Identifying Changes in the Cybersecurity Threat Landscape Using the LDA-Web Topic Modelling Data Search Engine

Noura Al Moubayed[1(✉)], David Wall[2], and A. Stephen McGough[1]

[1] School of Engineering and Computing Sciences, Durham University, Durham DH1 3LE, UK
{noura.al-moubayed,stephen.mcgough}@durham.ac.uk
[2] Centre for Criminal Justice Studies, School of Law, University of Leeds, Leeds, UK
d.s.wall@leeds.ac.uk

Abstract. Successful Cybersecurity depends on the processing of vast quantities of data from a diverse range of sources such as police reports, blogs, intelligence reports, security bulletins, and news sources. This results in large volumes of unstructured text data that is difficult to manage or investigate manually. In this paper we introduce a tool that summarises, categorises and models such data sets along with a search engine to query the model produced from the data. The search engine can be used to find links, similarities and differences between different documents in a way beyond the current search approaches. The tool is based on the probabilistic topic modelling technique which goes further than the lexical analysis of documents to model the subtle relationships between words, documents, and abstract topics. It will assists researchers to query the underlying models latent in the documents and tap into the repository of documents allowing them o be ordered thematically.

1 Introduction

Changes in the cybersecurity threat landscape, especially with regard to the impact of cloud technologies on cybercrime are generally very hard to detect and they are made more difficult by the fact that cybersecurity threat reports tend to focus upon proprietary information for which the source information is rarely if ever shared. An alternative method of detecting change in the modern cybersecurity threat landscape is to analyse contemporary news and information sources. Not just a few sources, but tens or hundreds of thousands over a period of time using advanced topic modelling techniques. Such an open source technique is neither exhaustive nor new, but this original take on the technique does provide new prospects for identifying thematic, quantitative and qualitative changes in the cybersecurity threat landscape that can be used to efficiently begin a research enquiry.

Rapidly developing technologies have generated huge benefits to businesses, society and organisations of all sizes. These include technologies such as the world wide web, social media, the internet and Cloud computing. They offer

T. Tryfonas (Ed.): HAS 2017, LNCS 10292, pp. 287–295, 2017.
DOI: 10.1007/978-3-319-58460-7_19

rapid access to information and a mechanism by which people can interact with each other, or with businesses, in a way which has hitherto been unimaginable. However, such an utopian change in the way we interact has brought with it major security threats such as fraud, network infrastructure attacks, and data breaches. Such problems are exasperated by the large volumes of available data. Which provides a greater incentive for the attackers but, more significantly for those who seek to identify such attacks, vastly increases the problem of sifting through all of the available data in order to identify those pertinent facts which are required to understand the situation. For example the UK Crime Statistics Website [1] shows an average of 500,000 reported crimes per month for England and Wales alone. If we seek to identify common attacks, be they against similar victims or similar modes of operation this becomes ever more complicated as no individual could hope to keep abreast of all this data.

Conventional approaches to this problem allow searching of the corpus of collected data for specific words which are likely to relevant. However, as much of the data is collected by many untrained operators who share no common grounding the words which are used to describe things need not be the same. An ontology, where words which can be used to mean the same thing are linked together, can be used to expand the search. Though the effectiveness of this approach is diminished by the fact that these ontologies are rarely complete and require significant effort to compile.

Instead, here, we propose the use of topics as a searching mechanism. Rather than identifying keywords, which are themselves a diminished view of the topic that a searcher is trying to find, we instead allow the searcher to identify information based on the topics that they are interested in. A user may provide a document, or small collection of words, which represent one or more topics. These topics can then be searched for within the corpus. As the construction of what the relevant topics are, and which words would be indicative of a particular topic, would be a time-consuming process for humans to perform we instead use computer software to automatically identify those topics which are present in the entire corpus. This process identifies those words which probabilistically are most likely to form a topic. The same process can then be used to identify the topics present within a new document. This can then be used to identify the other documents within the corpus which are most likely to share similar topics and hence be of interest to the searcher. It should be noted that the size of the document that is used for searching need not be extensive in size.

