Detecting Social Polarization and Radicalization

Pir Abdul Rasool Qureshi, Nasrullah Memon, Uffe Kock Wiil and Panagiotis Karampelas

***Abstract*—This paper proposes a novel system to detect social polarization and to estimate the chances of violent radicalization associated with it. The required processes for such a system are indicated; it is also analyzed how existing technologies can be integrated into the proposed system to fulfill the objectives. We propose a scalable design to enable integration and extension of future state of the art technologies into the system.**

***Index Terms*—DETECT, radicalization; social polarization, violence**

I. Introduction

The objective of the DETECT system is to detect trends and assess risks of social polarization and violent radicalization based on automated processing of online open source information combined with human intelligence. It undertakes modeling and analysis of the total causal environment behind social polarization, radicalization, and violent extremism. Accurate modeling of real time processed information will lead to early detection of polarization, allowing time for destabilization strategies to be devised in case the groups identified proceed to the path of extremism. Accurate prediction of violent events will enable the effective deployment of the forces of law to minimize or even remove the serious negative outcome of such events.

The DETECT system will mainly populate its model by deep processing of online open source information. Human intelligence will also be integrated with the online data in order to maximize accuracy and processing depth and to reduce the risk of missing some significant pieces of data.

The complexity of the proposed system is inherent since it requires understanding of human emotions, favor and dislike, reactions and behavior – and an estimation of the depth and extent of these. The presence of different elements, origins, cultures, religions, and ideologies further complicates the modeling of social situations. Our aim is to discover radical individuals and groups, to predict the emergence of radicalization, and particularly to predict violent actions resulting from such extremism. This requires discovery of indicators predicting the reactions of these humans to outside events such as new government policies, new articles in the media, or particular legal judgments. Predicting human reactions is always a challenge.

We approach the problem empirically, analyzing past incidents and their underlying causes, environments, and deriving indicators from the results to predict anomalies as shown in Figure 1.

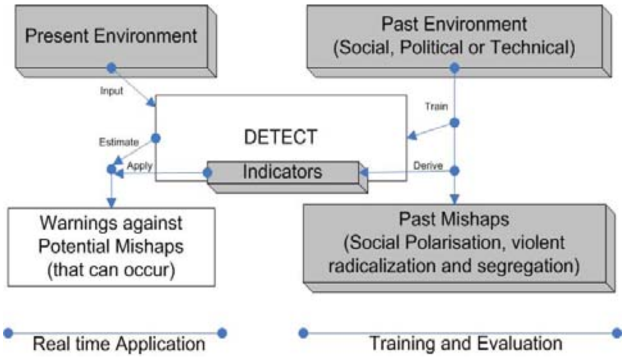
Present Environment
Past Environment (Social, Political or Technical)
Input
Train
Estimate
DETECT
Apply
Indicators
Derive
Warnings against Potential Mishaps (that can occur)
Past Mishaps (Social Polarisation, violent radicalization and segregation)
Real time Application
Training and Evaluation

Figure 1. The emperical approach towards the problem.

The DETECT system will be trained on such historical data, finding indicators which can be sought on the internet, predicting risk levels in real time. The approach involves collecting data sufficient to allow the system to build social network graphs for polarized groups and apply social network analysis to identify the key players, the presence of different communities within a network, and to find their preferences, skills, and the technologies they may have available. This will assist the relevant authorities at national and international levels in building effective strategies to curb, stop, or prevent social processes or individuals that encourage violent extremism or terrorism.

The sequence of processes that the system will undergo is as follows:

1. Acquire online data and normalize it;
2. Process the data in various ways to extract maximum information content;
3. Investigate the processed information for presence of identified indicators;
4. Construct social network graphs and uncover graph internals like key players, skills involved, or technologies used.

The following have been identified as key objectives of the DETECT system:

1. The system must be capable of collecting, integrating, and consolidating information obtained from open sources like blogs, discussion forums, published reports, news articles, and other classical intelligence sources (such as human intelligence).
2. The system must be able to investigate the most appropriate formal representation of the consolidated information. The formalization must be flexible, yet understandable to the machine, so that an environment can be extracted and mapped onto the rapidly changing behaviours of different communities within the society. By the term "environment" we mean the set of different actions taken by a particular group. For example, these may be the design and implementation of government

Manuscript received March 22, 2011.

policies or the set of behaviours of a particular social group.

