AFRL-RH-WP-TR-2020-0137

PERSONALIZED AUTONOMOUS AGENTS COUNTERING SOCIAL ENGINEERING ATTACKS (PANACEA)

Amir Masoumzadeh / Alan Zemel
RFSUNY, UAlbany
1400 Washington Ave, MSC100A,
Albany, NY 12222

Tomek Strzalkowski
Rensselaer Polytechnic Institute
SA Sage Bldg, 110 8th Street
Troy, NY 12180

Adam Dalton / Bonnie J. Door
Florida IHMC
40 S. Alcaniz St.
Pensacola, FL 32502

Samira Shaikh / Ehab Al-Shaer
UNC - Charlotte
9201 University City Blvd
Charlotte, NC 28223

OCTOBER 2020

FINAL REPORT

AIR FORCE RESEARCH LABORATORY
711th HUMAN PERFORMANCE WING
AIRMAN SYSTEMS DIRECTORATE
WARFIGHTER INTERACTIONS AND READINESS DIVISION
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the AFRL Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2020-0137 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//

TIMOTHY R. ANDERSON, DR-IV, Ph.D.
Work Unit Manager
Mission Analytics Branch
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

//signature//

WILLIAM P. MURDOCK, DR-IV, Ph.D.
Chief, Mission Analytics Branch
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

//signature//

LOUISE A. CARTER, DR-IV, Ph.D.
Chief, Warfighter Interactions and Readiness Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
10/31/2020	Final Report	09/21/2018 - 08/31/2020

4. TITLE AND SUBTITLE
Personalized AutoNomous Agents Countering Social Engineering Attacks (PANACEA)

5a. CONTRACT NUMBER
FA8650-18-C-7881

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
[1]Amir Masoumzadeh / Alan Zemel
[2]Tomek Strzalkowski
[3]Adam Dalton / Bonnie J. Dorr
[4]Samira Shaikh / Ehab Al-Shaer

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H0X8

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
[1]RFSUNY, UAlbany, 1400 Washington Ave, MSC100A, Albany, NY 12222
[2]Rensselaer Polytechnic Institute, SA Sage Bldg, 110 8th Street, Troy, NY 12180
[3]Florida IHMC, 40 S. Alcaniz St, Pensacola, FL 32502)
[4]UNC-Charlotte, 9201 University City Blvd, Charlotte, NC 28223

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711 Human Performance Wing
Airman Systems Directorate
Warfighter Interactions & Readiness Division
Wright-Patterson AFB OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-TR-WP-2020-0137

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
AFRL-2021-2062; Cleared 29 Jun 2021

14. ABSTRACT
We are reporting on the development of PANACEA, a system that supports natural language processing (NLP) components for active defenses against social engineering attacks. We deploy a pipeline of human language technology, including Ask andFraming Detection, Named Entity Recognition, Dialogue Engineering, and Stylometry. PANACEA processes modern message formats through a plug-in architecture to accommodate innovative approaches for message analysis, knowledge representation and dialogue generation. The novelty of the PANACEA system is that it uses NLP for cyber defense and engages the attacker using bots to elicit evidence to attribute to the attacker and to waste the attacker's time and resources.

15. SUBJECT TERMS
social engineering, natural language processing, active detection, active defense

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Timothy Anderson, Ph.D.
Unclassified	Unclassified	Unclassified	Unclassified	37	19b. TELEPHONE NUMBER *(Include area code)*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

We report on the development of Personalized AutoNomous Agents Countering SocialEngineering Attacks (PANACEA), a system that supports natural language processing (NLP) components for active defenses against social engineering attacks. We deployed a pipeline of human language technology, including Ask and Framing Detection, Named Entity Recognition, Dialogue Engineering, and Stylometry. PANACEA processes modern message formats through a plug-in architecture to accommodate innovative approaches for message analysis, knowledge representation and dialogue generation. The novelty of the PANACEA system is that it uses NLP for cyber defense and engages the attacker using bots to elicit evidence to attribute to the attacker andto waste the attacker's time and resources.

2.0 INTRODUCTION

PANACEA actively defends against social engineering attacks. *Active* defense refers to engaging an adversary during anattack to extract and link attributable information while also wasting their time and resources in addition to preventing the attacker from achieving their goals. This contrasts with *passive* defenses,which decrease likelihood and impact of an attack [1] but do not engage the adversary.

2.1 Threat Intelligence and Analysis

PANACEA's active defenses were built on top of a mature threat intelligence architecture. Social engineering attacks are formidable because intelligent adversaries exploit technical vulnerabilities toavoid social defenses, and social vulnerabilities to avoid technical defenses [2]. A system must besocially aware to find attack patterns and indicators that span the socio-technical space. PANACEA approaches this by incorporating the Find, Fix, Finish, Exploit, Analyze, and Disseminate (F3EAD) threat intelligence cycle [3]. The *find* phase identifies threats using language-based and message security approaches. The *fix* phase gathers relevant and necessary information to engagethe adversaries and plan the mitigations that will prevent them from accomplishing their maliciousgoals. The *finish* phase performs a decisive and responsive action in preparation for the *exploit* phase for future attack detection. The *analysis* phase exploits intelligence from conversations withthe adversaries and places it in a persistent knowledge base where it can be linked to other objects and studied additional context. The *disseminate* phase makes this intelligence available to allcomponents to improve performance in subsequent attacks.

PANACEA's makes threat intelligence available to every component, including improving NLP capabilities for cyber defense, through a plug-in design that enables purpose-guided access. Figure 1 illustrates PANACEA's active defense in the form of conversational engagement, diverting theattacker while also delivering a link that will enable the attacker's identity to be unveiled.

2.2 Social Engineering (SE) Lexicon

SE refers to sophisticated use of deception to manipulate individuals into divulging confidential or personal information for fraudulent purposes. Standard cybersecurity defenses are ineffective because attackers attempt to exploit humans rather than system vulnerabilities. Accordingly, we have built a *user alter-ego* application that detects and engages a potential attacker in ways that expose their identity and intentions.

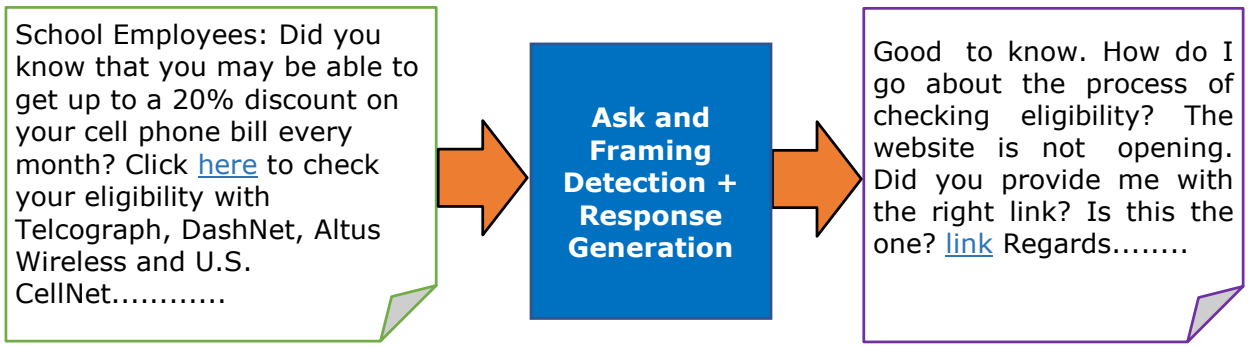

Figure 1. Active Defense Against SE: Attacker's Email (Left) Yields Bot's Response (Right)

Our system relies on a paradigm for extensible lexicon development that leverages the central notion of *ask*, i.e., elicitation of behaviors such as PERFORM (e.g., clicking a link) or GIVE (e.g., providing access to money). This paradigm also enables detection of risk/reward (or LOSE/GAIN) implied by an ask, which we call *framing* (e.g., *lose your job*, *get a raise*). These elements are used for countering attacks through bot-produced responses and actions. The system is tested in an email environment, but is applicable to other forms of online communications, e.g., short message service (SMS).

Table 1. Lexical Conceptual Structure (LCS) + Ask/Framing Output for Three SE Emails

Email	Ask	Framing
(a) It is a pleasure to inform you that you have won 1.7Eu. Contact me. (jw11@example.com)	PERFORM contact (jw11@...)	GAIN won (1.7Eu)
(b) You won $1K. Did you send money? Do that by 9pm or lose money. Respond asap.	GIVE send (money)	LOSE lose (money)
(c) Get 20% discount. Check eligibility or paste this link: http.... Sign up for email alerts.	PERFORM paste (http...)	GAIN get (20%)

More formally, an *ask* is a statement that elicits a behavior from a potential victim, e.g., *pleasebuy me a gift card*. Although asks are not always explicitly stated [4, 5], we discern these through navigation of semantically classified verbs. The task of ask detection specifically is targeted event detection based on parsing and/or Semantic Role Labeling (SRL), to identify semantic class triggers [6]. *Framing* sets the stage for the ask, i.e., the purported threat (LOSE) or benefit (GAIN) that the social engineer wants the potential victim to believe will obtain through compliance or lack thereof. It should be noted that there is no one-to-one ratio between ask and framing in the ask/framing detection output. Given the content, there may be none, one or more asks and/or framings in the output.

Our lexical organization is based on LCS, a formalism that supports resource construction and extensions to new applications such as social engineering detection and response generation. Semantic classes of verbs with similar meanings (*give*, *donate*) are read-ily augmented through adoption of the STYLUS variant of LCS [7] and [8]. We derive LCS+ from asks/framings and employ Categorial-Variation (CATVAR) Database [9] to relate word variants (e.g., *reference* and *refer*). Table 1 illustrates LCS+ Ask/Framing output for three (presumed) social engineering emails: twoPERFORM asks and one GIVE ask.[1] Parentheses () refer to ask *arguments*, often a link that the potential victim might choose to click. Ask/framing outputs are provided to downstream responsegeneration. For example, a possible response for Table 1(a) is *I will contact asap.*

A comparison of LCS+ to two related resources shows that our lexical organization supports refinements, improves ask/framing detection and top ask identification, and yields qualitative improvements in response generation. LCS+ is deployed in a social engineering detection and response generation system. Even though LCS+ is designed for the social engineering domain, the approach to development of LCS+ described in this paper serves as a guideline for developing similar lexica for other domains. Correspondingly, even though development of LCS+ is one of the contributions of this paper, the main contribution is not this resource but the systematic and efficient approach to resource adaptation for improved task-specific performance.

