Digital Food Chain Education, Research, Development, and Innovation Institute

Automated News Screening for Emerging Infectious Disease Risk Identification in the Lake Victoria Basin

By

Zorka Mwendwa Rakonczay

Supervisor:

Ákos Jóźwiak DVM, PhD.

BUDAPEST, HUNGARY
2022

Table of Contents

1 Abstract

Recent years have shown the importance of being able to predict and track outbreaks of infectious diseases with the aim of preventing major incidents that can cause vast disruptions to society. Predicting disease emergence can be aided by the study of the drivers and trends of infectious disease emergence. This thesis describes a proof of concept for the use of automated news screening as a tool to use in as part of an early warning system to identify the emergence of infectious disease, using the Lake Victoria Basin as the subject for this exploration.

The method described in thesis uses the open-source text mining and data analysis tools of KNIME and R's tidyGraph, iGraph and visNetwork packages to breakdown and examine digital news articles to create an easily visualizable summary of news articles relevant to news on a topic of choice, in this thesis, disease outbreaks.

2 Introduction

2.1 Emerging Infectious Diseases

Many factors, often complex, lie behind the emergence of infectious diseases [1]. Also known as drivers, these factors provide the conditions that enable a select pathogen to encroach on and adapt to a new environment, or transition into a newly identifiable pathogen. These pathogens are commonly named in literature as “neglected or emerging infectious diseases” or “neglected/emerging zoonotic diseases”, based on their location and manner of spreading between populations and species [1, 5–9]. There are various available characterisations for drivers, in this thesis, drivers are defined as antecedent events to infectious disease emergence; forces operating at various scales, and categorised as societal, environmental, technological, political, and economic in nature [3, 10].

Regions where drivers are most densely aggregated are where outbreaks of emerging infectious diseases are more likely to occur and are known as emergence hotspots [11]. Drivers manifest as trends, with multiple drivers able to concurrently cause or affect a trend. Conversely multiple trends can be traced back to a single driver [12]. Examples of drivers and trends include climate change (a driver) and its consequences as seen in population migration of malaria mosquitoes (a trend) [13].

Being able to link trends to their drivers and developing ways of tracking their occurrences is an important element in the study of disease outbreak patterns[14]. Knowledge of these forces and their distribution across various pathogens, geographical regions and production practices allows for the emergences of risks to be typified, making it possible to make predictions about where and how future outbreaks could occur.[15]

The analysis of digital news media is a useful tool for this as a fast and cost-effective method for monitoring drivers and trends [2]. For this thesis, the aim is to use digital news articles as the source of the input data to be analysed and to anticipate emergence risk. Digital news surveillance allows for the observation of changes in reported incidence, whether it relates to an acute outbreak or a long-term trend. Diseases do not emerge in a vacuum, and news articles contain contextual information that when analysed alongside many of others can reveal connections between seemingly unrelated issues that correlate with, and with further investigation may prove to be a causation of, an outbreak [4].

With just a handful of languages accounting for most content on the Internet, only a few key languages require translation to capture most content, namely English, Spanish, Chinese and French [2]. Furthermore, language processing and text mining tools such as the commercial software NetOwl, and the open-source software of KNIME text processing extensions, are getting better at analysing reports for contextual data making the search for relevant information more efficient [2, 4].

This method has drawbacks including the irregularity of digital information, with articles being liable to be taken down, prone to link rot and their sources, when analysed en mass, hard to verify. Curated datasets and media aggregating sites, such as Google News and the Europe Media Monitor (EMM) Newsbrief, are useful as they allow one to use one search to comb through articles of various sources. These are limited in number, and only a few are accessible for public use. Digital media can also exclude nondigital news reports and details due to language or source (e.g., print and audio) limitations [3]. The separation of original content from re-reporting (possibly erroneously and/or appearing in the wrong region) poses additional challenges[2–4]. While valid, these issues fall outside the scope of this thesis.

The scope of this project is to create a prototype of a potential early warning system based on the work done by the European Food Safety Authority's DEMETER project[16]. Utilising text mining and data analysis algorithms, the aim is to condense news articles into an easily visualisable summary of news events over an up to fourteen-day period of choice. This would allow those interested in the tracking of disease emergence events to get a useful overview of emerging trends and patterns going on in the news while grouping them based on topics in common. This method results in providing new insights that can guide to more robust investigations in disease monitoring. This thesis intends to demonstrate this method to be a helpful means in outbreak detection, tracking and eventually prevention.

The Lake Victoria Basin in East Africa was chosen as the area of to test the method, as it is a known hotspot, as visualised in Toph Allen's article on global hotspots (Figure 1), with many emergence drivers present [1, 9].

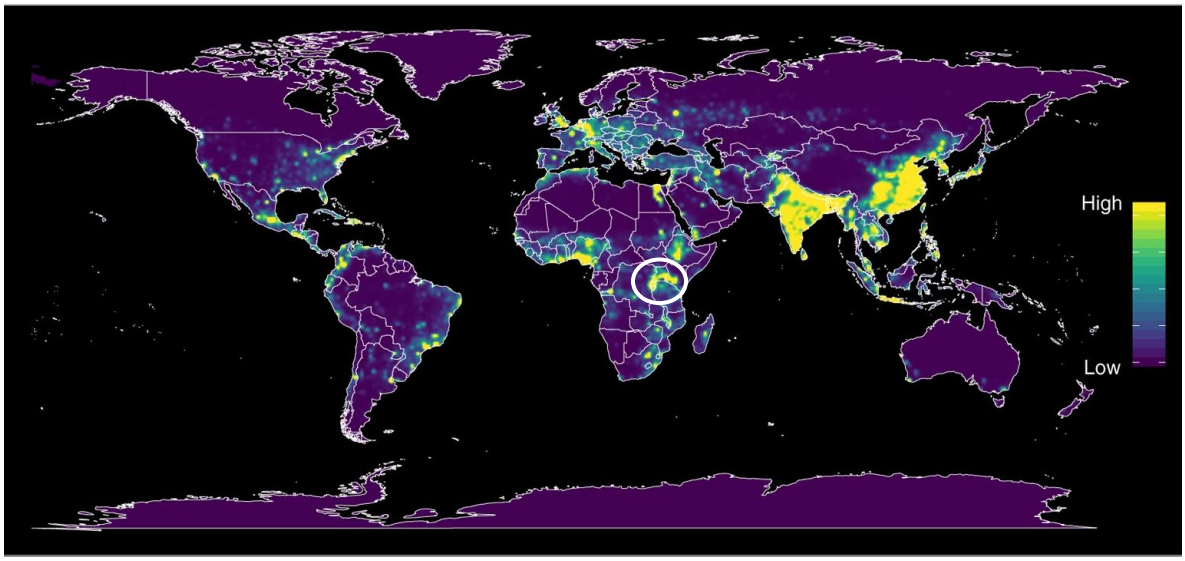

Figure 1:

Map of world emerging disease hotspots [9]*, with the Lake Victoria Basin circled in white*

Divided among the countries of Burundi, Kenya, Rwanda, Tanzania, and Uganda, the Lake Victoria Basin is densely populated, and its international nature leads to the presence of multiple drivers related to the diversity of social and institutional circumstances[17] [9]. There is a high level of movement, human and animal, between the countries which all have different rules, regulations, and disease monitoring systems in place. Economically, the Lake provides a source of income for many and a recent boom in the economy, has drawn in a lot of seasonal labourers living in shanty towns, with very low sanitation and few services [18, 19]. The permanent population is largely agriculturalist, mainly working on small independent farms, most of them livestock[20]

From an environmental point of view, the Lake is surrounded by a variety of biomes (Figure 2), with more wetland areas in Rwanda and Uganda and dryer arid climates in Kenya and Tanzania [17]. The region is greatly affected by climate change, with increased droughts and more intensive rainy seasons. [21] This has led to a change in the distribution of disease vectors such as mosquitoes as well as an increased number of people moving to the Basin due to being displaced from increasingly inhospitable areas further North [22]. The commercial use of the Lake and its surroundings has also affected the local ecosystem and has lead to a rise in arthropod vectors such as mosquitoes[17, 18, 21]. Managing the

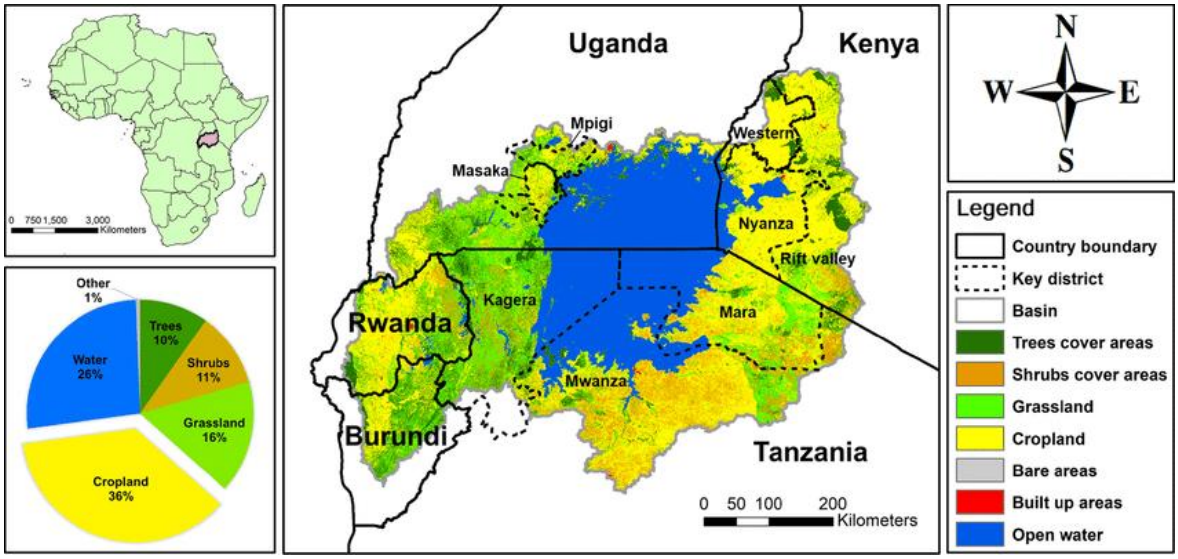

Figure 2

Biomes of the Lake Victoria Basin [27]

consequences of this is highly contentious due to the political and economic interest within and between the bordering nations.

This thesis aims to break down the various digital news articles related to the Basin and create a clearer picture of what is going on in the region, to demonstrate the useability of the method it proposes.

2.2 The DEMETER Project

The Determination and Metrics for Emerging Risks Identification (DEMETER) project was a project funded by the European Food Safety Authority (EFSA) [16]. The objectives and research proposed in the DEMETER project were specifically designed to support current (and future) EFSA procedures for emerging issue and risks identification by providing community resources to allow EFSA and EU Member State authorities to share data, knowledge, and methods on emerging risks identification in a rapid and effective manner through a digital platform.

In the framework of the project, different data analytical workflows were developed and deployed, as the trending topics in news based on text mining and network analysis, and patent network analysis.

The current bottlenecks in predictive sciences are not how to find data but how to make sense of them. There is a large amount of information and knowledge hidden in various news sources and extracting value from those is not an easy task: the amount of data and noise is so huge, that deriving meaningful insight is only possible with using computational science methodologies. As an attempt to capture important news items, a network analysis approach was used in the DEMETER project, using EMM Newsbrief as input source. The task of the data retrieval and analysis workflow developed was to collect news, then perform a text-mining and network analysis of the information embedded in the food safety news.

The main advantage of using such workflow is that it makes a quick analysis and visualization of the trending news possible. In a conventional emerging issue identification process large amount of news should be screened manually. This process could be speeded up by using simple automated text mining methods like word cloud or word count, however, these purely focus on quantifying the appearance of words and lack semantic notion. Using network analysis and visualizing the co-occurrence of words and identifying clusters in this network capture a semantic perspective and makes news topics pronounced and easily identifiable. With this, the workflow makes a quick screening of trending news topics possible from large amount of news corpora, speeding up the emerging issue identification process.

3 Objectives

This thesis aims to produce a proof of concept for using open-source text mining and data analysis tools for monitoring the drivers and trends of infectious diseases. Utilising the DEMETER Project's work as a template, their KNIME workflow was adapted to screen for infectious diseases in the Lake Victoria Basin. The workflow (Figure 5) was changed to incorporate a translating step as the original program only utilised English texts, and the text mining algorithm was adapted towards looking for emerging diseases occurring in the Basin.

KNIME, the Konstanz Information Miner, is a free and open-source data analytics, reporting and integration platform which supports easy integration of new algorithms, data manipulation and visualization methods as new modules or nodes. Its modular environment enables easy visual assembly and interactive execution of a data pipeline [23].

The goal is to highlight and visualise topics of interest to those involved with tracking disease emergence, by creating a user-friendly summary of articles broken down into their key words and clustered into groups based on the relations between their component words.

The workflow can handle thousands of articles sourced from a seven day period within a matter of minutes. It selects the top 50 words chosen based on how frequently they appear in correlation with one another based on our choice of filters while maintaining their link to their texts of origin. The end result appears as a colour-coded "word-network" (figure10), with the 50 keywords displayed in a network of links showing how they relate to one another and grouped by frequency of co-occurrence within the texts.