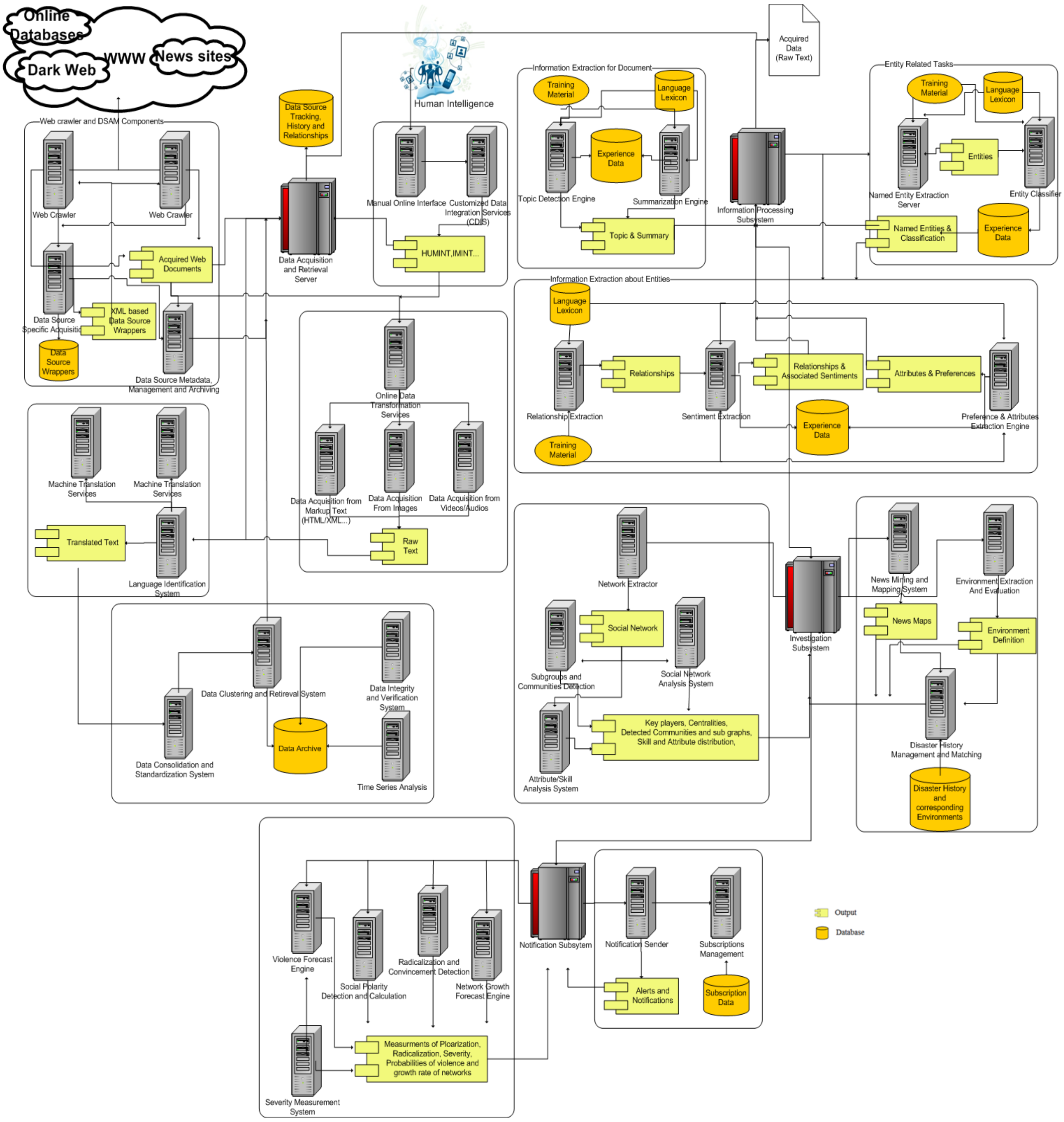
Online Databases
Dark Web
WWW
News sites
Web crawler and DSAM Components
Web Crawler
Web Crawler
Acquired Web Documents
Data Source Specific Acquisitio
XML based Data Source Wrappers
Data Source Wrappers
Data Source Metadata, Management and Archiving
Data Source Tracking, History and Relationships
Data Acquisition and Retrieval Server
Human Intelligence
Manual Online Interface
Customized Data Integration Services (CDIS)
HUMINT,IMINT...
Online Data Transformation Services
Data Acquisition from Markup Text (HTML/XML...)
Data Acquisition From Images
Data Acquisition from Videos/Audios
Raw Text
Machine Translation Services
Machine Translation Services
Translated Text
Language Identification System
Data Clustering and Retrieval System
Data Integrity and Verification System
Data Consolidation and Standardization System
Data Archive
Time Series Analysis
Acquired Data (Raw Text)
Information Extraction for Document
Training Material
Language Lexicon
Experience Data
Topic Detection Engine
Summarization Engine
Topic & Summary
Information Processing Subsystem
Entity Related Tasks
Training Material
Language Lexicon
Entities
Named Entity Extraction Server
Entity Classifier
Named Entities & Classification
Experience Data
Information Extraction about Entities
Language Lexicon
Relationships
Relationships & Associated Sentiments
Attributes & Preferences
Relationship Extraction
Sentiment Extraction
Preference & Attributes Extraction Engine
Experience Data
Training Material
Network Extractor
Social Network
Subgroups and Communities Detection
Social Network Analysis System
Key players, Centralities, Detected Communities and sub graphs, Skill and Attribute distribution,
Attribute/Skill Analysis System
Investigation Subsystem
News Mining and Mapping System
Environment Extraction And Evaluation
News Maps
Environment Definition
Disaster History Management and Matching
Disaster History and corresponding Environments
Violence Forecast Engine
Social Polarity Detection and Calculation
Radicalization and Convincement Detection
Network Growth Forecast Engine
Measurments of Ploarization, Radicalization, Severity, Probabilities of violence and growth rate of networks
Severity Measurement System
Notification Subsystem
Notification Sender
Subscriptions Management
Alerts and Notifications
Subscription Data
Output
Database

