Detecting Terrorist Activities via Text Analytics

Eric Atwell, Language Research Group
I-AIBS: Institute for Artificial Intelligence and Biological Systems

Overview
DTAct EPSRC initiative
Recent research on terrorism informatics
Ideas for future research

Background: EPSRC DTAct
EPSRC: Engineering and Physical Science Research Council
Detecting Terrorist Activities – DTAct
A joint “Ideas Factory” Sandpit initiative supported by EPSRC, ESRC, the Centre for the Protection of National Infrastructure (CPNI), and the Home Office to develop innovative approaches to Detecting Terrorist Activities; 3 projects to run 2010-2013

DTAct aims
“…Effective detection of potential threats before an attack can help to ensure the safety of the public with a minimum of disruption. It should come as far in advance of attack as possible…Detection may mean physiological, behavioural or spectral detection across a range of distance scales, remote detection; or detection of an electronic presence. DTAct may even develop or use an even broader interpretation of the concept. Distance may be physical, temporal, virtual or again an interpretation which takes a wider view of what it means for someone posing a threat to be separated from his or her target. …Effective detection of terrorist activities is likely to require a variety of sensing approaches integrated into a system. Sensing approaches might encompass any of a broad range of technologies and approaches. In addition to sensing technologies addressing chemical and physical signatures these might include animal olfaction; mining for anomalous electronic activity; or the application of behavioural science knowledge in detection of characterised behavioural attributes. Likewise, the integration element of this problem is very broad, and might encompass, but is not limited to: hardware; algorithms; video analytics; a broad range of human factors, psychology and physiology considerations (including understanding where humans and technology, respectively, are most usefully deployed); or operational research, analysis and modelling to understand the problem and explore optimum configurations (including choice and location of sensing components.)…”

How to use text analytics for DTAct?
Terrorists may use email, phone/txt, websites, blogs …
…to recruit members, issue threats, communicate, plan…
Also: surveillance and informant reports, police records, …
So why not use NLP to detect “anomalies” in these sources?
Maybe like other research at Leeds:
• Arabic text analytics
• detecting hidden meanings in text
• social and cultural text mining
• detecting non-standard language variation
• detecting hidden errors in text
• plagiarism detection

Recent research on DTAct
Engineering devices to detect at airport or on plane – too late?
Terrorism Studies, eg MA Leeds University (!)
…political and social background, but NOT detection of plots
Research papers with relevant-sounding titles
…but very generic/abstract, not much real NLP text analysis
Some examples:
Fienberg S. Homeland insecurity: Datamining, Terrorism Detection, and Confidentiality.

MATRIX: Multistate Anti-Terrorism Information Exchange system to store, analyze and exchange info in databases – but doesn’t say how to acquire DB info in the first place

TIA: Terrorist Information Program – stopped 2003

PPDM: Privacy Preserving Data Mining – “big issue” is privacy of data once captured, rather than how to acquire data


“… we explore an integrated approach for identifying and collecting terrorist/extremist Web contents … the Dark Web Attribute System (DWAS) to enable quantitative Dark Web content analysis.”

Identified and collected 222,000 web-pages from 86 “Middle East terrorist/extremist Web sites”… and compared with 277,000 web-pages from US Government websites

BUT only looked at HCI issues: technical sophistication, media richness, Web interactivity.

NOT looking for terrorists or plots, NOT language analysis

Last M, Markov A, Kandel A. Multi-lingual detection of terrorist content on the Web

Aim: to classify documents: terrorist v non-terrorist

Build a C4.5 Decision Tree using “word subgraphs” as decision-point features.

Tested on a corpus of 648 Arabic web-pages, C4.5 builds a decision tree based on keywords in document:

“Zionist” or “Martyr” or “call of Al-Quds” or “Enemy” → terror
Else → non-terror

NOT looking for plots, NOT deep NLP (just keywords)

Chen H, Reid E, Sinai J, Silke A, Ganor B (eds). 2008. TERRORISM INFORMATICS: Knowledge Management and Data Mining for Homeland Security

Methodological issues in terrorism research (ch 1-10);

Terrorism informatics to support prevention, detection, and response (ch 11-24)

Silke: U East London, UK; BUT sociology, not IS

57 co-authors of chapters! Only 2 in UK: Horgan (psychology), Raphael (politics)

Several impressive-sounding acronyms …

Abouzakhar N, Allison B, Guthrie L. Unsupervised Learning-based anomalous Arabic Text Detection

Corpus of 100 samples (200-500 words) from Aljazeera news
Randomly insert sample of religious/social/novel text
Can detect “anomalous” sample by average word length, average sentence length, frequent words, positive words, negative words, …
Problems in Text Analytics for Detecting Terrorist Activities

Not just English: Arabic, Urdu, Persian, Malay, ...