This will allow users focus on the topics highlighted and visualise how they may be related to a broader issue that may have otherwise been overlooked had one had to read them all of individually. To access more information about topic highlighted by the network, one can hover their cursor over the links making up the word network, to be shown the hyperlink of their source articles, which can then be accessed with a single click.

4 Material and Methods

The project involved three key steps

- Gathering news media as input data
- Modifying and running the KNIME workflow
- Evaluations of the data gathered

These steps were repeated as needed to get the most refined result possible.
The workflow was changed to incorporate a translating step, and the text mining was geared towards looking for emerging diseases.

4.1 Gathering the News Media

The input digital news media was gathered from various locations. Bearing in mind that the workflow can only display the top 50 findings, news sources were selected to achieve a greater focus on our topics of interest.

The European Media Monitor (EMM) News Brief news aggregating was selected as main source of data as it one of the few publicly accessible digital news curating sites. EMM's primary goal is to oversee a curated set of electronic news media from around the world, reducing the information flow to controllable amounts by grouping related news and sorting articles. The system continuously monitors around 8 000 HTML pages and RDF Site Summary (RSS) feeds in over 70 different languages to find new articles published on the Internet (~300 000 articles daily) [24]. RSS feeds are a kind of data format, also known as web feeds, that allows users and applications to access website updates and collect data in a standardised and computer-readable form.

EMM Newsbrief's advanced search features (Figure 3) were used to filter subjects applicable to infectious disease. Thirty-seven diseases were selected to examine for this thesis. Thirty-six of these are listed as being infectious zoonotic diseases of interest (Table 1) in the five countries of the Lake Victoria Basin, and Kenya in particular, by the Global Disease Detection Program of the United States Centres for Disease Control and Prevention and the Zoonotic Disease Unit of the Kenyan State Department of Veterinary Services [25]. A similar project was conducted in Uganda resulting in a shorter list of diseases [26]. Neither study included Malaria as it is not a particularly zoonotic disease, but it was added to the selection for this thesis due to its importance to human health.

Table 1

List of the 36 Diseases of Interest

Anthrax	Trypanosomiasis	Rabies	Brucellosis
Rift Valley Fever	Echinococcosis	Marburg	Q-Fever
Influenza	Cysticercosis	Dengue	Mycobacteria
Leptospirosis	Schistosomiasis	Yellow Fever	Rickettsiosis
Taeniasis	Sarcoptic Mange	Cryptosporidiosis	Leishmaniasis
Ebola	Non-Typhi Salmonellosis	Crimean-Congo Haemorrhagic Fever	Antimicrobial Resistance
Dermatophilosis	Cryptococcosis	Listeriosis	Aspergillosis
MERS-CoV	Plague	Chikungunya	West Nile Virus
Histoplasmosis	Diphyllobotriosis	Hanta Virus Fever	Lassa Fever

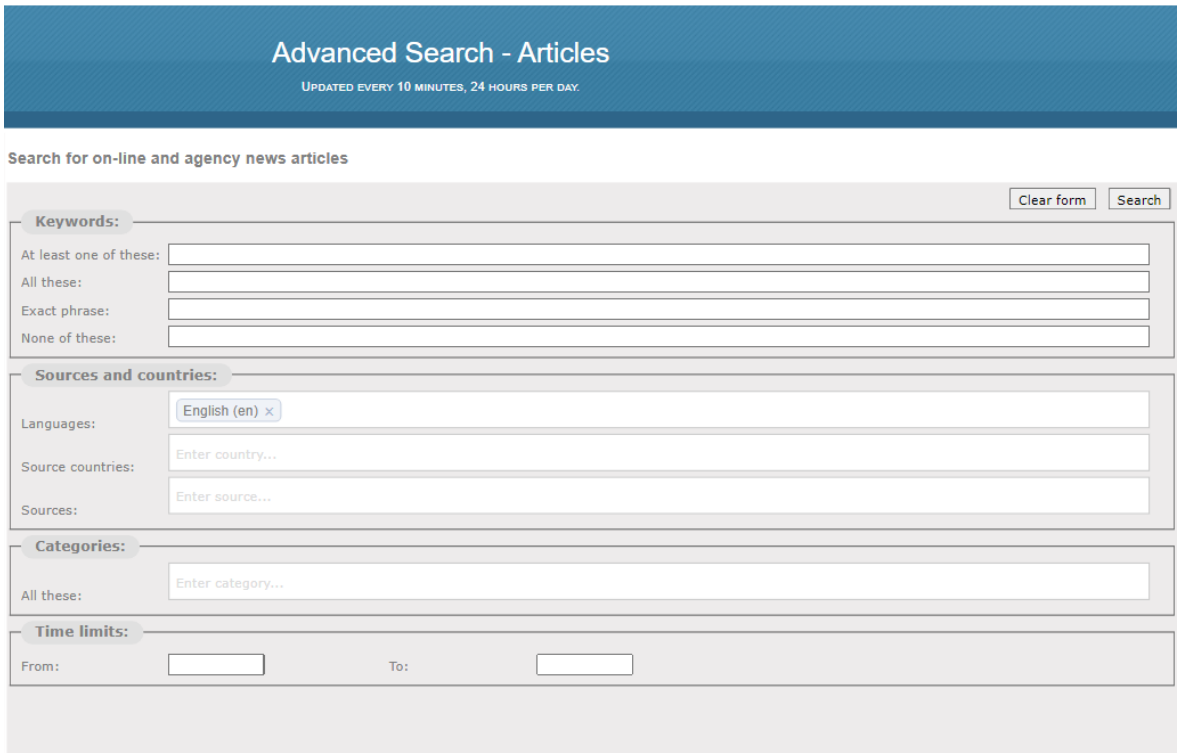

Figure 3

Europe Media Monitor advanced search layout

Deviating from the DEMETER Project's exclusive use of English language articles, this thesis selected "All" in the language filter, to broaden the range of retrieved articles relevant to the region. The searches can be broken down as follows:

1. Using EMM's built-in "categories" filters, comprising approximately seventy-five options, the most relevant ones to the topic of infectious disease were selected. These were "Animal Health", "Food Safety"; "Communicable Disease"; "Food Security". Two searches were conducted based on the "source countries", resulting in eight searches:
   1.1. Source: "all countries", using the five Basin countries in the "keywords" filter: "Burundi", "Kenya", "Ruanda", "Tanzania" and "Uganda"
   1.2. Source: the five Basin countries "KE","TZ","RW","BI","UG"
2. Searching for news on the merged list of the thirty-seven diseases, with the addition of key words on the topic of anti-microbial resistance and synonyms for diseases that have multiple names (e.g.: drug resistance as a synonym for antimicrobial resistance). This was an "OR" search, meaning any article mentioning one or more of the selected terms would appear in our results. A single search was conducted based on the "source countries": the five Basin countries "KE","TZ","RW","BI","UG"

