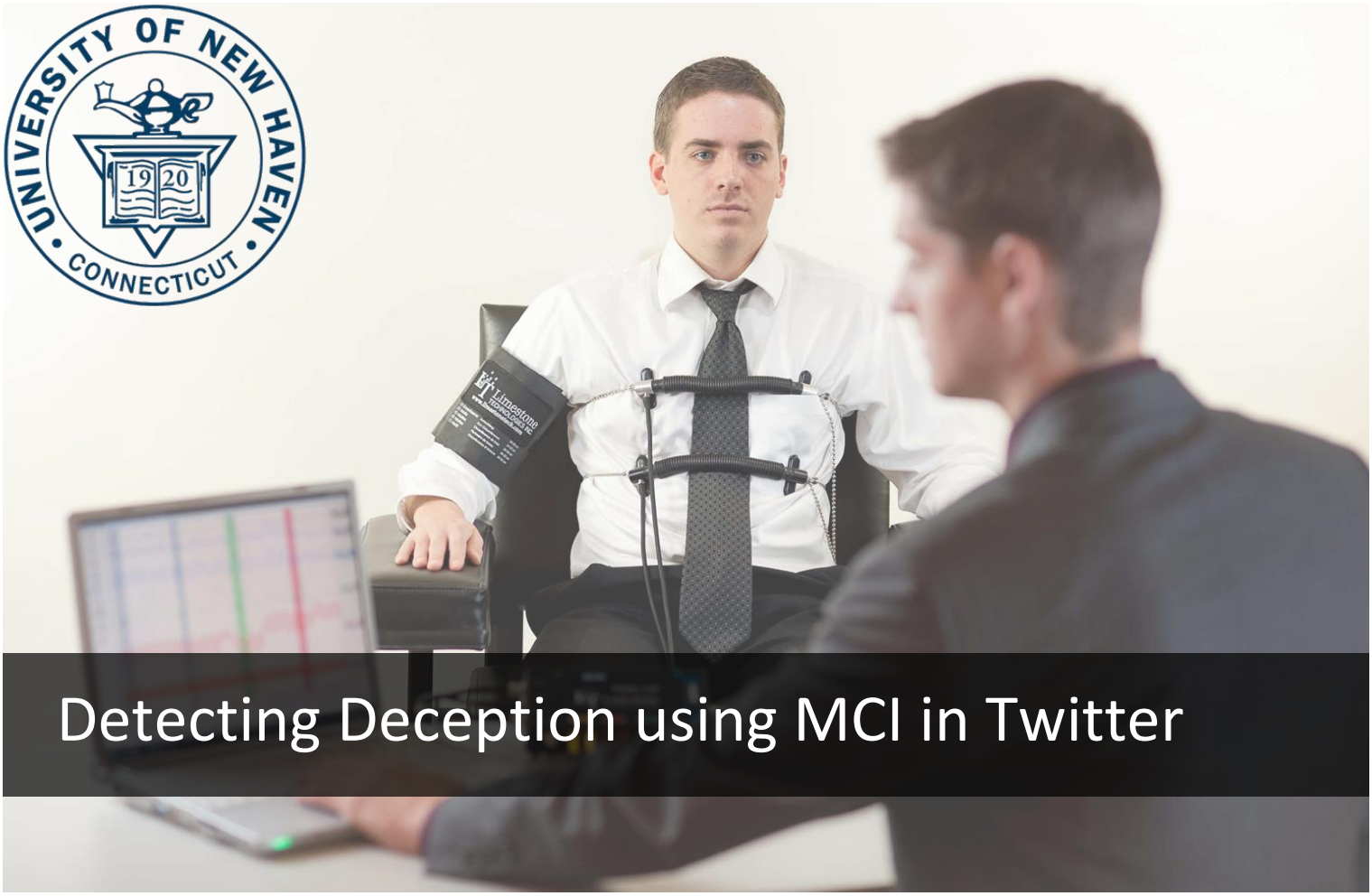

Detecting Deception using MCI in Twitter

Summary

The purpose of this study is to test whether Modified Cognitive Interviewing (MCI) is an effective method for detecting deceptive human eyewitness accounts in computer-mediated communication with a limited text space (such as in Twitter).

Bachelor Thesis

Author:	Jip Barthen	1st Reader:	Ms. E. Makri
Student nr.:	305662	2nd Reader:	Ms. S. Stronks
Study:	Security Management B.A.	Supervisor:	Dr. C.A. Morgan III M.D.
Saxion University of Applied Sciences		Date:	06/20/2016

University of New Haven – Connecticut (U.S.A.)

Detecting Deception using MCI in Twitter

A study about the Modified Cognitive Interviewing technique performed in the limited text based environment of Twitter, approached from two different perspectives regarding human rater judgment, and computer-processed text analysis.

"No man has a good enough memory to be a successful liar."

- Abraham Lincoln, 16th President
of The United States of America.

Colophon

Client:	University of New Haven (UNH)
Department:	Henry C. Lee College of Criminal Justice & Forensic Sciences, Department of National Security
Location:	University of New Haven
Author:	Jip Barthen (305662)
Co-author:	Stefan Aurelio Mol (325593)
Supervisor:	Charles A. Morgan III, M.D., M.A
For the purpose of:	Bachelor thesis Security Management, Saxion University of Applied Sciences Apeldoorn
Datum:	June 20, 2016
Status:	Final
Version:	1.0
1st reader:	Ms. E. Makri
2nd reader:	Ms. S. Stronks

Preface

Presented here is the thesis “Detecting Deception using MCI in Twitter”, the final product of our research project, commissioned by Dr. C.A. Morgan III. It is written in order to be published in The United States of America, and to fulfill the last requirements of the bachelors study “Security Management” for Saxion University of Applied Sciences in The Netherlands. The thesis project took place from March till July 2016, in West Haven Connecticut (U.S.A.).

From August 2015, we were part of an exchange program for the duration of one semester at The University of New Haven (UNH). During this semester we came in contact with Dr. Morgan, who is an associate professor at UNH. He offered to supervise a research project that we were able to conduct. Together we formulated a problem statement, in order to perform a research that benefits both parties. The research was in the field of forensic psychology with links to National Security, a discipline we had only basic knowledge of. Because we formulated the problem statement around two months before we started, we were able to read relevant studies and books in advance.

Some aspects of the research were really challenging, but fortunately we were always able to consult our supervisor Dr. Morgan. We would like to thank him for the guidance that he gave regarding the methods of Cognitive Interviewing and on how to perform proper academic research. We would also like to thank him for giving us the opportunity to work on this research. Furthermore, we are really thankful to the University of New Haven for sponsoring our visas, and for the facilitation during the research project.

Finally, we would like to thank our supervisor from Saxion University, Ms. Makri, for all the guidance and support during the graduation program.

The thesis is a result of four months hard work and dedication. We hope the result meets your expectations.

Jip Barthen and Stefan Mol

West Haven CT (U.S.A.), June 20, 2016

Abstract

The purpose of the study is to test whether Modified Cognitive Interviewing (MCI) is an effective method for detecting deceptive human eyewitness accounts in a computer-mediated communications with limited text space (such as in Twitter). The study is based on a previous study from Morgan, Christian, Rabinowitz, Palin and Kennedy (2015), where MCI has proven to be an effective method, with the use of face-to-face interviews.

In total 44 college students participated in this study, where they either had to perform, or pretend that they had performed a cognitive task. 15 students had to perform a task, and were instructed to answer completely honest when interviewed (truthful group), 16 students performed the task, and when interviewed they had to deny that they performed it, and instead make up a story (deceptive group), and 13 students read the instructions of the task, and had to claim that they performed the task, without actually having done it (false claim group).

The interview was held in a Twitter format, where six questions were sent out, and participants answered them one by one, with a maximum of 140 characters per answer. Every interview took place in a supervised space, where participants were able to fully concentrate. The questions (tweets) were the six prompts as known in the MCI method. Further explanation on these six prompts is given in Appendix B.

After the information gathering, the tweets were first used as input for a survey. The tweets processed in the survey were rated by (former) law enforcement professionals, and people with expertise in lie detection. The experts had to rate the first tweet and the whole Twitter conversation, and had to give their confidence level for every rating. The rater's ability to correctly distinguish truthful participants from deceptive participants for only the initial prompt in MCI resulted in 48%, while the rating of the whole interviews resulted in 54% of success. The 54% success rate of the whole interviews went along with a confidence level of 3.6 out of 5 (72%), this means that the human raters overestimated their abilities to discriminate truthful from deceptive participants.

Other than for the surveys, the tweets were also used as input for the computer-processed text analysis. After eliminating all fill words and repeating sentences in the tweets, the analysis determined the Response Length (RL), Unique Word count (UW), and Type Token Ratio (TTR). These so called "speech-content variables" were compared with each other, related to the different groups of participants (truthful, deceptive, and false claim). The outcome is that the computer-processed analysis, by using the speech-content variables RL and UW, is effective in lie detecting through computer-mediated communication with a limited text space, such as in Twitter (RL: 70% and UW: 79%).

As a comprehensive answer to the problem statement, we can conclude that the use of MCI is an effective method of detecting deception in a limited text based environment, such as Twitter. However, this only applies when using computer-processed analysis, with the speech-content variables RL and UW as the leading factors.

Table of Content

Glossary

Cognitive Interview (CI) A proven technique used to enhance retrieval of information from memory. Used in both eyewitness and suspect interrogations.

Computer-Mediated Communication (CMC) Communication within a digital environment, where there is a sender and a receiver. For example, Social media (Twitter, Facebook, Snapchat), instant messaging, and emailing.

Computer Processing Running data through a computer program (i.e., SPSS Statistics), and performing various statistic tests.

Deception Intentionally manipulating, or hiding the truth in order to deceive someone.

Deceptive group The group of participants that were requested to deny that they ever did the cognitive task. In other studies also called the "denial group".

False claim group The group of participants who had to pretend to have performed a cognitive task, while only having read the instructions.

Human raters (former) law enforcement professionals and experts trained in detecting deception with extensive knowledge of Cognitive Interview methods.

Law Enforcement Agencies Government organizations such as FBI, NSA, CIA , and local Police; tasked with enforcing the US law.

Likert scale A one to five based scale widely used to scale responses in surveys. In this case, it was used to measure the confidence level of the human raters.

Modified Cognitive Interview (MCI) A shorter more formal version of the Cognitive Interview method, used in various studies.

N Scientific notation, in this study representing the number of participants (44 in total).

Speech content variables Used in this study in order to analyze textual response: Response Length (RL), Unique Word count (UW), and Type Token Ratio (TTR).

Tweet 140-character message, used in Twitter.

Twitter A social media platform that lets users send and read short 140-character messages called "tweets".

Type Token Ratio The Response Length (Type) divided by the Unique Words (Token), also known by the name "Lexical Density".

Introduction

The increase in interactions via computer-mediated communication (CMC) media is resulting in interest in the dynamics of deception in online environments. For example, hundreds of millions of people are using the social network Twitter. More and more communication goes through online social platforms. Given the increased use of CMC, there has been an increased interest on the part of national security professions to detect deception in CMC environments. Deceptions in this kind of online environments are reflected in a small but growing body of literature (Burgoon, Blair, Qin and Nunamaker, 2003; Zhou, Burgoon, Twitchell, 2003; Zhou, Twitchell, Qin, Burgoon & Nunamaker, 2003b; Zhou, 2005; Hancock, Curry, Goorha and Woodworth, 2005). The focus of this literature has been to determine whether or not the same verbal attributes indicative of deception that have been isolated in face-to-face communication studies exist in CMC environments as well. However, no studies have been published on deception in the more limited form of CMC communication known as Twitter. The environment of Twitter restricts communications to a length of 140 characters. Given the fact that most studies on deception have shown that truthful people tend to have more to say than liars (Morgan et al., 2015), it is possible that the limited space allowed in Twitter may render previously validated methods for detecting deception ineffective. In our research, the limited text message as in Twitter (140 characters) has been used to examine whether or not proven methods in the real world can be used for detection deceptions in an online environment. The method used to detect deception is the Modified Cognitive Interview (MCI), an interrogation method used by law enforcement professionals to detect deception in suspect and eyewitness interrogations. The drive towards automated deception detection has resulted in CMC research that has focused primarily on the verbal behavior of deceptive message senders and the isolation of linguistic cues that can be detected by computers (Burgoon et al., 2003; Zhou et al., 2003; Zhou, et al., 2003b; Zhou, 2005).

We hypothesized that truthful individuals, compared to deceptive, type more characters, and provide more unique works. Furthermore, this research examines the possibility of automated deception detection in limited text messages, such as Twitter, rather than the detection of deception being accomplished by trained message receivers.

The objective of this study is to find out if Modified Cognitive Interviewing is useful in a computer-mediated communication (CMC) environment with a limited answering space (140 characters). The results of this research are a substantial contribution because such an approach could potentially be used to create new methodologies in suspect or eyewitness interrogation. This could improve the level of National Security, and increase the number of solved crimes.

1. Personal and General Information

This chapter provides the personal information of the author and supervisor. Also, the institute and the mission of the institute are described.

1.1 Personal Information

Name:	Jip Barthen
Student number:	305662 (Dutch) 00594064 (US)
Email address:	jipbarthen@gmail.com
Phone Number:	06-15658317 (Dutch) 1-475-201-4803 (US)
Study Program:	Security Management

1.2 About the Research

Thesis name:	Detecting Deception in Computer Mediated Communication Using Twitter.
Institute name:	University of New Haven New Haven, Henry C. Lee College of Criminal Justice & Forensic Sciences
Supervisor:	Charles A. Morgan III, M.D., M.A

1.3 About the Supervisor

Name:	Charles A. Morgan III, M.D., M.A
Job:	Associate Professor
College:	Henry C. Lee College of Criminal Justice & Forensic Sciences
Department:	National security
Phone:	203-932-1154
Email:	CMorgan@newhaven.edu
Office:	South Campus Hall 004

1.4 About the Institute

1.4.1 The University of New Haven

"The University of New Haven is a private, top-tier comprehensive institution recognized as a national leader in experiential education.

Founded in 1920 on the campus of Yale University in cooperation with Northeastern University, UNH moved to its current West Haven campus in 1960 and opened its Orange Campus in January, 2014. The University operates a satellite campus in Tuscany, Italy, and offers programs at several locations throughout Connecticut and in New Mexico. UNH provides its students with a unique combination of a solid liberal arts education and real-world, hands-on career and research opportunities.

