Detecting Trending Terms in Cybersecurity Forum Discussions

Jack Hughes[1] **Seth Aycock**[2] **Andrew Caines**[1] **Paula Buttery**[1] **Alice Hutchings**[1]

[1] Computer Laboratory, University of Cambridge, U.K.
firstname.lastname@cl.cam.ac.uk
[2] Theoretical & Applied Linguistics, University of Cambridge, U.K.
seth@manx.net

Abstract

We present a lightweight method for identifying currently trending terms in relation to a known prior of terms, using a weighted log-odds ratio with an informative prior. We apply this method to a dataset of posts from an English-language underground hacking forum, spanning over ten years of activity, with posts containing misspellings, orthographic variation, acronyms, and slang. Our statistical approach supports analysis of linguistic change and discussion topics over time, without a requirement to train a topic model for each time interval for analysis. We evaluate the approach by comparing the results to TF-IDF using the discounted cumulative gain metric with human annotations, finding our method outperforms TF-IDF on information retrieval.

1 Introduction

Underground hacking forums contain a large collection of noisy text data around various topics, with misspellings, changing lexicons, and slang phrases. The evolving domain-specific lexicon includes homonyms, where "rat" may be identified as an animal by off-the-shelf tools, but is typically defined as a "remote access trojan" in this context, a type of malware used to gain access to a victim's computer.

We work with texts from the HackForums site[1], the largest English-language hacking forum, with multiple bulletin boards arranged around various topics, and many active users submitting thousands of new posts every day. The dataset contains over a decade of text data, but detecting trends is non-trivial, due to the informal language used by members, not only technical terms such as "rat", but also misspellings, slang, orthographic variation, and acronyms.

For instance, the following texts demonstrate how posts are structured into threads on given topics, and how users both deliberately and accidentally use noisy language[2]:

User1	Ransomware infects hospitals all over UK: `link`
User2	anyone think they made some money from this?
User1	They might of done but idk they'll get caught eventually, it's stupid to commit crimes like this
User3	Who tf targets hospitals for ransomeware
User1	I dont believe they actually went for the nhs.. the ransom would be more $$$ lol
User4	I looked up a few btc addresses and can confirm they made money

Researchers interested in analysing hot topics on the forum will find it hard to gain a clear perspective on this due to the volume of data going through the forum every day. Therefore, an overview of trending topics with natural language processing and statistical techniques is useful for identifying what may be of interest to security researchers. We propose a tool to identify tokens from trending topics, by pre-tokenising post data, followed by adapting a statistical technique for measuring changes, which can be used to scan across the dataset.

The tool builds upon a weighted log-odds ratio (Monroe et al., 2008) with an informative Bayesian prior (Silge et al., 2020), used to compare differences in two corpora. In our case the corpora represents two distinct time periods of interest within the same subforum[3]. For known events, one period can be a set of texts preceding the event (the

[1] https://hackforums.net

[2] The texts are fabricated so as to preserve user anonymity, but they are based on real ones we have encountered in the database.

[3] A forum is the whole site, and a subforum or bulletin board is a page on the site, dedicated to a given general topic and created by the administrators. Subforums contain member–created threads consisting of an ordered set of posts typically focused on a single topic.

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 107–115

prior) and the other period can be texts following the event (the target). Also, the tool can be used for live listings of trending terms in the present day, by comparing new posts against some fixed prior.

Our method identifies the relative importance of tokens to each time period. The log-odds ratio indicates whether terms are more likely to appear in a given corpus over others. A log-odds score is higher for terms that are both unique and more frequent to a given period. Other NLP methods require the removal of pre-defined stopwords. However, for our approach, as stopwords have a similar distribution across both time periods, they will have a low log-odds score and rank.