Sources more tailored to the area of interest were also used. Gathering them was more challenging, as many relevant news sites could not generate RSS links. Ten news sources devoted to the topics of health and agriculture (due to the zoonotic nature of many of the diseases of interest) were eventually short-listed. These were:

1. The Organic Farmer (*theorganicfarmer.org*)
2. East-African Agrinews (*eastafrican-agrinews.com*)
3. Smart Farmer Kenya (*smartfarmerkenya.com*)
4. African Farming (*africanfarming.net*)
5. EA Agribusiness (*ea-agribusiness.com*)
6. African Journals Online (*ajol.info*)
7. Africa Health (*africa-health.com*)
8. African Journal of Laboratory Medicine (*ajlmonline.org*)
9. Swara Magazine (*swara.co.ke*)
10. The East African (*theastafrican.co.ke*)

Once the searches were made, they were saved as RSS links. Where possible, this was done using each site's built-in RSS feed generators. If that option was not available, these were created using the RSS feed generator *feedly.com*. These feeds then had to be altered into a form useable by KNIME, our text mining program (Figure 4), resulting in twenty RSS feeds used for sourcing the input data for the project. Each individual RSS feed can retrieve up to a hundred articles.

Search #1

any language

Kenya OR Tanzania OR Uganda OR Burundi OR Rwanda

Category: Animal Health and Welfare

original RSS link retrieved:

feed:https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&country=ET%2CKE%2CTZ&dateto=2021-08-29T23%3A59%3A59Z&datefrom=2021-08-23T00%3A00%3A00Z&category=AnimalHealth

- Delete 'feed:'
- Substitute the date with a variable where a new a new date will be injected in KNIME

link suitable for KNIME:

https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&country=ET%2CKE%2CTZ&dateto=end_dateT23%3A59%3A59Z&datefrom=start_dateT00%3A00%3A00Z&category=AnimalHealth

Figure 4:

Breakdown of the steps to taken to alter original RSS links to ones suitable for KNIME

The final list of all feeds used in the workflow can be found in Appendix B.

4.2 KNIME Workflow

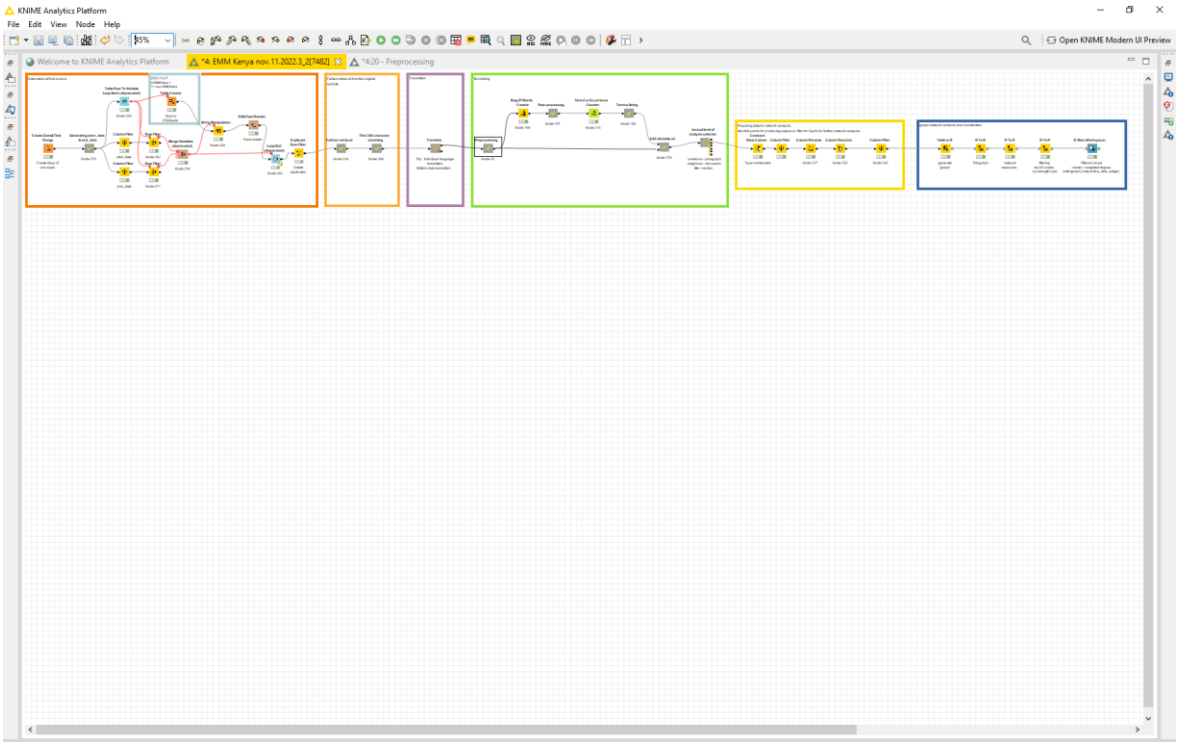

Figure 5:

Screenshot of entire KNIME workflow created for this project

KNIME, the Konstanz Information Miner, is a free and open-source data analytics, reporting and integration platform which supports easy integration of new algorithms, data manipulation and visualization methods as new modules or nodes. Its modular environment enables easy visual assembly and interactive execution of a data pipeline [23]. All these features combined make it an excellent tool to be handled by multiple people and was a factor behind why it was chosen for the project.

Our workflow (Figure 5) is made up of nodes, small programs with a specific task to accomplish. Their function can be deduced from their name and pictogram, as well as the "KNIME node repository", where a full description of their function can be found. Nodes can be grouped together to perform a large complex task such as translating. These groups are called "meta nodes" and they appear as larger grey nodes in the workflow. When clicked on, their constituent nodes can be seen and configured individually. Meanwhile, component-nodes are created by users of the platform to oversee a more complicated task. They are

cross-checked by other users and the KNIME platform before they become available to download by the public, one of the advantages of the open-source nature of this tool.

Both meta-nodes and component-nodes are useful to organise large workflows. One can identify isolated blocks of logical operations in a workflow and include them inside either a meta- or a component-node. This means the workflow will appear neat and tidy with less fewer words cluttering the user interface.

The workflow can be broken down into 5 major steps which are as follows: 1. Data retrieval; 2. Text retrieval; 3. Translation; 4. Text mining; 5. Network analysis; iGraph network analysis and visualization.