Each document within the corpus will, in general, relate to more than one topic, with the proportion of the document relating to each exhibited topic being identifiable. Likewise each word which is used, anywhere within the corpus, will have an associated probability for being part of each topic. By this means we can take any new document and obtain the probability of each of the pre-identified topics being present within this new document. Using this we can rank the topics present in the document. It is then a simple mapping exercise to identify those existing documents within the corpus which share the most similar topic make-up as the new document. Likewise we can search on specific topics within the corpus.

As well as identifying other documents within the corpus which are similar to the current document we can also perform topic filtering across the corpus by which topics which we have identified to not be of interest can be removed. In which case any document in the corpus which consists of more than a pre-defined proportion of that topic can be removed. Thus allowing a reduction of the corpus.

This provides a powerful toolkit for the cybersecurity expert allowing them to rapidly track changing information and follow important leads through the ability to search, summarise and understand large volumes of data.

Our work is developed around a novel search engine based on probabilistic topic modelling to help gain an insight and retrieve information from cybersecurity textual datasets. Unlike traditional search engines our tool is customised to train a probabilistic model that fits the dataset. The model transforms the data from the textual space to the more abstract topics' space. This translates into the ability to probe the dataset by the more abstract concepts rather than key words. Each topic within the dataset is colour coded which can be used to categorise and directly access the documents involved. The topic itself is represented as a word cloud – in which the more commonly used words within the topic are displayed in larger sizes – facilitating the understanding of what constituents the concept of that topic. By grouping the documents by relevance the search engine facilitates navigating through large datasets of documents through added functionalities such as finding similar documents, discovering trends, identifying anomalies and finding most/least common topics.

Although our work here is focused on the analysis of cybersecurity and criminal documents this approach is not in any way dependant on the underlying type of data being processed and can therefore be re-applied to any other corpus of data or even applied to non-textual data.

The rest of this paper is structured as follows. In Sect. 2 we discuss the probabilistic topic modelling approach used in this work. Section 3 presents an overview of the framework we have developed for our tool. We conclude the ideas and consider future directions in Sect. 4.

2 Methods

Topic modelling is an established method for text mining that goes beyond the traditional lexical level of text analysis to try to understand the concepts which the text is conveying. These models are usually built as generative probabilistic models within the framework of Bayesian networks [2]. In this framework the model describes how observed variables (words in our case) can be generated by realisation of random variables arranged in a certain network. This allows for the modelling of mixture of models, i.e. each document is assigned a probability of it belonging to a topic, and a document may indeed contain more than one topic. A word can belong to more than one topic and a topic will contain more than one word. The model defines the probabilities governing the relationships between words, documents, and topics.

Latent Dirichlet Allocation (LDA) [3] is the most commonly used method for probabilistic topic modelling. Intuitively, LDA identifies the topic structure in

documents using the co-occurance structure of words/terms. LDA assumes that there are k underlying topics responsible for generating the documents in the dataset, and that each topic is represented as a multinomial distribution over the words in the vocabulary. Each document is assumed to be generated by sampling a finite mixture of these topics and then sampling words from each of these topics.

Each document is assigned a feature set where each value in the feature set represents the probability that the document contains a given topic. These features help group and identify the documents and facilitates the ease of use for the end user. To visualise the topic model and to present it in a user-friendly manner to the end user several visualisation systems for topic modelling have been built [4–6,8]. However, these systems focused on browsing the model and demonstrating the inter-connections amongst the documents, topics, and words. The systems, however, are heavily focused on the model or the metadata of the documents in the corpus without any facilities to search, rank, or access the documents within a topic in a direct and accessible manner. LDAVis [7] is a visualisation tool which provides a compact representation of the probabilistic topic model.

In this work we refocus the visualisation effort from the structure of the LDA model to accessibility of information to the end user. The emphasis is then not on how the user navigates the model but rather how the end-user can capitalise on the modelling capabilities of LDA to find efficiently related documents to a topic of interest, which is particularly useful for a cyber-security expert given the size of the data and the urgency to minimise the response time to a given threat. To achieve this goal a (topic modelling) search engine is built to categorise the documents based on their topics. The end users submit their queries to the search engine which in return compares the documents in the dataset with the query based on the topic features extracted from the input. The search engine is built within a web-based framework which provides tools to easily navigate through the entire corpus of documents easily.