Figure 2. System architecture of the DETECT system

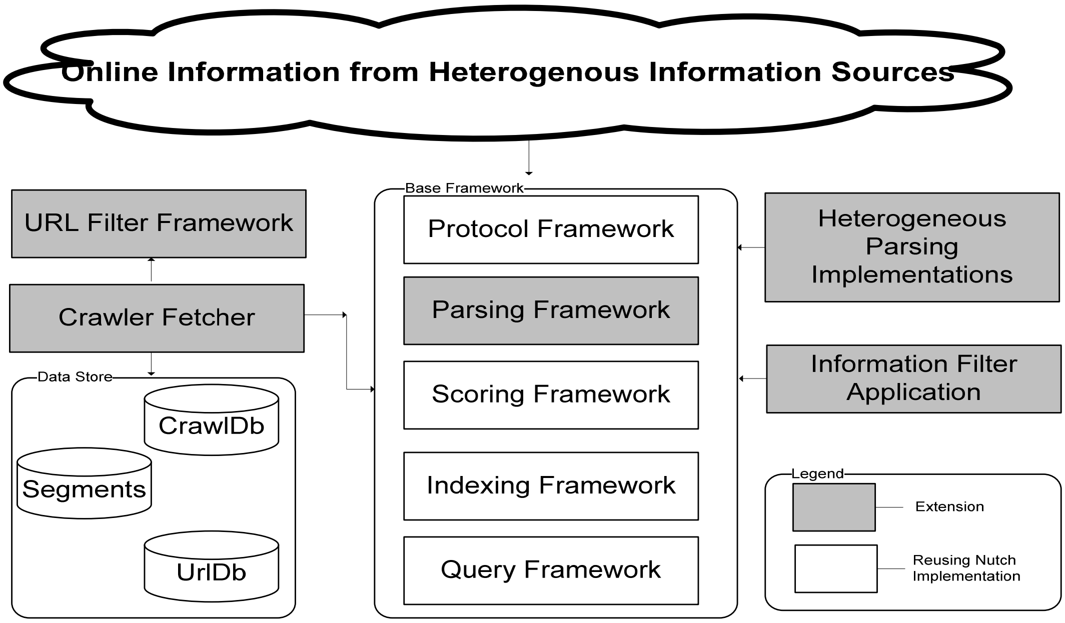

Figure 3: Architectural overview of Nutch based data acquisition and retrieval subsystem.

3. The system must be capable of carrying out social network analysis (SNA) and applying advanced technologies of investigative data mining (IDM) to detect social networks and determine their strength. It is crucial to determine the extent of polarization and the key nodes in networks.
4. The system must be equipped with statistical inference and forecast engines capable of modeling historical data: past environments and their impact on different historical social groups, and results such as violence, radicalization or social unrest. Such a system can then be used to extrapolate the polarization into the future, given the contemporary environment(s) of the group being investigated.
5. The system must have capability to notify subscribed end users against such situations, if they occur.

The next section describes the solution proposed to achieve the aforementioned objectives. It also decomposes the system into subsystem and streamlines the key responsibilities of each subsystem. Section III describes the potential candidate technological solutions available, whereas Section IV identifies the necessary extensions in state of the art technologies to develop the desired system. Section V documents the technical and social impact of the proposed work and Section VI concludes the paper.

II. Proposed Solution

The proposed system is decomposed into four subsystems. The explanation and functionality of each of the subsystems is shown in Figure 2 and described below:

A. Data Acquisition and Retrieval Subsystem

This subsystem will be capable of working with almost all sorts of user generated media on the World Wide Web, and will accommodate very flexible data integration capabilities. The challenge is not only to integrate open source information gathered from resources like news websites, blogs, discussion forums, etc., but also to provide a flexible channel of incorporating manual information provided by analysts. The other challenge is to incorporate information from heterogeneous information sources providing information in different possible propriety formats such as Microsoft Word (.doc), Microsoft Power Point (.ppt), Portable Document Format (PDF), etc.

A formal representation and protocol to integrate open source information with information gathered from traditional intelligence methods is needed for the effective use of the consolidated information. To achieve this, the open source data acquisition system will take care of the retrieval of open source data available online. The subsystem uses an implementation of a web crawler for retrieving such information. There are many open source crawler implementations available like Apache Nutch [1, 2, 3, 4, 5].

Nutch is basically an open source Java implementation of a search engine with the capabilities of crawling. It provides all the needed tools to run a web search engine [4]. Nutch is built on top of Lucene [35, 36], which is an API for text indexing and searching that will be explained in the following section. Lucene provides a powerful indexing and search library which may be used as a base for online search engines, however on its own the library does not include any form of Web crawling nor parsing abilities. These features are necessary in order to create the software that will accomplish the goals of the presented system. Nutch was developed by creator of Lucene and was developed with the purpose to provide an open source search engine capable of indexing the World Wide Web as effectively as commercial search engines [36]. Anyone can download, modify, and use the search engine. The project started under SourceForge.net but was moved to Apache Software Foundation as a subproject of Lucene. Lucene is built with a plugin based architecture, i.e., the architecture accommodates unforeseen future tools and developments via implementation of plugins corresponding to that development. Inherently, Nutch is also scalable. Since it fulfills most of the requirements for the subsystem, we have extended Nutch for the implementation of the data acquisition and retrieval subsystem. The following are the main characteristics of Nutch, which influenced that decision:

- Integrated support for crawling (fetching Web content).
- Allows plugin development to support different protocols (like HTTP and FTP) and language identifiers.
- Allows plugin development to implement new parsers against to support propriety formats and file types.
- Integrated search engine in the shape of Lucene to fulfill needed information retrieval functionality.
- Indexing and clustering of results.