4.2.1 Data Input

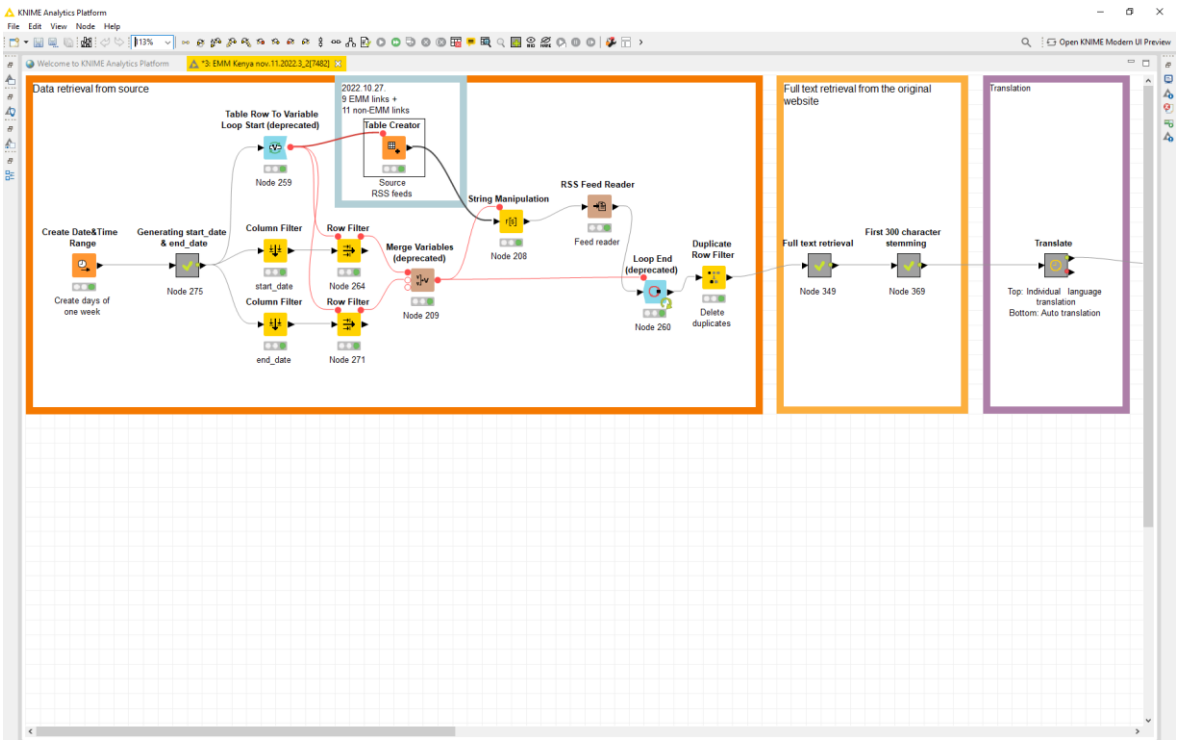

Figure 6:

Screenshot of data retrieval, text retrieval and translation sections of the workflow

- This first step of the workflow (Figure 6) is where all the input data is entered. By choosing a time interval of interest in the “date and time” node, articles can be selected from as far back as eighteen months, with the ability to select for precise dates and time intervals.
- Entering the selected RSS feeds in the “table creator” node, each individual RSS link can retrieve up to 100 articles.
- The rest of the nodes work on various steps of reading the RSS feeds and finding the articles from the selected time interval
- Experience showed it was best to limit searches to a maximum two-week window. More than that risked retrieving too much data that could overload the workflow, leading to the program crashing. After running the workflow more, this limit was changed to one week (seven days) due to the pressure placed on the “translator” meta-node

4.2.1 Text Retrieval

- In this step, the texts of the articles recovered during data retrieval are extracted by the “full text retrieval” meta node (Figure 6). Each article was limited to its first 300 characters by the “First 300-character stemming” meta node.
- The “Full text retrieval” meta node contains nodes using the Palladian KNIME extension, these extract the written content from each website (article) and arranges it in separate columns for the title and text of each.
- The character limit was arrived at as it was observed that most of the pertinent information in an article is found in its first 300 characters, and thus save computing time, as depending on the time windows, thousands of articles would be retrieved, and then translated.

4.2.2 Translation

Deviating from the original DEMETER Project, the searches used the “all languages” filter. This necessitated the innovation and inclusion of a translating step utilizing Google’s translating extension for KNIME (Figure 6).

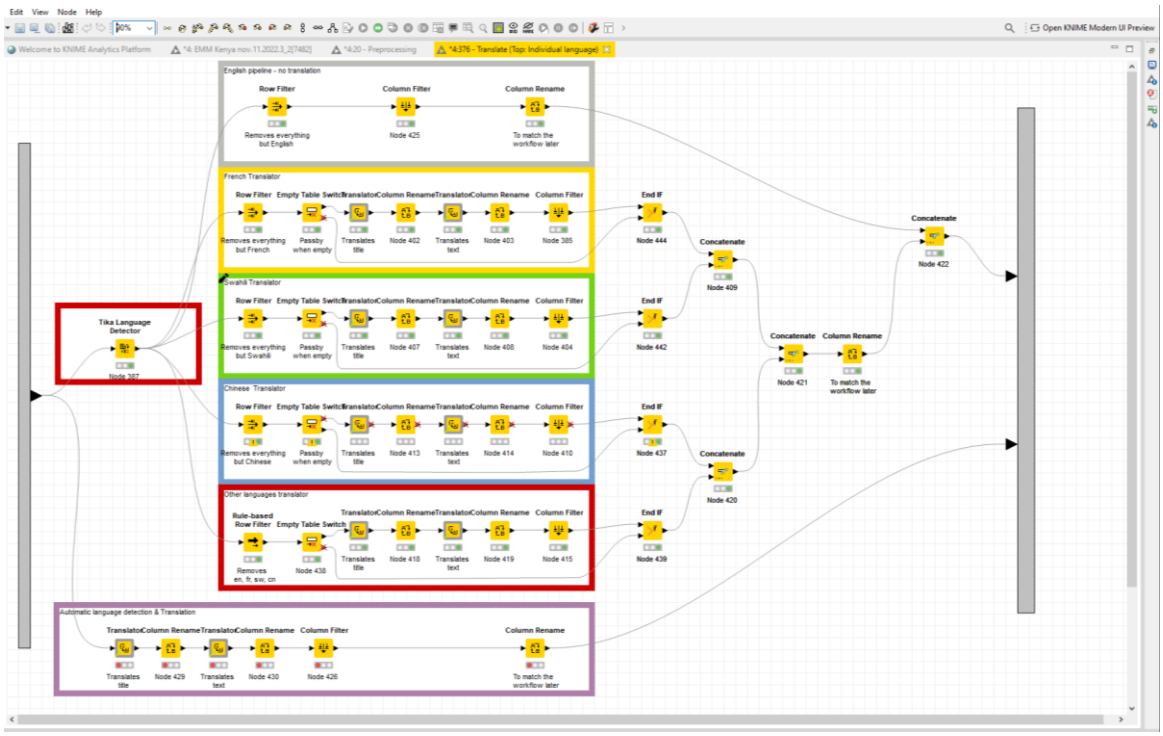

Figure 7:

Screenshot of components of the translation meta node

The translating can take place in one of two ways, visualized by two “Flows” with the input data able to go through either, depending on which one is tied in with the rest of the workflow (Figure 7). The “Translator” component-node is downloadable from KNIME’s Node repository. It can automatically detect and translate languages between supported languages by running Google’s Translator program in the background. It works best on small datasets, which is why the multiple flows were created to not overload the workflow should a greater number of articles appear for a particular period. This bottle-neck point requires further investigation to overcome, but for the current project, a seven day limit was set for the data retrieval so as not to overload this component of the workflow.

