Brian O'Connor
Address: Psychology Department, Irving K. Barber School of Arts and Science, University of British Columbia Okanagan, 1147 Research Road Kelowna, BC, V1V 1V7.
Email: brian.oconnor@ubc.ca, phone: (250)807-9636.

Co-Investigator (primary contact): Ran Wei
Address: Psychology Department, Irving K. Barber School of Arts and Science, University of British Columbia Okanagan, 1147 Research Road Kelowna, BC, V1V 1V7.
Email: ran.wei@ubc.alumni.ca, phone: (778) 821-0203.

What you will be asked to do in this study?
If you chose to participate in this study, we will ask you to watch four short videos (2-3 minutes). These videos depict real-life crimes and incidents of moderate violence (If any of these videos makes you uncomfortable, you may stop or withdraw from the study at any time). After watching each video, you will write a truthful statement and a deceptive statement about the content of the video. Each statement should be approximately 150 to 300 words. It will take approximately 2 hours of your time in total to complete this study. However, you don't need to complete all the writing tasks in one session. You will have the option to save your progress during the study and resume at a later time (you will need to complete the study before the end your semester in order to receive SONA credits).

Once you have agreed to the terms of this study, you will be instructed to create a unique participant ID. Please make sure you write down this ID or choose an ID that you are unlikely to forget (e.g., jd123). This ID will help us trace your responses if you wish to withdraw your responses after you have submitted them.

UBC Okanagan students may receive 2 SONA credits toward an eligible psychology course for completing the study. Please note, SONA regulations only allow 2nd, 3rd and 4th year students to earn a maximum of 1.5 credits from online studies. In order to receive the 2 credits in full, these students will need to ask special permission from their instructor. First year students are exempted from this rule.

In addition, the SONA credit award is conditioned upon a sincere attempt at completing the writing tasks. Providing statements that are too short (far below 150 words), not following study instruction, or providing nonsensical responses will result in deduction of the credits.

We will also award a $100 (Canadian) gift card to the participant who writes the most convincing statements. The winner of this competition will be selected by consulting both the results of our statistical analyses and the judgments of the researchers.

Who can participate in the study?

In order to participate you need to be 18 years of age or older, and a native English speaker. You also need to be able to see and hear details from Internet videos with ease.

Confidentiality

We do **NOT** require you to provide your name, address, or any personal identifiable information to participate in our study. However, please note that we cannot 100% guarantee your privacy given that data is being collected over the Internet. While you are not required to provide any identifying information, IP addresses may be recorded. This study is hosted using the UBC Survey Tool. The UBC Survey Tool is a cloud-based service provisioned by Qualtrics, a service provider contracted by UBC. The information collected using this tool is stored in Toronto, Ontario and backed up in Montreal, Quebec. The Survey Tool has completed UBC's Privacy Impact Assessment process, which assesses the privacy and security of UBC systems. Information collected using the Survey Tool is kept secure using measures including data encryption. There will be no physical copy of the data. The researchers will store an encrypted copy of the digital data offline. This data will be kept on a password protected flash drive, which will be stored at UBC-Okanagan campus for 5 years. Your decision to complete the survey indicates that you consent to this manner of data collection and storage.

This study is part of a doctoral dissertation, which will be published and made publicly available on cIRcle (https://circle.ubc.ca/). We also plan to submit the findings of this study to a scientific journal for publication. The data from all participants will be pooled and analyzed as a group, as the responses of any single individual are meaningful only in relation to the responses of others. This means that no firm conclusions can be drawn about the responses of individual participants. The results will only be reported for groups with no possibility of individual participants being identified.

If you would like to receive the aforementioned prizes, or a summary of the study results, you will need to provide us with your name and an Email address at the end of the study. This information will not be connected to your data or shared with any other party. It will be kept separately in an encrypted file, to which only the researchers of the study have access.

Possible risks and discomforts

No physical risks or discomforts are expected from participation in this study. However, some individuals may find the content of the videos disturbing. In the event that you no longer wish to participate for any reason, you may stop or withdraw from the study at any time.

Your identity will not be associated with your responses in the study.