The tool looks at "bursty" events: for a token to be trending, frequency of the token should be significantly different between the prior and target periods, and be more frequent than other terms in the target period. For identifying topics, our method uses a feature-pivot approach (a topic is a cluster of keywords) over a document-pivot approach (a topic is a cluster of documents). The latter may struggle with documents about multiple topics, whereas the former may incorrectly identify correlations between words as topics.

A major challenge in developing the tool is that it is to be used on a large dataset of noisy data, for exploring the evolution of underground hacking forums. We take mitigating steps such as storing the pre-tokenised and part-of-speech tagged text, to decrease computation time for longitudinal analysis. While we focus on a cybercrime context, we note this type of data has similarities to Twitter data: short posts, and informal language. However, while Twitter data has some minimal inter-tweet connections through hashtags, quoting comments and replies, forums have a rigid discussion-based structure set by the forum administrators.

Our contributions are:

- We adapt a technique used for capturing the linguistic changes between two corpora, to be used as a trending topics tool for temporal analysis of data.
- We show the application of this trending topics tool in the context of cybercrime research.

2 Related work

2.1 TF-IDF

Term-frequency inverse-document-frequency (TF-IDF) (Spärck Jones, 1972) identifies common terms in a document, but not common across all documents. This technique provides a mechanism for ranking tokens which are "important" to a document. However, forum text is noisy, with varying spelling of words and creative use of punctuation. While TF-IDF is a popular NLP technique, use on forum data would require stemming or lemmatisation, and defining a document either as individual posts, or a thread of posts, for best performance.

2.2 LDA

TF-IDF assumes each document is based on a single topic, although with forum data, posts and threads may discuss several topics. LDA (Blei et al., 2003) takes a different approach by assuming each document is built from a number of topics, with one primary topic, by learning a distribution of terms in topics. Similar to TF-IDF, this method also requires finding a suitable tokenisation approach and representation of a document. Also, while LDA learns a distribution of terms in topics, this is not as lightweight computationally as TF-IDF.

2.3 Trending Topic Techniques

TF-IDF and LDA are both commonly used, but these both have limitations, and improved models have been proposed.

Burst and dynamic topic models have been used for detecting trending topics, including a burst model proposed by Kleinberg (2003), and Takahashi et al. (2012); Koike et al. (2013) who combine Kleinberg's burst model with a dynamic topic model. While these approaches measure frequency changes over time to detect bursts, we use a different approach similar to "two-point trends" discussed by Kleinberg (2016) with "rising" and "falling" words. In addition, we use a Bayesian approach instead of measuring absolute change.

Aiello et al. (2013) explored common NLP methods for detecting trending topics on Twitter, related to major events which differ in time scale and topic churn rates, and suggest later work should look at topics evolving in parallel. They found n-gram co-occurrence (groups of words typically appearing in the same document), and DF-IDF$_t$ topic ranking (an adaptation of TF-IDF to look for common topics unique to a given time period in comparison to prior time periods) to perform the best. They also boosted the score of proper nouns in their approach, finding these are useful keywords for trending topics.

Follow-up work by Martin et al. (2015) detected bursts of phrases for a topic detection system, using DF-IDF$_t$ to group co-occurring bursty phrases, followed by topic ranking, using the apriori algorithm. They also look at windowing, where events which are focused on real-time activity (e.g sports) have a smaller window of activity, with greater topic recall than longer topics (e.g. politics) with discussions continuing after events. Super Tuesday (the Tuesday in which many US states hold their primary elections) performed better with fewer prior tweets as this was a longer event, than others which performed better with a longer window.

Previous research has focused on static snapshots of events, whereas Shamma et al. (2011) used temporal analysis to identify both peaky and persistent topics. Trending topics tools which are sensitive to noise may only detect peaky topics over persistent topics. They used normalised term frequency, with the number of tweets containing the word, rather than the number of times a word is used, and the peaks look at terms particular to an exact window of time. Persistence looks at peaks of normalised term frequency, assuming these terms have not been used before, and have been used more frequently afterwards.