The meta-node was split into two main flows, of which the first one is currently in use. Further research is required to explore which one is the better system, although both were found to be functional the scale of this project.

Flow 1:

- Using the "Tika language detector" node, the texts were filtered into one of five streams depending on the language detected: English, French, Chinese (simplified), Swahili (also known as Kiswahili) and "other". These were selected on the criteria of being the most used in the area and being available in the node's Menu. This menu is limited in local languages, so potentially many relevant articles may end up being overlooked.
- The TIKA Language detector assigns code to each language; this language detector is similar to Google's own, but it was chosen for this task because it can handle larger amounts of text, though further studies need to be done to observe how it compares with Google's detector in terms accuracy
    i. English is the most common language in the input data, so all texts flow through that stream without filtering, as the end analysis is in English.
    ii. French, Simplified Chinese and Swahili were selected as being the most relevant languages with respect to the region, but less frequent than English. To accommodate for this, the input texts are first filtered for each of these languages before they are translated to English
    iii. The "other" languages are those that are not detected as one of the previous four, so undergo automatic language translation, the same process utilized by Flow 2. These languages were either deemed not to be very important or undefined by TIKA language detector, but the information they convey is still potentially relevant.

- A series of joiner nodes is required to remove redundancies, stemming especially from the "other languages" filter, and to create a unified table of text again.

Flow 2 (purple):

- This flow utilizes Google's own language detection program (unlike Flow 1 that uses TIKA).

    i. The “Automatic language detection and translation” stream utilizes the translator plugin’s “all languages” setting. This works as a universal translator, and during the creation and fine tuning of this project, has not made any observable mistakes in translation.

- This Flow also uses the Translator component node, so the risk of overloading is still present.
- This flow should be able to do the entire task of translating alone, but to avoid overloading the workflow and for easier catching of errors, we opted to use Flow 1.

Further research would be required to compare Flow 2 with Flow 1, so that it may one day be used as a validating tool, or even take over completely. That step did not fit within the scope and timeline of this thesis.

4.2.3 Text Mining

The objective is the breakdown of the translated texts into individual words, while maintaining their associations with their original article and its weblink. The resulting output data only contains terms (words) and their link to their place within their text of origin as well as the link to the text's original location online. The work here can be divided into pre-processing, or the modification of the text into data usable by our text mining algorithms, and the text mining itself (Figure 8).

Pre-Processing

- Due to computational power limitations, only 50 nodes can be visualized in the final clustered network of words, so an accurate pre-processing work must precede it to avoid unnecessary filler words using up space (Figure 9). This process takes place in the "Pre-Processing" meta-node.
- This phase involves the manual selection of words and characters to be removed from the final product. Using the "Table Creator" node, the list was started off with prepositions, adjectives, numbers (written and characters), grammatical characters, pronouns and some frequently re-occurring non-topic specific words such as "news", "meeting", "observed" and additional filler words. With each run of the Workflow, this list was increased with the inclusion of words that proved to be either irrelevant, vague or so frequently used that they became superfluous to the task (eg: days of the week, months, "doctors", "hours", "government", "groups", "militia", "terrorist")

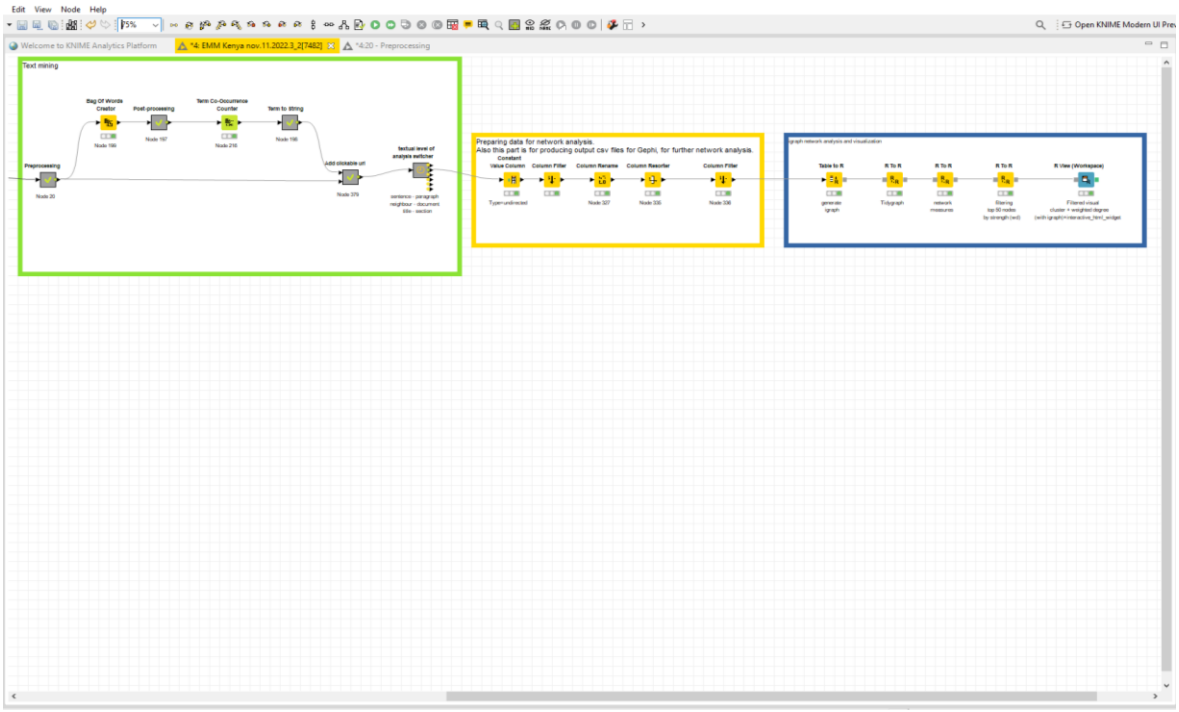

Figure 8:

Screenshot of the text mining and network analysis steps

- Of all the steps, this is the most subjective one as it is fully reliant on the user's interpretation of what they deem relevant.
- A post-processing meta-node was added after the "Bag of Words" creator. Its function is the same as that of pre-processing meta-node and its role is to filter out any artefacts left over from pre-processing as an additional fail-safe measure.