3 Framework

The framework starts by colour coding the topics and presenting them as squares on the top of the web page as demonstrated in Fig. 1. By hovering over the coloured topics the framework displays the constituent words as a word cloud with the size of the word representing the probability of it appearing in that topic. The user has the option of either entering a query in the form of free text or uploading a document; alternatively the user may choose to navigate through the topics. For the purpose of demonstrating the functionality of the framework we used a cybersecurity related dataset collected by experts in the School of Law at the University of Leeds.

The framework provides the following main use case scenarios:

– Search: The trained model extracts topic features from the user search query and compares those against the features in the dataset and retrieves the most relevant documents. In the results page, each retrieved document is presented

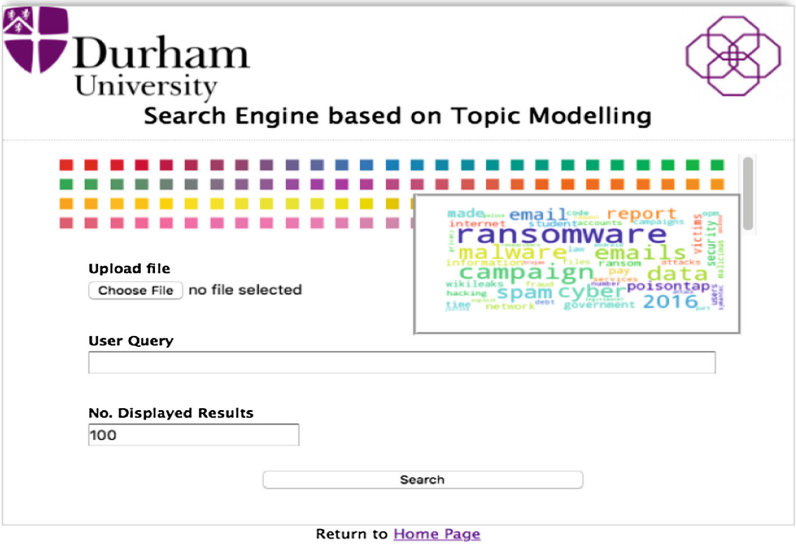

Fig. 1. The home page of the search engine. The topics are colour coded. Each square represents a topic. By hovering over a topic a word cloud is presented to demonstrate the main words within the topic. (Color figure online)

as a link to the original source, similarity measure to the query, colour coded topics that appear in the document, and a pie chart to demonstrate the relative contribution of each topic in the document (Fig. 2).

- Topic Navigation: By clicking on a topic square presented on the top of the page, the framework will display all the related documents ordered by relevance to the chosen topic (Fig. 3). The relevance is measured by the probability of the document being generated by that topic.
- Topic Filtering: to narrow the search space the user can eliminate documents that belong to one or more irrelevant topic(s). Figure 2 demonstrates the results after filtering two topics, which are marked by a red frame around their colour block on the top of the page.

All these scenarios are only possible thanks to the interactive web-based interface of the framework. They allow for direct and easy access to a huge dataset of the corpus collected from a wide range of variable sources. This is particularly interesting for cyber security experts who are dealing with continuously increasing datasets.

The framework also integrates the LDAVis tool [7], which helps visualising the inner contents of a given topic. After building an LDA model, it is passed to a custom webpage where LDAVis is utilised. Figure 4 demonstrate the first view of LDAVis, where the topics are plotted on the left panel as circles with the diameter of the circle is proportional to the probability of the topic to appear in any document. In the right panel the 30 most repeated words (terms) in all the documents are presented. To help navigate the LDA model, by clicking on