This allowed us to integrate support for different propriety formats. However, to integrate information filtering mechanisms, it does not provide any foundations. To cope with the situation, we have integrated information filtering facilities added to the content parsers. Another reason behind this is that information filtering is format dependent. For example, HTML content is more likely to contain navigational, commercials, and hypertext impurities as compared to PDF or Word documents. Figure 3 shows the modified architecture with the presented modifications represented by shaded boxes.

However, for frequent and popular web sources containing user generated information like Facebook, it accommodates specific data source wrappers determining the particular areas or DOM components in the web pages associated with that data source to locate particular information or metadata. The metadata of a page contains, but is not limited to, the hyperlinks it contains, the videos it references, any reference to news items or environmental variables, author name(s), when was last updated, created, number of posts, frequency of posts etc. Metadata about any page is archived and managed by the Metadata Management and Archiving component.

To incorporate the data collected from the analysts, the system provides extensible functionality of the Customized Data Integration Service (CDIS). Since the data provided by the analysts will be associated with less noise and high confidentiality, such data will be essential in determining how trustworthy any particular open source information source is. Since the open source information and the information collected via CDIS are different in nature, a protocol must be defined for their integration. We have shown the possibility of designing such protocol in Figure 2 (CDIS Protocol). The service will be made available online to enable an investigator working in the field to use it.

The subsystem is equipped with a group of components which are used for comparative analysis among different open source information sources or different posts on the same data source to identify the conflicts and level of support for each data source pair. As this group of components is comparing or mapping a particular data source over another, it is called Data Source Analysis and Mappings (DSAM). The DSAM also integrates some specialized routines to accommodate time series analysis by determining how the opinions of the users are changing over to time. Usually, this type of technique is used to express how a particular news story is progressing over to time. However, we intend to use it with all of the available data sources because change in the user opinion may also occur over time. If any information provided by the user posting or any opinion which is conflicting with the reality, then it is communicated to the system with the use of the Data Integrity and Verification System. Similar types of conflicts and similarities between the information provided by the different data sources is detected and tracked by the Cross Data Source Relationship Detection System and Data Source Tracking and Management System, respectively.

B. Information Processing Subsystem

Once the data sources have been identified and information has been gathered, it is now the time to process it. Information processing typically deals with the extraction of different named entities and relationships among them.

The component Named Entity Recognizer (NER) detects named entities such as organization, person, location, etc. There are quite a few existing NER available such as Stanford NER [7], Illinois NER [9], and Lingpipe NER [8]. We may use any of them depending upon their performance. It is however worthy to specify that according to our initial experiments none of them is performing satisfactory with English text containing Arabic names. The NER is supported by the Entity Classification System. This classification further categorizes the entity into more specialized subclasses. For example a person can be further specified as terrorist, victim, reporter, witness, etc. and location can be further classified as building, target location, place of occurrence, etc.

The Relation Extraction System detects the relationships among the named entities found. Existing relationship extraction systems are also available such as the Self Supervised Relation Extraction System [16] and the KnowItAll System [31]. Both systems are capable of finding simple and obvious relationships among the entities. However, the performance and accuracy of these systems for their use in a more sophisticated system like DETECT is yet to be investigated. Furthermore, for such application we also have to find the polarity of the extracted relationships to identify the conflicting or supporting relationships. The Sentiment Analysis System (SAS) detects the polarities and hidden sentiments in relationships. SAS is responsible for detecting the sentiments among the entities within the same document or across multiple documents.

The information processing subsystem also extracts the preferences of users, if available, within the posted text and suggests a suitable topic and summary of the contents with the help of state of the art technologies.

The information processing subsystem also deals with the consolidation of acquired information from multiple sources and even with traditional intelligence documents like stolen facts, secret reports, diplomatic statements, etc. This enables the effective merging of open source intelligence with the traditional intelligence sources such as Human Intelligence (HUMINT), Signal Intelligence (SIGINT), Imagery Intelligence (IMINT), and Measurement and Signal Intelligence (MASINT) [37, 38]. According to a report [39]:

"Collecting intelligence these days is at times less a matter of stealing through dark alleys in a foreign land to meet some secret agent than one of surfing the Internet under the fluorescent lights of an office cubicle to find some open source. The world is changing with the advance of commerce and technology. Mouse clicks and online dictionaries today often prove more useful than stylish cloaks and shiny daggers in gathering intelligence required to help analysts and officials understand the world. Combined with stolen secrets, diplomatic reports, and technical collection, open sources constitute what one former deputy director of intelligence termed the *intricate mosaic* of intelligence".

This emphasizes two important things. First, the importance of open source intelligence and, second, its integration and consolidation with the intelligence gathered from traditional intelligence sources. Traditional information processing and text mining frameworks are likely to include many algorithms that can provide a base to build classical Natural Language Processing (NLP) applications. However, for them to be effective in the domain of open source intelligence and, the aimed DETECT system, they need to be accommodated with information consolidation mechanisms, both with heterogeneous open sources and secret documents and diplomatic reports usually residing within the boundaries of law enforcement and secret service agencies. We propose the development of a framework, which besides supporting the classical predictive and graphical methods of information processing also supports information consolidation mechanisms. This will also help in carrying out the necessary detailed and effective investigation.

