Recognizing Radicalization Indicators in Text Documents Using Human-in-the-Loop Information Extraction and NLP Techniques

Benjamin W.K. Hung, Shashika R. Muramudalige, Anura P. Jayasumana
Electrical & Computer Engineering
Colorado State University
Fort Collins, Colorado, USA
{benjamin.hung, shashika.muramudalige, anura.jayasumana}@colostate.edu

Jytte Klausen, Rosanne Libretti, Evan Moloney, Priyanka Renugopalakrishnan
Western Jihadism Project
Brandeis University
Waltham, Massachusetts, USA
{klausen, roselib, emoloney, priyankarina}@brandeis.edu

***Abstract*—Among the operational shortfalls that hinder law enforcement from achieving greater success in preventing terrorist attacks is the difficulty in dynamically assessing individualized violent extremism risk at scale given the enormous amount of primarily text-based records in disparate databases. In this work, we undertake the critical task of employing natural language processing (NLP) techniques and supervised machine learning models to classify textual data in analyst and investigator notes and reports for radicalization behavioral indicators. This effort to generate structured knowledge will build towards an operational capability to assist analysts in rapidly mining law enforcement and intelligence databases for cues and risk indicators. In the near-term, this effort also enables more rapid coding of biographical radicalization profiles to augment a research database of violent extremists and their exhibited behavioral indicators.**

***Index Terms*—information extraction, natural language processing, human-in-the-loop, machine learning**

I. Introduction

Terrorism experts contend that despite ISIS's significant territorial losses in Iraq and Syria recently, the threat of jihadist attacks against the US and the West will nevertheless persist due to the organization's continued ability to attract followers and inspire or direct attacks [15], [16], [23]. In fact, the US House of Representatives Homeland Security Committee reported in October 2018 a 63% increase in the number of ISIS-inspired attacks within the last 2 years, just as the so-called Caliphate was dwindling in size [16]. It is evident that law enforcement and intelligence agencies must remain vigilant against future terrorist attacks by investigating and intercepting those on suspected radicalization pathways to violent extremism.

This work was supported by the U.S. Department of Justice, Office of Justice Programs/National Institute of Justice under Award #2013-ZA-BX-0005. Opinions or points of view expressed in this article are those of the authors and do not necessarily reflect the official position of policies of the U.S. Department of Justice.

Among the operational shortfalls that hinder these agencies from achieving greater success is the difficulty in dynamically assessing individualized violent extremism risk at scale given the enormous amount of primarily text-based records in disparate databases [28]. According to the FBI's 9-11 Review Commission in 2015, an "important question faced by all intelligence agencies" is "how to scan and assess voluminous amounts of collected information strategically and [to identify] valuable intelligence leads" [14].

Our research aims to support law enforcement and intelligence agencies by advancing the state-of-the-art in identifying domestic radicalization to violent extremism and preventing future extremist attacks. The term "radicalization" is commonly used but controversial. It is shorthand for "radicalization to violent extremism," which implies a process view of how individuals move from beliefs to actions [3]. For this reason, a recent guide issued by the Office of the Director of National Intelligence uses the term "mobilization" to indicate the overt behavioral dynamics associated with growing radicalization leading to terrorism-related actions [9]. The new approach to threat assessment signals a shift in the understanding of violent extremism, relies on behavioral evidence of growing extremism, and tracks cues to changes in an individual's behavior that suggest an increased concern in relation to progression towards committing an act of violent extremism.

Research has shown that violent extremist actions are the result of learned behaviors and, often, a protracted process of cognitive and behavioral adaptation. A study of 63 mass shooter incidents published by the FBI found that, on average, the perpetrators displayed 4 to 5 concerning behaviors that were observed by others prior to the shooting incident [36]. A growing body of social scientific research supports the view that the pathways to violent action or to criminal actions in