[1]To view our system's ask/framing outputs on a larger dataset (the same set of emails which were also used for ground truth (GT) creation described below), refer to https://social-threats.github.io/PANACEA-ask-detection/data/case7LCS+AskDetectionOutput.txt.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Threat Intelligence and Analysis

PANACEA's processing workflow is inspired by Stanford's CoreNLP annotator pipeline [], but witha focus on using NLP to power active defenses against SE. The F3EAD motivated phased analysisand engagement cycle is employed to conduct active defense operations. The cycle is triggered when a message arrives and is deconstructed into Structured Threat Information Expression (STIX) objects. Object instances for the identitiesof the sender and all recipients are found or created in the knowledge base. Labeled relationshipsare created between those identity objects and the message itself.

Once a message is ingested, plug-in components process the message in the *find* phase, yielding a response as a JavaScript Object Notation (JSON) object that is used by plug-in components in subsequent phases. Analyses performed in this phase include message part decomposition, named entity recognition, and email header analysis. The *fix* phase uses components dubbed *deciders*, which perform a meta-analysis of the results from the *find* phase to determine if and what type of an attack is taking place. *Ask detection* provides a fix on what the attacker is going after in the *fix* phase, if an attack is indicated. Detecting an attack advances the cycle to the *finish* phase, where response generation is activated. Each time PANACEA successfully elicits a response from the attacker, the new message is *exploited* for attributable information, such as the geographical location of the attack and what organizational affiliations they may have. This information is stored as structured intelligence in the knowledge base which triggers the *analysis* phase, wherein the threat is re-analyzed in a broader context. Finally, PANACEA disseminates threat intelligence so that humans can build additional tools and capabilities to combat future threats.

PANACEA's main components are (1) Message Analysis Component and (2) Dialogue Component. The resulting system is capable of handling the thousands of messages a day that would be expected in a modern organization, including failure recovery and scheduling jobs for the future. Figure 2 shows PANACEA throughput while operating over a one-month backlog of emails, SMS texts, and LinkedIn messages.

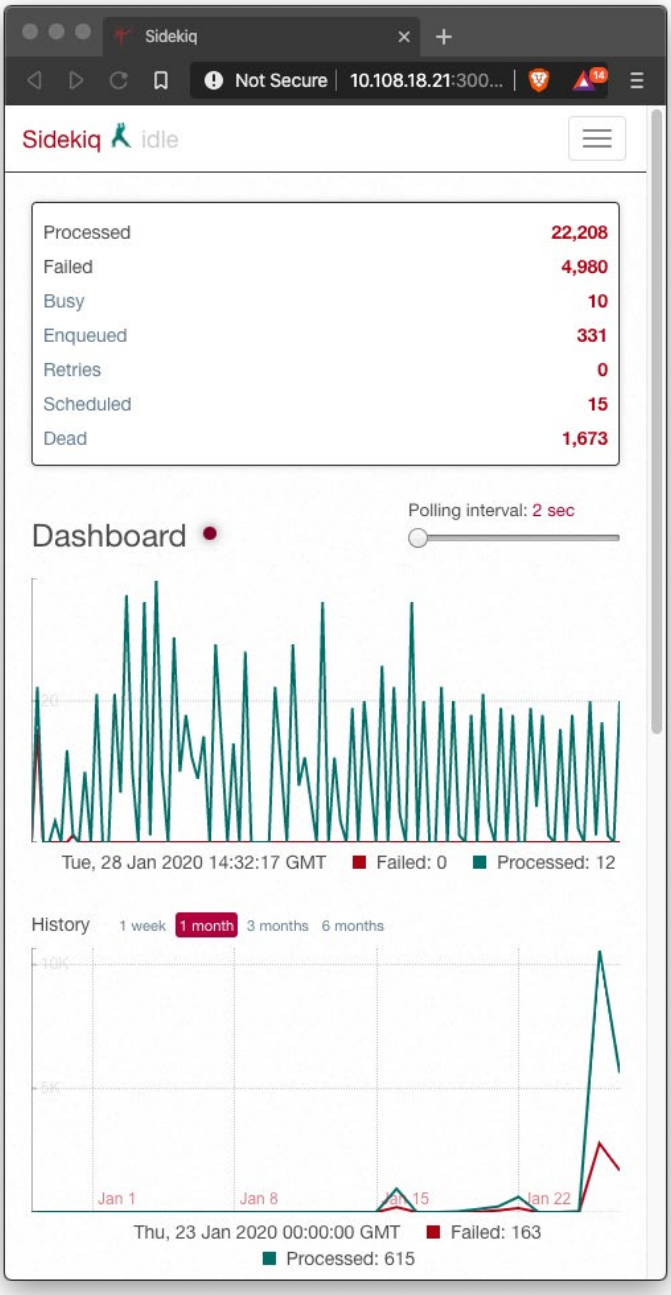

Figure 2. PANACEA Components Run Asynchronously in the Background for Scalability and Flexibility to Add or Remove Components Based on the Underlying Task.

3.1.1 Email Header Classification

When communication takes place over a network, metadata is extracted that serves as a user fingerprint and a source for reputation scoring. Email headers, for example, contain authentication details and information about the mail servers that send, re- ceive, and relay messages as they move from outbox to inbox. To distinguish between benign andmalicious emails, PANACEA applies a multistage email spoofing, spamming, and phishing detector consisting of: (1) a signature-based detector, (2) an active investigation detector, (3) a receiver- oriented anomaly detector, and (4) a sender-oriented anomaly detector.

3.1.2 Email Content Classification

Dissecting email headers is not enough for detecting ma- licious messages. Many suspicious elements are related to email bodies that contain user messagesrelated to a specific topic and domain. Analyzing email content provides valuable insight for de- tecting threats in conversations and a solid understanding of the content itself. PANACEA incorporates machine learning algorithms that, alongside of header classifiers, digest email exchanges:

3.1.2.1 Benign/Non-Benign Classifier

Word embedding vectors [11, 12] trained on email samples from different companies (e.g., Enron) are extracted using neural networks [13], i.e., back-propagationmodel with average word vectors as features. This classifier provides a binary prediction regardingthe nature of emails (friend or foe).

3.1.2.2 Email Threat Type Classifier

Spam, phishing, malware, social-engineering and propaganda are detected, providing fine-grained information about the content of emails and support for motive detection (i.e., attacker's intention).

3.1.2.3 Email Zone Classifier

Greetings, body, and signature are extracted using word embedding im- plemented as recurrent neural network with handcrafted rules, thus yielding senders, receivers andrelevant entities to enable response generation.

3.1.2.4 Gender Prediction

Users gender (male/female) is detected on email text body by using a bi- nary classifier based on the extraction of stylistic elements of text (punctuation, stopwords use, capitalize letters, etc.) as features in back-propagation model with average word vectors.

3.1.2.5 Human/Bot Classifier

Like with the gender classifier, detection of real users from bots is basedon the extraction of stylistic features of email texts but in this case there are added extra features associated to the use of language like the use of specific part-of-speech tags (e.g., use of nounsor adjectives) and detection of repetitive patterns. For this classifier it is used a recurrent neural network model that digest previous and current words in a email for detecting previous, current and future patterns.

The document collections used for training and testing the email content classifiers include be-nign and malicious email samples obtained from employees of public companies and government departments. Benign emails correspond to internal interactions among users on day-to-day work issues. On the other hand, most of suspicious emails are obtained from employees spam boxes and specific email threat repositories (like the Anti-Phishing Working Group (APWG) dataset).

Table 2 summarizes the key details of each collection. Enron and the Anti-Phishing Working Group

APWG (among other collections) are used for training purposes while non-public datasets called dry-run 1 and dry-run 2 are used for testing.

Table 2. Datasets Used for Training the Distinct Content Classifiers

Dataset name and/or type	Feature	Training	Testing
Enron [14]	Used for word embeddings:	✓	
	Collection type:	Publicly available	
Benign emails	Number of emails:	84111	-
APWG [15]	Used for word embeddings:	✓	
Phishing/Malware	Collection type:	Publicly available	
Non-benign emails	Number of emails:	30776	-
BC	Used for word embeddings:	✓	
	Collection type:	Publicly available	
Benign emails	Number of emails:	259	-
Phishing/non-phishing	Used for word embeddings:	✓	
	Collection type:	ASED program	
Non-benign emails	Number of emails:	5338	-
Malware/non-malware	Used for word embeddings:	✓	
	Collection type:	ASED program	
Non-benign emails	Number of emails:	2914	-
Propaganda/non-propaganda	Used for word embeddings:	✓	
	Collection type:	ASED program	
Non-benign emails	Number of emails:	261	-
Spam/non-spam	Used for word embeddings:	✓	
	Collection type:	ASED program	
Non-benign emails	Number of emails:	1294	-
social engineering/non-social engineering	Used for word embeddings:	✓	
	Collection type:	ASED program	
Non-benign emails	Number of emails:	1059	-
Reconnaissance/non-reconnaissance	Used for word embeddings:	✓	
	Collection type:	ASED program	
Non-benign emails	Number of emails:	173	-
Dry-run 1	Used for word embeddings:		
	Collection type:	ASED program	
Benign and non-benign emails	Number of emails:	-	1025
Dry-run 2	Used for word embeddings:		
	Collection type:	ASED program	
Benign and non-benign em ils	Number of emails:	-	3023

From the above table, it is important to highlight that the dry-run datasets comprise also email samples of day-to-day interactions in a work environment. These collections are not publicly available considering that there are utilized for evaluating the PANACEA system in the active social engineering program. Despite that, it can be mentioned that these datasets have an unbalanced nature with a proportion of 80% benign samples and 20% non-benign ones which is consistent with a real world scenario.