Termination

You may stop participating in the study at any time by exiting the study web page. Even if you do not complete all the questions we will still include your data in our analyses.

If you want to withdraw from the study completely after you have responded to all or some of the questions, please email Ran Wei using the email address provided above, and reference your participant ID. We will remove your data from the data pool.

Contact for concern about the rights of research participants

Any inquiries concerning the procedures should be directed to Ran Wei. His email address, and telephone number are listed in the "Contact Information" section of this form.

If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Services at 1-877-822-8598 or the UBC Okanagan Research Services Office at 250-807-8832. It is also possible to contact the Research Participant Complaint Line by email (RSIL@ors.ubc.ca). Please reference the study number (H18-00087) when contacting the Complaint Line so the staff can better assist you.

Consent

Your participation in our study is completely voluntary and you may refuse to participate or withdraw from the study at any time.

Please indicate below whether or not you consent to these terms.

Please click here to download a pdf version of this form. Once you click on the right arrow below to continue, you will not be able to return to this page

Make a choice between "Yes" and "No", then click on the right arrow to continue. By choosing "Yes" you are giving your informed consent to participate in this study; if you choose "No", your browser will lead you to exit the study.

Yes No

Appendix D List of Included LIWC Features

Word count	Hearing	Death
Summary Variables	Feeling	Informal Speech
Analytical thinking	Body	
Clout	Biological processes	
Authentic	Affiliation	
Emotional tone	Achievement	
Words per sentence	Power	
Words>6 letters	Reward focus	
Positive emotion	Risk focus	
Anxiety	Past focus	
Anger	Present focus	
Sadness	Future focus	
Social words	Motion	
Insight	Space	
Cause	Time	
Discrepancies	Work	
Tentativeness	Leisure	
Certainty	Home	
Differentiation	Money	
Seeing	Religion	

Appendix E NLTK Universal POS Tags

Tag	Definition	Example
CC	Coordinating conjunction	and, but, or
CD	Cardinal digit	one, two, three hundred
DT	Determiner	a, the
EX	Existential there	'there' is
FW	Foreign word	
IN	Preposition/subordinating conjunction	Until, before
JJ	Adjective	big
JJR	Adjective, comparative	bigger
JJS	Adjective, superlative	biggest
LS	List marker	1)
MD	Modal could,	will
NN	Noun, singular	desk
NNS	Noun plural	desks
NNP	Proper noun, singular	Harrison
NNPS	Proper noun, plural	Americans
PDT	Predeterminer	'all' the kids
POS	Possessive ending	parent's
PRP	Personal pronoun	I, he, she
PRP$	Possessive pronoun	my, his, hers
RB	Adverb	very, silently
RBR	Adverb, comparative	better
RBS	Adverb, superlative	best
RP	Particle	'to' fly
TO	To go	'to' the store.

UH	Interjection,	errrrrrrrm
VB	Verb, base form	take
VBD	Verb, past tense	took
VBG	Verb, present participle	taking
VBN	Verb, past participle	taken
VBP	Verb, present, non-3d	take
VBZ	Verb, present, 3rd person	takes
WDT	Wh-determiner	which
WP	Wh-pronoun	who, what
WP$	Possessive wh-pronoun	whose
WRB	Wh-abverb	where, when

Appendix F Confusion Matrices (Averaged)

Fixed-Effect Logistic Regression
Cross-validation: 10-Fold randomized

	PSYC				POS				PSYC + POS		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	43	22		Truth	42	24		Truth	42	22
	Lie	19	38		Lie	21	35		Lie	20	38

Cross-validation: 4-Fold Cross Context

	PSYC				POS				PSYC + POS		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	81	70		Truth	86	66		Truth	79	79
	Lie	68	85		Lie	61	91		Lie	72	74

Mixed-Effect LASSO Logistic Regression
Cross-validation: 4-Fold Cross Context

	PSYC				POS				PSYC + POS		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	102	53		Truth	89	61		Truth	95	63
	Lie	57	92		Lie	55	99		Lie	57	89

