Extracting Value from NATO Data Sets through Machine Learning and Advanced Data Analytics

Ivana Ilic Mestric, Arvid Kok, Giavid Valiyev, Michael Street, Peter Lenk
NATO Communications and Information Agency (NCIA), Data Analytics and Innovation
Oude Waalsdorperweg 61
2597 AK The Hague, NETHERLANDS
Ivana.IlicMestric@; Arvid.kok@; Giavid.Valiyev@; Michael.Street@; Peter.Lenk@ncia.nato.int;

Mihaela Racovita, Filipe Vieira,
Joint Analysis and Lessons Learned Centre
Avenida Tenente Martins
1500-589 Lisboa, PORTUGAL
Mihaela.racovita@jallc.nato.int;
Filipe.vieira@jallc.nato.int

ABSTRACT

Exercises and operations generate large quantities of data, but until recently this data was not exploited to enhance decision support and situational awareness, or for analysis and assessment. This paper describes the steps taken to collect, prepare and extract knowledge from such data collected at exercise Trident Juncture in 2018, using advanced analytics [1] and machine learning [2] to extract high level perspectives. Results show initial macro level assessments of exercise performance including response times; visualisation of information flows across locations and systems; and identification of other items of interest in this "big data" collection.

The paper also presents results using natural language processing [3] to identify situations and incidents of interest to the lessons learned community from exercises and operations. The data sources for this assessment are both the raw data from exercises as well as from lessons learned reports about exercises and operations.

The paper also describes the technical challenges of using advanced data analytics and machine learning on large, classified data sets, together with solutions for the hardware and software challenges encountered.

1.0 INTRODUCTION

The work described in this paper makes use of two distinct data sets; one from exercise Trident Juncture 2018 (TRJE 2018) and another of reports from the NATO Lessons Learned Portal (NLLP). During exercise Trident Juncture, NCIA stored and repatriated around 3.9 TB of data from six different locations operating on two security domains. This heterogeneous collection consists of data relating to:

- Collaboration services (over 23,000 chat conversations)
- Functional area services for: Land, Maritime, Air and Joint Command & Control systems, Planning Logistics, Intelligence, Geographic Information Systems, Incident management, Collaboration Core services: Exchange (Email), Tasker Tracker, Active Directory, SharePoint Portal, Document Handling System
- Exercise management: 1161 confirmed events & 373 injects

Figure 1 lists the file sizes and file counts which make up this data set.

FAS File Size (GB)

FAS	Size GB
lc2is	14,655.51
logfas	2,589.38
sql server	1,177.87
ncop scc	190.21
exchange	178.31
sharepoint	139.58
jocwatch	129.60
coregis	110.77
jchat	39.38
domain controller	36.96
intel-fs	8.67
niris	7.97
icc	3.60
Total	**19,267.82**

FAS File Count

FAS	Count
lc2is	1625361
jchat	1202972
logfas	288510
ncop scc	132434
coregis	73130
jocwatch	46193
domain controller	6311
sharepoint	4620
exchange	3778
sql server	1890
intel-fs	647
niris	420
icc	312
Total	**3386578**

Figure 1 File Sizes & Counts repatriated from TRJE 2018

The NLLP is a repository of over 400 documents which capture lessons learned, best practices identified and other assessments of NATO activities in operations and exercises. This repository also provides a rich dataset where data science can be applied to generate valuable insights, and enable topics and trends to be extracted.

Two toolsets were used for this work, in some cases both were applied to address similar problems, allowing a comparison to be made between the toolsets. One tool was IBM Watson Explorer; the other was a combination of the open-source KNIME analytics platform [4], in conjunction with Microsoft's PowerBI for data visualisation.

2.0 EXERCISE ANALYSIS

Initial analysis of data collected from Trident Juncture focussed on data generated by functional area services (FAS) used for intelligence (IntelFS) and for collaboration (JCHAT). Analysing this data at an exercise level revealed a high-level picture of how these tools were used during the exercise to support analysis and reporting.

2.1 Analysis of collaboration services

Analysis of all messages sent during the exercise using the JCHAT (joint tactical chat) service for collaboration showed that there were 23,123 unique group conversations, involving 693 unique users. Applying language recognition to the messages showed that while the majority of the messages were in English, a significant portion were in French, Norwegian, Italian and Swedish plus very limited amounts of other NATO and Partner languages.