Considering the above, all classifiers support active detection of malicious emails and helpin the engagement process of automated bots. Additionally, all trained models have an overall accuracy of 90% using a cross validation approach against the email collections presented before,which makes them reasonably reliable in the context of passive defenses.

3.1.3 Behavioral Modeling.

If an adversary is able to compromise a legitimate account, then the header and content classifiers will not be sufficient to detect an attack. The social engineer is able to extract contacts of the account owner and send malicious content on their behalf, taking advantage of the reputation and social relationships attributed to the hijacked account. Two dis- tinctive approaches address these issues: impersonation detector and receiving behavior classifier.Both of these classifiers were initially developed and tested based on the publicly-available Enron email dataset. The deployed classifiers were trained using historical email datasets that were available in the Active Social Engineering Defense (ASED) testbed.

3.1.3.1 Impersonation Detector

Sender entities are extracted from historical email messages in the system. A personalized profile is created for each sender based on communication habits (e.g., time of emails), stylometric features of messages, and social networks (i.e., other entities they arecommunicating with). The unique profiled model is used to assess whether this email has been written and sent by an account's legitimate owner. If a message arrives from a sender that does nothave a profile, PANACEA applies similarity measures to find other email addresses for the unknownentity. This serves as a defense against impersonation attacks where the social engineer creates anemail account using a name and address similar to the user of an institutional account for whicha model is available. If PANACEA links the unknown account to an institutional account, then that account's model is used to determine whether a legitimate actor is using an unknown account, or anefarious actor is attempting to masquerade as an insider in order to take advantage of the access such an account would have.

3.1.3.2 Receiving Behavior Classifier

Based on historical email messages in the system, a personalized profile model is built for the receiving behavior of each institutional account (how and who communicates with this account). Incoming emails are evaluated against the constructed models and anomalous messages are flagged as potential attacks by the classifier.

3.1.4 Deciders

PANACEA must have high confidence in determining that a message is coming from an attacker before deploying active defense mechanisms. A strategy-pattern approach fits different meta-classifiers to different situations. For example, a program manager who frequentlycorresponds with new people from outside their organization would require a different strategy than an office manager whose correspondences are overwhelmingly internal and from the same setof people. Four classification strategies, called *Deciders*, combine all component analyses after a message is delivered to an inbox to make the final *friend/foe* determination. The Decider API expects all component analyses to include a *friend/foe* credibility score using six levels definedby the Admiralty Code [16]. Deciders may be deterministic through the application of rule-based decision making strategies or they may be trained to learn to identify threats based on historical data. The four implemented and operated deciders are described below.

3.1.4.1 Fair Decider

This strategy aggregates the decisions from all active classifiers. Each classifier assessment is weighted only by its internal credibility score.

3.1.4.2 Fair Threat Decider

Threat deciders first check to see if the sender or the sender's organizationis a known threat actor. If they are not, then the decider falls back to another strategy. The fall backfor the Fair Threat Decider is the Fair Decider.

3.1.4.3 MetaLearner Decider

The MetaLearner strategy uses a SciKit support vector machine (SVM) to make an assessment of six components (plus six Individual label decisions from the Albany Email Labeler) to make a final decision. The MetaLearner is trained on organizational ground truth in order to create sender andrecipient profiles.

3.1.4.4 Forensic Decider

The Forensic Decider is an evolution of the Threat Decider approach. Senders who have not communicated with a recipient in the past or investigated for other communicationswithin the organization. If they do not have a history, then their communications are tracked untileither the Fair Threat Decider classifies them as a Threat Actor or their identity is verified. Identifyverification can either be delegated to a different strategy or use alternative methods such sendingmessage replies requesting material only legitimate actors would be able to provide.

3.1.5 Threat Intelligence

PANACEA stores component analysis results in a threat intelligence knowledge base for aggregation of attack campaigns with multiple turns, targets, and threads. Theknowledge base adheres to STIX 2.0 specifications and implements MITRE's Adversarial Tactics, Techniques and Common Knowledge (ATT&CK) frame- work [17] to enable attribution and anticipatory mitigations of sophisticated social engineering attacks. PANACEA recognizes indicators of compromise based on features of individual emails as well as historical behavior of senders and recipients. Intrusion sets and campaigns are thus constructed when malicious messages are discovered subsequently linked to threat actors based on attribution patterns, such as Internet Protocol (IP) address, message templates, socio-behavioral indicators, and linguistic signatures. This feature set was prioritized to work with Unit 42's ATT&CK Playbook Viewer. The knowledge base uses a PostgreSQL database backend with an application layer builtwith Ruby on Rails.

3.2 Social Engineering Lexicon

In our experiments described in Section 4.2, we compare LCS+, our lexical resource we developedfor the social engineering domain, against two strong baselines: STYLUS and Thesaurus.

3.2.1 STYLUS Baseline

As one of the baselines for our experiments, we leverage a publicly avail-able resource STYLUS that is based on LCS [7] and [8]. The LCSrepresentation is an underlying representation of spatial and motion predicates [18, 19, 20], such as *fill* and *go*, and their metaphorical extensions, e.g., temporal (the hour *flew* by) and possessional (he *sold* the book).[2] Prior work [21, 22, 23, 24, 25, 26, 27] has suggested that there is a close relation between underlying lexical-semantic structures of verbs and nominal predicates and theirsyntactic argument structure.

[2]LCS is publicly available at https://github.com/ihmc/LCS.

We leverage this relationship to extend the existing STYLUS verb classes for the resource adaptation to social engineering domain through creation of LCS+ whichis discussed below.

For our STYLUS verb list, we group verbs into four lists based on asks (PERFORM, GIVE) and framings (LOSE, GAIN). The STYLUS verb list can be accessed here: https://social - threats.github.io/PANACEA-ask-detection/resources/original_lcs _classes_based_verbsList.txt.

Examples of this classificationare shown below (with total verb count in parentheses):

- PERFORM (214): remove, redeem, refer
- GIVE (81): administer, contribute, donate
- LOSE (615): penalize, stick, punish, ruin
- GAIN (49): accept, earn, grab, win

Assignment of verbs to these four ask/framing categories is determined by a computational linguist, with approximately a person-day of human effort. Identification of genre-specific verbs isachieved through analysis of 46 emails (406 clauses) after parsing/part-of-speech (POS)/Semanic Role Labeling (SRL) is applied.

As an example, the verb *position* (Class 9.1) and the verb *delete* (Class 10.1) both have an underlying *placement* or *existence* component with an affected object (e.g., the cursor in *position your cursor* or the account in *delete your account*), coupled with a location (e.g., *here* or *from the system*). Accordingly, *Put* verbs in Class 9.1 and *Remove* verbs in Class 10.1 are grouped together and aligned with a PERFORM ask (as are many other classes with similar properties: Banish, Steal, Cheat, Bring, Obtain, etc.). Analogously, verbs in the *Send* and *Give* classes are aligned with a GIVE ask, as all verbs in these two classes have a sender/giver and a recipient.

Lexical assignment of framings is handled similarly, i.e., verbs are aligned with LOSE and GAIN according to their argument structures and components of meaning. It is assumed that the potential victim of a social engineering attack serves to lose or gain something, depending on non-compliance or compliance with a social engineer's ask. As an example, the framing associated with the verb *losing* (Class 10.5) in *Read carefully to avoid* ***losing*** *account access* indicates the risk of losing access to a service; Class 10.5 is thus aligned with LOSE. Analogously, the verb *win* (Class 13.5.1) in *You have* ***won*** *1.7M Eu.* is an alluring statement with a purported gain to thepotential victim; thus Class 13.5.1 is aligned with GAIN. In short, verbs in classes associated withLOSE imply negative consequences (Steal, Impact by Contact, Destroy, Leave) whereas verbs inclasses associated with GAIN imply positive consequences (Get, Obtain).

Some classes are associated with more than one ask/framing category: Steal (Class 10.5) and Cheat (Class 10.6) are aligned with both PERFORM (*redeem, free*) and LOSE (*forfeit, deplete*). Such distinctions are not captured in the lexical resource, but are algorithmically resolved during ask/framing detection, where contextual clues provide disambiguation capability. For example, *Redeem coupon* is a directive with an implicit request to click a link, i.e., a PERFORM. By contrast, *Avoid losing account access* is a statement of risk, i.e., a LOSE. The focus here is not on the processes necessary for distinguishing between these contextually-determined senses, but onthe organizing principles underlying both, in support of application-oriented resource construction.

3.2.1.1 LCS+ Resource for Social Engineering Adapted from STYLUS

Setting disambiguation aside, resource improvements are still necessary for the social engineering domain because, due to its size and coverage, STYLUS is likely to predict a large number of both true and false positives during ask/framing detection. To reduce false positives without taking a

hit to true positives,we leverage an important property of the LCS paradigm: its extensible organizational structure wherein similar verbs are grouped together. With just one person-day of effort by two computational linguists (authors on the paper; the algorithm developer, also an author, was not involved in this process), a new lexical organization, referred to as "LCS+" is derived from STY- LUS, taken together with asks/framings from a set of 46 malicious/legitimate emails.[3] These emails are a random subset of 1000+ emails (69 malicious and 938 legitimate) sent from an external red team to five volunteers in a large government agency using social engineering tactics. Verbs from these emails are tied into particular LCS classes with matching semantic peers and argument structures. These emails are proprietary but the resulting lexicon is released here: https://social-threats.github.io/PANACEA-ask-detection/resources /lcsPlus_classes_based_verbsList.txt.