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

VARIABLE	CUES	EXAMPLE
Information Seeking: This field tracks the dates on which an individual began actively seeking out information regarding radical Islamist or Jihadist beliefs and groups.	• Conducting internet research on Islamic fundamentalism • Seeking out Jihadist literature and doctrine • Dialogue with extremist figures • Seeking out a more radical place of worship	Bryant Neal Vinas became interested in Islam after meeting a Muslim woman at a mall before 9/11. After he converted into Islam in 2004 at a Queens mosque frequented by other extremists. As the US invasion of Afghanistan dragged on and Iraq slipped into a civil war, Vinas started visiting jihadist websites and watching al-Qaeda propaganda videos.
Lifestyle Changes: This field tracks the dates of an individual's personal changes in order to better fit with a doctrinal or group-enforced idea of 'proper' Islamic conduct.	• Growing a beard • Starting to wear trousers cut above the ankle • Abstention from food/substances deemed haram • Changing or showing new expressions of overt religious piety	Prior to leaving for Sweden to marry Yasin Mohammed, Ariel Bradley wore a hijab, but after her return to the States, she began wearing an abaya.
Convert: This field tracks the date an individual is known to convert to Islam. It does not indicate an individual's radicalization, but simply their conversion to Islam.	• Reciting Shahada • Taking on a new name	After only a few months of dating Muhammad Dakhlalla beginning in November of 2014, Jaelyn Young converted to Islam.
Seeking Out New Religious Authority: This field tracks the dates on which an individual began actively seeking out and abiding by doctrine created by more extreme figures of religious authority.	• Direct support via personal communication with extremist religious leaders • Attendance, reading, or listening to material produced by spiritual authority figures	In early 2013, Dzhokhar Tsarnaev began listening to sermons by Anwar al-Awlaki and reading radical publications that encourage jihad against the United States.
Peer Immersion: This field tracks the dates on which an individual began seeking out and associating with a group of like-minded individuals in equal or near-equal Jihadist social strata or with similar goals.	• Starting to spend all of their time with likeminded people and discontinuing social contact with old friends • Making a deliberate effort to enhance in-group cohesion; "brotherhood" • Engaging with others on social media and message boards to discuss Jihad or related topics	Before travelling to a Pakistani *Lashkar-e-Taiba* training camp, Yong Ki Kwon ordered and received matching jackets for him and his travelling companions to wear abroad.
Desire for Action: This field tracks the dates on which an individual expresses desire to take part in extremist action, speaking in broad terms that do not indicate formation of a concrete plan.	• Statements of desire to take part in extremist actions, including but not limited to: o Foreign fighting o Martyrdom or Domestic plot o Fundraising or Propaganda Creation	On January 1, 2012, Osmakac met "Amir Jones," an undercover FBI agent, in person and stated that he wanted to launch an attack "during the night." He said that "he wanted to get a hotel room, park the vehicle with the bomb in it at his target, leave the area, detonate the bomb in the car, and then 'go get the other stuff from the hotel.'"
Non-Violent Support: This field tracks the dates at which an individual lent tangible non-violent support to a terrorist or a terrorist group.	• Fundraising by credit fraud, theft, and other means. • Providing/smuggling material • Creating or translating propaganda, videos, or magazines for a terrorist group • Recruitment of foreign fighters	Mohammad Hassan Khalid solicited funds online to support Colleen LaRose in her attempt to travel and murder journalist Lars Vilks. Additionally, after she was questioned by the FBI, he sent emails to online forums asking that all of LaRose's posts get taken down.
Issues Threats: This field tracks the dates on which an individual issues violent threats against another individual or group of individuals, whether online or in-person.	• Threats against specific persons or groups • Soliciting others to commit acts of violence against specific persons or groups • Threats to large groups of people	On January 8th, 2015, Jalil Twitter, under the name @AnsaralUmmah8, to post "#KillAllKufar#KillAllKufar#KillAllKufar#KillAllKufar #KillAllKufar#KillAllKufar#KillAllKufar"
Steps Towards Violence: This field tracks the dates on which an individual began actively preparing to carry out a violent attack on behalf of an extremist organization or ideology.	• Acquiring weaponry or military gear • Selecting human targets for attacks • Conducting research on how to make bombs • Surveillance	Sami Osmakac met an undercover FBI agent in order to purchase weapons from him. Osmakac put down a $500 deposit on a fully automatic AK-47, six grenades, a handgun, and an explosive belt.
Joins Foreign Terrorist Organization: This field one of two things: 1) The dates in which an individual is invited or selected to participate in a training camp run by a foreign insurgency group, or 2) The dates in which an individual travels abroad in an attempt to join a foreign insurgency group.	• Arriving at foreign fighting locations • Registering membership in a terrorist organization • Attending a training camp • Receiving military training, such as shooting, bomb-making, etc. from a terrorist organization	A few weeks after Najibullah Zazi, Zarein Ahmedzay and Adis Medunjanin were denied entry to Afghanistan, they were invited by al-Qaeda operatives to a compound in Waziristan where they received necessary trainings to carry out an attack in the US. In an al-Qaeda compound in Waziristan, they learned how to make a bomb, using common items like nail polish remover, pipe cleaning chemicals, Christmas light and cooking oil.

Fig. 1. The 10 behavioral indicators (variables) and associated cues of violent extremist radicalization selected for automated information extraction.

support of violent action are relatively predictable, and that the individuals will display signature behaviors that are observed by bystanders [17].

The major project goals of the interdisciplinary team are to 1) produce a more reliable, empirically-tested dynamic radicalization risk assessment protocol, and 2) produce an associated technology based on that risk assessment protocol that can mine, monitor, and screen for the occurrence of indicators in large heterogeneous databases in order to provide early warnings of individuals or groups on behavioral trajectories toward extremist violence.

In this work, we undertake the critical task of employing natural language processing techniques and supervised machine learning models to classify textual data in analyst and investigator notes and reports for radicalization behavioral indicators. This effort will build towards an operational capability to assist analysts in rapidly mining law enforcement and intelligence databases, but also in the near term enables the more rapid coding of biographical radicalization profiles to augment a research database of violent extremists and their exhibited behavioral indicators.

II. Related Work

Natural language processing and machine learning algorithms in information extraction tasks have been used successfully in other domains such as understanding patient medical profiles in free-text clinical notes [1], [13], [26] and detecting cyberbullying [40], [41]. In the counterterrorism domain, nascent applications include detecting terrorist intentions [4] or determining a social media account's state of radicalization [27]. To our knowledge, there has not been heretofore an effort to classify text for the presence of distinct radicalization indicators in a manner that is consistent with a risk assessment protocol developed by terrorism experts [21], [25].

Moreover, we note that this information extraction effort is intended to support a growing body of work to develop a capability that assists analysts in rapidly mining law enforcement and intelligence databases for cues and risk indicators as well as dynamically assessing individualized violent extremism risk at scale through computational modeling [18]–[20], [30]. The underlying approach leverages advances in graph pattern matching over a heterogeneous knowledge graph in order to identify those on a trajectory of extremist violence according to a risk assessment protocol. Our work here supports the construction of such knowledge graphs by extracting structured information from law enforcement and intelligence reports.