Random Forest
Cross-validation: 10-Fold randomized

	PSYC				POS				POS-bi		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	46	23		Truth	40	27		Truth	39	24
	Lie	18	35		Lie	21	34		Lie	22	36

	POS-bi+				Unigram				Bigram		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	40	26		Truth	44	17		Truth	38	23
	Lie	24	32		Lie	15	45		Lie	19	42

	Bigram+				Trigram				Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	41	20		Truth	36	19		Truth	42	19
	Lie	18	42		Lie	22	44		Lie	18	43

	PSYC + POS				PSYC + POS-bi				PSYC + POS-bi^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	42	19		Truth	42	21		Truth	45	21
	Lie	20	40		Lie	19	39		Lie	20	36

	PSYC + Unigram				PSYC + Bigram^{+}				PSYC + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	42	19		Truth	40	20		Truth	43	19
	Lie	18	43		Lie	20	42		Lie	18	42

	POS + Unigram				POS + Bigram^{+}				POS + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	45	16		Truth	42	17		Truth	42	20
	Lie	16	46		Lie	19	43		Lie	18	41

	POS-bi^{+} + Unigram				POS-bi^{+} + Bigram^{+}				POS-bi^{+} + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	40	21		Truth	41	21		Truth	42	20
	Lie	22	39		Lie	21	39		Lie	22	40

Random Forest
Cross-validation: 4-Fold Cross Context

	PSYC				POS				POS-bi		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	78	73		Truth	86	67		Truth	81	69
	Lie	68	85		Lie	66	85		Lie	69	85

	POS-bi+				Unigram				Bigram		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	86	66		Truth	84	72		Truth	73	86
	Lie	67	85		Lie	69	79		Lie	79	66

	Bigram+				Trigram				Trigram^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	73	77		Truth	72	80		Truth	72	79
	Lie	72	82		Lie	77	75		Lie	72	81

	PSYC + POS				PSYC + POS-bi				PSYC + POS-bi^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	77	74		Truth	89	61		Truth	95	63
	Lie	57	89		Lie	55	99		Lie	57	89

	PSYC + Unigram				PSYC + Bigram^{+}				PSYC + Trigram^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	80	75		Truth	74	77		Truth	73	75
	Lie	70	79		Lie	71	82		Lie	71	85

	POS + Unigram				POS + Bigram^{+}				POS + Trigram^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	84	72		Truth	75	75		Truth	82	70
	Lie	69	79		Lie	72	82		Lie	69	83

	POS-bi^{+} + Unigram				POS-bi^{+} + Bigram^{+}				POS-bi^{+} + Trigram^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	87	63		Truth	87	66		Truth	86	68
	Lie	66	88		Lie	70	81		Lie	67	83

BiMM Forest
Cross-validation: 4-Fold Cross Context

	PSYC				POS				POS-bi		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	88	65		Truth	86	70		Truth	85	69
	Lie	55	96		Lie	65	83		Lie	62	88

	POS-bi+				Unigram				Bigram		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	88	67		Truth	95	57		Truth	77	69
	Lie	59	90		Lie	53	99		Lie	76	82

	Bigram+				Trigram				Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	89	63		Truth	85	73		Truth	83	76
	Lie	62	90		Lie	68	78		Lie	68	77

	PSYC + POS				PSYC + POS-bi				PSYC + POS-bi^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	92	60		Truth	91	60		Truth	96	63
	Lie	54	98		Lie	57	96		Lie	52	93

	PSYC + Unigram				PSYC + Bigram^{+}				PSYC + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	87	61		Truth	87	63		Truth	89	63
	Lie	57	99		Lie	60	94		Lie	55	97

	POS + Unigram				POS + Bigram^{+}				POS + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	93	58		Truth	90	61		Truth	87	70
	Lie	52	101		Lie	59	94		Lie	54	93

	POS-bi^{+} + Unigram				POS-bi^{+} + Bigram^{+}				POS-bi^{+} + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	91	62		Truth	82	67		Truth	85	70
	Lie	52	99		Lie	52	103		Lie	51	98