Two categories (PERFORM and LOSE) are modified from the adaptation of LCS+ beyond those in STYLUS:

- PERFORM (6 del, 44 added): copy, notify
- GIVE (no changes)
- LOSE (174 del, 11 added): forget, surrender
- GAIN (no changes)

Tables 3 and 4 show the refined lexical organization for LCS+ with ask categories (PERFORM,GIVE) and framing categories (GAIN, LOSE), respectively. Boldfaced class numbers indicate theSTYLUS classes that were modified. The resulting LCS+ resource drives our social engineering detection/response system. Each class includes italicized examples with boldfaced triggers. The table details changes to PERFORM and LOSE categories. For PERFORM, there are 6 deleted verbs across 10.2 (Banish Verbs) and 30.2 (Sight Verbs) and also 44 new verbs added to 30.2. ForLOSE, 7 classes are associated with additions and/or deletions, as detailed in the table.

3.2.1.2 Thesaurus Baseline

The Thesaurus baseline is based on an expansion of simple forms of framings. Specifically, the verbs *gain*, *lose*, *give*, and *perform*, are used as search terms to find related verbs in a standard but robust resource thesaurus.com (referred to as "Thesaurus"). The verbs thus found are grouped into these same four categories:

- PERFORM (44): act, do, execute, perform
- GIVE (55): commit, donate, grant, provide

[3]It should be noted that this resource adaptation is based on an analysis of emails not related to, and without accessto, the adjudicated ground truth described in Section 4.2 That is, the 46 emails used for resource adaptation are distinct from the 20 emails used for creating adjudicated ground truth.

Table 3: Ask Categories (PERFORM, GIVE) in Lexical Organization of LCS+.
Italicized Ex-emplars with Boldfaced Triggers Illustrate Usage for Each Class. Boldfaced Class Numbers Indicate Those STYLUS Classes That Were Modified to Yield the LCS+ Resource.

PERFORM

9.1 Put Verbs: ***Position*** *your cursor here*
10.1 Remove Verbs: ***Delete*** *virus from machine*
10.2 Banish Verbs→5 deleted (banish, deport, evacuate, extradite, recall): ***Remove*** *fee from your account*
10.5 Steal Verbs: ***Redeem*** *coupon below*
10.6 Cheat Verbs: ***Free*** *yourself from debt*
11.3 Bring and Take Verbs: ***Bring*** *me a gift card*
13.5.2 Obtain: ***Purchase*** *two gift cards*
30.2 Sight Verbs→1 deleted (regard), 44 added (e.g., check, eye, try, view, visit): ***View*** *this website*
37.1 Transfer of Message: ***Ask*** *for a refund*
37.2 Tell Verbs: ***Tell*** *them $50 per card*
37.4 Communication: ***Sign*** *the back of the card*
42.1 Murder Verbs: ***Eliminate*** *your debt here*
44 Destroy Verbs: ***Destroy*** *the card*
54.4 Price Verbs: ***Calculate*** *an amount here*

GIVE

11.1 Send Verbs: ***Send*** *me the gift cards*
13.1 Give Verbs: ***Give*** *today*
13.2 Contribute Verbs: ***Donate****!*
13.3 Future Having: ***Advance*** *me $100*
13.4.1 Verbs of Fulfilling: ***Credit*** *your account*
32.1 Want Verbs: *I* ***need*** *three gift cards*

- LOSE (41): expend, forefeit, expend, squander
- GAIN (53): clean, get, obtain, profit, reap

The resulting Thesaurus verb list is publicly released here:

https://social-threats .github.io/PANACEA-ask-detection/resources/thesaurus_based_verbs List.txt.

We also adopt categorial variations through CATVAR [9] to map between different parts of speech, e.g., *winner*(N) ⟶ *win*(V). STYLUS, LCS+ and Thesaurus contain verbs only, but asks/framingsare often nominalized. For example, *you can reference your gift card* is an implicit ask to examine a gift card, yet without CATVAR this ask is potentially missed. CATVAR recognizes *reference* asa nominal form of *refer*, thus enabling the identification of this ask as a PERFORM.

Table 4. Framing Categories (GAIN, LOSE) in Lexical Organization of LCS+.
Italicized Exemplars with Boldfaced Triggers Illustrate Usage for Each Class. Boldfaced Class Numbers Indicate Those STYLUS Classes That Were Modified to Yield the LCS+ Resource.

LOSE

10.5 Steal Verbs→11 added (e.g., forfeit, lose, relinquish, sacrifice): *Don't **forfeit** this chance!*
10.6 Cheat Verbs: *Are your funds **depleted**?*
17.1 Throw Verbs: *Don't **toss** out this coupon*
17.2 Pelt Verbs: *Scams **bombarding** you?*
18.1 Hit Verbs: *Don't be **beaten** by debt*
18.2 Swat Verbs: ***Sluggish** market getting you down?*
18.3 Spank Verbs: ***Clobbered** by fees?*
18.4 Impact by Contact: *Avoid being **hit** by malware*
19 Poke Verbs: ***Stuck** with debt?*
29.2 Characterize Verbs→16 deleted (e.g., appreciate, envisage): ***Repudiated** by creditors?*
29.7 Orphan Verbs→5 deleted (apprentice, canonize, cuckold, knight, recruit): *Avoid **crippling** debt*
29.8 Captain Verbs→35 deleted (e.g., captain, coach, cox, escort): ***Bullied** by bill collectors?*
31.1 Amuse Verbs→91 deleted (e.g., amaze, amuse, gladden): *Don't be **disarmed** by hackers*
31.2 Admire Verbs→26 deleted (e.g., admire, exalt); *Are you **lamenting** your credit score?*
31.3 Marvel Verbs→1 deleted (feel): *Living in **fear**?*
33 Judgment Verbs: *Need to remove **penalties**?*
37.8 Complain Verbs: *Want your **gripes** answered?*
42.1 Murder Verbs: *Debt **killing** your credit?*
42.2 Poison Verbs: ***Strangled** by debt?*
44 Destroy Verbs: *PC **destroyed** by malware?*
48.2 Disappearance: *Your account will **expire***
51.2 Leave Verbs: *Found your **abandoned** prize*

GAIN

13.5.1 Get: *You are a **winner** of 1M Eu.*
13.5.2 Obtain: *You can **recover** your credit rating*

3.3 Dialogue Engineering

3.3.1 NLU Using Asks And Framing

The representation we use to generate plans leverages *asks* and *framings* based on conversation analysis literature. An *ask* is closely related to the notionof a request. Perhaps most importantly, an ask elicits relevant responses from the recipient. *Framing* refers to linguistic and social resources used to persuade the recipient of an ask to comply andperform the requested social action. Put another way, an *ask* creates a social obligation to respond,while framing provides an adequate basis for compliance with the ask.

3.3.2 Method

Our goal is to generate an informative response to the input utterance by first generating an appropriate **Response Plan**. We train two components separately. In the ***Planning Phase***, we experiment with generating plans in three ways:

3.3.2.1 Symbolic Planner

Foremost, we need to extract plans automatically from utterances. To accomplish this goal, our symbolic planner adapts lexical representations previously used for language analysis to the problem

of constructing **Response Plans**. We use lexical conceptual structures andbasic language processing tools for parsing the input, identifying the main **action**, identifying thearguments (or **targets**), and applying semantic-role labeling.

Once response plans are identified for all utterances in a given corpus using the symbolic planner, we need to address *automated generation* of such plans. Using the asks and framings as annotated data for a "silver" standard, we report precision of 69.2% in detecting asks/framings. Wetrain models to learn to generate "Response Plans" that are encoded with the same representationformat used for asks/framings. We use the language modeling paradigm and use a large pre-trained model (GPT-2) with the transformer architecture and the self-attention mechanism. We fine-tune this language model with the constraint of the input utterance and the plan for this input utterance,and train it to produce the plan for the response utterance. We adopt the fine-tuning approach specified by Ziegler *et al.* and train two specific models (Context Attention Planner [CTX] and Pseudo Self-Attention [PSA]) described below.

3.3.2.2 CTX

Based on the encoder/decoder architecture. In this model,the decoder weights are initialized with the pre-trained weights of the language model. However, anew context attention layer is added in the decoder that concatenates the conditioning informationto the pre-trained weight. The conditioning information, in our case, is the plan for the input utterance.

3.3.2.3 PSA

Proposed by Ziegler *et al.*, PSA injects conditioning information from the encoder directly into the pre-trained self-attention (similar to the "zero-shot" modelproposed by Radford *et al*.

In the ***Realization Phase***, we generate responses by utilizing the response plan generated fromthe planning phase as well as the input utterance. We expect a more guided generation of responses that are constrained by the response plan. In this phase, we only experiment with the PSA model, based on Ziegler *et al.*, who demonstrate that PSA outperforms otherapproaches on text generation tasks. We use nucleus sampling to overcome some of the drawbacksof beam search.

3.3.3 Implementation

We implement the models using Open-Neural Machine Translation (NMT) and the PyTorch framework.[4] We use publicly available Generative Pre-Trained Transformer (GPT)-2 model with 117M parameters, 12 layers and 12 heads inour implementations. The input utterances and the plans are tokenized using byte-pair encoding toreduce vocabulary size. Both phases are trained separately. In the Planning Phase, the *plan for the input* utterance along with the input utterance is used to generate the *response plan* for the responseutterance; in the Realization Phase, the response plan and input utterance are input to the modelto generate the response. In both planning and realization phase, separation tokens are added (e.g., <plan>), as is common practice for transformer inputs. We use Adam optimizer with a learning rate of 0.0005 and β_1 = 0.9 and β_2 = 0.98. During decoding, we use nucleus sampling both in theplanning and realization phase. All models are trained on two TitanV Graphics Processing Unit (GPU) and take roughly 15 hours each to train the planner and realization component. The trained models and the codebase are available at https://github.com/sashank06/planning_generation

[4] https://pytorch.org/

3.4 PANACEA Efforts Related To Corona Virus Disease (COVID-19) Information in Social Media (Extension)

We have extended the use of PANACEA tools and techniques to other social media platforms (besides LinkedIn), including Twitter, Reddit and YouTube, and changed the focus from individual- oriented social engineering to mass-scale disinformation. The main focus was to analyze user attitudes towards trending topics associated with COVID-19 on Twitter and Reddit platforms since January 2020 and the spread of disinformation, fake news, and malicious contents (e.g., fraudulent offers of medicines or personal protective equipment [PPE]).