Our effort to use computational modeling to assist analysts and case workers responsible for sorting diverse pools of people thought to present a risk to public safety is inspired in part by comparable efforts in public health management

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

approaches to preventive intervention. Notable examples in this line of investigation is the use of NLP techniques for identifying high risk child abuse cases [5].

The research involved in fulfilling these assumptions is often underestimated. The use of digital technology and NLP techniques for risk assessment rests on two premises: 1) the risk factors and overt behaviors associated with the specific pathology have to be known, and 2) the ability of algorithms to sort case loads from data. Expectations of what machine learning technologies can do should be tempered by the caveat that research on the psycho-sociology of violent political extremism is itself work-in-progress [37], and the best methods for harnessing machine learning techniques to track complex human behaviors are in the early stages of development.

III. Dataset

The dataset of documents for this effort is a part of Klausen's Western Jihadism Database (WJDB) [22], a collection that includes information on approximately 6,600 individual jihadists of Western origin or residence who have engaged in criminal terrorist action. All the data derives from public sources ranging from court records, government press releases, and autobiographical statements made by the terrorism offenders themselves on social media or in jihadist forums. This work utilizes a subset of WJDB and the analysis in [24]. In the current project, detailed forensic biographies were developed for 122 homegrown terrorism offenders who radicalized between 2001 and 2018 and committed terrorism-related crimes in this period. Coders were trained to read a variety of publicly accessible documentation for evidence of the 24 distinct behavioral indicators theorized to be associated with radicalization, and instructed to record the dates at which such behaviors were publicly observed. This analysis enabled the retrospective estimation of timelines for the radicalization trajectories.

The coders manually extracted a core set of sentences and sentence fragments used to create a labeled dataset to train machine learning models. In order to reduce initial problem space, the team reduced the number of indicators (or variables) from 24 to a critical set of 10, which are described in detail in Fig. 1. To date, the coders have annotated over 1273 sentences or paragraph samples extracted from over 158 different primary sources and secondary sources such as the Western Jihadism Project codebook and the terrorist profiles summarized in [24]. Because a significant number of sentences or paragraphs refer to two or more indicators, the coders in fact constructed a training dataset of over 1619 labeled sentences or paragraphs. Fig. 2 depicts the count of occurrences for each of the 10 radicalization indicators in training dataset.

IV. Methods to Extract Radicalization Behavioral Indicators from Text

The team investigated three methods to extract radicalization behavioral indicators from text documents (named entity recognition (NER), rule-based matching, and multi-label text classification) and utilized various open-source software libraries for implementation (Table I).

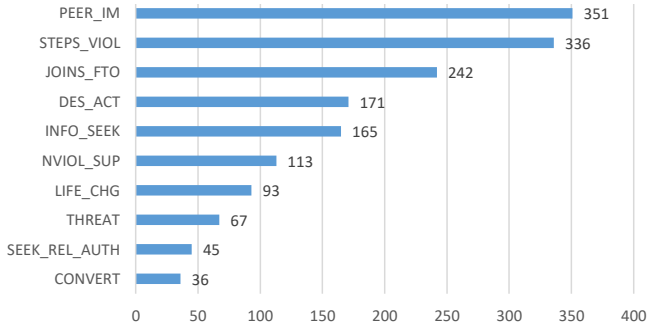

Fig. 2. The count of occurrences for each of the 10 radicalization indicators in training dataset.

A. NER Classifiers

NER classifiers seek to identify specific entities in a document based on an entity/object training dataset. Usually NLP libraries consist of default NER classifiers for entities such as person names, locations, date and time detection. In this specific application, experiments were conducted with keywords and keyword phrases serving as training data for each radicalization indicator (object) as well as a generic radicalization object to implement custom NER classifiers by highlighting phrases for coders. Major limitations with this method are 1) the inability to allow for certain keyword or keyword phrases to be associated with more than one indicator, and 2) the inability for NER classifiers to maintain the contextual information of the text. Such shortcomings led to critical errors in the entity detection and limited NER classifiers to the document coders.

TABLE I
Information Extraction Techniques

Technique	Software Library
Named Entity Recognition (NER)	spaCy [33], Stanford NLP [38]
Rule-Based Matching	spaCy
Multi-Label Text Classification	prodigy [32] and spaCy

B. Rule-Based Matching

Rule-based matching is an annotation method for finding specific patterns of tokens in text. It requires the manual production of an extensive set of rules based upon key word phrases from the training dataset. SpaCy provides a robust rule matching engine while allowing to implement any number of rules. The following rule example consists of 3 tokens.

eg : *[{'LEMMA': 'watch'}, {'POS': 'ADJ', 'OP': '*'}, {'LEMMA': 'video'}]*

The 'LEMMA' keyword provides the lemmatization and returns the base or dictionary form of a word. This enables the standardization of different tenses of verbs and plural words. The middle token expects an adjective and 'OP' key makes it

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

optional and allows an adjective zero or more times. The rule-based approach addresses the linguistic variation to a certain extent. However, the rule-based methods in spaCy were not viable for the identification of a numerous and complex set of linguistic markers because of the difficulty for the researchers to scale the creation of a correspondingly robust rule set.