Activity patterns were generated by comparing activity through the day and the week, as shown in Figure 2. This shows that while collaboration takes place throughout the week and largely reflects the exercise schedule with most activity taking place during normal working hours, but some coordination activity continuing into the night. This figure also shows the most active organisational elements participating in most JCHAT conversations.

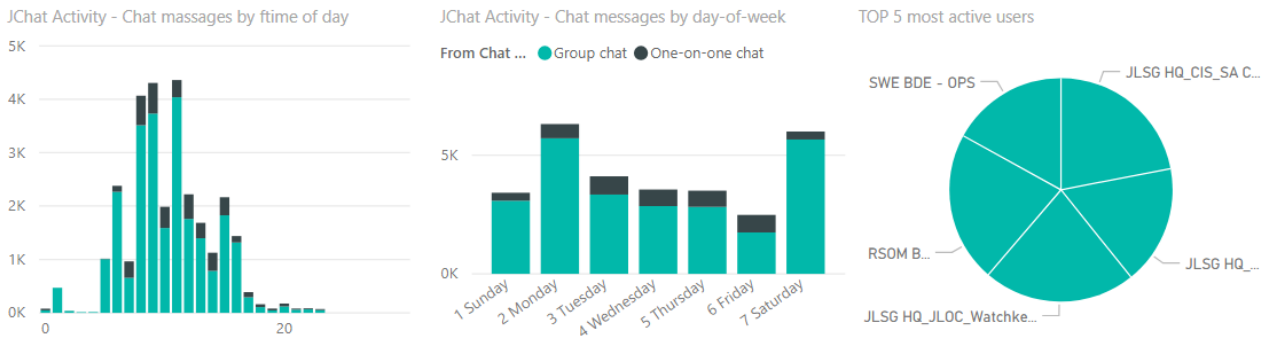

Figure 2: JCHAT activity

Figure 3 shows the level of JCHAT activity – for collaboration – throughout the exercise (upper plot) against the overall activity in other functional area services (lower plot). While there is some level of correlation, to be expected as coordination often precedes action, these results merit more detailed analysis.

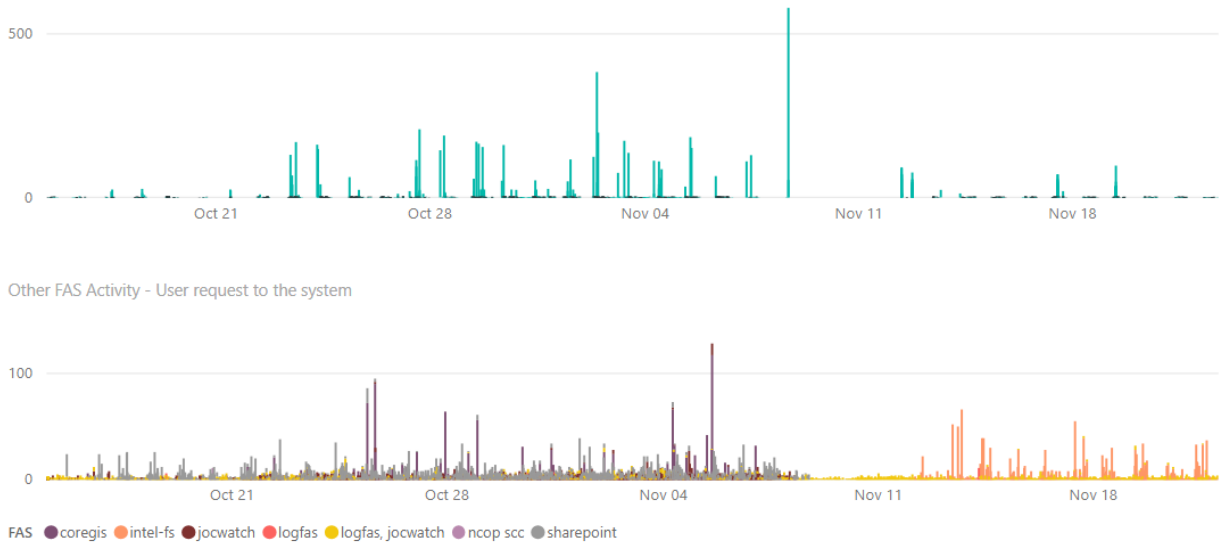

Figure 3: Activity levels for collaboration and other functions during the exercise

2.2 Analysis of information flows in IntelFS

Initial analysis of information flows in Intel FS focussed on requests for information and the associated response. Figure 4 shows the main sources of intelligence requests for information (RFI), the main responders and the routing. It is unsurprising that the main source of responses is the Joint Warfare Centre, which is responsible for running the exercise. Figure 4 also shows that the majority of requests are routed through the Joint Force Command in Naples (JFCNP), or through its Mission Information Room (JFCNP-MIR). On the right of Figure 4 are also shown the prime sources and consumers of RFIs.