ANN
Cross-validation: 10-Fold randomized

	PSYC				POS				POS-bi		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	42	19		Truth	40	20		Truth	38	21
	Lie	23	38		Lie	27	35		Lie	29	34
	POS-bi+				Unigram				Bigram		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	47	12		Truth	38	21		Truth	38	21
	Lie	18	45		Lie	18	46		Lie	21	42
	Bigram+				Trigram				Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	39	20		Truth	39	21		Truth	41	19
	Lie	12	51		Lie	19	43		Lie	15	47
	PSYC + POS				PSYC + POS-bi				PSYC + POS-bi^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	39	20		Truth	36	23		Truth	47	14
	Lie	25	38		Lie	23	40		Lie	18	43
	PSYC + Unigram				PSYC + Bigram^{+}				PSYC + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	43	16		Truth	40	19		Truth	40	19
	Lie	14	49		Lie	13	50		Lie	17	46
	POS + Unigram				POS + Bigram^{+}				POS + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	40	16		Truth	41	18		Truth	39	20
	Lie	20	45		Lie	16	47		Lie	16	47
	POS-bi^{+} + Unigram				POS-bi^{+} + Bigram^{+}				POS-bi^{+} + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	42	16		Truth	42	16		Truth	41	16
	Lie	20	43		Lie	18	45		Lie	19	46

ANN
Cross-validation: 4-Fold Cross Context

	PSYC				POS				POS-bi		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	73	69		Truth	63	79		Truth	79	73
	Lie	64	98		Lie	63	99		Lie	71	81

	POS-bi+				Unigram				Bigram		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	93	56		Truth	106	67		Truth	66	79
	Lie	53	102		Lie	71	60		Lie	71	88

	Bigram+				Trigram				Trigram^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	91	63		Truth	72	80		Truth	39	113
	Lie	75	85		Lie	65	87		Lie	43	109

	PSYC + POS				PSYC + POS-bi				PSYC + POS-bi^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	79	63		Truth	81	81		Truth	83	65
	Lie	66	96		Lie	52	90		Lie	49	107

	PSYC + Unigram				PSYC + Bigram^{+}				PSYC + Trigram^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	105	47		Truth	106	46		Truth	95	54
	Lie	93	59		Lie	94	58		Lie	78	77

	POS + Unigram				POS + Bigram^{+}				POS + Trigram^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	103	56		Truth	104	48		Truth	105	47
	Lie	86	61		Lie	87	65		Lie	88	64

	POS-bi^{+} + Unigram				POS-bi^{+} + Bigram^{+}				POS-bi^{+} + Trigram^{+}		
	Predicted				Predicted				Predicted		
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	103	60		Truth	110	54		Truth	102	64
	Lie	55	86		Lie	51	89		Lie	55	83

"Mixed" ANN
Cross-validation: 4-Fold Cross Context

	PSYC				POS				POS-bi		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	81	66		Truth	76	70		Truth	85	67
	Lie	58	99		Lie	55	103		Lie	54	98

	POS-bi+				Unigram				Bigram		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	92	52		Truth	107	45		Truth	59	93
	Lie	46	114		Lie	69	83		Lie	48	104

	Bigram+				Trigram				Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	105	52		Truth	67	85		Truth	88	64
	Lie	71	76		Lie	50	102		Lie	72	80

	PSYC + POS				PSYC + POS-bi				PSYC + POS-bi^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	81	71		Truth	78	74		Truth	102	41
	Lie	48	104		Lie	46	106		Lie	37	124

	PSYC + Unigram				PSYC + Bigram^{+}				PSYC + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	105	57		Truth	101	56		Truth	102	56
	Lie	71	71		Lie	64	83		Lie	70	76

	POS + Unigram				POS + Bigram^{+}				POS + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	104	52		Truth	106	53		Truth	97	55
	Lie	70	78		Lie	67	78		Lie	69	83

	POS-bi^{+} + Unigram				POS-bi^{+} + Bigram^{+}				POS-bi^{+} + Trigram^{+}		
		Predicted				Predicted				Predicted	
Actual		Truth	Lie	Actual		Truth	Lie	Actual		Truth	Lie
	Truth	108	54		Truth	112	45		Truth	105	50
	Lie	52	90		Lie	48	99		Lie	52	97