The University enrolls approximately 6,800 students, including nearly 1,800 graduate students and more than 5,000 undergraduates – the majority of whom reside in University housing. Through its College of Arts and Sciences, College of Business, Henry C. Lee College of Criminal Justice and Forensic Sciences, Lyme Academy College of Fine Arts, and Tagliatela College of Engineering, UNH offers over 80 undergraduate and graduate degree programs. UNH students have access to more than 500 study abroad programs worldwide and its student-athletes compete in 16 varsity sports in the NCAA Division II's highly competitive Northeast-10 Conference." (UNH, 2016)

1.4.2 Mission Statement

"The University of New Haven is a student-centered comprehensive university with an emphasis on excellence in liberal arts and professional education. Our mission is to prepare our students to lead purposeful and fulfilling lives in a global society by providing the highest-quality education through experiential, collaborative and discovery-based learning." (UNH Mission, 2016)

1.4.3 Place Within the Organization

The University of New Haven is divided into multiple colleges and departments. The Department of National Security belongs to the Henry C. Lee College of Criminal Justice and Forensic Sciences, and is part of the School of Public Services.

"The core academic programs in the School of Public Service prepare students for professional careers in areas such as Public Safety, Legal Studies, Intelligence Analysis, National Security, Emergency Management, Fire Protection Engineering and Fire Science. Each of these professional areas of study involve important aspects of public service and each prepares students for employment in organizations that provide for public protection and maintaining order in a civil society." (University of New Haven: School of Public Service, 2015).

Associate professor, and supervisor of this research, Dr. C.A. Morgan III acquired a grant from the US Airforce to conduct multiple research projects. One of these research projects is the one we designed and conducted. The grant is distributed by the Center for Research and Development, Inc. (CR&D). We work as independent consultants on this research with Dr. C.A. Morgan as supervisor and co-author. In this context, we had to sign a consultant and confidentiality agreement.

The University of New Haven will benefit from this research because it will be additional to current knowledge of detecting deception. Successful research projects allow universities (in our case UNH) to become reliable partners for organizations that provide grants for future research. Our research results may contribute in grants being easier issued for UNH, or being of greater volume. It also promotes the reputation of UNH as a high ranked and well respected university.

The United States Airforce (USAF), which is governed by the Department of Defense, benefits from this research because it expands the knowledge on the subject of detecting deception. This knowledge could potentially be used to apply in real world situation benefitting the overall National Security of the United States, which is one of the primary objectives of the USAF. The USAF is always working on research and development (U.S.Air Force – Mission, 2016). Part of this is providing grants to universities so that they are able to do relevant research. It should be noted that the USAF is also responsible for the sector Cyberspace. This means that the ability to detect deception in computer mediated communication is for them a valuable ability to possess

2. Goals

In this chapter the main goal of this research is described, and the problem and research questions are given.

2.1 Purpose

The over-arching objective of this study is to find out whether the Modified Cognitive Interviewing (MCI) method, is useful in a computer-mediated communication (CMC) environment with a limited answering space (140 characters). At present, professionals working in the area of Intelligence and National Security are confronted with the challenge of evaluating genuine from deceptive or fraudulent cyber communications proffering information 'of interest' to the US Government as well as detecting when extremists or terrorists are communicating deceptively in CMC environments (Associated Press, 2016; Irshaid, 2014). Currently, no standardized, scientifically validated techniques exist within the Intelligence community for the triage of cyber communications. By contrast, significant scientific progress in the area of detecting deception in the real world has been made over the past decade. It is our intent to provide information on the state of the science in this area so as to promote the development of better, scientifically valid and reliable techniques for the detection of deception in cyberspace communications relevant to National Security.

This study demonstrates that the use of Modified Cognitive Interviewing to detect deception in computer-mediated communication (e.g., Twitter) is effective in discriminating truthful from deceptive individuals. More specifically, and based on previous research using the MCI to detect deception (Morgan, Rabinowitz, Leidy, & Coric, 2014; Morgan et al., 2015), we hypothesized: A) response length (i.e., number of characters in Twitter) will be greater in truthful compared to deceptive individuals; B) unique word count will be greater in truthful compared to deceptive individuals; C) the Type Token Ratio (the ratio of unique word count against the response length) will be higher in Truthful compared to Deceptive individuals; D) human raters will score higher than chance, when discriminating truthful from deceptive individuals; and finally, E) these computer assessed variables will perform better than humans assessing the Twitter communications in detecting deception.

We tested the efficacy of the Modified Cognitive Interviewing method for discriminating between Deceptive and Truthful participants in computer-mediated communication by using the social media platform Twitter. In previous research modified cognitive interviewing has been used in a face-to-face situation yet never in a limited text message based environment (Morgan et. al., 2015). We have selected the medium Twitter because it is widely used by a large and diverse population, with over 310 million active users (Twitter Inc., 2015), which most likely includes persons of interest for professionals working in National Security.

2.2 Problem Statement

The main goal of this research is to answer the main problem statement:

"Is detecting deception possible using the Modified Cognitive Interview method in the computer-mediated environment of Twitter*?"*

2.3 Research Questions

The research questions, as displayed in Table 1, are needed to answer the problem statement.

1	*How effective is computer-processed data in detecting deception with the use of MCI in Twitter?*
2	*How effective is expert judgment in detecting deception with the use of MCI in Twitter?*
3	*Does MCI used in Twitter perform better than other methods: chance (i.e., 50%), and polygraph (i.e., 50%) in detecting deceptive from truthful participants?*
4	*How effective is MCI in detecting deceptive from truthful participants as compared with the results presented in the research of Morgan et al. (2015)?*

Table 1: Research Questions

2.4 Practical Implications

By performing this study, we showed that Modified Cognitive Interviewing is able to determine the difference between true and false statements about autobiographical experiences in an online platform with limited text space. Information regarding the effectiveness of integrating content analysis and actuarial data will aid professionals working in a variety of fields for example law enforcement, intelligence gathering and insurance companies. Geiselman (2012) made an adaptation of the original CI, which was mostly used for eyewitnesses. In this adaptation he made CI possible for interrogating suspects by law enforcement. It would be helpful for law enforcement agencies if they could add the use of CI in Twitter to their toolbox in order to gather relevant and reliable information from both suspects and eyewitnesses. When MCI is proven to be an effective method in Twitter, this could in a broader view lead to a higher level of security in the public and private sector because it could potentially lead to:

Public Sector

- Interrogation in high risk and remote areas (i.g., Iraq, Syria)
- Judging the level of trustworthiness in online communication with criminal and terrorist organizations through social media platforms
- Increasing the efficient use of resources in law enforcement agencies
- Extending the reach of intelligence agencies
- Modernization of investigative methods in both real-life and cybercrime

Private Sector

- Economize insurance companies, by allocating resources more effectively
- Extending the reach of investigation and intelligence gathering companies
- Increasing the effectiveness of pre and in-employment screenings

If the use of MCI in Twitter turns out to be ineffective, it could still prove to be advantageous in an adaption of MCI or on other social media platforms, because the characteristics of Twitter could be a limiting factor to the use of MCI.

2.5 Stakeholders

2.5.1 Directly Involved Stakeholders

The University of New Haven

This institute is the main client of the research. They provided the supervisor to oversee the project, and facilitated this research.

Research participants

The people recruited as the research participants are a crucial part of the research. They provided us with the research data and trusted us to handle the data carefully with respect to their privacy and all other matters as stated in the consent form.

Researchers

The completion of this research depends primarily on the commitment of the researchers. This research is used as the final thesis for the bachelors study Security Management at Saxion University of Applied Sciences.

Supervisor

The supervisor of this research, Dr. Charles Andy Morgan III, provided feedback and guidance to the researchers. He also assisted in writing and publishing the scientific article as a co-author.

Expert raters

The expert raters provided their expert opinion on the collected data. This data is analyzed and contributes to our final conclusions,

2.5.2 Indirectly Involved Stakeholders

CR&D Inc.

This company is in charge of the necessary funds to conduct the research.

United States Air Force (USAF)

The USAF provided the grant to make this research possible.

Twitter

The social media platform Twitter is used as a medium to conduct the experiment because it limits the user's text space to 140 characters, and is used by a large variety of people.

Saxion University of Applied Sciences

This research is conducted as final thesis, which is necessary to graduate from the study Security Management at Saxion Apeldoorn.

3. Literature Review

A large amount of research has been done about lying and truth telling. Although, the literature covers a wide variety of theories about lying and truth telling, our literature review will focus on three major themes. These themes are: The (Modified) Cognitive Interview, Detecting Deception in Computer-Mediated Communication and Detecting Deception in Twitter and other Social Media. This paper will primarily focus on the application of the covered researches to Modified Cognitive Interviewing in an online environment. All studies mentioned in the next chapters, are peer reviewed studies, published in various scientific journals. This fact increases the reliability because the studies have already been reviewed by people that are experts in their specific field.

3.1 The (Modified) Cognitive Interview

Cognitive interviewing (CI) is a method of detecting deceptions, widely used by law enforcement and medical professionals. CI is initially developed in the early 80s to help law enforcement professionals to overcome their inability to effectively gather intelligence from victims and eyewitnesses (Geiselman & Fisher, 1985). Geiselman and Fisher tested three methods of interviewing eyewitnesses. Eighty-nine subjects viewed police training films of simulated violent crimes and were questioned 48 hours later by experienced law-enforcement personnel. The cognitive procedures elicited a significantly greater number of correct items of information from the subjects than the standard interview did.

The three basic psychological processes of cognitive interviewing are organized around: memory and cognition, social dynamics, and communication. It is a systematic process in order to increase the amount of information, without diminishing decreasing the accuracy. CI is based on multiple analyses of law-enforcement interviews and scientifically derived principles of memory and communication theories (Geiselman & Fisher, 2014).

Since the development of cognitive interviewing it has been the subject of a number of studies. A meta data analysis of 53 studies has shown an increase of 34% in gathering relevant information with cognitive interviewing compared to other interview models (Köhnken, Milne, Memon & Bull, 1999). Wright and Holliday (2007) proved that cognitive interviewing is more effective on all age groups than regular interviewing techniques. They used Enhanced Cognitive Interviewing (ECI) and MCI and made participants from different ages watch a video of a crime (car break-in). Thirty minutes later they were interviewed using ECI or MCI. Both methods seemed to be very useful as an eyewitness interrogation method for all age groups. However, testimonies from 75-95 year olds were less complete and accurate than those of the 60-74 year olds, which were less complete and accurate than the testimonies of 17-31 year olds. For children, research shows (Verkampt & Ginet, 2009) that for between the age of 4-5 and 8-9 a shortened version of CI is more effective. Verkampt and Ginet (2009) interviewed 229 children using the standard interview protocol, CI or one of five variations of the CI. In all cases the children showed better results when the CI was used instead of a standard interview.

In 2012 Geiselman conducted a research using an adjusted version of CI so it could be used on suspects instead of eyewitnesses. This adjusted version added various prompt like drawing or sketching the situation. Geiselman's (2012) study showed the potential of using CI during investigative interviews.

Throughout the years, a large number of studies has been conducted to assess the efficacy of CI under a variety of conditions. Some have assessed efficacy under conditions of realistic stress (Morgan, Steffian & Hazlett, 2007), others have assessed its efficacy in cross-cultural settings with participants that had English as a second language (Morgan, Christian, Rabinowitz & Hazlett, 2009; Morgan, Colwell, Steffian & Hazlett, 2008; Morgan, Mishara, Christian & Hazlett, 2008b). All have found the CI to result in high classification accuracies (i.e., 82-85%). However, this is only scientifically proven to be true for a select few cultures because only a few cultures have been subjected to scientific study.

To address the limitation of the absence of realistic stress in previous researches, Morgan et al. (2007) tested the efficacy of speech content analysis methods in distinguishing genuine from deceptive reports of military personnel exposed to mock interrogation stress during military survival school training. Active duty military personnel were randomized to genuine or deceptive eyewitnesses groups. Genuine group members were exposed to interrogation stress at survival school; deceptive group members were not. Genuine eyewitness reported truthfully about their exposure to interrogation stress at Survival School whereas deceptive eyewitnesses lied and gave an account that was based on their study of transcripts of genuine eyewitnesses for 24 hours prior to being interviewed. Persons trained in MCI rated these interviews. The accuracy of lie detecting using MCI was 82%. The findings of this study suggested that exposure to stress did not disrupt the efficacy of forensic statement analysis techniques. However, only military personnel were tested in this study. The question is if they can represent the average participants. Military personnel might be better trained in handling of stressful situations than an average individual. In order to suggest that stress does not disrupt the efficacy of forensic statement analysis techniques, a new study has to be done with more diverse participants.

Finally, as noted in recent scientific publications, (Morgan, 2008; Morgan 2008b; Morgan et al., 2009, 2011; 2014) the accuracy rates for automated forensic statement analysis targeting deception in interviews that involve translated speech (Arabic, Vietnamese, Russian) are significantly better than judgments about deception, assessed by human raters. Because Arabic, Vietnamese, and Russian are the only languages that have been subjected to researches, these studies have the same limitations as the cross-cultural studies. Only a few languages have been subjected to scientific study, which means that it is not proven to be effective in every language.

The current research project is an adaption of a recent study by Morgan et al. (2015). This study was the first to use cognitive interviewing to directly assess and compare the nature of true claims, false claims and denials. Morgan et al. (2015) assessed 104 military personnel who had performed either a cognitive (making a puzzle) or manual task (flying a remote control helicopter) prior to being interviewed by someone who was blind to their activities. Participants were randomized to Lie or to tell the Truth. Of the 54 people assigned to the manual task: 17 truly performed the task and were truthful when interviewed about their activities; 18 performed the task and denied having performed the task when interviewed; and finally 19 read the instructions regarding the manual task and when interviewed falsely claimed to have performed the task. Transcripts of the interviews were assessed by human raters and also used for computer-based speech analysis. The computer-based analysis performed significantly better in detecting deception than the human raters (i.e., 80% vs. 62%, respectively). This data supports the view that MCI derived statement analysis methods are scientifically valid and can be used by professionals tasked with distinguishing between true claims, false claims and denials.

In the current research, it is even harder for experts to rate the speech content because of the limited character use. The raters only received limited text interviews (in total 840 characters) where they based their judgment on, instead of unlimited answer possibilities as in (Morgan et al., 2015). Also the

Response Length in comparison with Morgan et al. (2015) is very limited in this research. When limiting the text, it is common sense that results will be closer to each other.

3.2 Detecting Deception in Computer-Mediated Communication

In addition to studying deception in direct, face-to-face interviews, a number of research groups have evaluated detecting deception methods in Computer-Mediated Communications (CMC). Research to date in this arena has consisted of two general types: research examining what people say they do when being deceptive in CMC environments, and research directly examining human behavior related to deception in CMC environments.