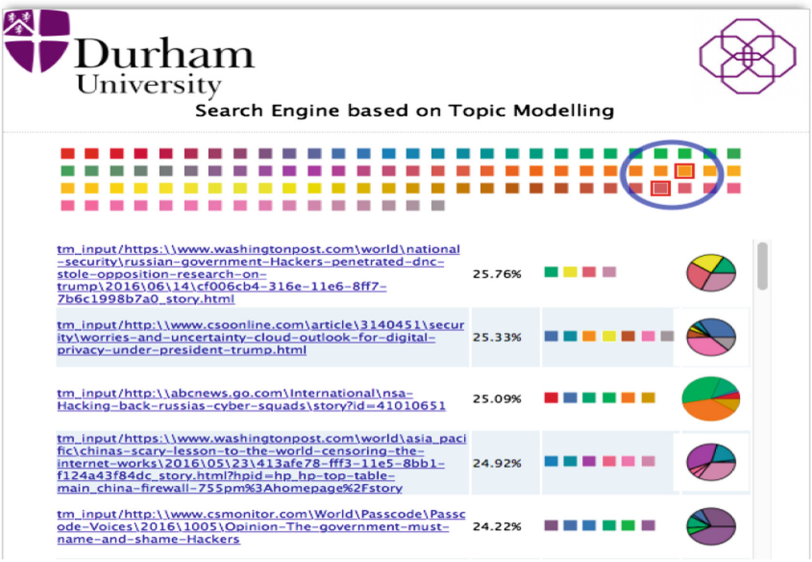

Fig. 2. Demonstration of topic filtering. Each document is associated with coloured squares representing topics. By clicking on a topic, all the associated documents are removed from the page and the topic is marked as deleted. (Color figure online)

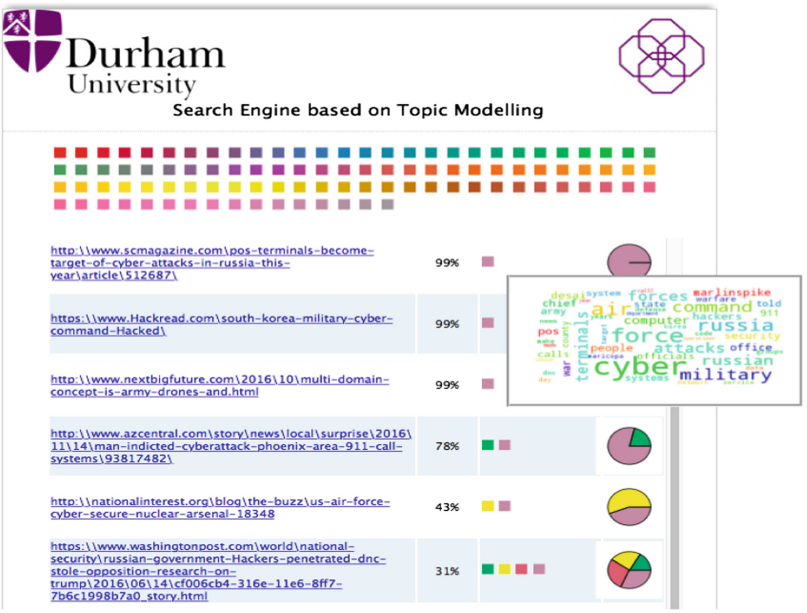

Fig. 3. Demonstration of Topic Navigation. When a topic square is clicked, all the associated documents are drawn. For every document the probability of it belonging to the topic is shown along with the other topics mentioned in the document.

