LinkAlong

White Paper

Event Detection in **Peek**

June 2022

Weak Signals and Early Warning

Organizations are today constantly facing uncertainty, risks, and crisis. To address these challenges, reliable information is essential. Open-source content from News, Web or Social Media is one potential way to obtain such information.

Today's current tools are good in finding well-known and frequent information, e.g., information about an organization that we know or a content that is highly popular.

Typically, the most valuable information is unknown in advance.

It is also sparse, i.e., a weak signal, as it is not widely known in general, or it needs to be detected at an early stage before it is widely known for early warning. Therefore, such information is hard to detect, especially with current technology.

The Case Study: identify cyberattacks on healthcare

We illustrate the benefit of using recent advances in artificial intelligence (AI) for text content analysis in detecting weak signals. We conduct a case study on efficiently identifying new cyberattack incidents on healthcare related organizations. To perform this case study, we use our **Peek platform** that exploits neural networks for text understanding.

The evaluation shows that with a modest human effort, new events can be comprehensively discovered due to optimized support using AI. Compared to a traditional approach, a human analyst can extract this information comprehensively in much shorter time than by using a traditional methods of document search and analysis. We estimate a speedup by a factor of 50.

Methodology

Using the Peek platform, we collect with Peek data from Social Media and Web sources using two approaches:

1. Documents that are mentioning a term related to cyberattacks (e.g. cyberattack or ransomware)

2. Documents from sources known to produce regularly information related to cyberattacks

Using this approach, we collect in Jan – Mar 2022 and select documents that mention a notion related to healthcare (e.g., hospital, patient, clinic etc.). This results in a total of 26'693 documents, containing 55'394 candidate sentences mentioning a cyberattack.

In a traditional document analysis system, the analyst would at this point inspect the resulting documents one by one to detect new cyberattack incidents. Performing this task daily, every day about 600 candidate sentences would have to be reviewed.

The candidate sentences contain both positive cases (i.e., mentions of a cyber-attack on healthcare) as well as negative cases (i.e., discussions related to cyber-attack on healthcare).

Goodman Campbell computer network attacked by hackers - Indianapolis Business Journal

Indiana National Guard seeks employers for cybersecurity career fair May 20, 2022

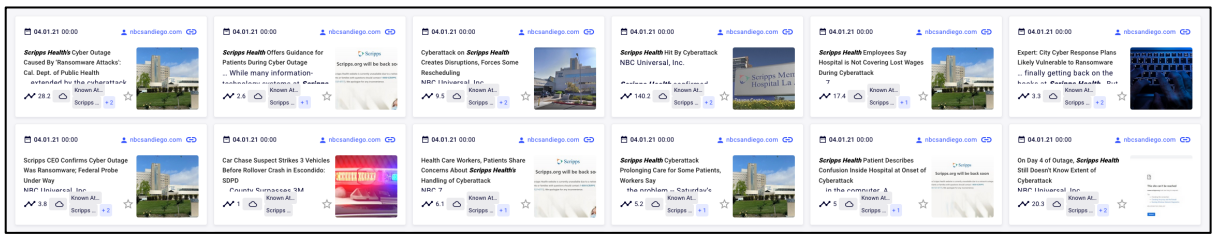
The computer network of Goodman Campbell Brain and Spine, a large, independent surgical group based in Carmel, has been hacked, compromising patient and employee data.

Example of a positive case

Cyberattacks on healthcare: Healthcare hit by 45% spike in e-virus cases | India Business News - World News Live https://t.co/qCS73v7hQX #Infosec #CyberSecurity #CyberAttack #Hacking #Privacy #Threat #Malware #Ransomware #Cyberwarning #Phishing #SpyWare #Tech #Technology https://t.co/xXk9blwAYj

Example of a negative case

Furthermore, many attacks are mentioned many times in the media, exposing the data analyst to an abundance of replicated information. For example, the attack on Scripps Health in 2021 produced more than 2000 articles.