With respect to what people have said they do in CMC environments, it appears that deceivers may feel that they have less time to plan and edit their responses (Burgoon, Blair, Qin, Nunamaker, 2003). In addition, research comparing deception rates in different conversational environments found that face-to-face and IM environments share a similar display of deception by message senders (Hancock, Thom-Santelli & Ritchie, 2004). In this research Hancock et al. (2004) participants recorded their social interactions and lies for about seven days. As a result of the study they concluded that people lied the most by using their phone and the least in e-mail interactions. Based on this study, Whitty et al. (2012) also performed a diary study with some additional questions. Participants had to keep track of their lies in all sorts of mediums. The study came up with a similar result as Hancock et al. (2004). Relatively seen people are lying the most with phone calls. In another study where participants had to keep track of their deceptive behavior in text messaging, Hancock, Smith, Reynolds and Birnholtz (2014) discovered that a large majority of them (77%) at least told one lie during the experiment. This indicates that more participants were deceptive than being totally honest during the whole experiment. Although, the vast majority did lie quite infrequent, there was only a small number of prolific liars.

With respect to direct assessment of human behavior and deception in CMC environments, Zhou's (2005) research findings support the idea that similarities exist between verbal behaviors in face-to-face (FTF) communication environments and verbal behaviors in CMC environments. Instant messaging (IM) for example is thought to have more characteristics in common with speech generated in face-to-face communication than any other CMC medium (Zhou, 2005).

The first experiment to focus exclusively on verbal behaviors in CMC was carried out by Zhou, et al. (2003). The researchers set up an experiment using an asynchronous text-based exchange, designed to replicate an email environment. 60 students were evenly divided into groups of senders and receivers with 16 of the senders assigned the task of deceiving their receivers and 14 of the senders instructed to be truthful with their receivers. The receivers were not aware of whether their sender was truthful or deceitful, and neither senders nor receivers were aware of the identity of the other. The sender/receiver groups were instructed to correspond with each other regarding a list of 10-12 items that they would most want to salvage should they be stuck in the desert. A computer software program designed to recognize 27 linguistics-based cues derived from previous deception research processed the resulting text.

Zhou (2005) conducted a study where behavioral indicators of detecting deception in a group IM setting are explored. As a result of this study three types of nonverbal, and three types of verbal behaviors were able to significantly differentiate deceivers from truth tellers (Zhou, 2005). IM shares the informal nature of face-to-face communications and tends to consist of brief interactions between conversants. Unlike email, where responses are not necessarily instantaneous and interactions can spread out over time, IM most closely mimics the immediacy and spontaneous nature of conversation.

The expectations of sender and receiver in IM resemble those in face-to-face conversations as the IM sender expects an instantaneous reply from the receiver (Zhou, 2005).

Given the similarities between face-to-face and IM, CMC researchers have investigated whether or not verbal behaviors (key stroke analyses) indicative of deception in face-to-face environments are present in CMC environments as well. In addition to incorporating the theoretical framework of cognitive load, CMC research has drawn on other theories initially constructed to explain deception dynamics to face-to-face communications. One such theory is Interpersonal Deception Theory from Buller and Burgoon (1996), which stated that there are three strategies of deception: "falsification, concealment, and equivocation." Interpersonal Deception Theory supports the cognitive load premise as it also theorizes that deception is a mentally demanding task for message senders (Buller, Burgoon, 1996). For example, in order to maintain a deception throughout an interaction, deceitful message senders must constantly alter their behavior in an attempt to ward off suspicion on the part of message receivers. In the course of this process, message senders may ultimately present an account to the message receiver that contains fewer details and is less spontaneous in nature (Buller & Burgoon, 1996).

As previously noted, Zhou, et al. (2003b) focused on deceptive cues used in textual CMC. As predicted in the initial hypothesis of Zhou, et al. (2003b), the computer analysis of the resulting transcripts revealed that deceivers displayed higher word count than truth tellers. This result runs contrary to findings in the face-to-face communication literature, which suggests that deceivers use fewer words than truth tellers (Vrij, Edward, Roberts, Bull, 2000). In a later publication on the same topic, Zhou, Burgoon, Nunamaker and Twitchell (2004) attributed this difference to the fact that deceivers use more words in an attempt to convince receivers that they are truthful. Zhou, Burgoon and Twitchell (2003) supported the findings of Newman, Pennebaker, Berry and Richards (2003) that deceivers use more negative emotion words and fewer first person singular pronouns in their discourse. Deceptive senders in the experiment exhibited less lexical diversity, or use of unique words, than truth tellers. As for detail inclusion, Zhou, et al. (2003b) found that there was little difference in language specificity, as measured in spatio-temporal and perceptual details, between deceptive and truthful accounts. In a later publication of the same study results, Zhou et al. (2004) attributed this lack of difference in detail inclusion in the messages of deceivers and truth tellers to the fact that they had a limited dictionary of terms in their automated word detection program.

Burgoon et al. (2003) followed up on the research of Zhou, et al. (2003) and tested the ability of a computer software program to search text for linguistic cues and combinations of cues. Based on the cues Zhou, et al. (2003) found to be indicative of deception, Burgoon et al. (2003) created an experimental software program that could measure quantity (number of words and sentences), vocabulary complexity (number of syllables per word), grammatical complexity, as well as "specificity and expressiveness" as measured by the number of adjectives and adverbs in the text. Similar to the experimental setup of Vrij, et al., (2000), a mock theft was staged. Half of the subjects were instructed to pretend to steal a wallet and the other half were told a theft would occur at some point. Subjects were later questioned by interviewers using what was referred to in the study as a "standardized Behavioral Analysis Interview format that is taught to criminal investigators" (Burgoon et al., 2003). The interview was conducted either in a text chat or audio conferencing environment. The text resulting from these interviews was transcribed and analyzed by the experimental software program.

In both the text chat and audio conferencing mediums, deceptive messages were briefer, less complex and, in the text chat medium, contained fewer details. Burgoon et al. (2003) posited that the lower number of details in deceptive messages in this experiment versus that carried out by Zhou, et al. (2003) was due to lack of time for rehearsal and response in the synchronous text chat medium versus

the asynchronous email medium. Although this experiment used an interview technique designed for law enforcement officials, the study description did not make clear what the objectives of the format were and how the receivers and senders interacted with each other in the course of the interview.

Zhou, et al. (2003b) sought to examine how the textual language of deceivers changes over time and at what time in the communication change occurs. As was done in the research of Zhou, et al. (2003), this study was conducted in an asynchronous environment designed to mimic that of email. Zhou, et al. (2003b) used the results of the study conducted by Zhou, et al. (2003) to select the linguistic variables. The researchers divided the asynchronous interaction into three time segments. Study subjects were divided in to deceitful and truthful senders and receivers and again instructed to discuss which items from a list they would keep to survive in the desert. A computer software program analyzed the text from the interactions. The researchers found that there was a high level of deception cues in "time" 1 of the interaction, with the highest level occurring in "time" 2 and then the lowest number of cues in "time" 3. The researchers speculated that these results were supported by Interpersonal Deception Theory (Buller & Burgoon, 1996), in which deceivers alter their deceptive strategy over time in response to receiver reactions.

Hancock et al. (2005) expanded upon the research of Zhou, et al. (2003b) and Burgoon et al. (2003) by examining linguistic deception cues in the synchronous CMC environment of IM. Hancock et al. (2005) set up 66 students into sender and receiver dyads, with senders and receivers anonymous to each other. They were asked to converse with each other on five topics, senders being instructed to be deceptive on some of the topics and truthful on others. Transcripts were analyzed using the Linguistic Inquiry Word Count. Hancock et al. (2005) corroborated the findings of Zhou, et al. (2003b) that deceivers use more words than truth tellers and reemphasized the explanation of Zhou, et al. (2003b) that deceivers use more words in order to construct a believable story. Hancock et al. (2005) also supported the face-to-face communication findings of Newman et al. (2003) and CMC findings of Zhou, et al. (2003) that deceptive accounts contain fewer first person pronouns and more negative emotion words. Hancock et al. (2005) also found that lying accounts had more sense words (that relate to one of the five senses) in them than those of truth tellers. The presence of more sense words in deceptive accounts runs contrary to findings in the face-to-face communication research (Vrij, Edward, Roberts, Bull, 2000). Hancock et al. (2005) raised the possibility that this could either be due to differences in face-to-face and IM mediums or could be due the fact that in their experiments subjects were asked to describe unverifiable opinion and in face-to-face communication experiments, subjects described verifiable events, such as mock crimes.

Whereas the research of Zhou, et al. (2003b) did not seek to isolate specific receiver behaviors in their study, Hancock et al. (2005) explored how receiver behavior is affected by sender behavior. Hancock et al. found that receivers who were being deceived by their senders asked more questions of them. Hancock et al. (2005) speculated that this might be an attempt on the part of the receiver to probe the story of the deceptive sender by peppering their responses with questions. This speculation is in keeping with the theoretical framework of Interpersonal Deception Theory (Buller & Burgoon, 1996), as these questions indicated suspicion on the part of the receiver, thus encouraging the sender to work harder to maintain the deception.

Zhou (2005) also constructed a study using a synchronous IM environment. She set up teams of three students in which two were truth tellers and one was a deceiver. As with the CMC experiments of Zhou, et al. (2003; 2003b) the subjects were instructed to send messages to each other discussing which items on a list would be most useful if they were stranded in the desert. Some of the senders were instructed to be deceptive and some were instructed to be truthful. As analysis of group dynamics in deceptive situations was the primary focus of the study, few linguistic cues were

examined. Zhou (2005) did not replicate the findings of Zhou, et al. (2003) that truth tellers display a higher degree of lexical diversity, as the count of unique words did not vary among deceivers and truth tellers.

Computer-mediated communication consists of multiple modalities, that all can be used to detect deception. A research done by Zhou and Zhang, (2007) shows the effectiveness of three different modalities: “typing”, “messaging” and “chatting”, using an IM platform. Typing is the non-verbal behavior of the sender (e.g., keyboard behavior), messaging is a couple of messages presented in order of being received, and chatting is a combination of both. The result is that messaging and chatting modalities are more helpful for detecting deception than a typing modality. The chatting modality is the best choice in order to detect deception.

Researches of detecting deception in CMC and FTF are primarily focused on western countries, while nowadays there are no borders in the digital world. The question is if there are differences in cross-cultural communication, and more specifically: are deceivers from different parts of the world behaving differently? Lewis and George, (2008) did a study to understand the role of culture in deception for both FTF and computer-mediated communication. American and Korean participants were used in order to test the hypotheses. As a result, they found out that in general deceptive behavior was greater regarding face-to-face communication than it was for CMC. There were no differences between the Korean and American participants in the CMC groups. However, the face-to-face results showed significantly different scores; the Koreans scored higher on the deceptive behavior variables (Cross-cultural deception in social networking sites and face-to-face communication). This study only shows the comparison between two particular cultures. It should be noted that America does not represent the total of the so called western countries, and Korea does not represent the total of Asian cultures.

Galanxhi and Fui-Hoon Nah (2007) did a study in the relation between the use of avatars and deception in CMC. They found that deceivers are more likely to choose avatars that differ from themselves. A more relevant finding is that the use of an avatar in IM shows no increase or decrease in the perceived trustworthiness of the message sender in both the deceptive and truthful condition (Galanxhi & Fui-Hoon Nah, 2007). Hooi and Cho (2013) on their turn found out that people who have more similarities with their avatars, will be more self-aware. Similarities in appearances are following this research leading to less deception (Hooi, Cho, 2013).

Taken together, the literature on Deception Detection in CMC supports the view that A) some deception is detectable in CMC; B) the patterns in CMC appear to be at odds with face-to-face data. For example, in CMC Deceptive people type more; in face-to-face deceptive people write less and speak less and C) very different methods have been used. So the differences in the findings could be due to the medium (CMC/FTF) or the methods (i.e., the type of interviewing and interaction). To date, no studies have examined, and directly compared the use of the MCI in CMC environments to MCI used in FTF environments.

3.3 Detecting Deception in Twitter and other Social Media

Regarding the increased use of Social Media in the last decade, it is important to develop methods for detecting deception in both face-to-face communication and in CMC. Lying rates in face-to-face communication and in IM are about the same (Hancock et al. 2004). Our question is if this also applies for a one-sided interview with the use of limited texts such as in Twitter. Texts in Twitter and in IM are both limited, but people who are using IM do have a conversation instead of commenting on texts. Because we are using multiple tweets in our research, the previous resembling could also apply for interactions in Twitter. However, answering tweets is not the same kind of conversation as is that of

one associated with IM, because of a lack of intensity and expectations. The conversation we tend to create in twitter is more like a one-sided interview, where senders have a limited space to express themselves. The research on detecting deception in Twitter is limited. Alowibdi, Buy, Yu, Ghani and Mokbel (2015) did a study in detecting deceptive profiles in Twitter. In the study they used algorithms, and were able to be very accurate in detecting fake profiles. However, this study only addressed the use of profiles in Twitter, it is about deception but in a different way than in the present study.

At present no studies have been conducted in Twitter using validated interviewing methods that have been used in face-to-face or non-Twitter CMC environments. It would be useful to know whether and to what degree a validated method for detecting deception (i.e., the MCI) would be effective when used in Twitter environments. It is possible that the limited text space prevents the MCI from working effectively.

The present study is designed to assess how well the MCI could be used to distinguish between truthful and deceptive tweets about one's autobiographical experiences.

4. Research Methods

The literature review sheds light on the problem statement and research questions. In this chapter the method to perform this research, as well as the operationalization of the research questions are illustrated. The data acquisition, written in co-operation with Dr. Morgan, is based on his previous study. In order to obtain similar and comparable results, the method of data gathering corresponds to Morgan et al. (2015).

4.1 Data acquisition

In this study, 44 students at the University of New Haven participated in determining whether or not MCI is an effective method in Twitter. Each participant was given an oral briefing about the project, and each provided written informed consent prior to participation in the study.

This study design consisted of two phases: in Phase One, participants engaged in, or only read about, a cognitive task; in Phase Two, interviewers questioned the participants about their declared activities on Twitter. When interviewing participants, the interviewers used a Modified Cognitive Interview (MCI) (Morgan et. al., 2011; 2014; 2015). The questions were adapted to fit into one tweet each (140 characters). Experts rated the interviews from transcripts, and the computer analysis used speech content variables. Morgan et al. 2015, as well used both methods to understand which one is the most effective in detecting deception.

4.1.1 Phase One: Task Exposure

Randomized participants engaged in, or only read about the cognitive task: all 15 truthful persons completed the task; 16 deceptive participants assigned to the "denial" group also completed the task, and 13 deceptive participants assigned to the "fabrication" group were only permitted to read the instructions of the task.

The task involved participants in a series of timed trials during which they had to make use of a set of shapes (i.e., using a commercially available game called Tangoes©) to construct an image that matched the figure shown to them by the instructor administering the task.

After completing their task, participants assigned to the Truthful condition were told that they would be interviewed through Twitter, about how they had spent their time. They were instructed to respond openly and honestly about the nature of their activities. Conversely, after completing their task, participants assigned to the "deceptive" condition were told that they could not report on their activities; so were instructed to lie when interviewed. Participants assigned to the "false claim" deceptive condition were given written detailed instructions about the task. They had 10 minutes to study the materials. They were told when given the instructions that they would have to lie and claim that they had actually performed the task when interviewed.