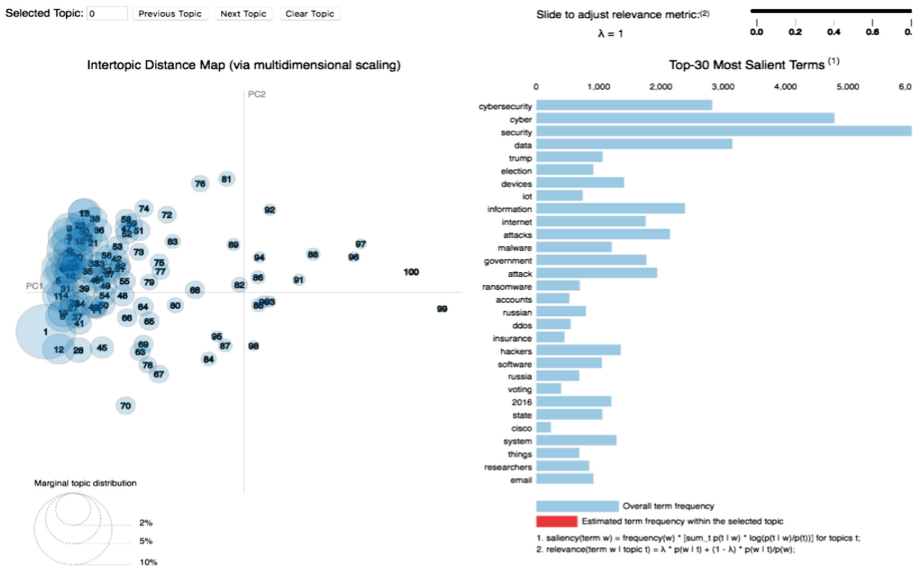

Fig. 4. The layout of LDAVis, with the global topic view on the left, and the 30 most common terms in all the corpus on the right presented as bar charts.

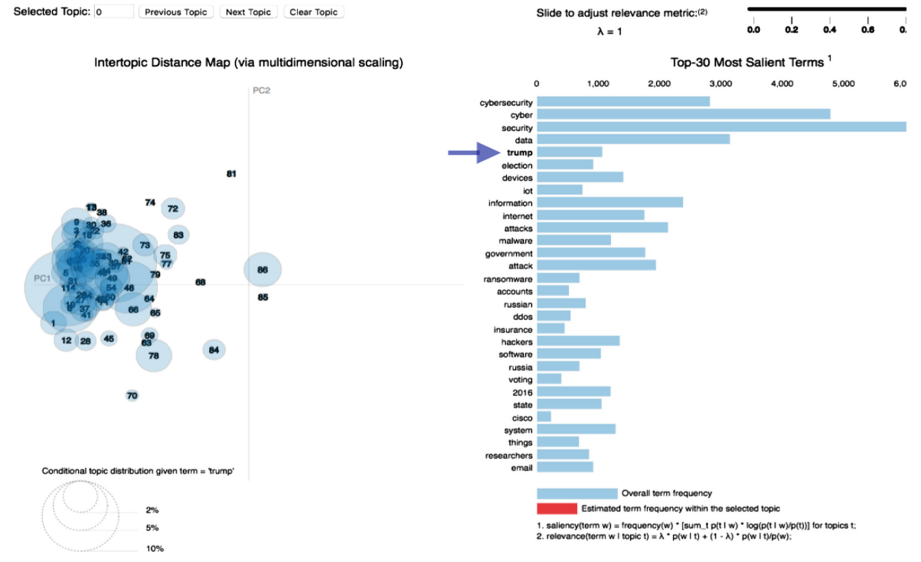

Fig. 5. Once a term is selected on the right panel, all the unrelated topics are removed from the left panel.

a word all the topics which are not related to this word, i.e. the word does not appear in those topics, are removed from the left panel (Fig. 5). By selecting a topic from the left panel the top repeated words in this topic are presented on