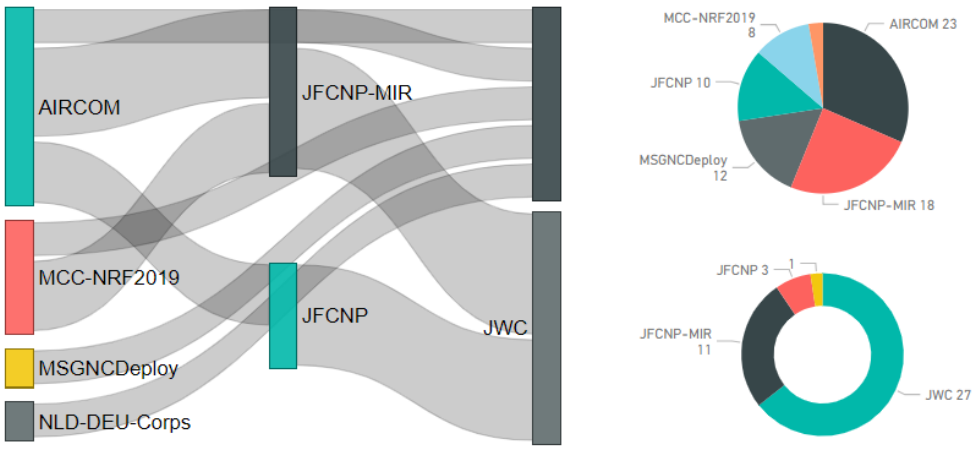

Figure 4: Information flows in IntelFS

Natural language processing (NLP) and machine learning (ML) techniques were used to process a number of sources of text data including intelligence reports and requests for information (RFIs), JChat conversations and Exercise events. NLP was able to extract entities e.g. locations, people, organisations and times. From these entities, independently extracted from exercise information in many different systems, it was possible to recreate events which occurred during the exercise and which included the same entities (location) and actors (individuals, organisations) all being reported within the same time period. These are visualised in a dashboard of Figure 5, where source reports are shown in the top left, content of a selected report is shown in the top right, relationships between entities – such as the location and people – are shown in the bottom left, linked by bars which are proportional to the strength of relationship identified in the data. The bottom right corner shows all entities extracted from the selected documents. In Figure 6 we can see the same entity (Arvidsjaur) extracted from conversations in JChat collaboration service during the same time period.

Figure 5 shows entities identified by entity extraction and machine learning techniques as being relevant at a particular time. At this time, the exercise featured an event at Arvidsjaur airport involving a number of key actors shown in the diagram. This demonstrates the ability of machine learning to find entities and relationships of relevance to assess – or conduct – the exercise.

Figure 5: Named entity extraction Intelligence system

Figure 6 Named Entity extraction from collaboration service (JChat)

3.0 MINING TEXT FOR LESSONS LEARNED

Manually extracting lessons learned from such a huge exercise is a resource intensive activity. Data science tools offer the possibility to handle datasets with large volume and variety, to speed up the process and to sift through large quantities of material, highlighting items worthy of further investigation by lessons learned analysts.

In addition to the FAS data described above, exercises generate large quantities of text, from formal documents, email and text collaboration tools. Text data from Trident Juncture has also been analysed, taking data from a number of sources:

- Documents (from SharePoint)
- Emails and email attachments
- Skype for Business (chat) conversations
- JChat conversations

Results from analysis of classified data are not suitable for publication in this format. But similar text analytics has been performed on unclassified text documents from the NLLP.

3.1 Extracting Lessons Learned from reports

NCIA, in conjunction with the JALLC, compared two toolsets for text mining, analysis and visualisation. These were IBM's Watson Explorer (WEX) and KNIME Analytics platform coupled with Microsoft Power BI. The dataset for this activity comprised 397 lessons learned reports from the unclassified NATO Lessons Learned Portal (NLLP). The majority of these documents were PDF reports.