Tangoes is mentally challenging and as suggested in previous studies (Morgan et al., 2011; 2014) requires significant mental effort for the participant to complete in an accurate manner. The task was considered "complete" when participants completed the task or when the time expired.

4.1.2 Phase 2: The Modified Cognitive Interview

The interviews, both questions and answers, were conducted in the social media platform Twitter. Participants were assigned a pre-registered Twitter account, in which they had to answer the MCI questions. The interviews were conducted with multiple participants simultaneously, ranging from groups of four to eight participants interviewed at the same time. Each answer could only consist of one tweet. This means that they only had 140 characters to formulate their answer. Once the

participants had answered the six tweets, their participation in the study was completed. They were debriefed about the study before they were free to leave.

4.1.3 Polygraph Test and Chance

Multiple studies are suggesting that the polygraph test, also known as the lie detector test, has a 65% success rate (Vrij, 2008; Morgan et al., 2007; 2008 ;2008b; 2009; 2014; 2015). The success rate of chance is 50% because the participant is either truthful or deceptive. To compare the polygraph test, and chance with the MCI used in Twitter, the success rate of both methods was put next to each other, as well as the computer-processed analysis as the human ratings. The success rate of the following methods are entered in a percentage scale, in order to display how effective every method is compared to each other:

- MCI in real life interviews
- MCI in Twitter expert judgment
- MCI in Twitter computer-processed data
- Chance
- The polygraph test

4.1.4 Variables

The Modified Cognitive Interview (MCI) is a practical adaptation of the cognitive interview. The reason to use MCI over CI is that because MCI is simpler and more structured. The six prompts of MCI (Full Recall, Visual, Auditory, Emotional, Temporal, and Detailed) are displayed in Appendix B.

For processing the data through a computer, we used three common variables: Response Length, Unique Word count, and Type Token Ratio. Morgan et al. (2015) has shown that those variables can be highly effective in detecting deceivers. They used these variables to assess whether someone was deceptive or truthful. However, the current research goes beyond the findings of Morgan et al. (2015), and is focused on achieving high rates of accuracy in detecting truthful from deceptive participants in an online environment with limited text space.

Response Length (RL)

In real life, people who lie use fewer words in their answers during cognitive interviews. However, Zhou et al. (2005) already showed that in a computer-mediated environment deceptive people use more words to describe their lies. According to Zhou et al. (2005) this is because deceptive people want to make the story more believable. Response Length can be measuerd by simply counting all the words that are used.

Unique Word Count (UWC)

The text was analyzed through a computer program, and all the unique words were counted. In general, deceptive people maintain a lower unique word count, and tend to reproduce the story multiple times in almost the same way. This means that words that were used multiple times only count once.

Type Token Ratio (TTR)

Type Token Ratio is the ratio of the number of unique words (types) against the total number of words (tokens). The outcome of this ratio is a percentage. The more words that are used, the lower the percentage is expected to be. So, in shorter messages, there are normally more unique words in relation to the total word count. For processing the data through an expert analysis, we used (former) law enforcement professionals and MCI experts.

Truthful or deceptive

(Former) law enforcement professionals and MCI experts rated every Twitter conversation as either truthful or deceptive. The results of this rating were used to determine if human raters accurate detect deception on Twitter. The results of this variable were compared to the results of the computer-processed data.

Level of Confidence

After rating the tweets as deceptive or truthful, the raters determined their level of confidence of their judgment, measured by using a scale from 1 to 5. This variable was used to determine if the confidence level of a human rater correlates to his or her ability to correctly judge the tweets.

4.2 Computer-Processed analysis

For the computer-processed data, the tweets were analyzed twice with the use of *Using English*, an online text-analyzer tool (Text Content Analyser, 2016). At the first analysis, the entered tweets were left unchanged, and at the second analysis all fill words that did not contain memory were taken out. Some examples are words in the context of *just, like and only,* and repeating sentences such as *if you would have been there with me, you would have heard*, or *you would have seen*. Those kinds of words and sentences do not say anything about the actual memory. The outcome: Response Length (RL), Unique Words (UW), and Type Token Ratio (TTR), were used as input for SPSS. Statistical Package for the Social Sciences (SPSS) is a computer program that can analyze statistics.

Table 2 gives an overview of the variables. Separate variables exist in Prompt Total 1-5 because Prompt 6 (did you leave anything out?) is experimental, so it is uncertain at providing reliable answers containing memory. To prevent outliers, and therefore systematic errors, there are additional "Tweet Total Variables" made. Since the separate prompts are very limited (140 characters), the main focus is on the prompt totals.

Prompt 1			Prompt (memory only)		
RL	UW	TTR	RL	UW	TTR
Prompt 2			Prompt (memory only)		
RL	UW	TTR	RL	UW	TTR
Prompt 3			Prompt (memory only)		
RL	UW	TTR	RL	UW	TTR
Prompt 4			Prompt (memory only)		
RL	UW	TTR	RL	UW	TTR
Prompt 5			Prompt (memory only)		
RL	UW	TTR	RL	UW	TTR
Prompt 6			Prompt (memory only)		
RL	UW	TTR	RL	UW	TTR
Prompt Total			Prompt (memory only)		
RL	UW	TTR	RL	UW	TTR
Prompt Total 1-5			Prompt 1-5 (memory only)		
RL	UW	TTR	RL	UW	TTR

Table 2: Overview of the Prompts related to the Speech Variables

Different analyses in SPSS are used to normalize the data of detecting deception with computer-processed analysis (e.g., log10 and square root), and to compare the outcome of analyses between the variables related to the different groups. In order to do that, a variable named "groups" is made, with values 1, 2, and 3. Those values stand for the truthful (1), deceptive (2), and false claiming (3)

participants. The purpose of all the comparisons between the variables related to the groups, was to show which variables are the most effective in detecting deception in Twitter. An example of a conclusion could be that RL gives a similar outcome between the groups, but TTR is very different. Another example can be that using the tweets with only memory would make the standard deviations more overlapping, which means that it is harder to say who is lying or telling the truth. In the Descriptive Analysis the total number of participants, N, minimum, maximum, mean, and standard deviation is calculated.

The purpose of the analysis in this research is to compare absolute differences between the means of the edited and unedited data. After this analysis it can be concluded whether a large amount of fill words and repeating sentences are used, which are not considered "memory", and whether using the edited data would result in a different outcome. After the determination of what data would be used, histograms were created for every prompt, and the total of the prompts related to the groups, for both the edited and unedited data. The histogram-overview in Figures 2, 3, 4, and 5 shows where the data of the groups are similar and where it discriminates whether or not the edited data with only memory or the non-edited data is more useful.

The "Explore" function in SPSS showed the whole edited dataset. SPSS created a table to display the differences in mean and standard deviation for the prompts 1 – 5, and the total of the prompts of the truthful, deceptive, and false claim groups. The display was to conclude which speech content variables related to a prompt would be the most useful in detecting deception, and therefore further analyzed.

The next step was to perform T-Tests to compare means of truthful and deceptive groups related to prompt variables. A t-test is a statistical hypothesis test in which the test statistic follows t-distribution under the null hypothesis. It can be used to determine if two sets of data are significantly different from each other. A t-test is applied when the test statistics are normally distributed, if the value of a scale in the test statistic is known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's t distribution. The T-tests used in this study, "Independent Samples T-Tests", compared two groups with each other. Every T-Test had two hypotheses, "0" (no significance) and "1" (significance).

To assess whether RL, UW, and TTR differed between truthful and deceptive participants, a Multivariate Analyses of Variance (MANOVA) was performed, using Group (truthful, deceptive, and false claim) as the independent variable and the different speech content variables (i.e., RL, UW, TTR related to each prompt of the MCI) as the dependent variables. The MANOVA compares multivariate sample means, and is used when there are two or more dependent variables. Tukey Post Hoc Tests were used to evaluate how speech content variables differed among the three groups (truthful, deceptive, and false claim). Here the significance of multiple variables were being compared. An example is given in Table 3, where the groups truthful, deceptive, and false claim are compared among each other in relation to the Unique Word count in "Tweets Total Edited". The * symbol is used to point out if there is a significant difference. So for example in the table below, truthful participants significantly differs with deceptive participants regarding the use of unique words in the total Twitter interview. Also, the significance number (.012) is close to zero, what should mean that there is a significant difference.

Multiple Comparisons

Dependent Variable: Unique Word Count Tweets Total Edited

	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Truthful	Deceptive	13,01*	4,307	,012	2,54	23,49
		False claim	8,43	4,541	,164	-2,61	19,47
	Deceptive	Truthful	-13,01*	4,307	,012	-23,49	-2,54
		False claim	-4,58	4,475	,566	-15,46	6,30
	False claim	Truthful	-8,43	4,541	,164	-19,47	2,61
		Deceptive	4,58	4,475	,566	-6,30	15,46

Based on observed means.
The error term is Mean Square(Error) = 143,638.

Table 3: Example Post Hoc Tukey

In order to find out on what exact point the chance exists to detect truthful from deceptive participants, a Receiving Operating Characteristic (ROC) curve is used. The ROC analyses provided evidence that a number of the speech content variables could be useful when trying to distinguish truthful from deceptive accounts. However, for both tasks the variable Prompt Total appeared to generate the most area under the curve.

In the example given in Figure 1, there are three curves. The worthless curve (blue) has a 50% chance to be true positive or false positive. At the good curve (purple) the groups are some further separated, and it is more clear who is deceptive and who is truthful. However, there is still a significant marge of error on the left side of the curve. This means that you can only be sure about a select amount of people. The most ideal curve is the yellow one, which shows us that the groups are totally separated. This means that for example, looking at Response Length, almost all the truth tellers have way more to say than almost all the deceivers. This splits the group in two, which leads to a very high accuracy in determining who is telling the truth. In this research the Response Length is limited because of the maximum length of the tweets. It is therefore important to analyze different variables such as Unique Words and Type Token Ratio. In order to prove that MCI in Twitter performs better than chance, and/or polygraph test, the area under the curve must be larger than 50% (chance) or 65% (polygraph).

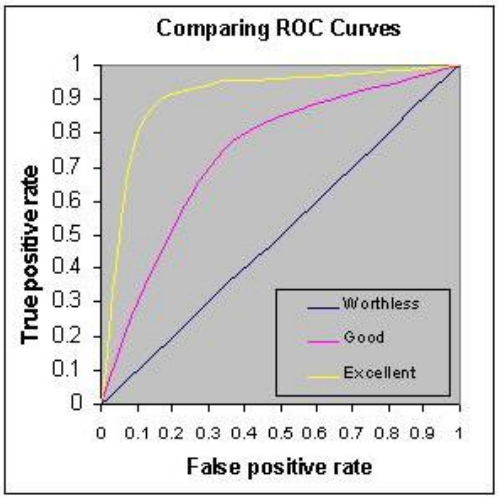

Figure 1: Comparing ROC Curves

4.3 Human Rater Judgment

The first step was creating a survey, using the online platform *TypeForm* (as seen in Appendix B) to find out if the MCI used in Twitter will perform better than chance (e.g., 50%). Seven (former) law enforcement professionals, or people trained in the use of Cognitive Interviewing performed the survey. The questions every rater had to answer per participant were:

1A: Do you think the tweet below is deceptive or truthful?

1B: On a scale from 1 to 5 how confident are you about your choice?

Question 1A and 1B are related to the first tweet of the Twitter interview. The raters had the choice to assign the tweet as deceptive or truthful, and rated the level of confidence about their choice.

1C: Do you think the tweets below are deceptive or truthful?

1D: On a scale from 1 to 5 how confident are you about your choice?

Question 1C and 1D are related to the whole Twitter interview. The rater had the choice to assign the whole conversation as deceptive or truthful and rated the level of confidence about their choice.

Each expert independently reviewed the transcripts of the MCI. After reading the transcript, each expert rendered a judgment about a participant's status (truthful/deceptive) based on the personal experience of each rater. If a rater judged a participant to be deceptive, their judgment was coded as a "1"; if truthful, their judgment was coded as a "0".

In SPSS the status of each participant is displayed as a "1" or a "0". Every truthful participant is assigned as a "0" and every deceptive participant as a "1". This variable is then compared to the answer from the rater. If the two variables match then the rater made a correct judgment. If they did not match then the judgment of that rater was incorrect.

Individual cross-table analyses were performed using the variables status of the participant (truthful "0" and deceptive "1") and the individual judgment scores from the rater (truthful "0" and deceptive "1"). These individual cross-tables were made for every rater on their judgment of the first tweet (Table 5) and the Whole Conversation (Table 6). The table has four possible outcomes as displayed in Table 4.

		Rater Judgment	Rater Judgment
		Truthful	Deceptive
Participant Status	Truthful	**True Positive**	**False Positive**
Participant Status	Deceptive	**False Negative**	**True negative**

Table 4: Cross-Table Outcomes

True Positive:
The participant status is truthful and the rater judged the participant as truthful, thus being correct.

True Negative:
The participant status is deceptive and the rater judged the participant as deceptive, thus being correct.

False Positive:
The participant status is truthful and the rater judged the participant as deceptive, thus being incorrect.

False Negative:
The participant status is deceptive and the rater judged the participant as truthful, thus being incorrect.

This resulted in cross-tables like the one presented in Table 5. Table 6 shows that on the first tweet, rater 4 had 14 True Positives, 1 False Positive, 26 False Negatives and 3 True Negatives. The True Positives and the True Negatives added together is the total of correct answers. In Table 6 the rater had 17 out of 44 correct answers when only the first tweet was given. In Table 6 the rater had 27 out of 44 correct answers when the whole Twitter interview was given.

Status Participant Truthful vs Deceptive * Rater 4 tweet 1 Judgment Crosstabulation

Count

		Rater 4 tweet 1 Judgment		Total
		Truthful	Deceptive	
Status Participant Truthful vs Deceptive	Truthful	14	1	15
	Deceptive	26	3	29
Total		40	4	44

Table 5: Cross-Table Example 1

Status Participant Truthful vs Deceptive * Rater 4 Whole Conversation Judgment Crosstabulation

Count

		Rater 4 Whole Conversation Judgment		Total
		Truthful	Deceptive	
Status Participant Truthful vs Deceptive	Truthful	10	5	15
	Deceptive	12	17	29
Total		22	22	44

Table 6: Cross-Table Example 2

The success rate of the rater is calculated by number of correct answers times 100 divided by the total amount of participants. For tweet one, this results in:

$$(17 \times 100) : 44 = 38{,}64\%$$

For the whole conversation the success rate is:

$$(27 \times 100) : 44 = 61.36\%$$

To prove that human raters perform better than chance (e.g., 50%) when rating MCI in Twitter, the success rate of all raters is calculated for both the first tweet and the whole conversation. To find the mean of all rater judgment the success rate of tweet one from every rater is added up and divided by the total number of raters. The same is done with the results from the whole Twitter conversation. If this percentage is higher than 50% then human raters perform better than chance when using the MCI in Twitter.