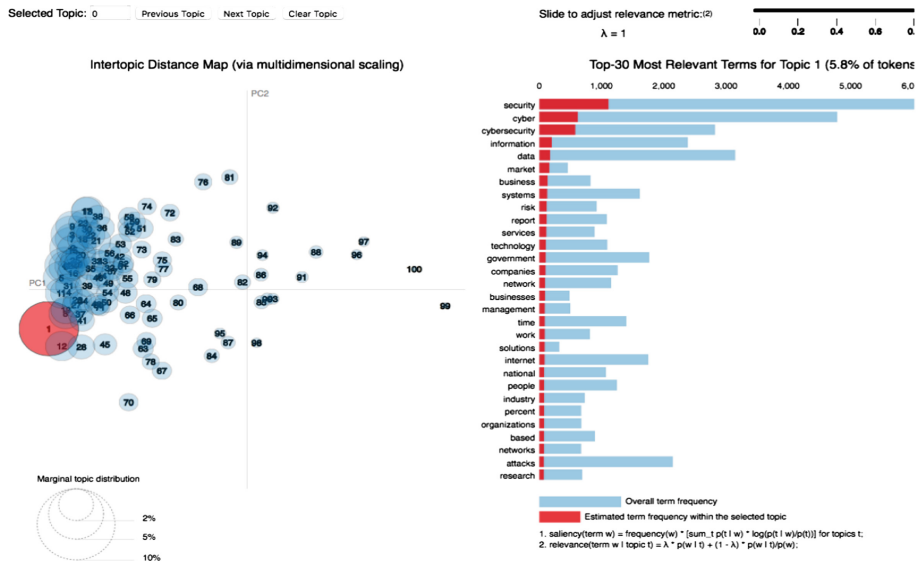

Fig. 6. If a topic is chosen on the left panel, the top 30 words in the topic are plotted on the right panel. The red bar is the number of times the word appeared in the documents within the selected topic, while the blue bar is the overall appearance in all the other topics. (Color figure online)

the right panel but now with added red bar to show the repetition of every word within the documents in the selected topic versus the appearance of the word in all the documents in the corpus – the blue bar (Fig. 6).

The combination of the dynamic search engine capabilities provided by the framework and the static visualisation of the LDA model using LDAVis, gives the security expert a great deal of in-depth knowledge to navigate through the documents easily and efficiently to find the necessary information.

4 Conclusion

The LDA-Web modelling tool technique provides a way to detect changes in the cybersecurity threat landscape, not least by it detecting anomalies in the flow of information. It facilitates the categorisation and summarisation of large unstructured cybersecurity text data. The tool is based on a text mining approach which models complex inter-relationships between words, documents, and abstract topics. The tool aims at providing security experts with full accessibility and functionality and helping them in their investigations.

LDA-Web facilitates the categorisation and summarisation of large unstructured cybersecurity text data. The tool is based on a text mining approach which models complex inter-relationships between words, documents, and abstract topics. The tool aims at providing security experts with full accessibility and functionality and helping them in their investigations.

Topic modelling using Latent Dirichlet Allocation (LDA) is used to model the documents within a large cyber-security related database of unstructured text documents obtained from a wide range of resources. The generated model is presented through our web-based interface and search engine which facilitate easy navigation through the topics and their constituent words. The search engine have the ability to find documents based on keywords, or similarity to an uploaded document. The results are ranked according to their similarity to the queried topic(s).

A third party tool, LDAVis, is also incorporated to provide in-depth understanding of the inner structure of the LDA model. The tool provides easy and user-friendly navigation functinalities to show the words within a topic and the relationship between topics and words. This is a very helpful tool for the security expert to be able to evaluate the quality of the built model and to fine tune their search queries passed to the search engine.

The future work will focus on evaluating the framework by cyber security experts and the feedback will be analysed using the A/B testing methodology. We are also working on collecting more cyber-security related data in order to make our model and search engine more accurate.

We see this as an excellent tool for the analysis of vast corpuses of documents in order to identify interesting and relevant data in a way hitherto difficult to do based on normal searching methods. As such we aim to apply this technique to different datasets in order to identify interesting phenomena which we can then work with subject specific experts in order to identify the significance of the data observed.

Acknowledgment. This work is part of the CRITiCal project (Combatting cRiminals In The Cloud - EP/M020576/1) funded by the Engineering and Physical Sciences Research Council (EPSRC).

References

1. Crime states. http://www.ukcrimestats.com/National_Picture/
2. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Chaney, A.J.B., Blei, D.M.: Visualizing topic models. In: ICWSM (2012)
5. Chuang, J., Ramage, D., Manning, C., Heer, J.: Interpretation and trust: designing model-driven visualizations for text analysis. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 443–452. ACM (2012)
6. Gardner, M.J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., Seppi, K.: The topic browser: an interactive tool for browsing topic models. In: NIPS Workshop on Challenges of Data Visualization, vol. 2 (2010)
7. Sievert, C., Shirley, K.E.: Ldavis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 63–70 (2014)
8. Snyder, J., Knowles, R., Dredze, M., Gormley, M.R., Wolfe, T.: Topic models and metadata for visualizing text corpora. In: HLT-NAACL, pp. 5–9 (2013)