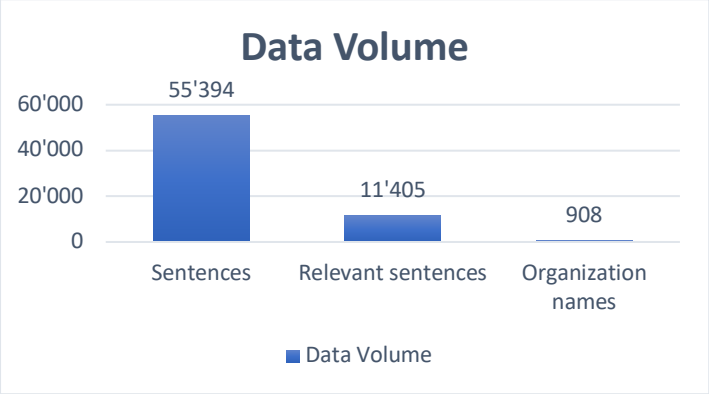


Examples of mentions of the attack in Scripps Health

To reduce the amount to detect relevant documents, we are using neural network classifiers for natural language understanding. In a first step, we can distinguish text that is mentioning a concrete cyberattack from other discussions on cyber risks in healthcare. This reduces the amount of data to be inspected to 20% of the original volume. In a second step, we extract names of the health organizations potentially affected by the attacks. Therefore, the analyst will inspect those organization names, before deciding to further analyze the related documents. This reduces the work in two ways:

1. The number of organization names is much lower than the number of texts. In particular, repeated mentions of the same organization names are clustered.
2. Reading extracted organization names is much faster than reading complete texts.

As a result, the analyst must inspect a number of names that is 50 times lower than the initial number of candidate sentences.

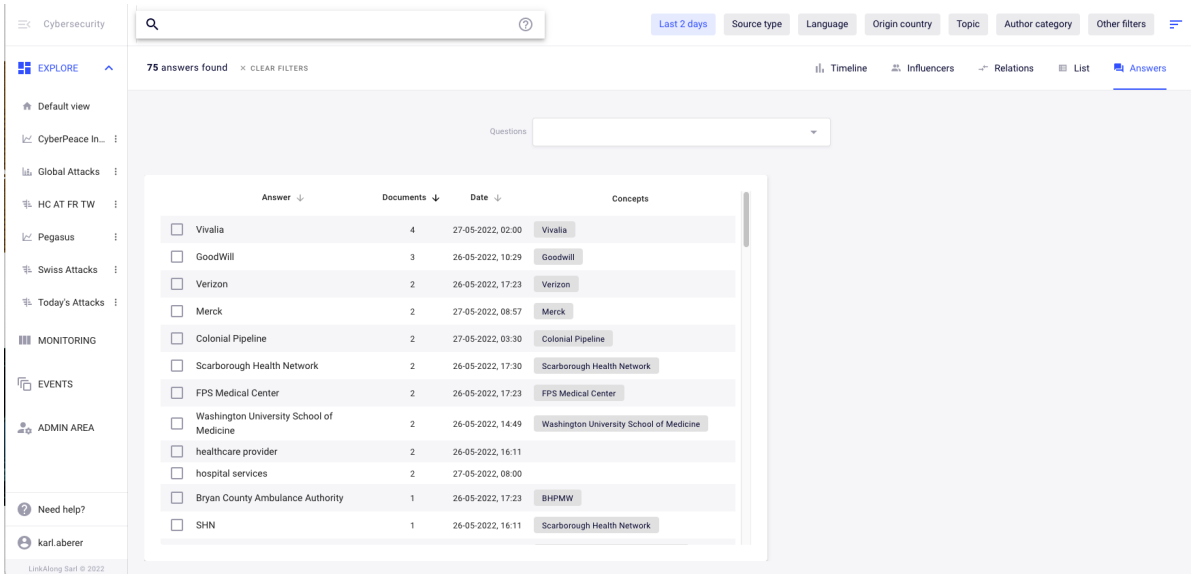


We performed a detailed evaluation on how many relevant incidents occur in the data. We find that about 25% of the organization names correspond to new incidents, whereas about 45% refer to known incidents and the others are not relevant. Therefore, with our method the data analyst needs to inspect about 10 organization names per day to detect the 1-2 new incidents that typically occur daily.

Assuming the inspection of 1 data element takes about 10-20 seconds an analyst will perform the task with our method in a matter of a few minutes per day. This compared to 1-2 hours that would be required to sift through the data produced by a system that relies on conventional keyword-based pre-filtering of documents.

Platform

The technology evaluated in this study is fully available as part of the Peek platform in the Answer module. In addition to extracting potential names of events, it provides a full set of filtering and document inspection capabilities to check every candidate in detail.



Contact: If you are interested in exploring this fascinating new way of weak signal detection please contact us: contact@linkalong.com

Linkalong Sarl

EPFL Innovation Parc, Bâtiment C, 1015 Lausanne - Switzerland