If these two results are compared to each other, there can be two different outcomes. If the first success rate is higher than the second success rate, it means that it is not more effective to see the whole MCI interview in Twitter, thus making it not an effective method of detecting deception in Twitter when judged by human raters. If the second success rate is higher than the first success rate it means that the MCI method expands the memory of the participant, which makes it easier for

human raters to determine if a participant is deceptive or truthful. Therefore making the MCI method more effective than only reading the initial tweet.

However, this does not determine the complete effectiveness of MCI in Twitter. The overall effectiveness of MCI in Twitter judged by human raters is determined by comparing the success rate to other detecting deception methods.

Not only the success rate was, but also the confidence level was rated. As a follow up question to the truthful or deceptive question, the raters were asked about their confidence level of their answer regarding the previous question. The confidence level is measured in a variation of the 1 to 5 Likert scale. The scale used is displayed in Table 7. This ordinal scale captures the range of how confident the participant feels about their choice.

Question: *On a scale from 1 to 5 how confident are you about your choice?*

Possible answers: Table 7

Level	Definition
5	Very confident
4	Confident
3	Not sure
2	Not quite confident
1	Not confident

Table 7: Possible Answers

By using the comparing means function in SPSS the mean of the confidence level is calculated. The overall mean of both the confidence level of tweet 1 and the whole Twitter conversation are calculated by adding the mean of every rater and dividing the result by the total number of raters. Thereafter, the overall confidence level is made into a percentage and is compared to the overall success rate of the raters. This comparison shows if the raters are either overconfident, realistic, or insecure about their performance.

- When the confidence level is higher than their success rate, it means that the raters are overconfident about their performance of detecting deception in Twitter.
- When the confidence level is the same as their success rate, it means that the raters are realistic about their performance of detecting deception in Twitter.
- When the confidence level is lower than their success rate, it means that the raters are insecure about their performance of detecting deception in Twitter.

Finally, the raters were asked to elaborate on the method they used to determine if a participant was deceptive or truthful. The question they got asked was:

Can you elaborate on the method you used to rate the tweets? What made you decide if they seemed truthful or deceptive?

The purpose of this question is to gain insight in the method the raters used. If one rater scored significantly better than another rater, it is interesting to know if it was because they used a different method or if it was due other influences. This can also be helpful to determine a possible effective method of detecting deception in Twitter by human raters.

4.4 Comparison with Previous Study

To determine how effective the MCI method of detecting deception is compared to the results from the study conducted by Morgan et al. (2015), the results from both researches are put next to each other in a percentage scale. Both the human raters and the computer-processed data are compared in the results. Due to the differences in studies, it is not valid to compare all tests and results directly with each other. For example, in Morgan et al. (2015) the participants have an unlimited Response Length, while in this research the responses are very limited. This means that in the result section (Chapter 5.2) the comparison is mainly focused on the end results instead of comparing individual tests with each other. However, some tests can be directly compared, such as the success rate of the human raters and the ROC curve.

By carrying out a number of tests as described in Chapter 4 (e.g., Cross-Table Tests, T-Tests, ROC Curves, MANOVA) the most effective method of detecting deception in Twitter is determined in the results. This method is compared to the most effective method of Morgan et al. (2015). Morgan et al. (2015) used the ROC curve, general linear model multivariate analyses of variance, cross tables and calculated the speech variables.

5. Results

In this chapter the results of the applied research methods are discussed. This relates to raw results without interpretation. The interpretation of the results and answering the research questions can be found in the conclusion and recommendation chapters.

5.1 Computer-Processed Analysis

In this chapter the overall effectiveness for the use of MCI in Twitter with computer-processed data is analyzed. Different variables and the separate prompts are compared with each other, and the results consist of the most discriminated data.

The collected data is edited as described in Chapter 4.2. In the edited data only words and sentences that fall under "memory" are present. To see if the edited data displays any notable differences with the unedited data, a descriptive statistics analysis is performed, using both the edited and unedited datasets. The descriptive statistics analysis in Table 8, shows a significant difference between the means of the edited and unedited data. For RL tweet Total (M=123.64, SD=27.08), the mean is notably higher than RL tweet Total Edited (M=93.82, SD=23.74). Where unedited data contains a large amount of filler words and repeating sentences (which are not memory), and therefore the mean of the edited data is notably lower than the mean of the unedited data. As to Unique Word Count tweet Total (M=74,9 and SD=15,08), the mean is also notably higher than Unique Word Count tweets Total Edited (M=57,98 and SD=12,96).

	N	Minimum	Maximum	Mean	Std Deviation
Response Length tweet Total	44	32	163	123,64	27,076
Response Length tweet Total Edited	44	12	128	93,82	23,740
Unique Word Count tweet Total	44	28	106	74,95	15,083
Unique Word Count tweets Total Edited	44	11	80	57,98	12,964
Valid N (listwise)	44				

Table 8: Descriptive Statistic Analysis

5.1.1 Histogram-Overview

After differences between the means and standard deviations of edited and unedited data has been determined, it is established how this affects the distribution of the different groups related to the variables. To see the difference in distribution of the different groups, both datasets are made into histograms in Figures 2-5. As shown in these figures, there is a notable difference between UW and RL in edited and unedited data. As shown in the histograms below the UW (x-axis Figure 2 and 3) and the RL (x-axis Figure 4 and 5) are set against the number of participants (Frequency).

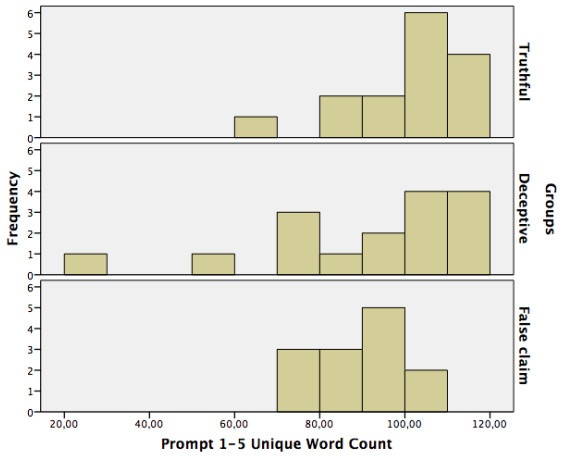

Figure 2: Unique Word Count Unedited Data

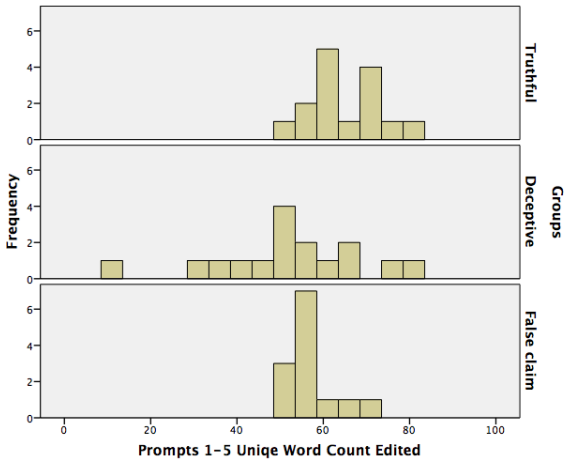

Figure 3: Unique Word Count Edited Data

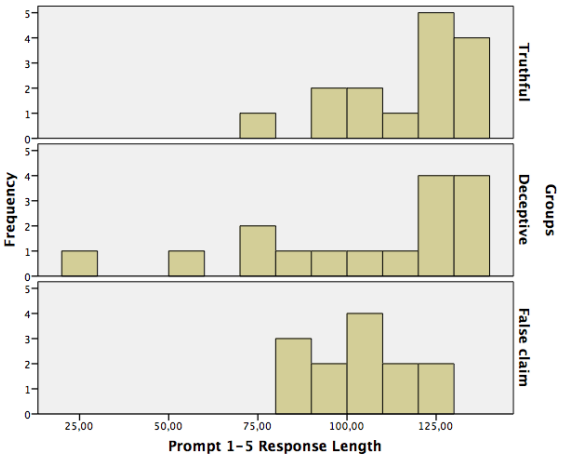

Figure 4:Response Length Edited Data

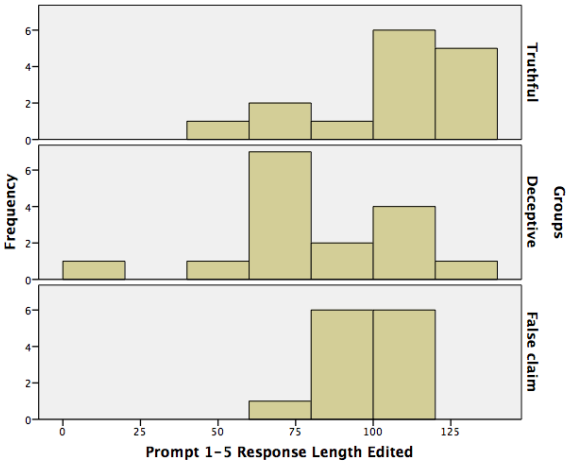

Figure 5: Response Length Unedited Data

As seen in Figure 2, there is a large amount of overlap between the different groups. The majority of the respondents in all groups had a UW count between 80 – 120.

In the edited data, there is a larger spread between the groups. The largest amount of respondents in the truthful group has a UW count between 60 and 80, while the largest amount of respondents of the deceptive and false claim groups are respectively set between 30 – 60 and 50 – 60 unique words.

The edited data of the RL displays also less overlap, although in a less significant way. The majority of the truthful group has a RL between 100 – 130 words, while the RL of the deceptive and false claim groups are respectively set between 60 – 120 words and 80 – 120 words. Because the edited data shows less overlap than the unedited data, the best results can be realized for the use of MCI in Twitter by using the edited data.

Analyses of the MCI derived speech content variables (Type-Token Ratio (TTR), Response Length (RL) and Unique Words (UW)) are noted in Table 10. This data was analyzed through the “Explore” function in SPSS, and after that compared with the use of a Post Hoc Tukey Analysis. The table shows that there are only small differences in TTR, which makes it an ineffective speech content variable in detecting deception.

5.1.2 T-Test

To see if there is significance between truthful and deceptive groups, a T-test is carried out. The T-test for the speech content variables RL Total and UW Total has been performed individually for every variable, with the hypotheses:

H0: *There is no significant difference between the truthful and deceptive group with RL/UW/TTR as a variable.*

H1: *There is a significant difference between the truthful and deceptive group with RL/UW/TTR as a variable.*

The T-test indicated if there is significance between the truthful and deceptive groups, with respect to the variables Response Length and Unique Word count. The results show a significant difference in the scores related to Response Length for truthful (M=103.87, SD=20,636) and deceptive (M=88.62, SD=23.886) conditions; t(42)=2.10 , p=0.042.

For Unique Word count it also resulted in a significant difference in the scores related to Unique Word count for truthful (M=65.20, SD=8.064) and deceptive (M=54.24, SD=13.535) conditions; t(42)=2.87 , p=0.006.

In the Levene's test, as shown in Table 15 of Appendix A, the significance (Sig.) between the truthful and the deceptive group related to Response Length and Unique Word count is respectively 0.688 and 0.367. These numbers are both higher than 0.05, what means that there is equality in the variance, so the equal variance assumed row should be used. 0.05 is the statistical standard value to determine if significance exists. The next step is to look at the 2-tailed significance (Sig. 2-tailed), which in this case is for Response Length 0.042 and for Unique Word count 0.006, thus both being smaller than 0.05, which means that there is a significant difference between the truthful and the deceptive groups. The significance difference between truthful and deceptive groups for both the RL and UW count means that for both variables hypothesis 1 is valid, and 0 is invalid.

For TTR it also resulted in a significant difference in the scores related to Unique Word count for truthful (M=1.59, SD=0.25) and deceptive (M=1.62, SD=0.24) conditions; t(42)=-0.46 , p=0.65. Because the 2-tailed significance is higher than 0.05 there is no statistically significant difference between the conditions.

5.1.3 Multivariate Analysis of Variance

The Multivariate Analyses of Variance indicated the presence of significant differences between the three groups of participants with respect to the different variables (Table 10). With respect to the variable Response Length, significant differences were noted for: Prompt One (F (2,41) = 2.36; p= 0.107), Prompt Two (F (2,41) = 3.83; p=0.030); Prompt Four (F (2,41) = 2.64; p=0.084); Memory Prompt (F (2,41) = 3.60; p=0.036), and the Prompt Total (F (2,41) = 3.06; p=0.058)

Significant differences between the groups of participants were also noted for the variable Unique Word count in response to, Prompt Two (F (2,41) = 3.55; p=0.038); Prompt Three (F (2,41) = 2.36; p=0.107); Prompt Four (F (2,41) = 3.28; p=0.048); Memory Prompt (F (2,41) = 3.83; p=0.030), and the Prompt Total (F (2,41) = 4.66; p=0.015)

Finally, except for Prompt One (F (2,41) = 4.13; p=0.023) no significant differences between the groups of participants were noted for the speech content variable TTR in: Prompt Two (F (2,41) = 1.55; p=0.225); Prompt Three (F (2,41) = 0.24; p=0.789); Prompt Four (F (2,38) = 0.11; p=0.892); Prompt Five (F (2,40) = 1.54; p=0.228), and the Prompt Total (F (2,41) = 0.62; p=0.543).

The primary ways in which the truthful, false claims and deceptive groups differ can be seen in Figures 6-9 below. Truthful, deceptive and false claiming participants had an equal amount to say in response to the first MCI prompt. Even though the table shows that there are significant differences, these are too small to make a reliable conclusion.

Prompt 2, 3 and 4 trigger the expansion the participant's memory. As displayed in Figures 8 and 9, the truthful group displays in all prompts a significant difference between at least one other group in UW count. Also the RL shows a significant difference in prompt 2 and 4 with at least one of the other groups. In the Memory Prompt (prompt 2, 3 and 4 together) the RL and UW of the truthful group differs significantly from the deceptive and false claim group.

Just as with the first prompt, the participant had about an equal amount of words to respond in the fifth prompt. As mentioned before, the sixth prompt is left out because after editing the data, the sixth prompt became in most cases blank. Just like in the research of Morgan et al. (2015), the sixth prompt was an experimental prompt that did not improve the effectiveness of the MCI.

The Total Prompts, as displayed in Figure 6 and 7, show a significant difference between the truthful and deceptive groups in RL and UW count. Also, it should be noted that, although not significant, there is a notable difference in UW count between the truthful and false claim group.