Over the project extension, we collected millions of postings (mostly text) using available Application Programming Interfaces (API). As of August 2020, we collected more than 5 million posts from the Reddit platform and more than 100 million messages from Twitter. We used an evolving set of keywords to select relevant content, starting with a small list of keywords (e.g., coronavirus, Wuhan, COVID, social distancing, etc.) and expanding it based on occurrence of new frequent terms (e.g., N-95, hydroxychloroquine, sheltering, protests, reopening, etc.).

Specifically, Reddit, Twitter and YouTube were selected as suitable medial channels due to their open access nature as well as APIs availability and high-text interactions. Reddit provided dialogue conversations where people discuss news, opinions, and even theories associated to the COVID-19 outbreak. In Twitter posts, messages reflected more immediate, real-time user opinions and user interactions across the globe. Finally, YouTube provided additional insight into users opinions and visual data sharing.

3.4.1 Twitter Data

The dataset is a subset of the publicly available collection USC-COVID-19-Twitter that covers posts from 21 January to 6 May 2020 and contains 118,938,245 tweets selected from the 952,549,092 in Twitter feed during that period. The collection comprises tweets in different languages including English, Spanish, Portuguese, and French, although more than 70% of content is in English.

The COVID dataset tweets were extracted using a list of trending hashtags related to the pandemic such as the ones show in Table 5. For a complete list there is available the COVID-19 keyword collection that describe hashtags used for each month.

For each tweet on the collection, there are more than 150 associated properties, including users mentioned, profile data, among others. A complete list can be found in the Twitter API object description

Table 5. Sample Twitter Hashtags Related to Pandemic

Coronavirus	CDC	Wuhan	N95
Epidemic	outbreak	China	covid
pandemic	panicbuy	14DayQuarantine	chinesevirus
stayhome	lockdown	trumppandemic	covidiot
PPEshortage	quarentinelife	panic-buy	COVD
flattenthecurve	DuringMy14Day	Coronials	sars

The complete distribution of tweets in the COVID-related subset is shown below:

- January 01-21-2020 to 01-31-2020: 9,705,777 tweets.
- February 02-01-2020 to 02-29-2020: 27,855,635 tweets.
- March 03-01-2020 to 03-31-2020: 50,615,724 tweets.
- April 04-01-2020 to 04-29-2020: 29,224,631 tweets.
- May 05-01-2020 to 05-31-2020: 1,536,478 tweets.

The retrieved tweets include all metadata normally associated with the messages (such as poster name, ID, timestamp, etc.) except for geolocation elements which are restricted by Twitter API due to security considerations.

3.4.2 Reddit Data

This dataset is derived from the public Reddit API using similar keywords to those in selecting the Twitter subset. Two subcollections were obtained: (1) Reddit-light which comprises dialogue title, description and the main posts, but without the replies; and (2) Reddit-complete that adds all the replies.

In each dataset, we save metadata, including: the author, timestamp, utterance-order, and twelve other features that are needed to analyze interactions between the users (e.g., replies, responses to the replies, questions posed, etc.) In total, we collected 24,311 posts for the Reddit-lightand 3,436,864 for Reddit-complete from January to May. All Reddit posts that were collected are in English.

Posts include metadata of real users except for geolocation elements that are eliminated by Reddit API according to usage terms.

3.4.3 YouTube Data

YouTube data contains mostly video plus textual metadata including title, author, description, among others. This dataset is the smallest in the collection due to the severe restrictions imposed by Google API.

We collected 21,314 video and channel descriptions from January to June 2020. All data is in English and restricted content created in the US (one of the major distinctions from the other collections). Some geolocation information is preserved such as area/region where the video was created or the location of certain public figures shown or mentioned.

4.0 RESULTS AND DISCUSSION

Friend/foe detection (Message Analysis) and response generation (Dialogue) are evaluated for effectiveness of PANACEA as an effective intermediary between attackers and potential victims.

4.1 Threat Intelligence and Analysis

4.1.1 Message Analysis Module

The Defense Advanced Research Projects Agency (DARPA) Active Social Engineering Defense (ASED) program evaluation tests header and content modules against messages for friend/foe determination. Multiple sub-evaluations check system accuracy in distinguishing malicious messages from benign ones, reducing the false alarm rate, and transmitting appropriate messages to dialogue components for further analysis. Evaluated components yield 90% accuracy.

Components adapted for detecting borderline exchanges (*unknown* cases) are shown to help dialogue components request more information for potentially malicious messages.

4.1.2 Dialogue Module.

The ASED program evaluation also tests the dialogue component. In-dependent evaluators communicate with the system without knowledge of whether they are inter-acting with humans or bots. Their task is to engage in a dialogue for as many turns as necessary. PANACEA bots are able to sustain conversations for an average of five turns (across 15 distinct threads).Scoring applied by independent evaluators yield a rating of 1.9 for their ability to display human-like communication (on a scale of 1–3; 1=bot, 3=human). This score is the highest amongst all other competing approaches (four other teams) in this independent program evaluation.

4.2 SE Lexicon

Intrinsic evaluation of our resources is based on comparison of ask/framing detection to an adjudicated GT, a set of 472 clauses from system output on 20 unseen emails.These 20 emails are a random subset of 2600+ messages collected in an email account set up to receive messages from an internal red team as well as "legitimate" messages from corporate and academic mailing lists. As alluded to earlier, these 20 emails are distinct from the dataset used forresource adaptation to produce the task-related LCS+.

The GT is produced through human adjudication and correction by a computational linguist[5] of initial ask/framing labels automatically assigned by our system to the 472 clauses. System output also includes the identification of a "top ask" for each email, based on the degree to which ask argument positions are filled.[6] *Top asks* are adjudicated by the computational linguist once the ask/framing labels are adjudicated. The resulting GT is accessible here:

https://social-t hreats.github.io/PANACEA-ask-detection/data/.

The GT is used to measure the precision/recall/F of three of three variants of ask detection output (Ask, Framing, and Top Ask) corresponding to our three lexica: Thesaurus, STYLUS, and LCS+. LCS+ is favored (with statistical significance) against the two very strong baselines, Thesaurus and STYLUS. Table 6 presents results: Recall for framings is highest for STYLUS, but at the cost of higher false positives (lower precision). F-scores increase for STYLUS over Thesaurus, and for LCS+ over STYLUS.

[5]The adjudicator is an author but is not the algorithm developer, who is also an author.

[6]Argument positions express information such as the ask type (i.e., PERFORM), context to the ask (i.e. financial),and the ask target (e.g., "you" in "Did you send me the money?").

McNemar [28] tests yield statistically significant differences for asks/framings at the 2% level between Thesaurus and LCS+ and between STYLUS and LCS+.[7] It should be noted that not all clauses in GT are ask or framing: vast majority (80%) are neither (i.e., they are true negatives).

We note that an alternative to the Thesaurus and LCS baselines would be a bag-of-words lexicon, with no organizational structure. However, the key contribution of this work is the ease of adaptation through classes, obviating the need for training data (which are exceedingly difficult to obtain). Classes enable extension of a small set of verbs to a larger range of options, e.g., if the human determines from a small set of task-related emails that *provide* is relevant, the task-adapted lexicon will include *administer*, *contribute*, and *donate* for free. If a class-based lexical organization is replaced by bag-of-words, we stand to lose efficient (one-person-day) resource adaptation and, moreover, training data would be needed.

Table 6. Impact of Lexical Resources on Ask/Framing Detection: Thesaurus, STYLUS, and LCS+

Thesaurus	**P**	**R**	**F**
Ask:	0.273	0.042	0.072
Framing:	0.265	0.360	0.305
TopAsk:	0.273	0.057	0.094
STYLUS	**P**	**R**	**F**
Ask:	0.333	0.104	0.159
Framing:	0.298	**0.636**	0.406
TopAsk:	0.571	0.151	0.239
LCS+	**P**	**R**	**F**
Ask:	**0.667**	**0.411**	**0.508**
Framing:	**0.600**	0.600	**0.600**
TopAsk:	**0.692**	**0.340**	**0.456**

A first step toward *extrinsic* evaluation is inspection of responses generated from each resource's top ask/framing pairs. Table 1 (given earlier) shows LCS+ ask/framing pairs whose responding (T)hesaurus and (S)TYLUS pairs are:

(a) **T**: None, None
S: None, GAIN/won(1.7Eu)
(b) **T**: PERFORM/do(that), LOSE/lose(money)
S: GAIN/won(money), GIVE/send(money)
(c) **T**: None, GAIN/get(20%)
S: PERFORM/sign(http:..), GAIN/get(20%)

Below are corresponding examples of generated responses[8] for all three resources, based on a templatic approach that leverages ask/framing hierarchical structure and corresponding confidence scores. This module is part of a larger, separate publication.

[7]Tested values were TP+TN vs FP+FN, i.e., significance of change in total error rate
[8]For brevity, *excerpts* are shown in lieu of full emails.

(a) **T**: How are you? Thanks.
S: ...too good to be true. What should I do?
L+: I will contact asap.

(b) **T**: Thanks for getting in touch, need more info.
S: Nervous about this. Your name?
L+: I would respond,[9] but I need more info.

(c) **T**: What should I do now?
S: Website doesn't open, is this the link?
L+: Thanks, need more info before I paste link

There are qualitative differences in these responses. For example, in (a) Thesaurus (T) yields no asks/framings; thus a canned response is generated. By contrast, the same email yields a more responsive output for STYLUS (S), and a more focused response for LCS+ (L). Similar distinctions are found for responses in (b) and (c). Note that in the LCS+ condition, if there is no match foundusing LCS+, downstream response generation prompts the attacker (e.g., "please clarify") until an interpretable ask or framing appears. In this social engineering task, not all responses move the conversation forward. A central goal of the social engineering task is to waste the attacker's time, play along, and possibly extract information that could unveil their identity.