While much of the literature focuses on detecting English-language trending topics, many cybercrime forums are not English-speaking, which can add complexity into analysis. Also, there are some cases where topic modelling may produce poor quality results, and could be refined with user feedback, which is explored by Hu et al. (2014) with iterating models (hierarchical-LDA trees).

2.4 Fightin' Words paper

Monroe et al. (2008) introduced a method for comparing lexical tokens used by two political parties. This uses a model-based approach, modelling terms as a function of political party, to compute the likelihood of terms used by a political party as log likelihoods ("log-odds"). They used an uninformative Dirichlet prior.

We adapt this method to time-based analysis, modelling terms as a function of time. We use an informative Bayes prior, which was used in the R tidylo library by Silge et al. (2020). While this method was initially used to compare two distinctly different political party news corpora, we adapt this to examine a longitudinal dataset to explore how a particular corpus has changed over time.

2.5 Named Entity Recognition

Our method uses a Bayesian approach to identifying trending topics, with filtering by noun phrases using a part-of-speech (PoS) tagger. However, an alternative approach may use named entity recognition to detect trending topics, and later, for extracting events from text. However, Caines et al. (2018) note named entity recognisers are trained on well-formed English text, and their performance is degraded with noisy text.

There has been prior work in using NER on noisy text, including with a shared challenge at W-NUT 2017 (Derczynski et al., 2017). One approach by Aguilar et al. (2017) used a convolutional neural network with both character-level and word-level features combined with contextual information, input into a bidirectional LSTM, for this task. Jansson and Liu (2017) also used a bidirectional LSTM for word and character embeddings, but combined these with an LDA topic model.

Additionally, contextual data can be used to assist with this task. Xing and Paul (2017) combined word embeddings with Twitter network and geolocation data to improve the accuracy of NER. While we do not have access to this type of data about HackForums users, the forum structure provides hierarchy with administrator defined subforums, which could be used as a feature to combine with embeddings.

2.6 Cybercrime trending topics

Work into trending topics in cybercrime has focused on identifying new threats, using data from tweets, blogs, and underground forums. This includes the creation of large-scale frameworks, such as Sapienza et al. (2018) who detect emerging threats across datasets, although this depends on annotations of known keywords. This is problematic for cybercrime research, due to the constantly changing lexicon.

Behzadan et al. (2018) released a tool to assist annotators in exploring Twitter data, with an annotated dataset of 21,000 tweets on cyber threats. However, this still requires manual identification of new terms.

Once a trending topic is identified, topic ranking is needed, to avoid overwhelming a user. This is used to highlight current important topics, including Bose et al. (2019) who use this to detect and flag known serious threats.

Also, other approaches such as PoS-tagging

and sentiment analysis have been used to identify threats, such as work by Macdonald et al. (2015), however there is a range of jargon used on the forum, with spelling variations and changes to meaning over time, which models would need to handle. There have been other approaches to look at trends on forums and marketplaces, including Tavabi et al. (2019) who use a large topic model to map the evolution of different forums as they evolve.

These communities also evolve over time, with changing meanings of words, and an evolving lexicon, which should be taken into account with longitudinal topic modelling. Bhandari and Armstrong (2019) have looked at subforums of Reddit to explore the use of high affinity terms used by communities, looking at how the semantics of these have changed.

3 Method

3.1 Data

For our method, we use the CrimeBB dataset from the Cambridge Cybercrime Centre (Pastrana et al., 2018b), available for researcher use from the Cambridge Cybercrime Centre[4]. CrimeBB contains posts scraped from 27 underground and dark web forums related to cybercrime, with over 13 years of post data. The database contains English, Russian, and German-language forums. Each forum is structured by subforums, which are based on general topics e.g. hacking methods or marketplace, and are defined by the forum administrators. Each subforum contains threads, which are an ordered collection of posts focusing on a defined topic set by the first post in the thread, such as a particular tutorial the author is sharing. Later posts can be providing a reply to the original first post, a reply to a later post by another user, or new information on the topic. While threads are typically focused on a particular topic, longer threads may become off-topic.