Need a Gold Standard corpus of “terror v non-terror” texts

What linguistic features to use?

Terrorists may use covert language: “the package”

Problems with other languages

Arabic:
- Writing system: short vowels, carrying morphological features, can be left out, increasing ambiguity;
- Complex morphology: root+affix(es)+clitic(s)

Malay:
- Opposite problem – simple morphology, but a word can be used in almost any PoS grammatical function;

Few resources (PoS-tagged corpora, lexical databases) for training PoS-taggers, Named Entity Recognition, etc.

Terror Corpus

We need to collect a Corpus of “suspicious” e-text

Start with existing Dark Web and other collections

Human “scouts” look for suspicious websites, and

Robot web-crawler uses “seeds” to find related web-pages

MI5, CPNI, Police etc to advise and provide case data

Annotate: label “terror” v “non-terror”, “plot”, ...

Linguistic Annotation

We don’t know which features correlate to “terror plot”

So: enrich with linguistic features (PoS, sentiment, …)

Then we can use these in decision trees etc based on deeper linguistic knowledge

Covert language

If we have texts which are labelled “plot”, look for words which are suspicious because they are NOT terror-words

E.g. high log-likelihood of “package”

Text Analytics for Detecting Terrorist Activities: Making Sense

Claire Brierley and Eric Atwell: Leeds University

International Crime and Intelligence Analysis Conference
Manchester - 4 November 2011
What is “Making Sense”?  
- EPSRC consortium project in the field of Visual Analytics  
- Remit to create an interactive, visualisation-based decision support assistant as an aid to intelligence analysts  
- Target user communities are law enforcement, military intelligence and the security services  

1. Involves automated approaches to “gisting” multimedia content  
2. Integrating gists from different modalities: audio, visual, text  
3. Identifying links/connections in fused data  
4. Visualisation of results to support interactive query and search

Nature of intelligence material

**Task:**
- To identify “suspicious” activity via multi-source, multi-modal data  

**Issues of quantity and quality:**
- DELUGE of multi-source, multi-modal data for target user groups to make sense of and act upon  
- Deluge of NOISY data

**Nature of intelligence data and its critical features:**
- It may be unreliable.  
- The credibility of sources may be questionable.  
- It’s fragmented and partial.  
- Text-based data may be non-standard (e.g. txt messages)  
- It’s from different modalities, and there’s a lot of it!  
- So it’s easy to miss that “needle in the haystack”.

Text Extraction: methodologies available

There are various options for extracting “actionable intelligence” from text.

1. Google-type search and Information Retrieval (IR) to pull documents from the web in response to a query  
2. Query formulation is informed by domain expertise and human intelligence (HUMINT) – another approach  
3. Automatic Text Summarisation to generate summaries from regularities in well-structured texts  
4. Information Extraction (IE), focussing on automatic extraction of entities (i.e. nouns, especially proper nouns), facts and events from text  
5. Keyword Extraction (KWE) uses statistical techniques to identify keywords denoting the aboutness of a text or genre

What is Leeds’ approach?

“Making Sense” proposal:
...the gist of a phone tap transcript might comprise: caller and recipient number; duration of call; statistically significant keywords and phrases; and potentially suspicious words and phrases...

Why use Keyword Extraction (KWE)?
- It can be implemented speedily over large quantities of ill-formed texts  
- It will uncover new and different material, such that we can undertake content analysis

Making Sense: The Team

- Funded by EPSRC/ESRC/CPNI  
- Multi-disciplinary:  
  - Psychology  
  - Law  
  - Operations research  
  - Computational linguistics  
  - Visual analytics  
  - Machine learning and artificial intelligence  
  - Human computer interaction  
  - Computer science  
- Approximately 300 person months over 36 months (full economic cost: £2.6m).