C. Multi-Label Text Classification

The currently most promising approach for this particular information extraction problem is multi-label text classification, whereby text documents are potentially assigned one or more categories or labels. We implemented a workflow that used Prodigy, an text annotation tool which supports multi-label categorization of training data and seamlessly calls built-in recipes in spaCy for training convolutional neural network (CNN) models for text classification [6]. To best aid law enforcement and intelligence analysts with the appropriate amount of detail and classification granularity for specific behaviors, the researchers focused on performing *sentence-level* classification. For each sentence in a text document, the CNN model returns probability scores for the presence of each of the 10 indicator classes.

The training corpus, obtained from the political scientists on the research team, contains phrases, sentences or short paragraphs which can be manually labeled with up to 3 radicalization indicators. The corpus then undergoes standard pre-processing, which includes lemmatization, as well as the removal of special characters and stopwords.

V. Results

At present, the spaCy multi-label text classification neural network model trained on over 1273 labeled sentences and a 90%-10% training-testing split resulted in a model with 80% precision, 71% recall, 75% F-score, and 95% accuracy. Fig. 3 shows a sampling of the model output for scoring sentences from an document [2] with the appropriate radicalization indicator.

[4] A Virginia man who was allegedly attempting to travel to Syria to join Islamic State and a man accused of helping him have been arrested.
JOINS_FTO 0.996 INFO_SEEK 0.070 DES_ACT 0.060

[7] .Officials also arrested 25-year-old Mahmoud Amin Mohamed Elhassan, who they say drove Farrokh to Richmond.
PEER_IM 0.880 JOINS_FTO 0.055 DES_ACT 0.040

[15] .He later met with two other FBI informants he believed were people who could help him join Isis.
PEER_IM 0.887 JOINS_FTO 0.181 INFO_SEEK 0.046

[18] .He also allegedly said that he wanted to die a martyr but did ask if his wife and family could eventually join him in Syria.
JOINS_FTO 0.974 DES_ACT 0.940 LIFE_CHG 0.024

[21] .He asked the opinion of one of the FBI informants of his plan to buy a round-trip plane ticket and reserve a hotel room in Jordan, to minimize suspicion.
STEPS_VIOL 0.925 PEER_IM 0.090 SEEK_REL_AUTH 0.062

[26] When questioned by FBI agents after he dropped Farrokh off, Elhassan said Farrokh was traveling to California to attend a funeral and would be back in two weeks, court documents said.
STEPS_VIOL 0.669 JOINS_FTO 0.591 CONVERT 0.099

Fig. 3. Sample output of multi-label classification model.

It is clear that the model is able to detect sentences which concern behaviors of individuals attempting to join a foreign terrorist organization (sentences 4, 18) and even the potential presence of peer immersion when the offender reached out to others for help (sentence 15). However, with limited data, the model still lacks the desired generalization as uncovered by subsequent tests using unseen documents. The researchers also need to more rigorously analyze the probability scores for each indicator type because there is currently not an identifiable cut-off threshold by which they can confidently narrow the ground truth indicators present among the top 3 scores.

Moreover, the researchers observed that model has degraded precision due to false positives (mostly in the form of classifying sentences for radicalization indicators where there are none present) and therefore decided to implement a two-phased processing pipeline shown in Fig. 4. A separate model would screen sentences for the presence of any radicalization indicator cues, and then the second model would classify those sentences that passed the first phase for the actual type of radicalization indicator. The screening model itself needed some its own distinct training dataset- it included all 1273 sentences that were manually labeled as having an indicator present as well as over 575 sentences gathered during model testing that were deemed not relevant to radicalization. For this spaCy CNN screening model, we obtained 99% precision, 99% recall, 99% F-score,and 99% accuracy. To illustrate the two-phased pipeline, Fig. 5 shows 7 sentences of a Department of Justice public affairs statement which announcing the referral of charges on a suspected radicalized individual [10]. Fig. 5a shows the result of the indicator classification model of the sentences. It is apparent that the model will return classifications for sentences which are not relevant to radicalization behaviors (e.g., there are some statements made to make the public aware and describe the conduct of the investigation or the prosecution process). Fig. 5b shows the results of the screening model ('Y' means relevant to radicalization behaviors, and 'N' otherwise). Clearly, the screening model removes from consideration 5 sentences an analyst would not need to consider when attributing behaviors to a person of interest, and leaves only 2 sentences that provide important information.

VI. Human-in-the-Loop (HITL) Multi-Disciplinary Collaboration in Model Development

A critical factor in advancing this effort has been the multi-disciplinary collaboration between political scientists and computer scientists in mutually supporting roles. The HITL work process starts when the political scientists identify and extract text segments from the documents that represent behavioral cues to a variable used by the dynamic radicalization model to track growing extremism and mark the text cue as pertaining to a specific variable. The computer scientists add the labeled data to the training corpus and (re-)trains a model.

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

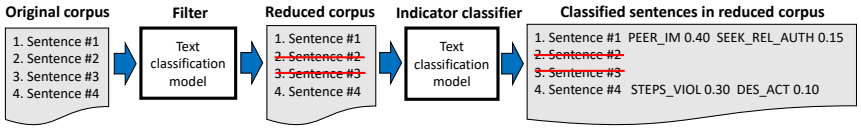

Fig. 4. Processing pipeline of the two-stage classification model. The first model is a binary classifier for relevant sentences to radicalization indicators in general, agnostic to any particular indicator. The second model is a multi-label classifier to identify the specific probabilities for each of the 10 indicators.