4.3 Experiments on COVID-19 Information in Social Media (Extension)

In the extended period of the project, using the collected data on COVID-19 in social media (as discussed in Section 3.4), we conducted a series of experiments to detect users' sentiment towards various COVID-related topics. We applied unsupervised machine learning to automatically detect topic-related terms using a Latent Dirichlet Allocation (LDA) algorithm. The resulting term sets were subsequently mapped into a pre-trained word embedding space, forming clusters of different shape. For each cluster we computed a centroid, and then selected the word closest to this centroid as the topic label. This process resulted in topics that were directly related to COVID-19 theme (e.g., case fatality rate (CFR), infection, etc.), as well as more "generic" topics such as "study" or "information" that captured certain activities associated with the pandemic.

Figure 3 shows a few topic clusters extracted from the Reddit dataset and their centroids. From the LDA process, we selected ten top words on average per cluster; embeddings of these words were then used to select the most appropriate label for each cluster. Overall, we obtained the set of 100 COVID-related topics.

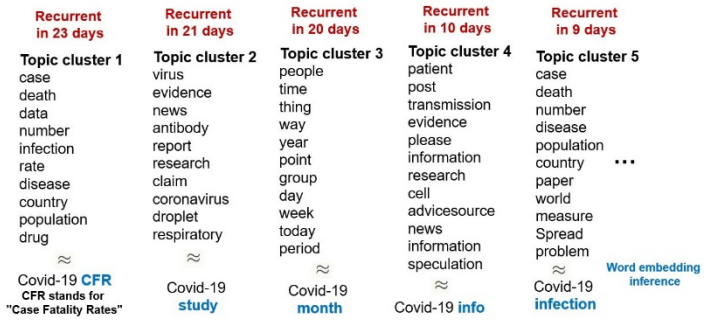

Figure 3. Examples of Generated Topics with Topic Centroids, i.e., Words Closest to the Cluster Centroid in an Embedding Space.

[9] LCS+ detects both GIVE/send and PERFORM/respond.

Given the extracted topics, we computed sentiment distribution in the dataset based on the prevailing sentiment polarity in messages that mention any of these topics. The goal was to see if the overall sentiment to all COVID-19 related matter changes over time as the pandemic progresses. Furthermore, we wanted to see if for some topics, such as "vaccine" or "infection", the sentiment follows a more distinctive pattern, e.g., increases or decreases in response to certain events. Sentiment was calculated using basic NLP tools, including Core NLP package and Natural Language Toolkit (NLTK)-sentiment.

The experiments for Reddit followed the processing pipeline shown in Figure 4 where we highlight each step associated with the detection of topics and sentiment that ultimately produces aggregate analyses of attitude trends in the population as related to the pandemic outbreak.

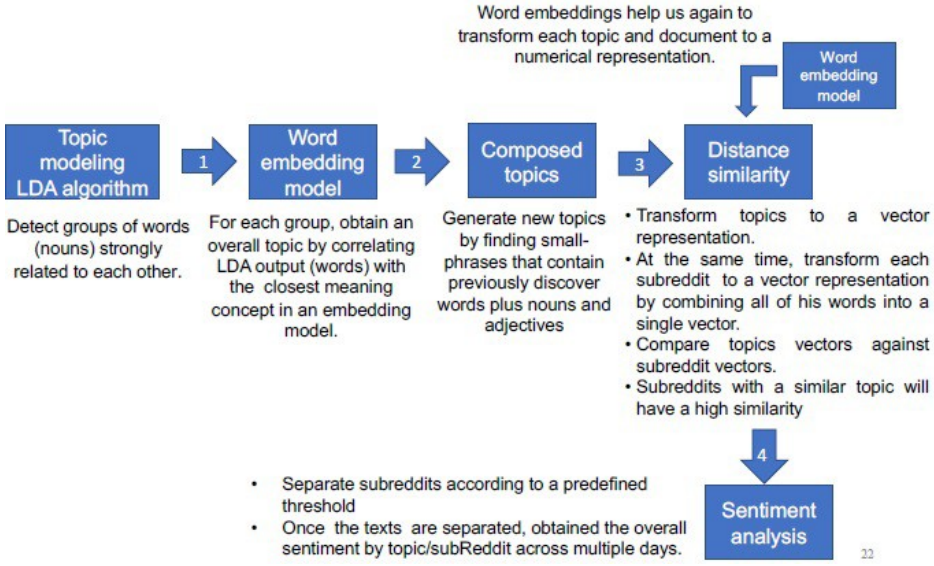

Figure 4. Experimental Processing Pipeline over Social Media Text Samples.

Figure 5 shows examples of prevailing sentiment towards selected COVID-19 topics (obtained through the LDA algorithm and word embeddings) in all Reddit discussion threads.

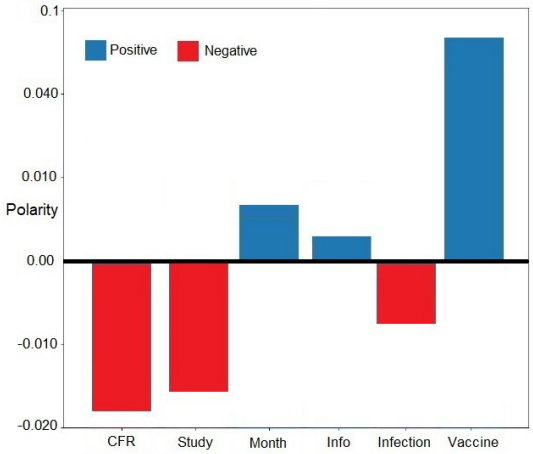

Figure 5. Sentiment across Selected Topic in Reddit Discussion Threads Averaged over the Experimental Period.

Figure 6 displays sentiment trend associated with the "vaccine" topic in the subreddit thread "Honest question: cost/benefit of shutdowns" that includes over 640 comments. We note that positive sentiment is dominating and growing in general; however, there are a few "hiccups" where it turns negative on a few dates. A manual analysis over the comments on these dates reveals these were the days when delays in the vaccine development and validation were announced in the media. This chart thus reflects the public sentiment associated with the topic.

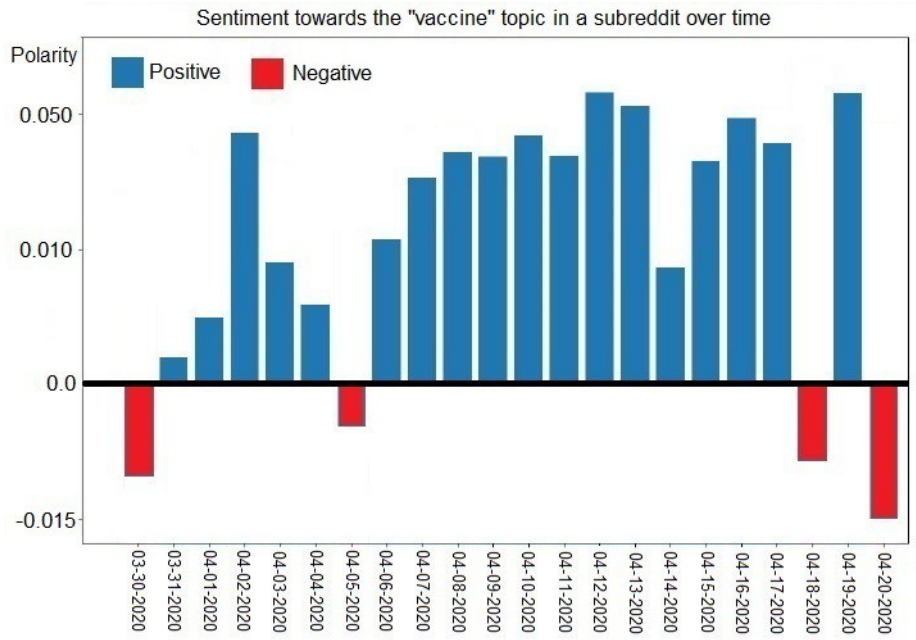

Figure 6. Fine-grained Analysis of the "Vaccine" Topic over Time within a Single Subreddit.

The above figures show sentiment analysis over a subset of 15,000 subreddits each with around 300 utterances per thread within a 24-day period from 3/30/20 through 4/20/20.

We ran similar experiments for Twitter data, but instead of generating topics through the LDA-word-embedding pipeline, we used the most frequent hashtags as proxies for the topics. Figure 7 shows average sentiment distribution across some selected topics/hashtags in the Twitter dataset.

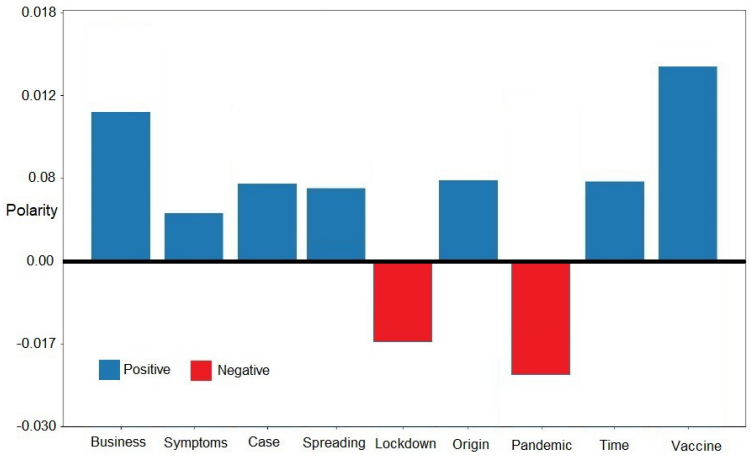

Figure 7. Twitter Topics Found in All Dates with Overall Sentiment.