Text Extraction

Newsreel word cloud

1980s BBC radio
**Newsreel word cloud**

1980s BBC radio

**Habeas Corpus?**

Text Analytics Research Paradigm:
- Uses a corpus of naturally-occurring language texts which capture empirical data on the phenomenon being studied
- The phenomenon under scrutiny needs to be labelled in the corpus in order to derive training sets for machine learning
- This labelled corpus constitutes a “gold standard” for iterative development and evaluation of algorithms

Therefore, our EPSRC proposal for Making Sense states that engagement with stakeholders and authentic datasets for simulation and evaluation are critical to the project.
Habeas Corpus?

Text Analytics Research Paradigm:
- Uses a corpus of naturally-occurring language texts which capture empirical data on the phenomenon being studied
- The phenomenon under scrutiny needs to be labelled in the corpus in order to derive training sets for machine learning
- This labelled corpus constitutes a "gold standard" for iterative development and evaluation of algorithms

Therefore, our EPSRC proposal for Making Sense states that engagement with stakeholders and authentic datasets for simulation and evaluation are critical to the project.

Problem: we do not have ANY data - never mind LABELLED data!

Survey Findings

- Gaining access to relevant data is generally raised as an issue in academic publications for intelligence and security research
- Relevant data is truth-marked data, essential to benchmarking
- Research time and effort is thus spent on compiling synthetic data
- So-called terror corpora have been compiled from documents in the public domain, often Western press
- Design and content of synthetic datasets like VAST and Enron email dataset assume an IE approach to text extraction
- Information Extraction is the dominant technique used in commercial intelligence analysis systems
- Only one (British) company is using KWE, which they say is "just as good a predictor [of suspiciousness] as IE"

Text Analytics: Style is countable

Text analytics is about pattern-seeking and counting things
1. If we can characterise, for example, stylistic or genre-specific elements of a target domain via a set of linguistic features...
2. ...then we can measure deviation from linguistic norms via comparison with a (general) reference corpus
3. Concept of KEYNESS: when whatever it is you’re counting occurs in your corpus and not in the reference corpus or significantly less in the reference corpus.

Leeds approach to genre classification and linking:
1. Derive keywords and phrases from a reliable "terror" corpus.
2. These lexical items can be said to characterise the genre and they also constitute suspicious words and phrases.
3. Compare frequency distributions for designated suspicious items in new and unseen data relative to their counterparts in the terror corpus.
4. Similar distributional profiles for these items, validated by appropriate scoring metrics (e.g. log likelihood), will discover candidate suspect texts.

Applying Text Analytics Methodology 1

- Leeds have been involved in collaborative prototyping of parts of our system with project partners Middlesex and Dundee for the VAST Challenges 2010 and 2011.
- VAST 2010: Keyword gists have been incorporated in Dundee "Semantic Pathways" visualisation tool.
- VAST 2011 Mini Challenge 3: Text Extraction has been useful in gisting content from 4474 news reports of interest to intelligence analysts looking for clues to potential terrorist activity in the Vastopolis region. Each news report is a plaintext file containing a headline, the date of publication, and the content of the article.
- VAST 2011 Mini Challenge 1: A flu-like epidemic leading to several deaths has broken out in Vastopolis which has about 2 million residents. Text Extraction has been useful in ascertaining the extent of the affected area and whether or not the outbreak is contained.

Mini Challenge 1: Tweet Dataset

- We’ve said that KWE can be implemented speedily over large quantities of ill-formed texts
- In this case, the ill-formed texts are tweets
- Problem with text-based data: different datasets need “cleaning” in different ways and tokenization is also problematic
- CSV format: ID, User ID, Date and Time, District, Message

<table>
<thead>
<tr>
<th>Tweet ID</th>
<th>User ID</th>
<th>Date and Time</th>
<th>District</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>11, 70840</td>
<td>30/04/2011 00:00</td>
<td>Westside</td>
<td>Be kind. If u step on ppl in this life u’ll probably come back as a cockroach in the next. #ummhhmm #karma</td>
<td></td>
</tr>
<tr>
<td>25, 177748</td>
<td>30/04/2011 00:00</td>
<td>Lakeside</td>
<td>August 15th is 2weeks away :/! That’s when Ty comes back! I miss him :)</td>
<td></td>
</tr>
</tbody>
</table>
| 44, 121322 | 30/04/2011 00:01 | Downtown | #NewTwitter #Rangers#TEAMfollowBACK#TFB #IReallyThink#becauseoftwitter #Mustfollow #MeMetiATerror #SHOUTOUT #justinbieber FOLLOW ME>