We selected HackForums from this dataset for our evaluation, which is an underground hacking forum discussing various aspects of hacking techniques. Our dataset contains over 190 administrator–curated subforums, with 4 million threads, and 42 million posts, created by over 630,000 members of the forum.

The method is selected due to the focus of the dataset: the data is "noisy", containing variations of spelling (e.g., "ransomeware" instead of "ransomware"), orthography (e.g., "NK" and "nk" for North Korea), and length of posts (ranging from short replies "pm me" to longer in-depth tutorials). In addition, due to the size of the dataset, our method requires a lightweight approach in order to measure the evolution of trends and topics over time.

3.2 Ethics

Ethics approval was granted from the department's ethics committee for this work. We used data collected from a publicly available forum, and could not gain informed consent from all members as this would be considered to be spamming. As we only analyse posts as a collective whole, rather than identifying individual posts, under the British Society of Criminology's Statement of Ethics, this falls outside of the requirement of informed consent. We also avoid publishing details that could identify individuals, including usernames and original post contents.

3.3 Tokenisation and pre-processing

We first remove chunks of the forum post text which are not the main content of posts, including quote, link, and code blocks. These are identified by using regular expressions to identify relevant markup blocks. This approach is specific to the dataset we use.

Secondly, we tokenise the lowercased forum post text, using TweetTokeniser in NLTK (Bird et al., 2009). This is suited to handling URLs and punctuation based emoticons in text. Note we do not remove a pre-defined list of stop words, however our Bayesian approach will decrease the relevance of a large number of very frequent words which appear equivalently in the prior and target texts.

Following this, we carry out PoS-tagging using spaCy (Honnibal and Montani, 2017) to identify nouns and noun phrases in posts, which we filter results by. Note that we do not apply this step before calculating log-odds, as this would change the distribution of tokens used in a period, affecting the quality of results.

We store both the token counts and set of nouns for each post in the forum. These are stored separately for each subforum in HackForums. Note that we do not attempt to merge terms which may vary in their orthographic form – for instance acronyms or abbreviations with their full forms, spelling errors, and casing differences. It remains a matter

[4] https://www.cambridgecybercrime.uk

for future investigation whether acronyms and abbreviations should always be associated with fully spelled-out forms, or whether they should be kept distinct because they represent different uses of the term. Secondly, we can introduce a spell-checker in future work to cluster misspelled words with their intended form, but this will need adaptation to the vocabulary of the cybersecurity domain. Finally, we do capture casing differences (e.g., "WannaCry" and "wannacry", and "NHS" and "nhs") because all texts are lower-cased before tokenisation.

3.4 Windowing: Prior and Target

The method requires the selection of two time windows: a prior and target period. The prior period is used to learn a distribution of terms used, as a comparator for the target period. The size and placement of windows can be varied depending on the desired results: long-term trend detection would have a longer, and more distant, target window than for short-term trend detection.

These windows should be selected depending on the dataset used and research questions. If the prior window and target window overlap the same event, then these terms will appear in both windows with a similar frequency, and will therefore have low log-odds. If the prior and target window are too far apart, then the prior may not be representative, leading to poor quality results. Also, if a topic is re-trending, and the previous trending period falls in the prior, then this may affect whether a term appears to be trending.

3.5 Overview of the log-odds method

Our approach uses a method implemented in the tidylo R library by Silge et al. (2020), which we have re-implemented in Python for compatibility with other tools. The tidylo R library uses an informative prior Bayesian approach, instead of the Dirichlet uninformative prior used by Monroe et al. (2008). A later version of the tidylo library added support for the uninformative prior. However, we chose to continue using the Bayesian approach as our time-based application of the tool is suited to using an informative prior.