In workshops held between JALLC analysts and NCIA data scientists, several data science techniques were identified to help analysts extract valuable information from this large document collection. These covered

NLP (Natural Language Processing), ML (Machine Learning) and advance analytics techniques, with the most useful being:

- Sentiment analysis [5] of documents – help analysts to identify Lessons Learned (strong negative sentiment) and Best Practices (strong positive sentiment)
- Topic and trend Analysis for selected entities
- Entity recognition – help analysts to identify the most frequent entities (people, organizations and locations) over period of time and documents
- Extraction of common phrases
- Relationship identification between entities and terms in documents

The initial task was to identify key topics and trends in lessons learned reports. This was done using a combination of entity extraction, custom dictionaries related to sentiment around lessons learned as well as organisations and locations of interest. Results were visualised to show prevalent terms over time; with the ability for lessons learned analysts to find individual references in documents in order to support their SME-driven assessment. Results from WEX are shown in **Error! Reference source not found.** and from KNIME, visualised in Power BI, in Figure 8.

Error! Reference source not found. shows the distribution of documents over a selected period of time for a list of topics which can represent entities, phrases, keywords and dictionary terms. The trend analysis shows sharp and unexpected increases in frequency of the topics over time. The trend index (the dotted line) measures the ratio of actual frequency of a topic compared to the expected average frequency based on the previous data. This expected change in frequency is estimated by using a modified Poisson distribution and can quickly highlight anomalies for a number of topic values on a single window, giving lessons learned analysts a clear indication of changes in importance of specific topic e.g. locations, activities, organisations etc.

Figure 7 Topic and trend analysis (IBM WEX)

In **Error! Reference source not found.** the grey bars show the actual frequency of a topic, unexpectedly high frequencies are highlighted in yellow. A given period can be selected (here highlighted in blue) and then

documents from that period will be displayed on the right of the screen, with the current topic/term highlighted (in green). In Figure 7 we can see described scenario for selected topic e.g. “deployment training”. This provides a rapid mechanism to identify periods of interest within the dataset for a given topic, and to rapidly home in on the relevant documents and text.

Figure 8 shows a trend analysis of terms associated with best practices (top) and lessons identified (bottom) across 15 years of lessons learned reports. The bar charts on the right of this figure show simplistic terms which indicate a lesson identified or a best practice to be adopted e.g. “maintain”, “facilitate”, “need” etc. The bar chart on the far right shows the documents with the greatest prevalence of such terms. Without data science tools and custom dictionaries tracking such trends and identifying key documents would be almost impossible on such a document set.

Although the visualisation appears superficially different, the more interesting aspect to both the data scientist and lessons learned analyst is the underlying information feeding that dashboards. Both tools were able to automatically extract valuable insights, such as entities, expressions and phrases accompanied with frequency and relevance measure. Apart from using custom dictionaries, which help analysts to focus on area of interest, tools were able to extract new knowledge and insights not yet known to analyst through functions for open information extraction such as subject-verb-objective analysis.

The ability to perform and display such analysis across a dataset of over 400 documents spanning many years is difficult for human analysts alone to perform. Data science tools provide an ability to search (and remember) volumes of material which humans alone cannot, to identify “hot topics” and to display trends. All of these functions bring additional value to the lessons learned process.