Mini Challenge 1: Collocations

- Used a subset of the dataset: start date/time of epidemic had already been established
- Each tweet had been tagged with its city zone, so created 13 tweet datasets, one for each zone
- Built wordlists for each zone and converted each wordlist into a Text object
- Then able to call object-oriented collocations() method on each text object to emit key collocations (bigrams or pairs of words) per zone
- The collocations() method uses log likelihood metric to determine whether bigram occurs significantly more frequently than counts for its component words would suggest
Mini Challenge 1: Collocations

```python
>>> smogtownTO.collocations()
Building collocations list
somewhere else; really annoying; getting really; stomach ache; bad
diarrhea; vomitting everywhere; sick sucks; extremely painful; can't
stand; terrible chest; feeling better; short breath; chest pain; every
minute; breath every; constant stream; bad case; flem coming; well
soon; anyone needs
```

```python
>>> riversideTO.collocations()
Building collocations list
declining health; best wishes; somewhere else; wishes going; can't
stand; terrible chest; atrocious cough; chest pain; constant stream;
flem coming; get plenty; really annoying; getting really; doctor's
office; short breath; every minute; office tomorrow; sore throat;
laying down.; get well
```

Mini Challenge 1: Keyword Gists

- Also computed keywords (or statistically significant words) per city zone
- Entails comparison of word distributions in 13 test sets (the tweets per
zone) with distributions for the same words in a reference set: all tweets
since start of outbreak
- Build wordlists and frequency distributions for test and reference corpora
- Apply scoring metric (log likelihood) to determine significant overuse in a
test set relative to the reference set

<table>
<thead>
<tr>
<th>City Zone</th>
<th>Stomach Frequency</th>
<th>Diarrhea Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plainville</td>
<td>1870.34</td>
<td>1771.62</td>
</tr>
<tr>
<td>Downtown</td>
<td>982.90</td>
<td></td>
</tr>
<tr>
<td>Uptown</td>
<td>606.52</td>
<td>646</td>
</tr>
</tbody>
</table>

Text Extraction:

Research question: Can keywords derived from training data which exemplifies a target concept be used to classify unseen texts?

Problems flagged up by survey:

- Non-availability of truth-marked evidential data is a problem in the intelligence and security domain
- No machine learning can take place without exemplars and yardsticks for the concept or behaviour being studied

Solution:

1. Simulate problem of “finding a needle in a haystack” on a real
dataset: English translation of Qur’an
2. Can annotate a truth-marked (labelled) subset of verses associated with
target concept via Leeds Qurany ontology browser
3. Target concept is NOT suspiciousness but is analogous in scope

Analogous in scope: skewed distribution

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Reference Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judgment Day</td>
<td>6236 verses</td>
</tr>
<tr>
<td>3680 words</td>
<td>164543 words</td>
</tr>
</tbody>
</table>

1. The subset represents roughly 2% of the corpus
2. Judgment Day verses are scattered throughout the Qur’an

Important finding:

The fact that the subset constitutes only 2% of the corpus has
implications for evaluation

- As many as 234 attribute-value sets (including class attribute)
- Prior probability for majority class: 0.98
- Prior probability for minority class: 0.02

Text Extraction: Qur’an-as-Corpus

Research question: Can keywords derived from training data which exemplifies a target concept be used to classify unseen texts?