We adapt this approach, created for comparing two corpora, to detect trending tokens. Instead of selecting corpora by pre-existing classes, we choose prior and target time windows, to find terms which are more likely to appear in the prior or target period. Each period is represented as a "bag-of-words", for all posts in the selected period.

This Bayesian approach is shown in the following series of equations, based upon the tidylo implementation.

For the corpus (combined set of posts in both periods) y, we define y_w as the frequency of token w, and y_{wi} as the frequency of the token w in period i. n is the sum of frequencies of tokens across all periods, and n_i is the sum of frequencies of tokens in the period i.

First, we calculate ω_{wi}, the odds of each token appearing in period i, and ω_w, the odds of each token appearing the corpus:

$$\omega_{wi} = \frac{y_{wi}}{n_i - y_{wi}} \tag{1}$$

$$\omega_w = \frac{y_w}{n - y_w} \tag{2}$$

Secondly, we calculate δ_{wi}, the log odds ratio to compare the usage of the token w in period i to the whole corpus:

$$\delta_{wi} = \log \omega_{wi} - \log \omega_w \tag{3}$$

Thirdly, we calculate the variance of our estimate, σ^2_{wi}:

$$\sigma^2_{wi} = \frac{1}{y_{wi}} + \frac{1}{y_w} \tag{4}$$

Finally, we calculate the log odds score ζ_{wi} for each token w in period i:

$$\zeta_{wi} = \frac{\delta_{wi}}{\sqrt{\sigma^2_{wi}}} \tag{5}$$

Depending on when the prior and target time windows occur, the tool will either pick up short or long term trending tokens.

4 Evaluation

We evaluate the results of the tool by carrying out an information retrieval task with human annotators. We compare the log-odds approach with TF-IDF using discounted cumulative gain and the human annotations as a ground-truth ranking of identified terms. We use both a known cybersecurity event to define our target window, as well as a randomly-selected target window.

4.1 Trending Event Selection

Within CrimeBB, we selected HackForums, as this is widely studied in prior cybercrime literature (Pastrana et al., 2018a,b; Bhalerao et al., 2019).

First, for the known event we selected the spread of WannaCry in the year 2017. WannaCry is a type of ransomware, which encrypts data until the victim pays a ransom. WannaCry spreads through vulnerable computer systems, instead of directly targeting specific entities, where these systems have not previously updated their systems to patch this issue. One of the largest organisations affected by this attack was the National Health Service (NHS), the universal public healthcare system in the UK. We selected this event as we anticipated it would have been extensively covered on the forum. Indeed, it was later revealed that the individual who was instrumental in stopping the spread of WannaCry had formerly been an active forum member (Krebs, 2017).

The incident within the NHS began on Friday 12 May 2017 (Smart, 2018), which we select as the start of the 7 day window for our analysis. We selected the "News and Happenings" subforum with a prior period of 2017-04-12 to 2017-04-18 and a target period of 2017-05-12 to 2017-05-18. The prior contains 404 posts, and the target contains 470 posts.

Secondly, we randomly selected a subforum, "Monetizing Techniques", and a random date range for the target (2016-12-23 to 2016-12-29 for the target, and a week in the previous month for the prior: 2016-11-23 to 2016-11-29). The prior contains 195 posts and the target contains 295 posts.

4.2 Log-odds and TF-IDF Results

We compare our approach to TF-IDF for topic ranking, using a similar approach to log-odds. This includes creating two TF-IDF "documents" as the set of posts for a given period (e.g. prior or target), as this is similar to the current method (frequent terms in the period but not frequent across all periods). We use the same tokenisation and pre-processing approach as the log-odds tool, to provide direct comparison. We selected TF-IDF, as it is a lightweight technique for topic ranking and detection.