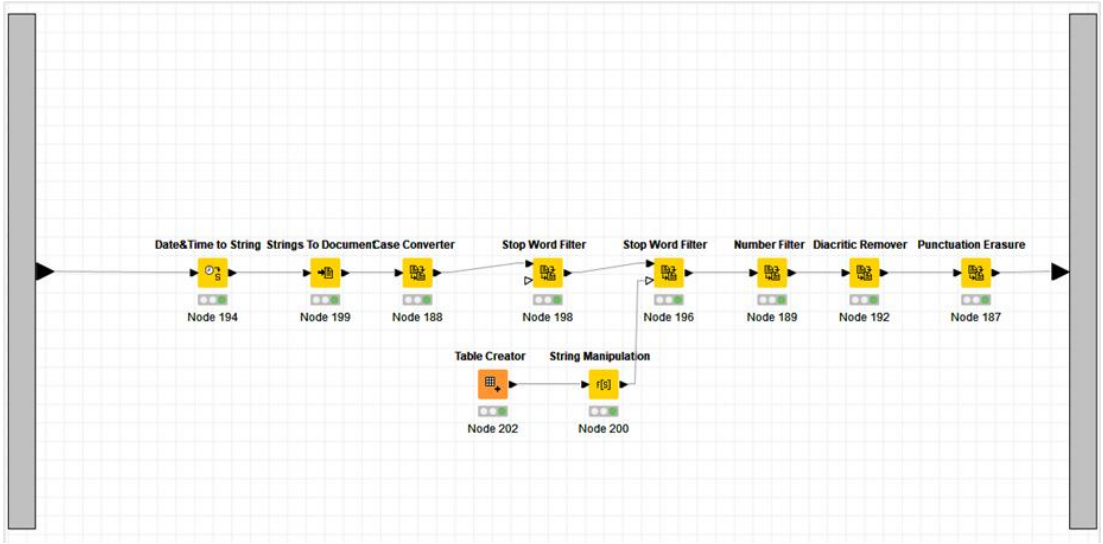

Figure 9:

Screenshots of the components of the pre-processing meta node

Bag of Words Creator

- Following pre-processing, this algorithm extracted and examined the words filtered from the articles while retaining the knowledge of their original place within their texts
- This process is important for the term co-occurrence counter, an algorithm that calculates the level of connection between words. Connection refers to the degree two words appear together in a text; how frequently they are used together in the same paragraph/ title/ sentence within a particular text (e.g.: “health” and “launch” appear x times at y level, while “symptomatic” and “infection” appear z times at w level). The algorithms chosen for this task can collect the information from all the articles, resulting in a table detailing how often two particular words appear together throughout all articles and arriving at the top fifty words that are the most linked.
- This is visualized in a table with each word listed alongside how many words it was associated with, and to what extent (same article/paragraph/sentence etc.)

4.2.4 Network Analysis and Visualization

The goal of this step is to create a graph that can visualize the connections between the words, clustering them into groups based on their levels of connection while preserving easy access to the words' text of origin.

All natural networks often show cluster characteristics. In the effort to demonstrate the link between trends, drivers and disease outbreaks, clusters are an ideal method to demonstrate connections that may not be otherwise easily noticeable. Some words appear in many texts, this may be because they are related to each other. Cluster of "malaria" vs "Salmonella" in the same country: same network because text of articles is connected to each other due to the appearance of similar words (e.g.: name of the country with the outbreak), but different clusters, because these are separate issues

This involves the filtering out the top fifty most connected words with the corresponding information, which requires a complex set of processes. KNIME has an available built-in algorithm, but these were found to have poor performance and were unreliable. Therefore, the packages created by the software company "R" for KNIME were selected for this process. This involved downloading their plugins (igraph, tidygraph and visNetwork) from KNIME's website as well as their separate software

Using the igraph, tidygraph and visNetwork packages of R, in that order, the fifty words and their corresponding data are

1. igraph (available in r and python, chose R), runs calculation in R and gives results in KNIME: this algorithm is able to generate the main graph itself, visualizing all the connections between the words; unfortunately it is unable to filter the results in a weighted list,this function requires tidyGraph.
    i. Clusters are defined by the igraph "optimal community structure" algorithm, based on the network properties of the words themselves (by maximizing the modularity measure over all possible partitions). This means that the clusters are formed around related stories/words. When this

works well, one can decipher a news story from reading the key words appearing in the word network.

2. tidygraph: calculates network measures, was selected partly due to its easy settings; it calculates values including the: degree; weighted degree, centrality; loops (words co-occurring with themselves) in network. This algorithm can cope with up to 50 nodes (due to performance power challenges), hence, and creates the end limit of available nodes.
3. visNetwork: clusters and visualizes the words by using iGraph's layout, therefore, to run it one need iGraph functioning in the background

4.3 Analyzing the Data Gathered

The result of the KNIME workflow is displayed as a network of words (Figure 10) in the computer's web browser. This works for Apple's Safari, Microsoft Edge and Linux systems., with or without an internet connection, as long as the text retrieval and translation has already been completed. This still allows for editing the pre-processing filters for clearer results even when on the move.