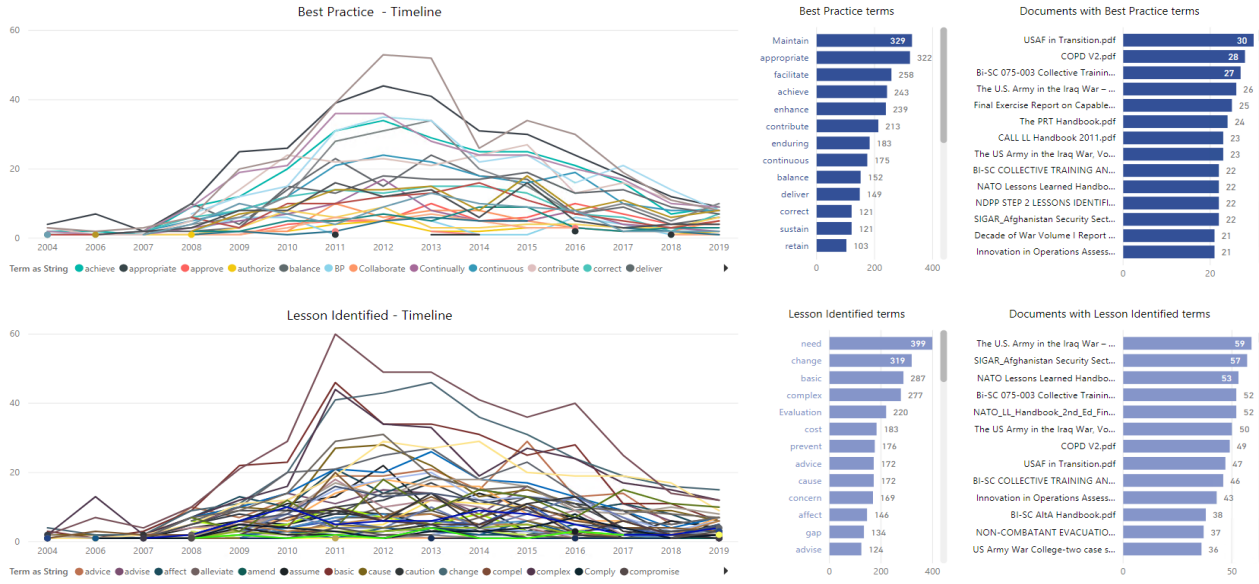

Figure 8: Topic and trend analysis (KNIME & MS Power BI)

5.0 CONCLUSION

Big data is often referred to as having four “V” characteristics, of having volume, variety, velocity and veracity, which are each too great for a human analyst to address. The datasets described in this work certainly demonstrate high volume and variety. The initial work described in this paper shows that data science tools and techniques have the ability to process large datasets – such as document sets spanning several years, or

exercise datasets of terabytes;- and to extract value from them – such as identification of trends and relationships of entities as well as numerical analysis at scale to show information flows.

Data collection and preparation has been a challenge in this area, as in almost every application of data science. But the time taken to prepare data for machine learning and analysis is low, compared to the impact on the lessons learned process. The ability of data science tools to process a variety of structured (databases), semi-structured (logfiles) and unstructured (email, documents) data and apply comparable techniques to them also provides a new approach to the extraction of value from large datasets.

Initial results of combining different data sources from an exercise (cross-domain analysis), result in more valuable insights and reveal connections and collaboration between different units which may be difficult to identify without big data analytics of all exercise data. Examples of this include the relationship between collaboration and intelligence activities in the exercise, or the automatic extraction of entities and their relationships which can reproduce events which occurred during the exercise.

Data science tools to identify and visualise key terms within large document sets have also been demonstrated to provide value to the lessons learned community, discovering and displaying trends in lessons identified and in best practices employed.

However, further work is needed in this area, for example the use of different languages for collaboration noted in section 2.1 does not itself prove the need for machine translation tools in NATO exercises unless further analysis of entities extracted from non-English conversations shows that the same exercise-related subjects are being coordinated in multiple languages. Additional work is also planned to extend the initial analysis of lessons learned material to operate on multi-word concepts rather than just on single words.

Further work is planned in this area as the authors continue to refine the data science techniques, tools and workflows needed to extract maximum value from large datasets in order to improve the effectiveness of NATO's existing activities and apply data science to better exploit NATO data.

ACKNOWLEDGEMENTS

This paper and the work within it would not be possible without the support of many colleagues at NCIA to collect the data from Trident Juncture, or the support of colleagues at JALLC, in particular Ms Jackie Eaton.

REFERENCES

[1] Street M., Lenk P., Ilic-Mestric, I. and Richter P,. "Lessons learned from initial exploitation of big data and AI to support NATO decision making", STO experts meeting on "Big data and AI to support military decision making", Bordeaux, 2018.

[2] M. Richter, M. Street & P. Lenk, "Deep Learning NATO document labels: a preliminary investigation", Int. Conf. on Military CIS, Warsaw, May 2018

[3] Kok A, Ilic Mestric I and Street M. "Named entity extraction in a military context", Int. Conf on Military CIS, Budva, May 2019.

[4] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kotter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, "KNIME: The Konstanz Information Miner," in Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 2007.

[5] R. Blunt, C. Riley, M. Richter, M. Street, D. Drabkin, "Using data analytics and machine learning to assess NATO's information environment", IST-160 specialists meeting on *Big data and artificial intelligence for military decision making*, Bordeaux, May 2018.