For each event and technique, we plotted the top 10 tokens for the prior and target periods. For the "WannaCry" event, Figures 1 and 2 show the top tokens and scores for the prior and target periods. The results of the log-odds tool for the target period all contain tokens related to the WannaCry ransomware event. While TF-IDF also includes tokens related to the WannaCry ransomware event, it additionally contains terms related to different events (e.g., "notebook", "pirates", and "sharing"). Figures 3 and 4 show the top tokens and scores for the randomly selected event.

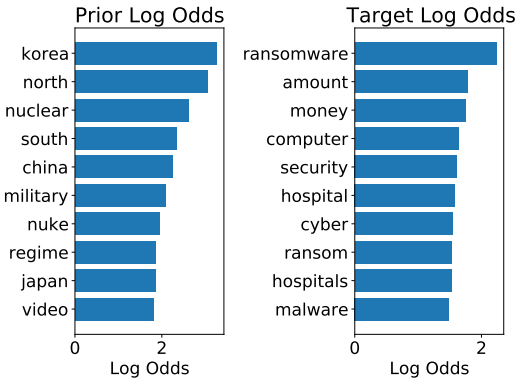

Figure 1: WannaCry Event with log-odds

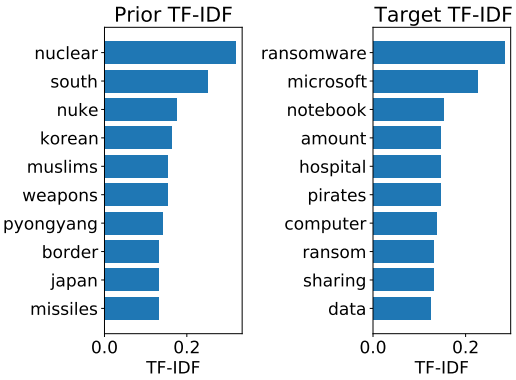

Figure 2: WannaCry Event with TF-IDF

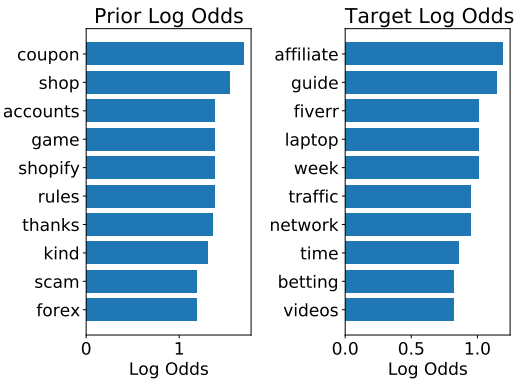

Figure 3: Random Event with log-odds

4.3 Annotation Task

First, we generated a list of ranked terms from both the tool and from TF-IDF, selecting the union of

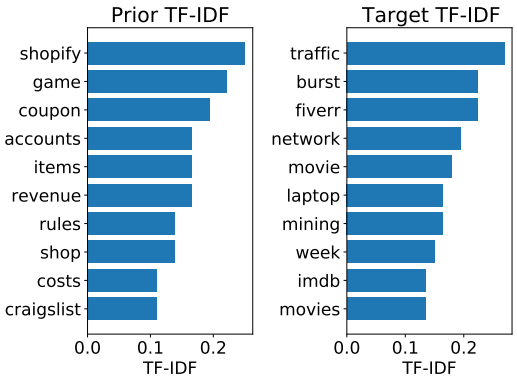

Figure 4: Random Event with TF-IDF

the top 10 terms of each event.

For the WannaCry event, these were: `amount`, `computer`, `cyber`, `data`, `hospital`, `hospitals`, `malware`, `microsoft`, `money`, `notebook`, `pirates`, `ransom`, `ransomware`, `security`, `sharing`.

For the randomly selected event, these were: `affiliate`, `betting`, `burst`, `fiverr`, `guide`, `imdb`, `laptop`, `mining`, `movie`, `movies`, `network`, `time`, `traffic`, `videos`, `week`.