C. Investigation Subsystem

The Investigation System typically deals with the transformation of extracted knowledge into a social network. The different networks of beliefs, entities, or both are formed. Key players are identified, measurements regarding densities of networks are made, and detection of subgraphs and communities then take place on the formed networks. It is also possible to measure the influence of one node over another or the centrality of a node in a network.

The networks extracted from news items and those from user generated media are also mapped. This is significant in developing the pairs of causes and effects and determining facts and figures which forms a sort of environment. The investigation system also discovers the true identification of the user, its aliases on different websites, its indirect connections and polarity of relations with other users on the basis of similarity of the ideas, its preferences, attributes, etc.

The investigation system inspects the history of past incidents in order to identify the conditions in which they occurred, i.e., the communities and groups of people and the polarities, sentiments, and information sharing among them. It keeps on matching the present environment with the environments in the past and learns from past events to predict the outcomes of the near future.

D. Notification Subsystem

This subsystem models the results of investigations in order to detect social polarization with the help of statistical modeling tools. It compares the produced models with the models that existed in the past. If the results of the comparisons reach a certain thresholds, it has detected potential social polarization. The thresholds and parameters corresponding to the model are extracted from past event. Thus the history of past incidents and their management are key responsibilities of such a system.

The notification subsystem also takes measures to estimate the changes in different salient features of the network like ratios of densities within communities and across the network, its evolving nature, and the changes in the roles of key players over a particular span of time. All these statistics are important in predicting the future of a social network and these facts and figures are also presented to the end users with corresponding notifications when these are generated. Then, the notifications are generated and reported to the subscribed end users and stake holders who can be related government institutes, research labs, intelligence and law enforcement agencies, authorized social scientists, etc.

III. State of the Art

Amongst the Web crawlers available today Apache Nutch [1, 2, 3, 4, 5] is perhaps the most flexible and widely used solution. Nutch provides the required functionality for web searching, grabbing, storing, and extracting text from web pages. It is equipped with a rule based engine to govern and fine tune the acquisition process. Nutch is also capable of handling RTF, XML, and PDF formats. It will be tested to identify the best possible settings and whether it can meet the needs of static web site monitoring. Independent of the particular crawler, a link analysis algorithm will be needed to track dynamically relocating sites and content.

Information retrieval software able to provide interface to social networking sites, to password protected online fora, and to real-time messaging systems like Twitter is yet to be developed.

Taxonomies and metadata indexing will be crucial to success in data management. There are two basic elements namely (a) a metadata schema and (b) a multilingual taxonomy. Flexibility is a key factor and the taxonomy will need to evolve as knowledge grows. This necessarily means that older legacy data will need continuous re-indexing. A goal of the proposed system is to accommodate live taxonomy updating and indexing, enabling human analysts to manage the knowledge base.

Extracting the information contained in a piece of text requires a proper understanding of the text content. Full semantic interpretation requires the identification of every individual conceptual component and the semantic roles it plays. But omissions and ambiguities are a common aspect of human communication, and most interesting sentences assume background knowledge. Any automatic system trying to understand a simple sentence will require, among other things, accurate capabilities for Named Entity Recognition and Classification (NERC), full Syntactic Parsing, Word Sense Disambiguation (WSD) and Semantic Role Labeling (SRL) [26].

Current baseline information systems are either large-scale, robust but shallow (standard IR systems), or they are small-scale, deep but ad hoc (Semantic-Web ontology-based systems). Furthermore, hardly any of the systems is multilingual, let alone cross-lingual and definitely not cross-cultural. WordNet [24] is an exception in being both multi- and cross-lingual [27, 28], so we believe it can be readily adapted to our needs.

The most widely used frameworks for information processing and extraction of information are GATE and UIMA [40]. General Architecture for Text Engineering (GATE) [41] is a popular open source framework built at the University of Sheffield. It offers extensive multilingual extraction techniques. It also provides standardized mechanisms for annotation and benchmarking. The framework is designed to successfully isolate low level tasks like data storage, data visualization, location, and loading of components from the data structures and algorithms that actually process the unstructured information. For processing unstructured text, GATE applies a queue based architecture (called pipeline in GATE architecture), containing Processing Resources (PRs). Each processing resource is a single algorithm to carry out a specific Information Extraction (IE) task or even more than one IE task, for example: recognizing Named Entities and Part Of Speech (POS) Tagging [42]. Each of the PRs is executed one by one in a controllable order and the output of a PR can be used as input to another PR. Each word in the input document is associated with different features (like POS Tag, its start position, etc.) can be consumed by a PR during execution of the underlying algorithms. The framework also provides channels to annotate the data and evaluate the performance of any PR against the preset Gold Standards and benchmarks. Although, it provides integration with relational databases, the implementation is space and time inefficient [43].

The Unstructured Information Management Architecture (UIMA) [44] offers to IBM customers the possibility of storing and analyzing massive amounts of unstructured or semi-structured information. The working of UIMA towards information extraction is somewhat similar to the working of GATE. For example, it implies recursive architecture for Text Analysis Engines (TAEs), which are analogous to PRs in Gate. Each TAE may include some specific algorithms and calculations to perform a specific NLP task. However, UIMA essentially provides crawler integration and acquisition facilities unlike GATE. Secondly, the Semantic Search Engine embedded into UIMA makes it much more efficient and usable than GATE in the domain of open source intelligence. UIMA uses Common Analysis Structure (CAS) [44] to encapsulate analysis results, which is similar to annotations in different architectures beginning with TIPSTER [45] and including GATE. UIMA works both on a single document and on collections of documents, similar to GATE, which works with a corpus (a set of documents) of one or more documents.