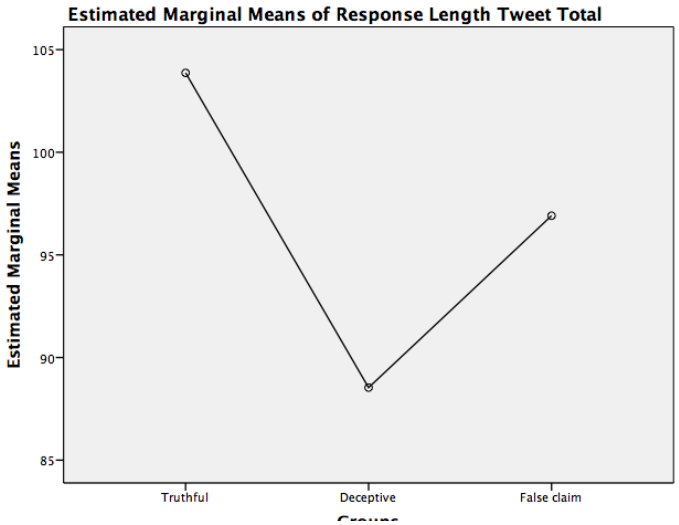

Figure 6: Means Response Length, tweets Total

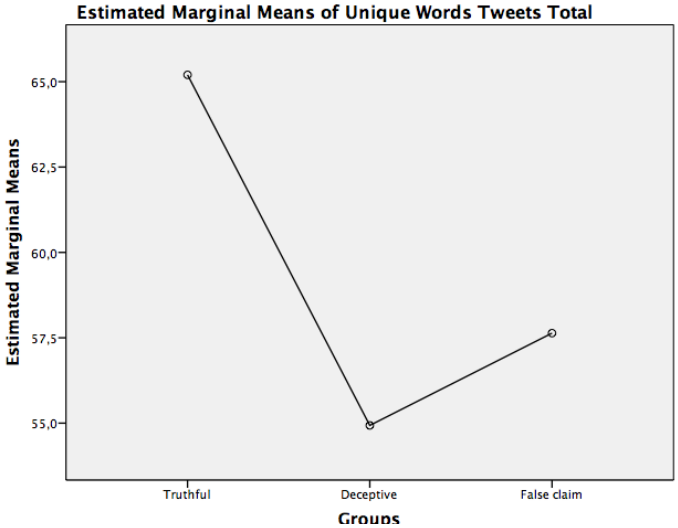

Figure 7: Means Unique Words, tweets Total

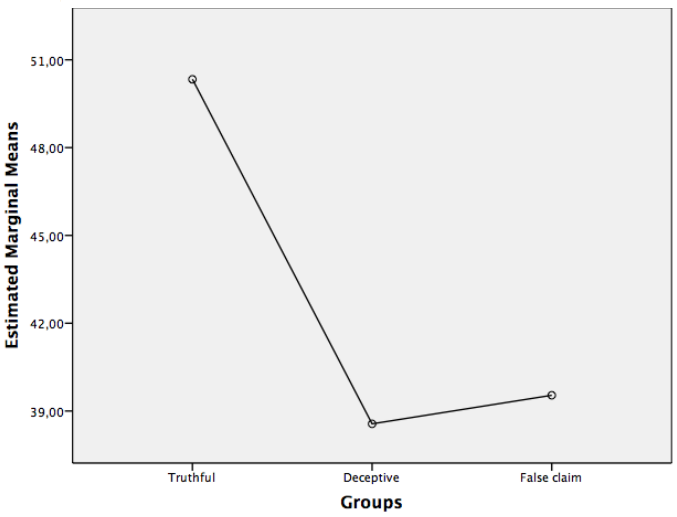

Figure 8: Means Unique Words, Memory Prompt

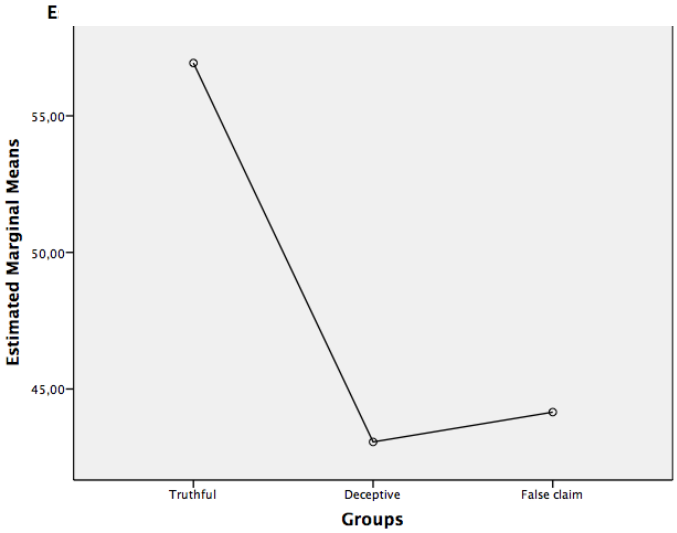

Figure 9:Means Response Length, Memory Prompt

5.1.4 Receiver Operating Characteristics (ROC)

After establishing that there is a significant difference between truthful and deceptive groups, for both RL and UW count, a ROC curve is performed. As noted in Chapter 4.2 the ROC curve can be used as a way to distinguish truthful from deceptive accounts. However, the variable Prompt Total appeared to generate the largest area under the curve. For example, if the UW data is used, the probability of being wrong (i.e., 1-Specificity in Appendix A-5) in judging a claim containing 75 unique words (or more) as "truthful," would be approximately 6.9%. Similarly, the probability of being wrong in making this judgment if the claim had 61 or 54 words would be respectively 24% and 59%.

The scale below is the classic classification of the area under the curve.

- .90-1 = excellent
- .80-.90 = good
- .70-.80 = fair
- .60-.70 = poor
- .50-.60 = fail

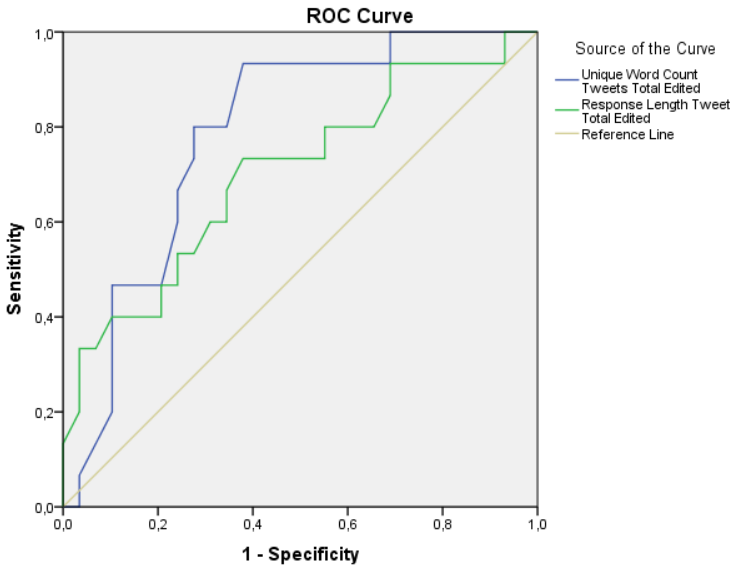

Figure 10: ROC Curve

Table 9 states that Area under the curve (1-Specificity) as seen in Table 16 is 0.702 (70.02%) for the RL, and 0.785 (78,5%) for the UW count. This means that if the RL or UW count is used as a variable there is an overall chance of respectively 70.02% and 78.5% that a truthful participant can be distinguished from "N" (total number of population). On the classification scale the percentages of both the RL and UW are classified as fair. The results (i.e., 70% and 79%) are higher than chance (i.e., 50%) and polygraph (i.e., 65%).

Area Under the Curve

Test Result Variable(s)	Area
Unique Word Count tweets Total Edited	,785
Response Length tweets Total Edited	,702

Table 9: Area under the Curve

Table 10: Speech Variables

Speech Variables for Truthful, False Claims and Deceptive Participants (Alpha = 0.10)

	Truthful	Deceptive	False Claim
Prompt One [Detailed Account]			
RL	24.27 (SD = 4.35)	22.19 (SD = 7.49)+	24.74 (SD = 3.10)*
UW	20.87 (SD = 3.07)	18.75 (SD = 5.64)	21.38 (SD = 1.94)
TTR	1.16 (SD = 0.08)+	1.17 (SD = 0.09)+	1.25 (SD = 0.09)*□□
Prompt Two [Visual Prompt]			
RL	20.40 (SD = 6.12)*	14.38 (SD = 6.59)□□	18.15 (SD = 5.51)
UW	17.80 (SD = 4.75)*	13.06 (SD = 5.47)□□	16.54 (SD = 5.09)
TTR	1.14 (SD = 0.11)	1.08 (SD =0.09)	1.10 (SD = 0.07)
Prompt Three [Auditory Prompt]			
RL	19.13 (SD = 5.71)	16.31 (SD = 6.70)	15.08 (SD = 4.54)
UW	17.03 (SD = 4.45)+	14.63 (SD = 5.51)	13.38 (SD = 3.36)□□
TTR	1.11 (SD = 0.10)	1.10 (SD = 0.08)	1.12 (SD = 0.08)
Prompt Four [Emotional Prompt]			
RL	17.40 (SD = 7.06)+	12.38 (SD = 8.85)	10.92 (SD = 7.75)□□
UW	15.47 (SD = 6.00)+	10.88 (SD = 7.01)	9.62 (SD = 6.27)□□
TTR	1.11 (SD = 0.09)	1.09 (SD = 0.13)	1.10 (SD = 0.11)
Prompt Five [Temporal Prompt]			
RL	24.47 (SD = 3.54)	20.38 (SD = 7.75)	23.85 (SD = 8.31)
UW	19.27 (SD = 2.55)	17.19 (SD = 6.32)	19.15 (SD = 2.54)
TTR	1.28 (SD = 0.16)	1.19 (SD = 0.13)	1.24 (SD = 0.15)
Total Prompts 1 – 5			
RL	103.87(SD = 20.64)*	83.75 (SD = 29.70)□□	94.62 (SD = 12.57)
UW	65.20 (SD = 8.06)*	52.19 (SD = 17.21) □	56.77 (SD = 6.70)
TTR	1.59 (SD = 0.25)	1.62 (SD = 0.22)	1.69 (SD = 0.25)
Memory Prompts 2 – 4			
RL	56.93 (SD = 15.72)*□	43.06 (SD = 18.16)□□	44.15 (SD = 11.98)□□
UW	50.33 (SD = 12.69)*+	38.56 (SD = 14.78)□□	39.54 (SD = 10.44)□□
TTR	1.12 (SD = 0.07)	1.10 (SD = 0.08)	1.11 (SD = 0.05)

* Variable differs significantly from Deceptive.
+ Variable differs significantly from False Claims Group.
□□Variable differs significantly from Truthful Group.

5.2 Human Rater Judgment

5.2.1 Success Rate

Conducting individual cross-tables analyses in SPSS resulted in the data shown in Table 11. It shows for every rater how many true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) they generated. The TP and TN added together, gives the number of correct answers. The percentage of the correct answers from the total population (44) states the success rate of the rater to discriminate truthful from deceptive participants. In the bottom row the means of all the results are shown. The means are rounded to two decimals, which means that the average rater would score, when only given the first tweet, 12.71 TP, 8.14 TN, 2.29 FP and 20.86 FN. When the whole Twitter conversation is given, the average rater would score 10.43 TP, 13.29 TN, 4.57FP and 15.71 FN. This means that the average rater was correct 20.86 times out of 44 when shown only the first tweet and 23.71 times out of 44 when given the whole Twitter conversation. The success rate is determined with the use of the formula described in Chapter 4.3. This resulted in a success rate of 47.82% when only the first tweet was given and an overall success rate of 54.04%.

Rater #	True positive	True negative	False positive	False negative	Correct	Success rate
1: Tweet 1	14	13	1	16	27	61.36%
1: Total Conv.	14	13	1	16	27	61.36%
2: Tweet 1	13	6	2	23	19	43.18%
2: Total Conv.	8	18	7	11	26	59.09%
3: Tweet 1	14	4	1	25	18	40.91%
3: Total Conv.	12	3	3	26	15	35.09%
4: Tweet 1	14	3	1	26	17	38.64%
4: Total Conv.	10	17	5	12	27	61.36%
5: Tweet 1	11	9	4	20	20	45.46%
5: Total Conv.	9	13	6	16	22	50.00%
6: Tweet 1	10	8	5	21	18	40.91%
6: Total Conv.	7	13	8	16	20	45.46%
7: Tweet 1	13	14	2	15	27	61.36%
7: Total Conv.	13	16	2	13	29	65.91%
Mean	11.57	10.71	3.43	18.29	22.29	50.72%
Mean T1	12.71	8.14	2.29	20.86	20.86	47.82%
Mean WC	10.43	13.29	4.57	15.71	23.71	54.04%

Table 11: Cross-Table Results

5.2.2 Confidence Level

By carrying out a means test in SPSS the means of every rater's confidence level is determined in Appendix A. To calculate the average confidence level of all raters, all means were added together and divided by the total numbers of raters. The outcome of this calculation, is an average confidence level of 3.39 out of 5 with an average standard deviation of 0.726 when shown only the first tweet, and 3.60 out of 5 with an average standard deviation of 0.865, when the whole Twitter conversation is given. When converted to a percentage, the results for the first tweet and the whole conversation are respectively 67.80% and 72%.

5.2.3 Human Rating Approach

Finally, the human raters were asked to elaborate about the method they used to determine if a tweet/Twitter conversation was truthful or deceptive. They gave a number of different explanations about the methods they used:

“For the most part if I could imagine what they were saying I rated them as truthful. However, I went with my gut on some.”

“Tried MCI, but difficult in limited character format.”

“The level of detail/specifics within each tweet, asking myself if someone would really do what the 'tweeters' claimed to be doing.”

“First I looked at each series of tweets to see if they provided more information as they went along, building the picture of what they did. Next I looked at how they performed -- a couple of them messed up details in the reverse-order question, such as the name of a building.“

“Use of past tense, sensory detail, additional detail added throughout the answers. Them actually answering the questions.”

“If the subject added greater or different (but not conflicting) detail to each prompt, he/she seemed truthful.”

The answers given by the participants about their method are further discussed in Chapter 6.2.

5.2.4 Human Raters vs. Chance and Polygraph

As described in Chapter 5.2.4, the success rate of detecting deception in Twitter by human raters is 47.82% when the first tweet is displayed, and 54.04% when the whole Twitter conversation is given. The rating from the first tweet is lower than chance (i.e., 50%) while the raters scored higher than chance when displayed the whole Twitter conversation. Regarding the polygraph, both ratings scored notably lower.

5.3 Comparison with Previous Study

Before starting with the comparison of the results, it should be mentioned that there are some major differences between the two studies. The most significant differences are (as displayed in Table 12) sample size, performed tasks, answering space, the method of interviewing, transcribing, number of human raters, interview scale, and the interview questions.