For each event, we presented the three annotators with each post from the prior and target periods with the accompanying tags. The annotators selected the most salient tags for each post, leaving posts not annotated if there were no suitable salient tags. We measured inter-annotator agreement using multinomial Krippendorff's alpha with the MASI distance metric of sets (Passonneau, 2006) for comparison, finding an overall agreement of 0.833.

4.4 Discounted Cumulative Gain

Using our annotations combined using majority voting, we compared the ranking of the log-odds tool against TF-IDF, using normalised discounted cumulative gain (Järvelin and Kekäläinen, 2002). This is a metric used to evaluate the usefulness of a ranking of a list, by measuring the quality (salience) of tokens returned from the tool. We use discounted cumulative gain with the annotations of salient tokens, as the metric increases the weight of errors towards the top of the ranked list, compared to other rank correlation measures, such as Kendall's tau. Additionally, we do not have ground truth information on the ordering of all tokens.

For the WannaCry event, our log-odds tool scored 0.979 compared to TF-IDF of 0.877. For the random event, the log-odds tool scored 0.978 compared to TF-IDF of 0.753.

For both events, the log-odds tool had a greater discounted cumulative gain score than the TF-IDF approach, finding the ranking of terms provided by the log-odds tool produced more relevant salient terms than the TF-IDF method, for our forum dataset.

5 Discussion

Detecting trending topics on noisy social media data is not a new problem for information retrieval and NLP. However, we believe our application of an existing statistical method onto a longitudinal dataset provides a novel lightweight approach to detecting trending terms, which returns terms of more relevance than TF-IDF, and remains computationally less expensive than topic modelling such as LDA.

This work provided an initial step towards detecting temporal linguistic changes over time, by preprocessing text data, followed by using a Bayesian approach with a moving prior and target window depending on whether a user is observing short or long term trends. While our method does not identify the relevant windows itself, the tool can be combined with trending topic detection techniques to identify lexically distinct events, where some terms may re-trend.

Having shown that the statistical model is strong, and using a Bayesian approach can support new and evolving slang in the dataset without fine tuning a language model, we recognise that there are ways to further improve the NLP of cybersecurity forum texts. For instance, we can improve preprocessing in order to better deal with noisy texts: this includes the detection of misspellings, orthographic variation, acronyms and abbreviation, and deliberate obfuscation such as leetspeak. In addition, we can start to incorporate the detection of multiword expressions and named entity recognition techniques for noisy language, since both are likely to be of interest to researchers analysing language use in cybersecurity forums.

In future work we aim to increase understanding of the evolution of forums, changing language over time, and the changing topics of discussion by forum members. We also aim to automatically detect and extract events in the CrimeBB dataset. Although we have focused on analysing forum data, the tool can be used to explore trends in other cor-

pora. In future work, we plan to use this approach to analyse how spam emails have changed following the COVID-19 pandemic.

6 Conclusion

In this work, we presented a new use-case for the log-odds tool introduced by Monroe et al. (2008) and implemented in the tidylo R library by Silge et al. (2020), for detecting trending terms in longitudinal historical noisy text data of an underground hacking forum. The tool can be used for both detecting short term and long term trends depending on the time windowing and separation of windows selected. Using annotations of salient terms during both discussion of WannaCry, and a randomly chosen duration, we found our approach to produce more relevant salient terms over TF-IDF.

Acknowledgments

We thank the Cambridge Cybercrime Centre for access to the CrimeBB dataset. We also thank our colleagues at the Cambridge Cybercrime Centre. This work was supported by the Economic and Social Research Council (ESRC), grant number ES/T008466/1. The third and fourth authors are supported by Cambridge Assessment, University of Cambridge.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Goker, Ioannis Kompatsiaris, and Alejandro Jaimes. 2013. Sensing Trending Topics in Twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282.

Vahid Behzadan, Carlos Aguirre, Avishek Bose, and William Hsu. 2018. Corpus and Deep Learning Classifier for Collection of Cyber Threat Indicators in Twitter Stream. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5002–5007, Seattle, WA, USA. IEEE.