Output File

[1] FOR IMMEDIATE RELEASE.Friday, August 30, 2019.Individual Who Planned Attack in Queens Charged with Attempting to Provide Material Support to ISIS.Awais Chudhary Recorded Videos and Took Photos of the Flushing Bay Promenade And the Worlds Fair Marina in Preparation for an Attack.
STEPS_VIOL 0.826 NVIOL_SUP 0.204 INFO_SEEK 0.113

[2] A criminal complaint filed today in federal court in Brooklyn charged Awais Chudhary, 19, a naturalized U.S. citizen born in Pakistan, with attempting to provide material support to the Islamic State of Iraq and al-Sham (ISIS), a designated foreign terrorist organization.
NVIOL_SUP 0.732 JOINS_FTO 0.605 INFO_SEEK 0.042

[3] Chudhary was arrested yesterday, and made his initial appearance this afternoon before United States Magistrate Judge James Orenstein.
LIFE_CHG 0.519 PEER_IM 0.103 INFO_SEEK 0.081

[4] The defendant allegedly planned to conduct a deadly attack in New York on behalf of ISIS, stated Assistant Attorney General for National Security John C. Demers.
STEPS_VIOL 0.986 DES_ACT 0.036 PEER_IM 0.027

[5] The National Security Division, working with our partners, will remain vigilant in our efforts to identify, disrupt, and hold accountable those who would conduct a terrorist attack on our soil.
STEPS_VIOL 0.874 NVIOL_SUP 0.135 CONVERT 0.041

[6] I want to thank the agents, analysts, and prosecutors who are responsible for this case and prevented this defendant from carrying out his deadly plans.
STEPS_VIOL 0.774 JOINS_FTO 0.278 DES_ACT 0.124

[7] As alleged, Awais Chudhary planned to kill innocent civilians on behalf of ISIS and record the bloodshed in the hope of inspiring others to commit attacks, stated United States Attorney Donoghue.
DES_ACT 0.493 THREAT 0.413 STEPS_VIOL 0.063

(a)

Corrected Output File

[1] FOR IMMEDIATE RELEASE.Friday, August 30, 2019.Individual Who Planned Attack in Queens Charged with Attempting to Provide Material Support to ISIS.Awais Chudhary Recorded Videos and Took Photos of the Flushing Bay Promenade And the Worlds Fair Marina in Preparation for an Attack.
Y 0.999 N 4.54e-05 STEPS_VIOL 0.826 NVIOL_SUP 0.204 INFO_SEEK 0.113

[2] ~~A criminal complaint filed today in federal court in Brooklyn charged Awais Chudhary, 19, a naturalized U.S. citizen born in Pakistan, with attempting to provide material support to the Islamic State of Iraq and al-Sham (ISIS), a designated foreign terrorist organization.~~
N 0.999 Y 0.001

[3] ~~Chudhary was arrested yesterday, and made his initial appearance this afternoon before United States Magistrate Judge James Orenstein.~~
N 0.929 Y 0.078

[4] ~~The defendant allegedly planned to conduct a deadly attack in New York on behalf of ISIS, stated Assistant Attorney General for National Security John C. Demers.~~
N 0.790 Y 0.210

[5] ~~The National Security Division, working with our partners, will remain vigilant in our efforts to identify, disrupt, and hold accountable those who would conduct a terrorist attack on our soil.~~
N 0.696 Y 0.315

[6] ~~I want to thank the agents, analysts, and prosecutors who are responsible for this case and prevented this defendant from carrying out his deadly plans.~~
N 0.900 Y 0.096

[7] As alleged, Awais Chudhary planned to kill innocent civilians on behalf of ISIS and record the bloodshed in the hope of inspiring others to commit attacks, stated United States Attorney Donoghue.
Y 0.999 N 3.84e-04 DES_ACT 0.493 THREAT 0.413 STEPS_VIOL 0.063

(b)

Fig. 5. (a) Output file of sample sentences scored using the multi-label classification model without screening, and (b) a corrected output file from the two-stage pipeline when the screening model filters for relevance, and then returns only the classification scores for those sentences which provide radicalization indicator information.

Term	Description
Relevance	Refers to the relevance of the sentence to radicalization trajectory. Relevance can be assessed as "Not relevant to radicalization trajectory," "Not relevant to ten variables, but otherwise indicative of extremism," "Subject Confusion," and "Relevant."
Ground Truth	Refers to whether the variable provided is accurate. A ground truth of "TRUE" means the variable is present in the sentence. A ground truth of "FALSE" means the variable is not present in the sentence
Assessment	Evaluates the score a variable is given. Assessment may change depending on the threshold you are testing on. Assessment can be correct (annotated as "Correct"), a false negative (annotated as "FN"), or false positive (annotated as "FP").
Relative Weight	Refers to the order of the variables. Relative Weight can be "Correct," "N/A," or "Incorrect."
Relatedness	Evaluates the applicability of the words in the sentence to radicalization. It is important to note that relatedness does not have to do with the subject of the sentence, but rather the keywords within the sentence. Relatedness can fall under three categories: related (annotated as "Related"), helpful in context (annotated as "Context"), and not related (annotated as "Not Related").

Fig. 6. Annotation/validation terms and descriptions.

The output file is returned to the political scientists listing text segments extracted by the machine learning model from the test text corpus with the top predicted radicalization indicator variables by probability score. It became quickly apparent that the different machine learning models could be trained: one for expediting the task of coders to sift through textual data to find related sentences to scale the annotation of labeled training data; and another for eventually replacing the coders in multi-label classification task. In order to address both aims, the political scientists codified in Fig. 6 the 5 dimensions necessary for validating model testing outputs: relevance, ground truth, assessment, relative weight, relatedness. The coders review the output file to assess the ground truths of the prediction (true/false), the relevance and relatedness of the particular text segments identified by the machine learning model to extremist radicalization, and also provide an assessment on the relative probabilities returned for each sentence. See Fig. 7 shows how model validation feedback is captured.