Differences	**Morgan et al. (2015)**	**This research**
Sample size	102 military personnel	44 UNH students
Performed task(s)	Cognitive and manual task	Cognitive task only
Answering space	Unlimited space to answer the questions	Text space limited to 140 per question
Method of interview	Spoken answers to the questions	Typed answers to the questions in Twitter
Transcribing	From speech to text	From Twitter timeline to text
Number of raters	Three students trained in MCI	Seven law enforcement professionals and people trained in the use of MCI
Interview scale	Individual interviews	Multiple participants at the same time
Interview questions	"Normal" MCI questions	MCI questions modified to fit in one *tweet* (140 characters)

Table12: Research Differences

In the study of Morgan et al. (2015) a cognitive and manual task were analyzed. For both studies, only results of the cognitive task are used in the comparison. In the study of Morgan et al. (2015), the human rater judgments and computer-processed speech analysis performed better than chance; computer based judgments were superior to the human judgments (i.e., 82% vs. 62%, respectively). Speech content variables derived from MCI differed significantly, and in different ways, between the truthful and false claiming participants and also between the truthful and denial (deceptive) type participants.

In our research the computer-processed analysis and human ratings also scored higher than chance (i.e., 79% vs. 54%). However, both scored lower than the research from Morgan et al. (2015), and there is more difference between the computer and human judgment, as seen in Figure 11.

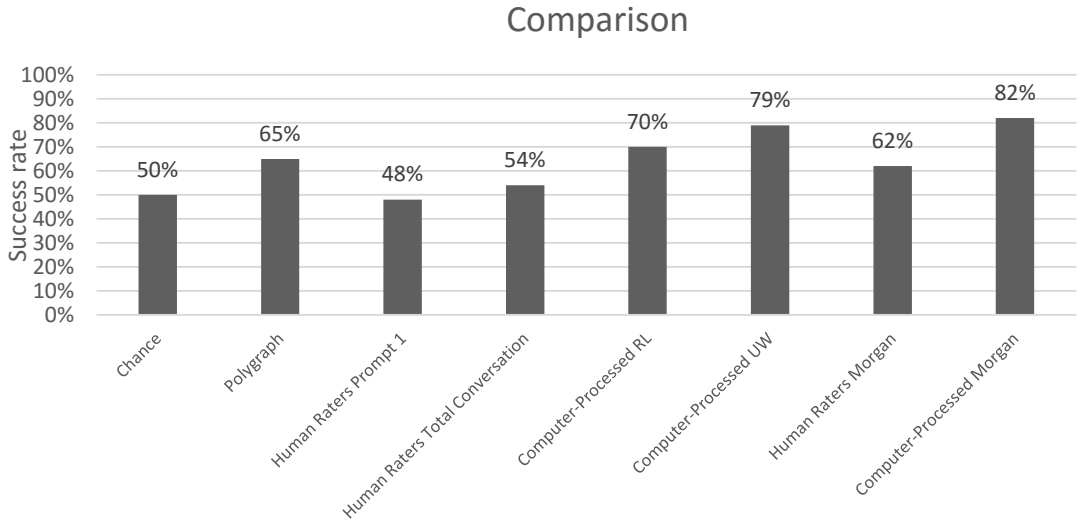

Figure 11: Comparison of Detecting Deception Methods

6. Conclusion and Discussion

In the conclusion and discussion, the answer to the problem statement is given and is further discussed in perspective to existing literature.

6.1 Conclusion

"Is detecting deception possible using the Modified Cognitive Interview method in the computer-mediated environment of Twitter?"

The problem statement is approached in two different ways, first by computer-processed analysis of tweets, with the speech content variable Response Length (RL), Unique Word count (UW), and Type Token Ratio (TTR), and second by proving the effectiveness by human rating judgment.

Regarding computer-processed analysis, a Descriptive Statistics Test is performed in order to determine if the edited or unedited dataset is more useful in our research. This showed that the edited data had a more positive impact on the used data, and that the original dataset contained a large amount of filler words and repeating sentences. After this determination, a T-Test was conducted to prove that there is a significance difference between the speech-content variables RL and UW. TTR was eliminated in the T-Test because it revealed that the text limitations of Twitter made the TTR an ineffective variable for this research.

By performing a Multivariate Analysis of Variance (MANOVA) significant differences among the groups (truthful, deceptive, and false claim) arose. The MANOVA enabled us to create an overview in Table 10 of the means combined with all significant differences among the groups. The memory prompt (prompt 2, 3, and 4 combined) presented compelling differences. This is a first indication that by poking the participant's memory, the groups responded differently. Also, regarding the Prompts Total, significant differences were measured. The Receiver Operating Characteristic (ROC) analysis showed us that the "Total Prompt UW" variable contained the largest area under the curve. By using the Total Prompt UW in this analysis, we can be 79% sure to distinguish a truthful participant from the total population.

As a second approach to the problem statement, expert ratings, confidence levels, and methods from human raters are collected. By comparing the rating results in cross-tables, the highest success rate of the human raters is determined. This shows that the human raters achieved a higher success rate when given the whole Twitter conversation, instead of only the first tweet (respectively, 54% vs. 48%). The 54% success rate went along with a confidence level of 3.6 out of 5 (i.e., 72%). This means that the human raters overestimated their abilities to discriminate truthful from deceptive participants.

When comparing our results with Morgan et al. (2015), we can conclude that human rater judgment is more effective in real life, than it is in computer-mediated communication with the use of limited text space. Human ratings for Twitter, in our research, scored only 4% higher than chance, which is not significant. In both studies there was a significant difference between the success rate of the computer-processed analysis versus the human rater judgment. The general conclusion is that the computer-processed analysis with the use of speech-content variables RL and UW, is both effective in real life and in communication through a digital platform with limited characters as in Twitter (i.e., 82% and 79%, respectively).

As a comprehensive answer to the problem statement, we can conclude that the use of MCI can be an effective method of detecting deception in a limited text based environment, such as Twitter. However, this only applies when using computer-processed analysis, with the speech-content variables RL and UW as the leading factors.

6.2 Discussion

Consistent with previous studies using the Cognitive Interview Method to detect deception, the present data analysis indicates that notable differences exist in the behavior of truthful and deceptive groups when exposed to cognitive interviewing techniques. Truthful people had more to say and used a wider variety of unique words than both deceptive groups, which corresponds with Morgan et al., (2014; 2015). These differences were seen when analyzing the transcripts and when assessing speech variables obtained by the use of MCI. However, in addition to what has previously been known about this issue, the present research indicates that even when exposed to MCI in the limited text environment of Twitter, MCI is still an effective method of detecting deception, although to a lesser extent than when the MCI method is used in a real life interview (78% vs. 82%).

Regarding computer-processed analysis, we hypothesized that RL, UW count, and TTR would be greater in truthful compared to deceptive individuals. Within the framework of this study, we can conclude that the RL and UW count is in fact greater in truthful than in deceptive individuals, which corresponds to previous studies considering the MCI method (Morgan et al., 2014; 2015). In contrast to previous studies (Morgan et al., 2014; 2015), the TTR seemed to have a very limited usability because of the limited text space. The TTR could not be used effectively to discriminate truthful from deceptive individuals, while in the previous mentioned studies, the TTR was a very useful indicator of distinguishing liars from truth tellers.

Contrary to Morgan et al. (2015), the answer to the first prompt was indistinguishable for all groups with respect to the variables RL, UW, and TTR. However, on the memory prompt, the truthful group had a greater RL and UW than the deceptive groups. Within this context, it is reasonable to hypothesize that this reduced responsiveness to the memory prompts by the deceptive groups, could be due to the increased cognitive load associated with lying as described in Vrij et al. (2009) and a desire to “tell their story” and stick to it so as to be believed. As stated by Morgan et al. (2015) this could be because in essence, they have memorized a story they want to “sell” to the interviewer and they tell the entire “memory” when exposed to the initial prompt. As a result, these types of liars have little to elaborate on when subsequently exposed to the memory prompts. The prompt total showed the same results as the memory prompt, although to a lesser extent because of the similarities among the groups in prompt 1 and 5.

A study by Zhou (2003), concluded that deceptive individuals display a higher RL than truth tellers in CMC with the use of instant messaging. With the differences among the research taken into account, this result runs contrary to our finding that deceivers have a lower RL than truth tellers, which is more in line with Morgan et al. (2015). The reason for these different results could be attributed to the fact that in our research, the data was stripped of unnecessary words that were not memory before being analyzed. However, just as in the study of Zhou (2003), our study showed that deceivers have a lower UW count than truth tellers. Overall, the findings in this study show results similar to other studies on the same subject.

Considering human rater judgment, the data indicates that human raters performed slightly better than chance (respectively 54% vs. 50%), and notably worse than the other methods addressed in this study, such as the polygraph test (65%) and face-to-face MCI (82%). This score could be explained by the lack of context and the limited text space. Multiple raters indicated that they struggled with these factors. The raters scored notably lower than in a previous study (Morgan et al. 2015). This could indicate the limit of human raters’ effectiveness in judging the MCI.

We also hypothesized that computer assessed variables perform better than humans assessing the Twitter communications in order to detect deception. Our results showed that this hypothesis was correct, and that the results of this study are similar to previous studies. The study of Morgan et al.

(2015) showed that computer assessed variables score higher than human raters (62% vs. 82%); our study showed similar results (54% vs. 79%).

By analyzing the ROC curve, we were able to demonstrate how data obtained from MCI in Twitter can be used in enterprises. When adopting a truthful database as a framework, MCI in Twitter could be used to discriminate truthful from deceptive individuals in a large population. This method may also work in other computer-mediated communications with a limited text space. MCI also is potentially useful for internal investigation within a company, or to reduce fraudulent insurance claims, which can help insurance companies to better allocate their resources.

In addition, the results of this study could be of value for law enforcement agencies. When further developing this method, MCI in Twitter could be used to discriminate truthful from deceptive eyewitness accounts and suspect accounts within a reliable framework.

Taking all that is previously mentioned in consideration, we can conclude that the MCI used in Twitter shows similarities with previous research, and is yet another effective way in which MCI can be used to detect deception.

7. Recommendations

In this chapter, recommendations for conducting further research regarding detecting deception in computer-mediated communication, using the Modified Cognitive Interview (MCI) method are discussed. The recommendations aim to indicate possible improvements that can be made in this study, and where further studies should focus on.

7.1 To Improve this Study

When using limited text-based analysis, the elimination of filler words and repeating sentences has a positive effect on the results. The limitation of characters in text answers causes an increase of similar results between the groups. By only keeping memory content, the differences between the groups will possibly be larger.

The use of the experimental prompt “did you leave anything out?” does only contain a small amount of memory content, and therefore does not give many positive or negative results. We recommend leaving this prompt out of future studies. The prompts 2, 3, and 4 contain the largest amount of memory content. Next studies should focus on these prompts.

Because of the limited Response Length (RL), the Type Token Ratio (TTR) is not very useful in this type of study. We recommend to mainly focus on Response Length and Unique Word count (UW). Especially UW is very effective when using limited text space responses.

Human rater judgment, by law enforcement professionals and lie detecting experts, are not very effective in the limited text space format of Twitter (54%). When only rating the first tweet, the results were even lower (48%), which means that human rater judgments are more effective when there is a larger amount of speech-content, such as in the previous study of Morgan et al. (2015). From this we can conclude that we reached the minimum number of words that can be used for human rater judgment.

7.2 Future Studies

Detecting deception in our research, with the use of a computer-processed text analysis, achieved a success rate of discriminating truthful participants from deceptive individuals of 79%, when using the UW related to prompts 1-5. If this percentage can be achieved in future studies, which should be individually designed for specific work fields, resources can be allocated significantly more effective. For example, in law enforcement agencies and insurance companies, determination of suspects and frauds can be performed with help of smart calculations, that can highly increase the effectiveness.

An important recommendation is to perform this study with participants, who are responding from their own device, with their own Twitter account, and from remote locations. This increases the validity of the study because it is closer to reality. It is also useful to see how participants are responding when using their own preferred device, instead of an unfamiliar PC or laptop.

As mentioned previously, a variety of applications can be realized by using this study for both the private and public sector. Further studies in this matter can eventually lead to multiple effective methods of detecting deception in an online environment. Examples of applications in both the public and private sector are listed below.

Public Sector

- Interrogation in high risk and remote areas (e.g., Iraq, Syria)
- Judging the level of trustworthiness in online communication with criminal and terrorist organizations through social media platforms
- Increasing the efficient use of resources in law enforcement agencies
- Extending the reach of intelligence agencies
- Modernization of investigative methods in both real-life and cybercrime

Private Sector

- Economize insurance companies, by allocating resources more effectively
- Extending the reach of investigation and intelligence gathering companies
- Increasing the effectiveness of pre and in-employment screenings

After performing further research on this topic, the aforementioned opportunities can be realized. In the public sector interrogations can take place from remote distances, and with criminals or terrorists who are normally unreachable. Responses can be evaluated, and digital investigation methods can be modernized. In the private sector intelligence companies can evaluate their findings on whether information is more likely to be truthful or more likely to be deceptive, and development of pre and in-employment screenings can provide more accuracy in assessing (future) employees. In both sectors, resources can be allocated more effectively, which provides significant lower costs in business operations.

8. Responsibilities

In this chapter the responsibilities of both researchers are explained in detail, containing all the tasks that are done individually. The methods used are theoretically justified, and supported by scientific references that indicate the selected methodologies as the most suitable ones.

In the preparation of the thesis we agreed to split the processing part in half. This way we were able to gather the data together, and write two different parts of the protocol, methods, results and conclusions. During the protocol I focused on everything that had to do with the computer processing of the data, while S. Mol (Stefan) focused on Modified Cognitive Interviewing in general. In addition, he also focused on the human rater judgments.

When all the data was collected from the participants, I made a dataset in SPSS for the computer processing, and Stefan made a survey for the raters and also a dataset for analyzing the data. We both entered our own data in SPSS. When making the methods chapter we both concluded that analyzing computer-processed data was way harder and far more work than the human rater judgments. While creating the methods section, we decided to work together in order to have an equal workload, and to finish the thesis project on time. For this reason it is difficult to describe individual tasks. In the following chapter (8.1), an explanation of individual work is given.

8.1 Individual Tasks

During the protocol I mainly focused on all the aspects of detecting deception through computer-mediated communication, and thereby the computer-processed analysis. I wrote the introduction, and in the literature review the second and third part, that were related to the computer-mediated communication. Also in the research plan, I was concentrating on the computer-processed analysis, while authoring this chapter together with S. Mol. In the method section, I explained all the variables I planned to use. We wrote the competences related to the research section together.