An application to extract entities from a text document, Named Entity Recognition, is a widely researched field at present. There are some very good implementations of Named Entity Recognisers (NER) available, such as Stanford NER [7], Lingpipe [8], KnowItAll [11], and Illinois Named Entity Tagger [6, 9].

To the best of our knowledge none of them performs satisfactorily with English text containing Arabic names, for example Abdul Aziz-Al Omari. This ability is absolutely critical for the proposed system. Lack of training on texts which contain Arabic names may be the root of a serious problem.

Self-supervised or semi-supervised Relation Extraction Systems (RES) will obviously be of more use to the project than any supervised RES. Most available RESs detect only binary relationships although higher-relation systems are being developed [12]. Most semi-supervised RES such as DIPRE [12], Snowball [13], and KnowItAll [11, 27] have predefined relation types; some e.g., TextRunner [15] discover relationships automatically.

The performance and accuracy of the presented systems in complex and sophisticated systems like DETECT has yet to be investigated. KnowItAll was developed for chemical and biomedical tasks, but has been shown to also work well when extracting Acquisition-Merger relationships [16].

Semiometrie is a quantitative research methodology that helps to reveal the emotions of groups of persons towards facts and opinions. This advanced technique uses a set of sensitive words to interpret emotions [10, 14] and is much used by media and advertising agencies to affect people's perceptions towards products or brands.

Available tools for identifying the sentiments associated with a textual sentence include: NaCTem Sentiment Analysis Tools [32], Buzzlogic [17], Alterian [33], and Nstein [25]. However, all of these detect the sentiments expressed by the author and induced in the reader. They do not consider sentiments existing between the different entities present within the text content. To illustrate: “A is an associate of B who was murdered by C” should ideally infer not only that the author supports B but also that A and B are positively connected but C is negatively connected to both A and B.

There are many existing social network analysis systems available such as Netminer [34], iMiner [18, 19, 20, 21], Coplink [29], i2 Analyst Notebook [22, 23], and CrimeFighter [30]. iMiner seems to have the best methods of detecting key players, using sophisticated techniques and novel algorithms alongside the usual social network analysis facilities [20]. Most of the social network analysis systems emphasize the use of centralities (degree, betweenness, and closeness) for detecting important and powerful nodes in a social network. However, for covert networks like Al-Qaeda network, the leaders and other important nodes are not the ones with high centrality values [19]. Thus, for covert networks, new measures are needed to assist the analysis. The software like iMiner, is packaged with some more interesting analysis tools like dependence centrality, identifying command structure, position role index, etc [19, 20]. The DETECT project will design and develop interfacing services to use the algorithms implemented in iMiner, but otherwise we expect to be able to use iMiner as it is. These services will be exploited by other components to feed their social network analysis needs.

The next section describes how the state of the art can be extended in order to solve the associated problems and most importantly assist towards the implementation of the proposed system.

IV. Extending the State of the Art

The implementation of the proposed system involves challenging extensions of existing technologies. A brief overview of the foreseen extensions is given below:

A. *Data Acquisition and Retrieval Subsystem*

- Link analysis algorithm to enable crawlers to track dynamically relocating sites and content.
- New statistical algorithms for alerting relevant new media content incorporation of entity/sentiment tracking into media monitoring methods of interfacing with social networking sites, password protected online fora, and real-time messaging systems like Twitter.
- Efficient algorithms for live taxonomy updating and indexing.
- New lexical resources, in terms of specific domain dictionaries, bilingual dictionaries, and new grammar rules.
- New specific methods/algorithms for semi-automatic updating of lexical resources.
- A special protocol enabling integration of reliable information sources with random information gleaned from the Web.

B. *Information Processing Subsystem*

- An extended Named Entity Recogniser which can cope with Arabic names in English text.
- An evaluation of available Relationship Extractors (RE) with regard to their use in the security informatics domain.
- The addition of a polarity extractor to a chosen RE.
- The development of a Sentiment Analysis System (SAS) that analyses page content, i.e., that gives the sentiments involved in the relationships found by our extended RE.
- The extension of this SAS to detect polarities and hidden sentiments in relationships.
- A further extension of the SAS to enable detection of sentiments among entities across multiple documents.
- A method of performing sentiment mining on syntactic and stylistic structures, rather than only on single words.
- A means of identifying where one person is using different aliases.

C. *Investigation Subsystem*

- A similarity detector for relationship networks.
- A similarity detector for the sentiments assigned to a Named Entity.
- A system for generating different English alternatives for Arabic names.
- A system to extract cause-effect pairs from news items and from user generated media.

D. *Notification Subsystem*

- Developing a suitable format and level of information for warnings to be passed to the appropriate end users and stake holders.

V. Impact

This project undertakes modeling and analysis of the total causal environment behind social polarization, radicalization, and violent extremism. Accurate modeling will lead to early detection of polarization, allowing time for destabilization strategies to be devised based on real and specific information extracted by our proposed system, in case the groups identified continue down to the path of extremism. Accurate prediction of violent events will enable the effective deployment of forces of law, to minimize or even remove the serious negative outcome of such events.