A few examples illustrate the complexity of provided machine-appropriate model validation feedback. In Fig. 7, line 4 reads: "Solano told the CHS [confidential human source, which typically is a co-conspirator who turns witness for the prosecution] that he wanted to join ISIS." The machine learning model predicted that 'Joins Foreign Terrorist Organization (FTO)' variable is present in the sentence with a score of 0.323, but the HITL considers the ground truth of this assessment "false" because the subject merely said he wanted to but did not actually join the FTO (ISIS). The assessment, meanwhile, is deemed "correct" because the model assigned a low probability value (0.323) to the statement below a certain threshold. The HITL also noted that the model missed how the subject expressed a 'desire for action,' an important precursory variable in a sociological model for radicalization trajectories. Nonetheless, the text segment was relevant to the radicalization model, but in connection with a different variable. To improve the model's performance, the HITL-reviewed test text documents are then incorporated in the training corpus with the correct labels, and the model is subsequently re-trained and re-tested against new test text documents.

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

		1				2					
ID	Sentence	Predicted Label	Score	Assessment	Ground Truth	Predicted Label	Score	Assessment	Ground Truth	Relative Weight	Relevance
1	Solano pleaded guilty yesterday to one count of attempting to provide material support to ISIS, in violation of Title 18, United States Code, Section 2339B(a)(1).	NVIOL_SUP	0.500	Correct	FALSE	JOINS_FTO	0.176	Correct	FALSE	N/A	Not Relevant to ten variables, but otherwise indicative of extremism
2	Solano faces a statutory maximum sentence of twenty years' imprisonment and a $250,000 fine.	THREAT	0.293	Correct	FALSE	STEPS_VIOL	0.271	Correct	FALSE	N/A	Not Relevant to radicalization trajectory
3	According to the stipulated factual basis filed with the Court, in early 2017, Solano told an individual, who later became a Confidential Human Source ("CHS") for the government, that he was upset with the United States and wanted to conduct an attack in Miami.	STEPS_VIOL	0.351	Correct	FALSE	SEEK_REL_AUTH	0.142	Correct	FALSE	Incorrect: Desire for Action missed and is most relevant	Relevant
4	Later, Solano told this CHS that he wanted to join ISIS.	JOINS_FTO	0.323	Correct	FALSE	NVIOL_SUP	0.119	Correct	FALSE	Incorrect: Desire for Action missed and is most relevant	Relevant
5	Solano planned to place and detonate an explosive device in a crowded area of a popular Miami mall.	STEPS_VIOL	0.372	FN	TRUE	JOINS_FTO	0.279	Correct	FALSE	Correct	Relevant
6	Solano discussed his plot with the CHS and two undercover FBI employees.	PEER_IM	0.612	FN	TRUE	STEPS_VIOL	0.436	Correct	FALSE	Correct	Relevant
7	According to the complaint, Solano provided three videos to the CHS, in which Solano makes pro-ISIS statements and expresses anti-U.S. sentiments.	NVIOL_SUP	0.349	FN	TRUE	SEEK_REL_AUTH	0.108	Correct	FALSE	Correct	Relevant
8	Just prior to his arrest, Solano took possession of what he believed was an explosive device, took steps to arm it, and walked toward a mall entrance in order to carry out his attack.	STEPS_VIOL	0.960	Correct	TRUE	JOINS_FTO	0.027	Correct	FALSE	Correct	Relevant

Fig. 7. Model output validation worksheet. The machine learning model produces indicator classifications for each sentence and the top 2 scoring classifications are presented to the subject matter experts (coders). The coders then use the annotation/validation construction and complete sections in orange as feedback to the model developers.

The political scientists have observed some systemic short comings with the text classification models. Subject confusion is a perennial problem. For example, the sentence "On [date], subject X drove associate Y to LaGuardia Airport, after which time X believed Y had boarded a flight and traveled to Turkey"[1] received a near perfect probability score for joining an FTO but was in fact false. The subject X did not fly out that day; someone else did.

Complex behaviors, linguistic winks, and slang present a significant challenges to machine learning modeling as well. For example, an individual described in [39] had tried to join a designated terrorist organization in Libya and used a text messaging app to communicate with someone he thought was a co-believer: "Getting rid of the device now...fo real. Gonna eat the Sim Card. Have a good day" [39]. The meaning of "eating a Sim card" with the intention of avoiding law enforcement detection is likely to defy many trained machine learning models with detecting a cue for Steps Towards Violence (STEPS_VIOL). Similar machine-logic defying expression include "get down with this ISIS shit" (Derrick Thompson [7], part of a text segment coded as Desire for Action (DES_ACT)); a reference to having "wiped clean" the internet history in preparation for doing something (Isse Aweis Mohamud [11], coded as Steps Toward Violence (STEPS_VIOL)); and the slang used by Nelash Mohammed Das who on his Twitter account expressed his exasperation that being on Twitter was lame: "Sitting on Twitter is not enough I envy seeing brothers getting shahada [martyrdom] n slaying kuffar [infidels] while im at home not getting any action"(Nicholas Rovinski [12], which was coded as Desire for Action (DES_ACT)).