Problems flagged up by survey:

- Non-availability of truth-marked evidential data is a problem in the intelligence and security domain
- No machine learning can take place without exemplars and yardsticks for the concept or behaviour being studied

Solution:

1. Simulate problem of “finding a needle in a haystack” on a real dataset:
   English translation of Qur’an
2. Can annotate a truth-marked (labelled) subset of verses associated with
   target concept via Leeds Qurany ontology browser
3. Target concept is NOT suspiciousness but is analogous in scope

Methodology: keyword extraction

- Build wordlists and frequency distributions for test and reference corpora
- Compute statistically significant words in the test set relative to the
  reference set

<table>
<thead>
<tr>
<th>Word</th>
<th>Subset Frequency</th>
<th>All Qur’an Frequency</th>
<th>Log likelihood statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>will</td>
<td>123</td>
<td>1973</td>
<td>94.82</td>
</tr>
<tr>
<td>together</td>
<td>25</td>
<td>87</td>
<td>77.03</td>
</tr>
<tr>
<td>gather</td>
<td>16</td>
<td>28</td>
<td>66.54</td>
</tr>
<tr>
<td>any</td>
<td>46</td>
<td>526</td>
<td>56.33</td>
</tr>
<tr>
<td>return</td>
<td>19</td>
<td>80</td>
<td>52.71</td>
</tr>
</tbody>
</table>
**Training instances: attribute-value pairs**

CSV format

| location, a, gather, burden, bearer, show, creation, back, one, brought, single, toget her, another, soul, trumpet, sepulchres, said, end, raise, laden, judgment, people, where on, day, excuses, call, exempt, marshalled, hidden, tell, be, good, return, truth, do, shall, gathered, toiling, ye, bear, you, observe, besides, graves, beings, with, response, original es, revile, sounded, this, goal, resurrection, originate, up, us, later, will, know, repeats, or, CountKws, countKeyBigrams, concept |

**Skewed Data Problem**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature Set</th>
<th>Success Rate %</th>
<th>Recall minority class</th>
<th>Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>OneR</td>
<td>63</td>
<td>98.20</td>
<td>0.09</td>
<td>TP 10, FN 103, TN 6111, FP 12</td>
</tr>
<tr>
<td>J48</td>
<td>63</td>
<td>98.41</td>
<td>0.27</td>
<td>TP 30, FN 83, TN 6107, FP 16</td>
</tr>
<tr>
<td>NB</td>
<td>63</td>
<td>93.41</td>
<td>0.66</td>
<td>TP 74, FN 39, TN 5751, FP 372</td>
</tr>
</tbody>
</table>

**Extra Metrics: BCR and BER**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature Set</th>
<th>Success Rate %</th>
<th>Recall minority class</th>
<th>Confusion Matrix BCR BER</th>
</tr>
</thead>
<tbody>
<tr>
<td>OneR</td>
<td>63</td>
<td>98.20</td>
<td>0.09</td>
<td>TP 10, FN 103, TN 6111, FP 12, 0.54, 0.46</td>
</tr>
<tr>
<td>J48</td>
<td>63</td>
<td>98.41</td>
<td>0.27</td>
<td>TP 30, FN 83, TN 6107, FP 16, 0.63, 0.37</td>
</tr>
<tr>
<td>NB</td>
<td>63</td>
<td>93.41</td>
<td>0.66</td>
<td>TP 74, FN 39, TN 5751, FP 372, 0.80, 0.20</td>
</tr>
</tbody>
</table>

**Applying Text Analytics Methodology 2**

Leeds have used KWE Text Analytics methodology to:
- identify verses associated with a given concept in the Qur'an
- ascertain extent of spread of a flu-like epidemic from a (synthetic) corpus of tweets
- gist the contents of (synthetic) news reports for intelligence analysts looking for clues to potential terrorist activity

We are planning to use it in Health Informatics, with real datasets:
- to classify cause of death in Verbal Autopsy reports
- to derive linguistic correlates from free text data such as clinicians’ notes for automatic prediction of likely outcome of a given cancer patient pathway at a critical stage
- to assist in recommending optimal course of action for patient: transfer to palliative care or further treatment
- entails careful scaling up via iterative development of clinical profiling algorithms

**Collaboration**

We are keen to collaborate on other projects!
- Corpus of text messages etc generated during the recent UK riots is a potentially interesting dataset?
- KWE extraction algorithms need fine-tuning so that they run in real time
- We need labelled examples in the dataset of the phenomenon/behaviour of interest in order to develop and evaluate machine learning algorithms

**Summary**

DTAct EPSRC initiative
Recent research on terrorism informatics
Ideas for future research

IF YOU HAVE ANY MORE IDEAS, PLEASE TELL ME!