When starting on the thesis project, I first only concentrated on the computer-mediated communication part. When I entered all my data in SPSS, and started to make analyses, I needed help because it was very challenging. From that moment, Stefan took over the lead of the ROC analysis, while I was still in charge of the other analyses. By taking the lead, and not totally individually performing tasks, we tended to achieve higher quality in our study, and gaining a broader knowledge. Instead of only learning about the behavior of human raters, or computer-processed analysis only, we also learned from each other. By taking the lead of the most part of the computer-processed analysis, I performed the analysis explained below.

Descriptive Statistic Analysis

This analysis provided an overview of the means and standard deviations. Based on this analysis I was able to decide which dataset was possibly more effective (edited or unedited), and should therefore be used.

Histograms

The histogram-overview showed which speech-content variables related to which prompts were the most discriminating within the groups (truthful, deceptive, and false claim). By creating the histograms, a raw observation could be made.

T-Test

The (independent-samples) T-Tests provided results that showed whether there are significance differences between truthful and deceptive groups, related to various prompts.

Multivariate Analysis of Variance (MANOVA)

The MANOVA showed where significant differences were for all the individual groups (i.e., truthful, deceptive, and false claim). Based on these methods, conclusions could be made.

The explanation of the previously mentioned methods, results, conclusions, and discussions were all part of my individual tasks. The reason for choosing those analyses was because they were also used in the research of Morgan et al. (2015). This research is a follow-up, and in order to compare results, it is important to use similar methods. Also, Dr. Morgan gave advice on which methods are the most effective to use.

References

About UNH. (2016). Retrieved April 08, 2016, from http://Twitter.newhaven.edu/about/

About - Mission Statement UNH. (2016). Retrieved April 08, 2016, from http://Twitter.newhaven.edu/about/mission-statement/

Abraham, L. (n.d.). BrainyQuote.com. Retrieved June 14, 2016, from BrainyQuote.com Web site: http://www.brainyquote.com/quotes/quotes/a/abrahamlin105816.html

Alowibdi, J.S., Buy, U.A., Yu, P.S., Ghani, S., Mokbel, M. (2015). Deception Detection in Twitter. *Social Network Analysis and Mining, 5, 1-16*

Associated Press, Daily Mail UK. (2016). Retrieved June 12, 2016, from http://www.dailymail.co.uk/sciencetech/article-3434121/Twitter-moves-actively-seek-terrorist-supporters.html

Buller, D.B., Burgoon, J.K. (1996). Interpersonal deception theory. *Communication Theory, 6*(3), 203-242.

Burgoon, J. K., Blair, J. P., Qin, T., Nunamaker, J. F. (2003). Detecting deception through linguistic analysis. *Intelligence and Security Informatics, 2665,* 91-101.

Irshaid, F., BBC News. (2014). Retrieved June 12, 2016, from http://www.bbc.com/news/world-middle-east-27912569

Galanxhi, Fui-Hoon Nah, F. (2007). Deception in cyberspace: A comparison of text-only vs. avatar-supported medium. *International journal of human-computer studies, 65(9), 770-783*

Geiselman, R. E. (2012). The Cognitive Interview for Suspects. *American Journal of Forensic Psychology, 30*

Geiselman, R. E., Fisher, R. P. (1985). Eyewitness Memory Enhancement in the Police Interview: Cognitive Retrieval Mnemonics Versus Hypnosis. *Journal of Applied Psychology*, *70*, 401-412

Hancock, J. T., Curry, L., Goorha, S., Woodworth, M. (2005). Automated linguistic analysis of deceptive and truthful synchronous computer-mediated communication. *Proceedings of the 38th Hawaii International Conference on System Sciences,*

Hancock, J. T., Smith, M. E., Reynolds, L., Birnholtz, J. (2014) Everyday Deception or A Few Prolific Liars? The Prevalence of Lies in Text Messaging. *Computers in Human Behavior*, 41

Hancock J. T., Thom-Santelli, J., Ritchie, T. (2004). Deception and design: The impact of communication technology on lying behavior. *Proceedings, Conference on Human Factors in Computing Systems,* Vienna, Austria. *, 6*(1) 130-136.

Hooi, R., Cho, (2013). Deception in avatar-mediated virtual environment. *Computer in Human Behavior, 29(1), 276-284*

Köhnken, G., Milne, R., Memon, A., Bull, R. (1999). A meta-analysis on the effects ofthe Cognitive Interview. *Special Issue of Psychology, Crime, & Law*, 5, 3–27.

Lewis, George, J. (2008). Cross-cultural Deception in Social Networking sites and Face-to-Face Fommunication. *Comput. Hum. Behav.*, 24(6), 2945– 2964

Morgan III CA, Steffian G, Hazlett G (2007): Efficacy of Forensic Statement Analysis in Distinguishing between Truthful and Deceptive Eyewitness Accounts. Journal of Intelligence Community Research & Development (JICRD), July 2007.

Morgan III C. A., Colwell, K., Hazlett, G. A. (2011). Efficacy of Forensic Statement Analysis in Distinguishing Truthful from Deceptive Eyewitness Accounts of Highly Stressful Events. *Journal of Forensic Sciences, 56(5)*

Morgan III, C. A., Christian J, Rabinowitz, Hazlett G. (2009). Detecting Deception in Vietnamese: Efficacy of Forensic Statement Analysis when Interviewing through an Interpreter. Manuscript submitted and in review.

Morgan III, C. A., Christian J, Rabinowitz TWITTER, Palin, B., Kennedy, K. (2015). Who should you Trust? Discriminating Genuine from Deceptive Eyewitness Accounts. The Open Criminology Journal, 8

Morgan III C. A., Colwell, K., Steffian, G., Hazlett, G. (2008). Efficacy of Verbal and Global Judgments in the Detection of Deception in Moroccans Interviewed Via an Interpreter. Journal for Intelligence Community Research and Development (JICRD)

Morgan III CA, Mishara A, Christian J, Hazlett, G. (2008). Detecting Deception through Automated Analysis of Translated Speech: Credibility Assessments of Arabic Speaking Interviewees. JICRD, August 2008.

Morgan, III C.A., Rabinowitz ,Y.G., Leidy, R., Coric, V. (2014). Efficacy of combining interview techniques in detecting deception related to bio-threat. *Behav Sci Law,* 32, 269-285.

Newman, M.L., Pennebaker, J., Berry, D.S., Richards, J.M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*(5), 665-675.

Text Content Analyser (2016). Retrieved 12, June 2016 from, http://www.usingenglish.com/resources/text-statistics.php

Thomas G. Tape, MD,(n.d.) Retrieved June 10, 2016. From http://gim.unmc.edu/dxtests/roc3.htm

Twitter Inc. (2015). Company | About. Retrieved from Twitter.com: https//about.twitter.com/company

University of New Haven: School of Public Service (2015). Retrieved June 12, 2016, from http://www.newhaven.edu/lee-college/departments/school-of-public-service/

U.S.Air Force – Mission (2016). Retrieved 12, June 2016 from, https://www.airforce.com/mission

Verkampt, F., Ginet, M. (2009). Variations of the cognitive interview: Which one is the most effective in enhancing children's testimonies? *Applied Cognitive Psychology, 24(9), 1279-1296.*

Vrij, A., Edward, K., Roberts, K. P., Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24*(4), 239-263.

Wright, A.M., Holliday, R.E., (2007). Enhancing the recall of young, young-old and old-old adults with cognitive interviews. *Applied Cognitive Psychology*, 21(1), 19-43.

Zhou, L. (2005). An empirical investigation of deception behavior in instant messaging. *IEEE Transactions on Professional Communication, 48*(2), 147-160.

Zhou, L, Burgoon, J.K., Nunamaker, J.F., Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13*, 81-106.

Zhou, L., Burgoon, J. K., Twitchell, D. P. (2003). A longitudinal analysis of language behavior of deception in e-mail. *Intelligence and Security Informatics*, *2665,* 102-110.

Zhou, L., Twitchell, D. P., Qin, T., Burgoon, J. K., Nunamaker, J. F. (2003b). An exploratory study into deception detection in text-based computer-mediated communication. *Proceedings of the 36th Hawaii International Conference of System Sciences.*

Zhou, L., Zhang, D. (2007). Typing or messaging? Modality effect on deception detection in computer-mediated communication. *Decision Support Systems, 44(1), 188-201.*

Appendix

A. Tables and Figures

A-1: Means Analysis Rater Judgment tweet One

Report

	Rater 1 tweet 1 Confidence	Rater 2 tweet 1 Confidence	Rater 3 tweet 1 Confidence	Rater 4 tweet 1 Confidence	Rater 5 tweet 1 Confidence	Rater 6 tweet 1 Confidence	Rater 7 tweet 1 Confidence
Mean	3,86	3,07	3,89	1,82	3,39	3,75	3,95
N	44	44	44	44	44	44	44
Std. Deviation	,852	,255	,722	,691	,579	1,332	,645

A-2: Means Analysis Rater Judgment Whole Conversation

Report

	Rater 1 Whole Conversation Confidence	Rater 2 Whole Conversation Confidence	Rater 3 Whole Conversation Confidence	Rater 4 Whole Conversation Confidence	Rater 5 Whole Conversation Confidence	Rater 6 Whole Conversation Confidence	Rater 7 Whole Conversation Confidence
Mean	3,91	3,55	3,70	2,77	3,45	3,41	4,36
N	44	44	44	44	44	44	44
Std. Deviation	,858	,504	1,069	1,138	,627	1,245	,613

A-4: Group Statistics T-Test

Group Statistics

	Groups deceptive and truthful	N	Mean	Std. Deviation	Std. Error Mean
Response Length Tweet Total Edited	Truthful	15	103,87	20,636	5,328
	Deceptive	29	88,62	23,886	4,435
Unique Word Count Tweets Total Edited	Truthful	15	65,20	8,064	2,082
	Deceptive	29	54,24	13,535	2,513

A-3: T-Test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Response Length Tweet Total Edited	Equal variances assumed	,164	,688	2,098	42	,042	15,246	7,268	,578	29,914
	Equal variances not assumed			2,199	32,357	,035	15,246	6,933	1,131	29,361
Unique Word Count Tweets Total Edited	Equal variances assumed	,832	,367	2,873	42	,006	10,959	3,814	3,262	18,655
	Equal variances not assumed			3,358	40,999	,002	10,959	3,264	4,367	17,550

A-5: ROC

Coordinates of the Curve

Test Result Variable(s)Unique Word Count Tweets Total Edited

Positive if Greater Than or Equal To[a]	Sensitivity	1 – Specificity
10,00	1,000	1,000
22,00	1,000	,966
33,50	1,000	,931
38,50	1,000	,897
44,00	1,000	,862
47,00	1,000	,828
49,50	1,000	,690
50,50	,933	,690
51,50	,933	,655
52,50	,933	,621
54,00	,933	,586
55,50	,933	,448
56,50	,933	,379
57,50	,800	,345
58,50	,800	,276
59,50	,733	,276
61,00	,667	,241
62,50	,600	,241
64,00	,467	,207
65,50	,467	,172
66,50	,467	,138
67,50	,467	,103
68,50	,400	,103
69,50	,333	,103
71,00	,267	,103
72,50	,200	,103
74,50	,133	,069
77,50	,067	,034
79,50	,000	,034
81,00	,000	,000

The test result variable(s): Unique Word Count Tweets Total Edited has at least one tie between the positive actual state group and the negative actual state group.

B. MCI Questions and Rater Instructions

Hi, welcome to this survey.
This research survey is about detecting deception in the social media platform Twitter using the cognitive interview method. We asked 46 research participants each six questions, which they had to answer either truthfully or deceptively. If they were assigned to the truthful group they had to answer every question completely honest, if they were assigned to the deceptive group they had to lie in all their answers. The questions were about what they did in a certain timeframe. We would like your expert opinion on whether you think the participant answered truthful or deceptive, and how confident you are about your answer.

Before you start the survey, we would like to give you the definitions of some words that should be known in order to fully understand the survey.

What is Twitter?
Twitter is a free social networking microblogging service that allows registered members to broadcast short posts called 'tweets'. Twitter members can broadcast tweets and follow other users' tweets by using multiple platforms and devices. A tweet consists of a maximum of 140 characters. Twitter is also known for using the so-called hashtags (#). This is used to either tag your tweet to a certain subject, or to make a statement separate from the original text. For example, “My sports team is #winning the game #Yankees”. In this case #winning is used as a statement and the #Yankees is used to attach this tweet to all tweets that also use #Yankees.

What is the cognitive interview method?

The cognitive interview method is originally used for memory retrieval in law enforcement investigations. It can be used for both eyewitness- and suspect interrogation. The cognitive interview method we used in this research consists of six questions:

Question 1, Full recall:
Tell me everything you remember about what you did between (start timeframe) and (end timeframe). Be as detailed as possible. Do not leave anything out.

Question 2, Visual:
If I had been there with you, what would I have seen from the beginning to the end of that timeframe?

Question 3, Auditory:
If I had been there with you, what would I have heard from the beginning to the end of that timeframe?

Question 4, Emotional:
What was the experience like for you?

Question 5, Temporal:
Start at the end and tell me, in reverse order, everything you remember about the event.

Question 6, Detailed:
Did you leave anything out that might be important or do you want to add something to your story?

A PDF file is Attached to the email, with also the link to this survey. This document includes the same six questions as mentioned above. **Please print the document, or have it on a second screen to use it as a reference during the survey.**

The research participants were only able to answer every question in a total of one tweet, consisting of a maximum of 140 characters. This means that every answer, is 1 tweet. The survey consists of 4 questions per research participant, which we want you to answer.

- The first question will display the first tweet from the participant (the full recall). You have to answer if you think the participant is truthful or deceptive.

- In the second question you will be asked how confident you were about your rating of only the first tweet, on a scale from 1 to 5, with 1 being not confident and 5 being really confident.

- In the third question you will see all the six tweets together, and you are asked again if you think the participant is either truthful or deceptive.

- In the fourth question you will be asked how confident you were about your rating of all the six tweets together, on a scale from 1 to 5, with 1 being not confident and 5 being really confident.

If you don't know the answer just answer what you think is the best answer or guess.

Now you are ready to start the survey. **Make sure you have the questions from the PDF printed out or on a second screen to use as reference during the survey.**

C. Invitation for Participants

Dear students,

Would you like to participate in a National Security Research bout lying and deception? Do you want to play a fun game and do you think you are a good liar? The research takes about 45 minutes and will be held in room **2.10 in the back of the library** on **Tuesday May 3rd** and Wednesday **May 4th.** If you have a laptop, please bring it with you, if not we have a couple available.

The times are:

Tuesday May 3rd: **3:45 – 4:30 PM**

Tuesday May 3rd: **4:45 – 5:30 PM**

Wednesday May 4th: **3:45 – 4:30 PM**

Wednesday May 4th: **5:00 – 5:45 PM**

Send us a text with your name, the selected time and with how many persons you are coming.

Phone number: 475-201-4803

The reward will be: **FREE PIZZA**