Figure 8 displays sentiment distribution for the topic “symptoms” and Figure 9 display sentiment distribution for the topic “Lockdown” in Twitter messages. While the latter is understandably negative, even as it eases off over time, the first graph is harder to explain: it may simply indicate better public understanding of what are and aren’t COVID symptoms, and thus reduced associated anxiety this moving the sentiment towards the neutral. All above Twitter figures show sentiment distribution for the topic “Lockdown” in Twitter messages. While the latter is understandably negative, even as it eases off over time, the first graph is harder to explain: it may simply indicate better public understanding of what are and aren’t COVID symptoms and thus reduced associated anxiety this moving the sentiment towards the neutral.

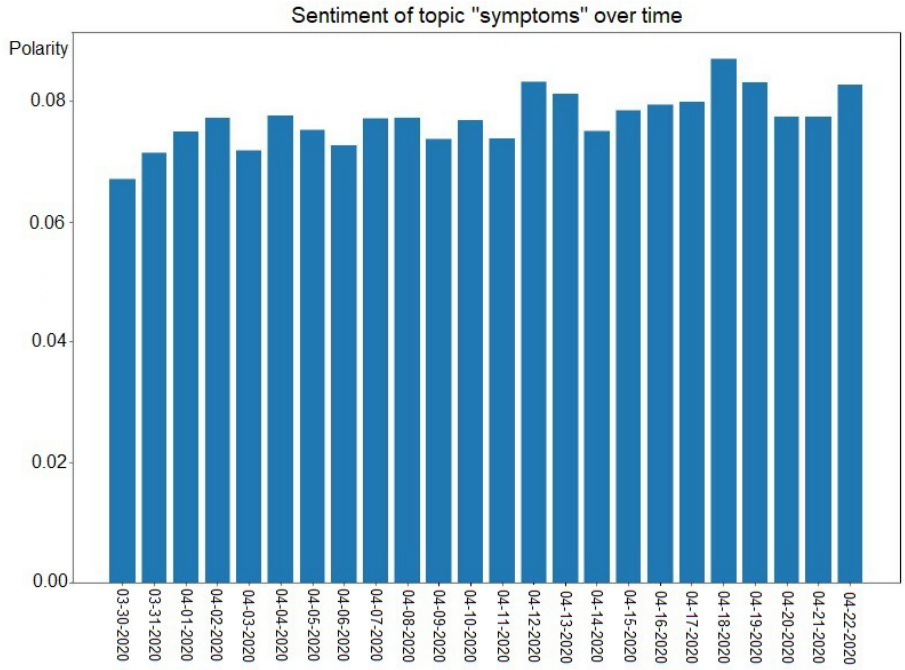

Figure 8. Twitter Topic “Symptoms” across Multiple Dates.

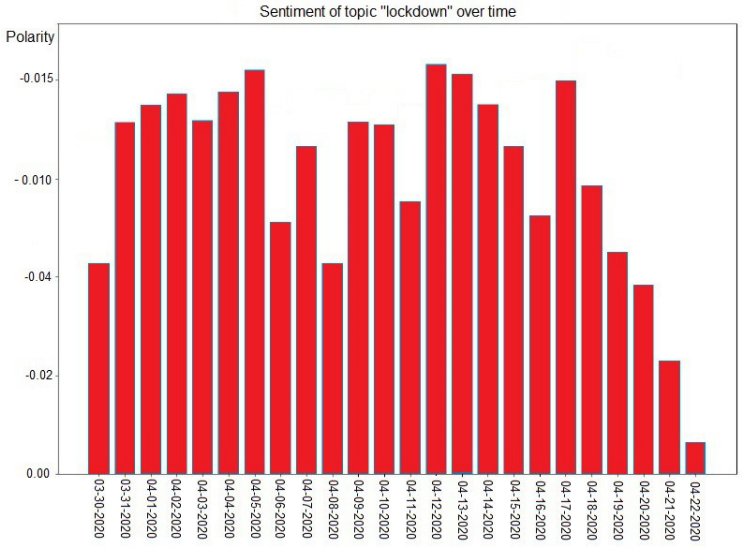

Figure 9. Twitter Topic “Lockdown” Across Multiple Dates

All above Twitter figures show sentiment analysis over a subset of 10,000 tweets within the 24-day period 3/30/20 through 4/22/20.

We also applied PANACEA content-classifier tools to Twitter messages in an attempt to detect messages that may include malicious content, i.e., messages that make COVID related claims that appear to be scams of some kind. We did not attempt to verify whether a message content is true or not; rather, we focused on the form of message to detect what might be intentionally misleading information (as opposed to simply ignorant, etc.) Specifically, we used the benign/non-benign classifier developed for emails (see Section 3.1.2) and adapted it to tweets. In this experiment we replaced the word embedding model based on email-content with one based on COVID-19 Twitter texts.

The resulting classifier was found to be promising in that the majority of intercepted messages appeared to be indeed malicious misinformation. We were unable to perform a thorough evaluation due to the lack of a proper ground truth – i.e., a large human-annotated subset of tweets. However, using a small sample of manually annotated tweets (100 tweets, 50-malicious/50-not-malicious) we were able to obtain accuracy of 72% with 78% precision and 86% recall of malicious content detection. These numbers require further verification, given the small size of the test data and its limited reliability (the data was score by a single annotator).

Below are a few examples of suspicious posts captured in our filter. Note that some posts were caught because they invite users to click on a link – a standard social engineering trick deployed in emails.

- "Apparently, a cure has been developed! #WuhanPneumonia #WuhanCoronavirus <URL>"
- "Curious about the coronavirus? Please click the link below to learn more: <URL>"
- "I have discovered an instant COVID-19 test which is 99%+ accurate. It is free and can be done remotely. Click here <URL> to get your free test."
- "Click here <URL> if you oppose the decision made by the Department of Veterans Affairs (VA) to remove the physician supervision of nurse anesthetists during the COVID-19 pandemic."

The above experiments were conducted by adapting the PANACEA tools that were developedto detect scams and other malicious content in email and LinkedIn in-mail messages. These are of course markedly different from the public posts on social media, which are not typically directed at anyone in particular. Thus, a more careful adaptation is necessary. Nonetheless, these initial experiments are encouraging as we were able to extract a variety of commercial scams related to PPE and antiviral drugs, and other remedies.

The initial results obtained during the project extension demonstrated the utility of word embedding for more accurate topic modeling than the basic LDA algorithm. More detailed experiments are needed to further improve topic and sentiment detection; for example, by exploiting ontology resources (e.g., Wikipedia) to take advantage of semantic relationships between words. In the future experiments, we plan to align the observed shifts in public attitude towards some controversial topics with the volume of online disinformation about these topic (caused by deliber ate disinformation campaigns, or by temporal virality of certain memes). The objective will be to determine the impact of such disinformation on the population attitudes towards e.g., vaccinations, or mask wearing, etc.

Future experimentation based on PANACEA extension could provide valuable and insight into public attitude towards certain social issues and their compliance with the regulations imposed by the authorities. This is of critical interest during the COVID pandemic, since public behavior likely correlates with the rate of disease spread and the number of infections. Malicious or otherwise delusional information campaigns in social media, e.g., those suggesting that COVID is not dangerous, may contribute negatively to public response, thus inflicting irreversible harm on the population. Unlike the SE scams in office email, the potential damage could besignificantly greater. We believe that this preliminary work has shown how population-level effects may be detected. Further research is required to understand how it can be prevented.

5.0 CONCLUSIONS

5.1 Threat Intelligence and Analysis

PANACEA is an operational system that processes communication data into actionable intelligence and provides active defense capabilities to combat SE. The F3EAD active defense cycle was chosen because it fits the social engineering problem domain, but specific phases could be changed to address different problems. For example, a system using the PANACEA processing pipeline could ingest academic papers on a disease, process them with components designed to extract biologicalmechanisms, then engage with paper authors to ask clarifying questions and search for additionalliterature to review, while populating a knowledge base containing the critical intelligence for thedisease of interest.

Going forward, the plan is to improve PANACEA's plug-in infrastructure so that it is easier to add capability without updating PANACEA itself. This is currently possible as long as new components use the same REST API as existing components. The obvious next step is to formalize PANACEA'sAPI. We have found value to leaving it open at this early state of development as we discover newchallenges and solutions to problems that emerge in building a large scale system focused on the dangers and opportunities in human language communication.

5.2 SE Lexicon

Both STYLUS and LCS+ support ask/framing detection in service of bot-produced responses. Intrinsically, LCS+ is superior to both STYLUS and Thesaurus when measured against human-adjudicated output, verified for significance by McNemar tests at the 2% level. Extrinsically, STYLUS supports more responsive bot outputs and LCS+ supports more focused bot outputs.

A more general advantage of adapting LCS+ to the social engineering domain is that it can act as a guideline for developing similar resources for other domains which will similarly support focused outputs appropriate for particular domains. The main contribution of this paper is not development of a particular task-specific resource, nor to suggest that LCS+ is a generic resource for many tasks, but to present a systematic, efficient approach to resource adaptation technique that can generalize to other tasks for improved task-specific performance, e.g., understanding viewpoints in social media or detecting motives behind activities of political groups. We acknowledge that our extrinsic evaluation is limited. While we have demonstrated the efficacy of ask detection approaches on a set of representative emails, a quantitative evaluation is required to test the statistical significance of our extrinsic observations. Future work is planned to conduct experiments with crowd-sourced workers judging the efficacy and effectiveness of generated responses.

5.3 Dialogue Engineering

Our key finding through two separate human crowdsourced studies is that decoupling realization, and planning phases outperforms an end-to-end No Planner system across three metrics (Appropriateness, Quality, and Usefulness).