R. Bhalerao, M. Aliapoulios, I. Shumailov, S. Afroz, and D. McCoy. 2019. Mapping the underground: Supervised discovery of cybercrime supply chains. In *2019 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–16.

Abhinav Bhandari and Caitrin Armstrong. 2019. Tkol, httt, and r/radiohead: High affinity terms in Reddit communities. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 57–67, Hong Kong, China. Association for Computational Linguistics.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Avishek Bose, Vahid Behzadan, Carlos Aguirre, and William H. Hsu. 2019. A novel approach for detection and ranking of trendy and emerging cyber threat events in twitter streams. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 871–878, New York, NY, USA. Association for Computing Machinery.

Andrew Caines, Sergio Pastrana, Alice Hutchings, and Paula J. Buttery. 2018. Automatically identifying the function and intent of posts in underground forums. *Crime Science*, 7(1):19.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Patrick Jansson and Shuhua Liu. 2017. Distributed representation, LDA topic modelling and deep learning for emerging named entity recognition from social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 154–159, Copenhagen, Denmark. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Jon Kleinberg. 2003. Bursty and Hierarchical Structure in Streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Jon Kleinberg. 2016. *Temporal Dynamics of On-Line Information Streams*, pages 221–238. Springer Berlin Heidelberg, Berlin, Heidelberg.

Daichi Koike, Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, and Noriko Kando. 2013. Time series topic modeling and bursty topic detection of correlated news and twitter. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 917–921, Nagoya, Japan. Asian Federation of Natural Language Processing.

Brian Krebs. 2017. Who Is Marcus Hutchins? https://krebsonsecurity.com/2017/09/who-is-marcus-hutchins/.

Mitch Macdonald, Richard Frank, Joseph Mei, and Bryan Monk. 2015. Identifying Digital Threats in a Hacker Web Forum. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, pages 926–933, Paris, France. ACM Press.

Carlos Martin, David Corney, and Ayse Goker. 2015. Mining Newsworthy Topics from Social Media. In *Advances in Social Media Analysis*, volume 602 of *Studies in Computational Intelligence*, pages 21–43. Springer International Publishing, Cham.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. 2018a. Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum. In *Research in Attacks, Intrusions, and Defenses*, volume 11050, pages 207–227. Springer International Publishing, Cham.

Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. 2018b. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In *Proceedings of The Web Conference 2018*, Lyon, France.

Anna Sapienza, Sindhu Kiranmai Ernala, Alessandro Bessi, Kristina Lerman, and Emilio Ferrara. 2018. DISCOVER: Mining Online Chatter for Emerging Cyber Threats. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 983–990, Lyon, France. ACM Press.

David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. 2011. Peaks and persistence: modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, page 355, Hangzhou, China. ACM Press.

Julia Silge, Alex Hayes, and Tyler Schnoebelen. 2020. tidylo: Weighted Tidy Log Odds Ratio. https://github.com/juliasilge/tidylo.

William Smart. 2018. Lessons learned review of the WannaCry Ransomware Cyber Attack. Technical report, Department of Health and Social Care.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, Noriko Kando, Tomohiro Fukuhara, Hiroshi Nakagawa, and Yoji Kiyota. 2012. Applying a burst model to detect bursty topics in a topic model. In *Advances in Natural Language Processing*, pages 239–249, Berlin, Heidelberg. Springer Berlin Heidelberg.

Nazgol Tavabi, Nathan Bartley, Andres Abeliuk, Sandeep Soni, Emilio Ferrara, and Kristina Lerman. 2019. Characterizing activity on the deep and dark web. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 206–213, New York, NY, USA. Association for Computing Machinery.

Linzi Xing and Michael J. Paul. 2017. Incorporating metadata into content-based user embeddings. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 45–49, Copenhagen, Denmark. Association for Computational Linguistics.