The DETECT system will provide the law enforcement agencies with an automatic data processing tool, that enables media monitoring, processing the text, images and videos; and finally producing the alerts when something suspicious is found. This will decrease the response time of the corresponding organizations and improve their chance of defusing a violent situation in its early stages.

The DETECT project involves the development of many independent components (described in Section II and III) integrated into a single architecture for the achievement of the desired goals. Although, state of the art systems such as the aforementioned components are present, they are not adequate. The possible extensions to enhance effectiveness, usability, or scalability against each of the component are described in Section IV. It shows that the project yields positive scientific impact and a vital extension to the state of the art technologies.

VI. Conclusion

In this paper, we have described the functionality and design of a novel system capable of detecting social polarization and estimating its chances of shaping into violent radicalization. We discussed how existing technologies and systems can be integrated or extended to implement the desired detection and estimation processes. The necessity for the development of such a solution and its positive social and scientific impact has also been addressed in this paper.

References

[1] M. Cafarella and D. Cutting “Building Nutch: Open Source Search”. Queue 2, (2), 54–61, 2004.

[2] Konwinski, M. Zaharia, R. Katz, and I. Stoica “X-tracing Hadoop”. Hadoop Summit, Mar 2008.

[3] D. Borthakur “The Hadoop Distributed File System: Architecture and Design”.hadoop.apache.org/core/docs/current/hdfs design.pdf.

[4] R. Khare, D. Cutting, K. Sitaker, and A. Rifkin “Nutch: A Flexible and scalable open-source web search engine”. Technical report, CommerceNet Labs, 2004.

[5] Apache Nutch: www.nutch.org/

[6] L. Ratinov and D. Roth “Design challenges and misconceptions in named entity recognition”.Proc. 13th Conf. on Computational Natural Language Learning (CoNLL ’09). Association for Computational Linguistics, Morristown, NJ, USA, 147–155, 2009.

[7] Stanford Named Entity Tagger: nlp.stanford.edu/ner/index.shtml

[8] Lingpipe: alias-i.com/lingpipe/index.html

[9] Illinois Named Entity Tagger: cogcomp.cs.illinois.edu/page/software view/4

[10] Camillo, F.; Morace, F. and Traldi, T., “From marketing to ”societing”. Reading ethnographic material through the use of digital matrix and Semiometrie”. In Innovation - The best in innovation from around the world - ESOMAR, PARIGI, Esomar, pp. 130 - 145, 2005

[11] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D.S. Weld, A.Yates, “Unsupervised named-entity extraction from the Web: An experimental study”, Artificial Intelligence 165 (1), 91–134, 2005.

[12] S. Brin, R. Motwani, L. Page, and T.Winograd, “What can you do with aWeb in your pocket?” Data Engineering Bulletin 21, (2), 37-47, 1998.

[13] Eugene Agichtein and Luis Gravano, “Snowball: extracting relations from large plain-text collections”. In Proceedings of the fifth ACM conference on Digital libraries (DL ’00). ACM, New York, NY, USA, 85–94, 2000.

[14] Camillo, F., Tosi, M. and Traldi, T., “Semiometric approach, qualitative research and text mining techniques for modelling the material culture of happiness”. In Knowledge Mining, HEIDELBERG, Springer, pp. 230 255, 2005.

[15] A. Yates, M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland “TextRunner: open information extraction on the web”. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations on XX (NAACL ’07). Association for Computational Linguistics, Morristown, NJ, USA, 25–26, 2007.

[16] B. Rozenfeld and R. Feldman, “Self-supervised relation extraction from the Web”. Knowl. Inf. Syst. 17, 1 (October 2008), 17–33, 2008.

[17] O. Ryan “The buzz around Buzz.” Fortune, 155 (5), 46–48, 2007.

[18] N. Memon, U.K. Wiil, R. Alhajj, C. Atzenbeck, and N. Harkiolakis “Harvesting Covert Networks: The Case Study of the iMiner Database”. Int. J. Networking and Virtual Organizations (IJNVO) Special Issue on Open Source Intelligence and Web Mining (OSINT-WM), 8(1/2), pp.10–74, 2011.

[19] D. Hicks, N. Memon, J.D. Farley, T. Rosenorn “Mathematical Methods in Counterterrorism: Tools and Techniques for a New Challenge”. In Memon N. et al (eds.) Mathematical Methods in Counterterrorism, Springer Verlag, 2009.

[20] N. Memon “A first look on iminers knowledge base and detecting hidden hierarchy of Riyadh bombing terrorist networks”, Proceedings of IMECS 2007, Hong Kong, Lecture Notes in Computer Science and Engineering, pp.705–715, 2007.

[21] N. Memon, and H.L. Larsen “Practical approaches for visualization, analysis and destabilizing terrorist networks”, Proceedings of The First International Conference on Availability, Reliability and Security (ARES), IEEE Computer Society, pp.906–913, 2006.

[22] Analyst Notebook: www.i2group.com/uk/products–services/analysis-product-line/analysts-notebook

[23] P. Klerks, “The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators?” Recent developments in the Netherlands. Connections 24, (3), 53–65, 2001.

[24] WordNet: http://wordnet.princeton.edu/

[25] Nstein: www.nstein.com/en/products-and-technologies/text-mining-engine/

[26] Carreras, X., Mrquez, L.: “Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling”. In CoNLL-2005, Ann Arbor, MI USA, 2005.

[27] P. Vossen, “EuroWordNet: a multilingual database with lexical semantic networks”, Kluwer Academic Publishers, Norwell, MA, 1998.