In this work, we have taken an initial step towards the goal of replicating human language generation processes. Thorough and rigorous evaluations are required to fully support our claims, e.g., by including additional metrics and more diverse corpora. In this work, we limit the types to GIVE, GAIN, LOSE, and PERFORM. However, we do not restrict the ask action and target at all. Also, since our symbolic planner can be used to obtain silver standard training data, straightforward changes like adding additional lexicons would enable us to generalize to other corpora as wellas include additional ask types in our pipeline. Another natural extension would be to explore training

the planning and realization phases together in a hierarchical process. This would, in principle, further validate the efficacy of our approach. Further details about the dialogue efforts in PANACEA are described in the Empirical Methods in Natural Language Processing (EMNLP) 2020 paper [29].

6.0 REFERENCES

[1] D. E. Denning. “Framework and principles for active cyber defense”. In: *Computers & Security* 40 (2014), pp. 108–113.

[2] C. Hadnagy and M. Fincher. *Phishing Dark Waters*. Wiley Online Library, 2015.

[3] J. A. Gomez. “The targeting process: D3A and F3EAD”. In: *Small Wars Journal* 1 (2011), pp. 1–17.

[4] P. Drew and E. Couper-Kuhlen. *Requesting in social interaction*. John Benjamins Publishing Company, 2014.

[5] A. Zemel. “Texts as actions: Requests in online chats between reference librarians and library patrons”. In: *Journal of the Association for Information Science and Technology* 67.7 (2017), pp. 1687–1697.

[6] B. Dorr, A. Bhatia, A. Dalton, B. Mather, B. Hebenstreit, S. Santhanam, Z. Cheng, S. Zemel, and T. Strzalkowski. “Detecting Asks in Social Engineering Attacks: Impact of linguistic and Structural Knowledge”. In: *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence 2020*. 2020.

[7] B. Dorr and C. Voss. “STYLUS: A Resource for Systematically Derived Language Usage”. In: *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 57–64.

[8] B. J. Dorr and M. B. Olsen. “Lexical Conceptual Structure of Literal and Metaphorical Spatial Language: A Case Study of Push”. In: *Proceedings of the First International Workshop on Spatial Language Understanding*. 2018, pp. 31–40.

[9] N. Habash and B. J. Dorr. “A Categorial Variation Database for English”. In: *In Proceedings of the Human Language Technology and North American Association for Computational Linguistics (NAACL) Conference*. 2003, pp. 96–102.

[10] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Association for Computational Linguistics (ACL) System Demonstrations*. ACL, 2014, pp. 55–60.

[11] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. “Neural Probabilistic Language Models”. In: *Innovations in Machine Learning: Theory and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 137–186. DOI: 10.1007/3- 540-33486-6_6.

[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Curran Associates Inc., 2013, pp. 3111–3119.

[13] A. Sherstinsky. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network”. In: Cornell University, 2013, pp. 1–39.

[14] B. Klimt and Y. Yang. “The Enron Corpus: A New Dataset for Email Classification Research”. In: *Machine Learning: ECML 2004*. Springer Berlin Heidelberg, 2004, pp. 217–

226. DOI: 10.1007/978-3-540-30115-8_22.

[15] A. Oest, Y. Safei, A. Doupe, G.-J. Ahn, B. Wardman, and G. Warner. "Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis". In: *Proceedings of the 2018 APWG Symposium on Electronic Crime Research, eCrime 2018*. IEEE Computer Society, 2018, pp. 1–12. DOI: 10.1109/ECRIME.2018.8376206.

[16] JDP 2-00. *Understanding and intelligence support to joint operations (JDP 2-00)*. Ministry of Defence (UK), 2011.

[17] B. E. Strom, J. A. Battaglia, M. S. Kemmerer, W. Kupersanin, D. P. Miller, C. Wampler, S. M. Whitley, and R. D. Wolf. "Finding cyber threats with ATT&CK-based analytics". In: *The MITRE Corporation, Tech. Rep.* 1.1 (2017).

[18] R. Jackendoff. *Semantics and Cognition*. Cambridge, MA: MIT Press, 1983.

[19] R. Jackendoff. *Semantic Structures*. Cambridge, MA: MIT Press, 1990.

[20] B. J. Dorr. *Machine Translation: A View from the Lexicon*. Cambridge, MA: MIT Press, 1993.

[21] R. Jackendoff. "The Proper Treatment of Measuring Out, Telicity, and Perhaps Even Quantification in English". In: *Natural Language and Linguistic Theory* 14 (1996), 305–354.

[22] B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, 1993.

[23] M. B. Olsen. "The Semantics and Pragmatics of Lexical and Grammatical Aspect". In: *Studies in the Linguistic Sciences* 24.1–2 (1994), 361–375.

[24] S. C. Chang, R. C. Shahani, D. J. Cipollone, M. V. Calcagno, M. J. B. Olsen, and D. J. Parkinson. *Linguistic Object Model*. 7,171,352. Jan. 2007.

[25] S. C. Chang, R. C. Shahani, D. J. Cipollone, M. V. Calcagno, M. J. B. Olsen, and D. J. Parkinson. *Lexical Semantic Structure*. 7,689,410. Mar. 2010.

[26] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. "A Large-scale Classification of English Verbs". In: *Language Resources and Evaluation*. 2007.

[27] M. Palmer, C. Bonial, and J. D. Hwang. "VerbNet: Capturing English Verb behavior, Meaning and Usage". In: *The Oxford Handbook of Cognitive Science*. Ed. by Susan Chipman. Oxford University Press, 2017.

[28] Q. McNemar. "Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika* 12.2 (June 1947), pp. 153–157.

[29] S. Santhanam, Z. Cheng, B. Mather, B. J. Dorr, A. Bhatia, B. Hebenstreit, A. Zemel, A. Dalton, T. Strzalkowski, and S. Shaikh. "Learning to Plan and Realize Separately for Open-Ended Dialogue Systems". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2020.

7.0 LIST OF ABBREVIATIONS, ACRONYMS AND SYMBOLS

API	Application Programming Interfaces
APWG	Anti-Phishing Working Group
ASED	Active Social Engineering Defense
ATT&CK	Adversarial Tactics, Techniques and Common Knowledge
CATVAR	Categorial-Variation
CFR	Case Fatality Rate
COVID	Corona Virus Disease
CTX	Context Attention Planner
DARPA	Defense Advanced Research Projects Agency
EMNLP	Empirical Methods in Natural Language Processing.
F3EAD	Find, Fix, Finish, Exploit, Analyze and Disseminate
GPT	Generative Pre-Trained Transformer
GPU	Graphics Processing Unit
GT	Ground Truth
IP	Internet Protocol
JSON	JavaScript Object Notation
LCS	Lexical Conceptual Structure
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NMP	Neural Machine Translation
PANACEA	Personalized AutoNomous Agents Countering Social Engineering Attacks
POS	Part-of-Speech
PPE	Personal Protective Equipment
PSA	Pseudo Self-Attention
SE	Social Engineering
SMS	Short Message Service
SRL	Semanic Role Labeling
SVM	Support Vector Machine
VA	Veteran Affairs

APPENDIX A - Publications and Presentations

The project has produced the following publications and presentations:

[1] S. Santhanam, Z. Cheng, B. Mather, B. J. Dorr, A. Bhatia, B. Hebenstreit, A. Zemel, A. Dalton, T. Strzalkowski, and S. Shaikh. "Learning to Plan and Realize Separately for Open-Ended Dialogue Systems". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2020.

[2] B. J. Dorr, A. Bhatia, A. Dalton, B. Mather, B. Hebenstreit, S. Santhanam, Z. Cheng, S. Shaikh, A. Zemel, and T. Strzalkowski. "Detecting Asks in Social Engineering Attacks: Impact of Linguistic and Structural Knowledge". In: *Proc. Thirty-Fourth AAAI Conferenceon Artificial Intelligence (AAAI 2020)*. AAAI Press, 2020, pp. 7675–7682.

[3] A. Dalton, E. Aghaei, E. Al-Shaer, A. Bhatia, E. Castillo, Z. Cheng, S. Dhaduvai, Q. Duan,
B. Hebenstreit, M. M. Islam, Y. Karimi, A. Masoumzadeh, B. Mather, S. Santhanam, S. Shaikh, A. Zemel, T. Strzalkowski, and B. J. Dorr. "Active Defense against Social Engi- neering: The Case for Human Language Technology". In: *Proceedings for the First Inter- national Workshop on Social Threats in Online Conversations: Understanding and Man- agement*. Marseille, France: European Language Resources Association, May 2020, pp. 1–8.

[4] A. Bhatia, A. Dalton, B. Mather, S. Santhanam, S. Shaikh, A. Zemel, T. Strzalkowski, and
B. J. Dorr. "Adaptation of a Lexical Organization for Social Engineering Detection and Re- sponse Generation". In: *Proceedings for the First International Workshop on Social Threatsin Online Conversations: Understanding and Management, STOC@LREC 2020, Marseille,France, May 2020*. Ed. by A. Bhatia and S. Shaikh. European Language Resources Associ-ation, 2020, pp. 9–14.

[5] A. Dalton, A. Zemel, A. Masoumzadeh, A. Bhatia, B. Dorr, B. Mather, B. Hebenstreit, E. Al-Shaer, E. C. J. Ellisa Khoja, L. Bunch, M. Vlahovic, P. Liu, P. Pirolli, R. Shah, S. Carta- cio, S. Shaikh, S. Santhanam, S. Dhaduvai, T. Strzalkowski, and Y. Karimi. "Modeling So-cial Engineering Risk using Attitudes, Actions, and Intentions Reflected in Language Use".In: *Proceedings of the Thirty-Second International Florida Artificial Intelligence ResearchSociety Conference, Sarasota, Florida, USA, May 19-22 2019.* 2019.

[6] E. Castillo, S. Dhaduvai, P. Liu, K.-S. Thakur, A. Dalton, and T. Strzalkowski. "Email Threat Detection Using Distinct Neural Network Approaches". English. In: *Proceedings for the First International Workshop on Social Threats in Online Conversations: Under- standing and Management*. Marseille, France: European Language Resources Association,May 2020, pp. 48–55.