The examples make the subject look pitifully foolish (which some of them are). Extremist ideology is rule-bound and the performative scripts associated with the ideology are often incomprehensible to outsiders. The terrorist radicalization process is symbolically governed by these scripts, encoded by the ideology, and advanced by extremist recruiters. The scripts mold the individual to conform with a new social order and govern the behaviors of would-be terrorists. The prevalence of rule-bound behaviors is the reason why these behaviors may be captured by a machine-learning model.

[1]Turkey was a common entry point for traveling on to join ISIS.

VII. Conclusion & Future Work

The classification of textual data in analyst and investigator notes and reports for radicalization behavioral indicators is a significant challenge but promises to greatly increase the ability of law enforcement and intelligence agencies to investigate and intercept those on suspected radicalization pathways to violent extremism. This work demonstrates progress in the application of NLP techniques and machine learning in addressing this challenge.

In future work, we intend to improve model accuracy by incorporating transfer learning with a new NLP library called the Bidirectional Encoder Representations from Transformers (BERT) [8]. BERT is a generic NLP module, pre-trained on large text corpus and customizable to develop various NLP techniques. Moreover, we also intend to carry-on and link this information extraction work to the larger effort of automatically generating structured knowledge graphs consisting of persons of interest and their timestamped behavioral indicators. The application of advanced graph pattern matching technologies such as INSiGHT [20] and the PINGS graph database library for Neo4j [30] over such knowledge graphs will enable analysts to rapidly mine law enforcement and intelligence databases for cues and risk indicators and dynamically assess violent extremism risk at scale.

References

[1] Amazon Web Services, "Amazon Comprehend Medical," 6 May 2019. [Online]. Available: https://aws.amazon.com/comprehend/medical/. [Accessed 6 May 2019].

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

[2] Associated Press, "Virginia man arrested after alleged attempt to join ISIS in Syria" [Online]. Available https://www.theguardian.com/us-news/2016/jan/16/virginia-isis-suspect-arrest-joseph-hassan-farrokh. [Accessed September 16, 2019].

[3] R. Borum, "Radicalization into Violent Extremism I: A Review of Social Science Theories," *Journal of Strategic Security*, Vol 4, No. 4, 2012.

[4] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Martenson and P. Svenson, "Harvesting and analysis of weak signals for detecting lone wolf terrorists," Security Informatics, vol. 2, pp. 11-26, 2013.

[5] B Castellani, F. Griffiths, R. Rajaram, J. Gunn, "Exploring comorbid depression and physical health trajectories: A casebased computational modelling approach" *Journal of Evaluation in Clinical Practice* vol. 24 no. 8, pp. 1293-1309.

[6] D. Campion, "Text Classification: Be lazy, use Prodigy" [Online]. Available https://medium.com/@david.campion/text-classification-be-lazy-use-prodigy-b0f9d00e9495. [Accessed August 1, 2018].

[7] M. Cassidy, "Phoenix terror suspect pleads guilty" [Online]. Available https://www.azcentral.com/story/news/local/phoenix/2018/01/05/phoenix-terror-suspect-pleads-guilty/1008964001/. [Accessed September 18, 2019].

[8] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,", CoRR, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805.

[9] Director of National Intelligence, "Homegrown Violent Extremist Mobilization Indicators," 2019. [Online]. Available https://www.dni.gov/files/NCTC/documents/news_documents/NCTC-FBI-DHS-HVE-Mobilization-Indicators-Booklet-2019.pdf.[Accessed September 16, 2019].

[10] DOJ Public Release, "Individual Who Planned Attack in Queens Charged with Attempting to Provide Material Support to ISIS" [Online]. Available https://www.justice.gov/opa/pr/individual-who-planned-attack-queens-charged-attempting-provide-material-support-isis. [Accessed September 16, 2019].

[11] DOJ Public Release, "KC Man Sentenced for Passport Fraud," [Online]. Available https://www.justice.gov/usao-wdmo/pr/kc-man-sentenced-passport-fraud. [Accessed September 15, 2019].

[12] DOJ Public Release, "Rhode Island Man Sentenced for Conspiring to Commit Acts of Terrorism to Support ISIS," [Online]. Available https://www.justice.gov/usao-ma/pr/rhode-island-man-sentenced-conspiring-commit-acts-terrorism-support-isis. [Accessed September 15, 2019].

[13] C. Friedman, T. C. Rindflesch and M. Corn, "Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine," *J. of Biomedical Informatics*, vol. 46, pp. 765-773, 2013.

[14] B. Hoffman, E. Meese III and T. J. Roemer, "The FBI: Protecting the Homeland in the 21st Century," Federal Bureau of Investigation, Washington, D.C., 2015.

[15] B. Hoffman, "ISIS' Shifting Focus," 23 April 2019. [Online]. Available: https://www.thecipherbrief.com/column_article/isis-shifting-focus. [Accessed 30 April 2019].

[16] Homeland Security Committee, "Terror Gone Viral- Overview of the 243 ISIS-linked incidents targeting the West," US House of Representatives, Washington, D.C., 2018.

[17] J. Horgan, "From Profiles to Pathways and Roots to Routes: Perspectives from Psychology on Radicalization into Terrorism," *The Annals of the American Academy of Political and Social Science*, Vol. 618, July 2008, pp. 80-94.

[18] B. Hung, A. Jayasumana, and V. Bandara, "INSiGHT: A System to Detect Violent Extremist Radicalization Trajectories in Dynamic Graphs," *Data & Knowledge Engineering*, Vol. 118, pp. 52-70, 2018.