[28] P. Vossen, F. Neri, R. Raffaelli et al.: “KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures”, Proc. LREC 2008, Marrakech (MO), 2008.

[29] H. Chen, D. Zeng, H. Atabakhsh, W. Wyzga, and J. Schroeder “COPLINK: managing law enforcement data and knowledge”. Commun. ACM 46, (1), 28–34, 2003.

[30] U. K. Wiil; J. Gniadek; N. Memon. CrimeFighter Assistant: A Knowledge Management Tool for Terrorist Network Analysis. In proceedings of the International Conference on Knowledge Management and Information Sharing (KMIS 2010). Institute for Systems and Technologies of Information, Control and Communication, 2010. pp. 15-24

[31] KnowItAll: www.knowitall.com/academic/welcome.asp

[32] NaCTem : www.nactem.ac.uk/software.php

[33] Alterian: socialmedia.alterian.com/

[34] NetMiner: www.netminer.com/NetMiner/home 01.jsp

[35] Apache Lucene: http://lucene.apache.org/

[36] J. F. Whissel,"Information Retrieval Using Lucene and WordNet", Thesis, University of Akron, Computer Science, 2009

[37] R. D. Steele, “The Failure of 20th Century Intelligence.” [www.oss.net/FAILURE]

[38] P. A. R. Qureshi, N. Memon and U. K. Wiil, “EWAS: Modeling Application for Early Detection of Terrorist Threats”, From Sociology to Computing in Social Networks, Part III, pp. 135-156, 2010, [DOI: 10.1007/978-3-7091-0294-7_8]

[39] R. J. Smith, "The Unknown CIA: My Three Decades with the Agency", Washington, DC: Pergamon-Brassey’s, 1989, 195

[40] Hamish Cunningham, "Software Architecture for Language Engineering", PHD Thesis, University of Sheffield, 2000 [http://gate.ac.uk/sale/thesis/]

[41] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, “GATE: an architecture for development of robust HLT applications”, In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 168-175, 2002 [DOI: 10.3115/1073083.1073112]

[42] K. Toutanova and C. D. Manning “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger” In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70. 2000

[43] Hamish Cunningham, "Software Architecture for Language Engineering", PHD Thesis, University of Sheffield, 2000 [http://gate.ac.uk/sale/thesis/]

[44] D. Ferrucci and A. Lally, “Building an Example Application with the Unstructured Information Management Architecture,” IBM Systems J.,vol.43,no.3,2004 [www.research.ibm.com/journal/sj/433/ferrucci.html]

[45] R. Grishman, "Tipster architecture design document version 2.2. Technical report", Defense Advanced Research Projects Agency, 1996

Pir Abdul Rasool Qureshi is working as Research Assistant at the Maersk Mc-Kinney Moller Institute, University of Southern Denmark. He is working on the design and development of an early warning system to detect terrorist threats. He has currently published more than a hand full of research papers and has vast experience in development of web harvesting and investigative data mining software.

Nasrullah Memon is an Associate Professor at the Maersk Mc-Kinney Moller Institute, University of Southern Denmark. He received a Master’s Degree in Applied Mathematics from the University of Sindh, Pakistan, and a Master’s Degree in Software Development from the University of Huddersfield, UK. He holds a PhD in Investigative Data Mining from Aalborg University, Denmark. He is also affiliated with Mehran University of Engineering and Technology, Jamshoro, Sindh, Pakistan, and Hellenic American University, Athens, Greece. He is Editor-in-Chief of *International Journal on Advances in Social Network Analysis and Mining*, Springer-Verlag. He has published more than 80 research papers in international conferences and journals. His current research is on knowledge management, mathematical methods in counterterrorism, information extraction and investigative data mining.

Uffe Kock Wiil is a Professor of Software Engineering and Technology at the Maersk Mc-Kinney Moller Institute, University of Southern Denmark. He holds a M.Sc. degree in Computer Engineering (1990) and a Ph.D. degree in Computer Science (1993) both from Aalborg University, Denmark. His research interest includes knowledge management, hypertext, computer supported cooperative work, software technology, and distributed systems. These research interests are currently being applied in three overall areas: counterterrorism, healthcare, and planning. He has published more than 150 research papers. His research papers have been cited more than 1200 times. He is currently serving on the editorial boards of the Elsevier Journal on Network and Computer Applications and the Springer Journal on Social Network Analysis and Mining and on the advisory board of the Springer Journal on Security Informatics.

Panagiotis Karampelas holds a PhD in Electronic Engineering from the University of Kent at Canterbury, UK. He also holds a Master of Science from the Department of Informatics, Kapodistrian University of Athens and a Bachelor degree in Mathematics from the same University. He has worked for 3½ years as an associate researcher in the Foundation for Research & Technology-Hellas (FORTH), Institute of Computer Science, and several years as a user interface designer and usability expert in several IT companies designing and implementing large-scale information systems. He has also participated in many European research projects and published a number of articles in his major areas of interests namely, Human Computer Interaction and Visualization, Social Network Analysis and Mining and Power Systems Simulation and Forecasting. He serves as an associate editor in the Social Network Analysis and Mining journal and reviewer in various scientific journals and conferences in his fields of interests. Panagiotis Karampelas also teaches information technology courses in Hellenic American University, ASPETE and Hellenic Air Force Academy.