[19] B. Hung, A. Jayasumana, and V. Bandara, "INSiGHT: Detecting the Radicalization Trajectories of Homegrown Violent Extremists with Dynamic Graph Pattern Matching," *IEEE Homeland Security Technologies (HST) Symposium 2017 Conference Proceedings*, 2017.

[20] B. Hung, A. Jayasumana, and V. Bandara, "Finding Emergent Patterns of Behaviors in Dynamic Heterogeneous Social Networks," to appear in *IEEE Transactions on Computational Social Systems*, 2019.

[21] J. Klausen, "A Behavioral Study of the Radicalization Trajectories of American "Homegrown" Al Qaeda-Inspired Terrorist Offenders, 2001-2015," Inter-University Consortium for Political and Social Research, Ann Arbor, MI, 15 December 2016. [Online]. Available: http://doi.org/10.3886/ICPSR36452.v1.

[22] J. Klausen, "Jytte Klausen's Western Jihadism Project: Data Collection." [Online]. Available https://www.brandeis.edu/klausen-jihadism/data-collection.html. [Accessed September 15, 2019].

[23] J. Klausen, "Why Jihadist Attacks Have Declined in Europe," *Foreign Affairs*, 19 December 2018.

[24] J. Klausen, S. Campion, N. Needle, G. Nguyen and R. Libretti, "Toward a Behavioral Model of 'Homegrown' Radicalization Trajectories," *Studies in Conflict and Terrorism*, vol. 39, no. 1, pp. 67-83, 2015.

[25] J. Klausen, R. Libretti, B. Hung, and A. Jayasumana, "Radicalization Trajectories: An Evidence-Based Computational Approach to Dynamic Risk Assessment of "Homegrown" *Jihadists*", *Studies in Conflict and Terrorism*, 2018.

[26] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug and T. Botsis, "Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review," *J. Biomedical Informatics*, vol. 73, pp. 14-29, 2017.

[27] R. Lara-Cabrera, A. G. Pardo, K. Benouaret, N. Faci, D. Benslimane and D. Camacho, "Measuring the Radicalization Risk in Social Networks," IEEE Access, vol. 5, pp. 10892-10900, 2017.

[28] MITRE, "Person-Centric Identity Management- Rapidly Assimilating Data," January 2017. [Online]. Available: https://www.mitre.org/publications/technical-papers/person-centric-identity-management-rapidly-assimilating-data-about-a. [Accessed 1 March 2017].

[29] "Maryland resident charged with supporting ISIS, allegedly plotted to kill member of U.S. military," *The Washington Post*, October 3, 2016, [Online]. Available http://wapo.st/2d9AbJg?tid=ss_tw. [Accessed September 16, 2019].

[30] S.R. Muramudalige, B. Hung, A. Jayasumana, and I. Ray, "Investigative Graph Search using Graph Databases," *Graph Computing* (GC 2019), Laguna Hills, CA, 2019.

[31] Neo4j, "Neo4j: The Leading Graph Database." [Online]. Available https://neo4j.com/. [Accessed September 16, 2019].

[32] Explosion, "Prodigy." [Online]. Available https://prodi.gy/. [Accessed September 15, 2019].

[33] Explosion, "SpaCy: Industrial-Strength Natural Language Processing in Python." [Online]. Available https://spacy.io/. [Accessed September 15, 2019].

[34] B. Schuurman, L. Lindekilde, S. Malthaner, F. O'Connor, P. Gill, N. Bouhana, "End of the Lone Wolf: The Typology that Should Not Have Been," *Studies of Conflict and Terrorism*, Vol. 42, No. 8, 2019, pp. 771-778.

[35] J. Silver, J. Horgan, P. Gill, "Foreshadowing Targeted Violence: Assessing Leakage of Intent by Public Mass Murderers," *Aggression & Violent Behavior*, Vol. 38, 2018, pp. 94-100.

[36] J. Silver, A. Simons, S. Craun, "A Study of the Pre-Attack Behaviors of Active Shooters in the United States Between 2000 and 2013," Federal Bureau of Investigation, U.S. Department of Justice, Washington, DC, 2018. [Online]. Available https://www.fbi.gov/file-repository/pre-attack-behaviors-of-active-shooters-in-us-2000-2013.pdf [Accessed September 16, 2019].

[37] A. Smith, "How radicalization to terrorism occurs in the United States: What research sponsored by the National Institute of Justice tells us; U.S Department of Justice, Office of Justice Programs, National Institute of Justice, 2018. [Online] Available https://www.ncjrs.gov/pdffiles1/nij/250171.pdf. [Accessed September 19, 2019].

[38] Stanford NLP Group, "Stanford RegexNER." [Online]. Available https://nlp.stanford.edu/software/regexner.html. [Accessed September 15, 2019].

[39] "Police officer stays silent in landmark terrorism trial," *The Washington Post*, December 16, 2017, [Online]. Available http://wapo.st/2zhf3hD?tid=ss_tw. [Accessed September 16, 2019].

[40] J.M. Xu, K.S. Jun, X. Zhu and A. Bellmore, "Learning from Bullying Traces in Social Media," in *Conference of the North American Chapter of the Association for Computational Linguistics*, Montreal, Canada, 2012.

[41] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth and E. Dillon, "Cyberbullying detection with pronunciation based convolutional neural network," in *IEEE International Conference on Machine Learning and Applications*, Anaheim, CA, 2016.

978-1-7281-5092-5/19/$31.00 ©2019 IEEE