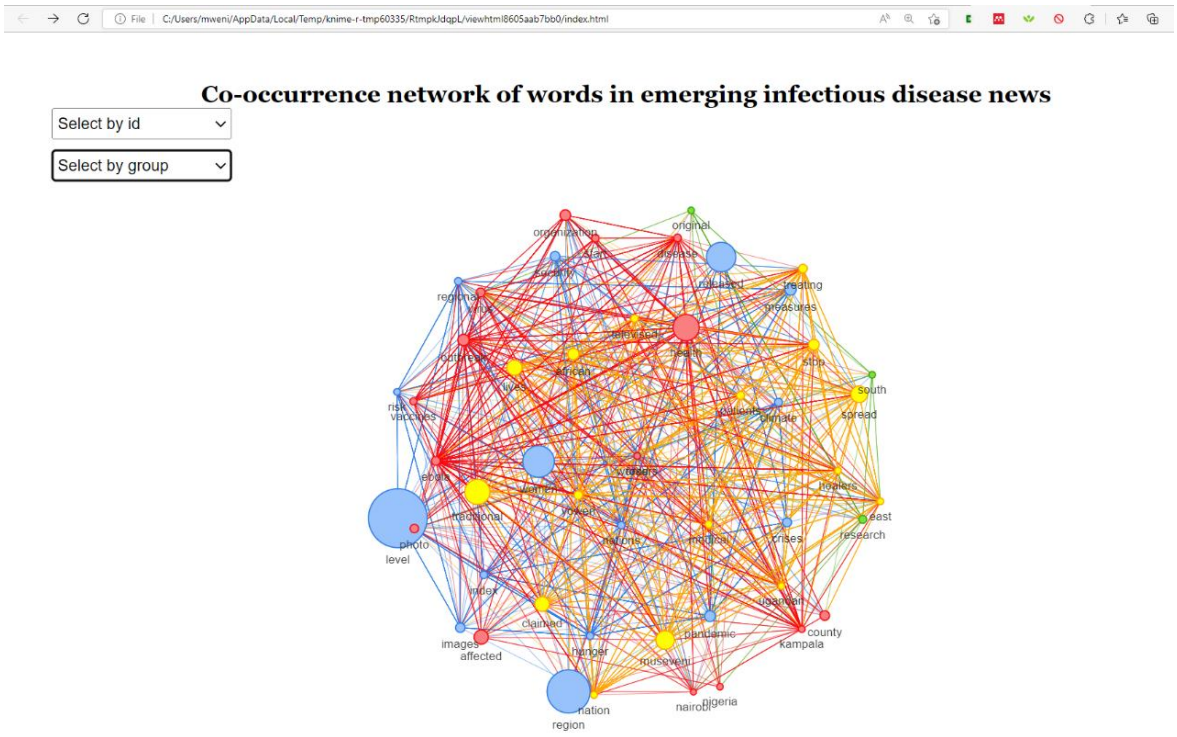

Figure 10:

Screenshot of the initial word network as seen in a Microsoft Edge web browser

Each word in the network is symbolised as a circle of varying size, connected to other words by lines, with the lines being illustrations of hyperlinks of the news articles the words were sourced from. By placing the computer's cursor on the lines connecting the words, one can see the link to the word's article of origin.

The size of the words' circle is representative of that word's frequency of appearance in various articles and therefore its level of connection, with the words having more connections being lager. Each circle is coloured depending on which cluster it belongs to with the links continuing the colouration. While working on this projects, no words were

observed to be of dual colouration, however some words may link to others outside their cluster if they have a strong enough connection to be visualised.

Two forms of clusters are created: those based around words, and those linked by shared topic.Only the clusters grouped by topic are of interest for this thesis. These were created based on the clustering algorithm used in the Network analysis step of the workflow. The user interface shows clusters numbered from 2 upwards. The word-based clusters are of equal number as there are words in the network, and are linked by the appearance of the particular word in any given article.

This visualisation enables users to look over thousands of articles and see a common theme between them, freeing them from having to comb through all of them individually. If the network shows an increase in interesting results, all one needs to do is click on the links of interest highlighted by their clusters to read the source articles to get a greater understanding of the situation.

5 Results

The final product is a KNIME workflow able to process, sort and analyse imported news media gathered from a period of up to one week from as far back as 18 months based on shared topics in under 5 minutes. The resulting word network can be broken down into clusters, based on their connection of a mutual subject.

The final KNIME workflow can be accessed at the following links:
https://dfi.univet.hu/wp-content/uploads/2022/11/EMM-Kenya-nov.11.2022.3_2.knwf_.zip
https://dfi.univet.hu/en/activities/education/automated-news-screening-for-emerging-infectious-disease-risk-identification-in-the-lake-victoria-basin/

To demonstrate the method in practice, an example run of the workflow was made. The period selected for analysis was the week of October 31st to November 6th, 2022. Around 420 articles were retrieved by the data retrieval nodes. The word network produced can be seen in Figure 10 and its component clusters in Figures 11-14.

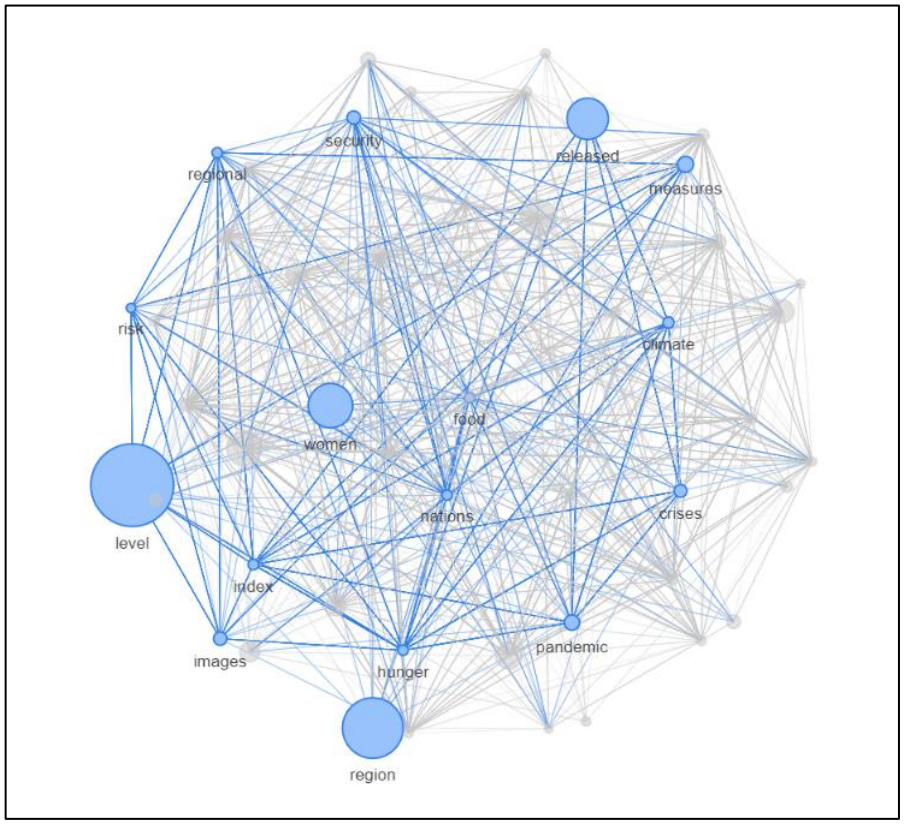

Figure 11

View of Clusters 2, highlighted in blue, taken from Microsoft Edge Web Browser

Examining Figure 10, one can see the circles representing the words highlighted in four different colours (red, yellow, blue, and green) depending on the cluster they belong to. One can select a cluster to examine by clicking on the “select group” button on the side of the screen. The cluster selected is then shown individually with the other ones dimmed in the background. The results of the example workflow run are as follows:

Cluster 2 (Figure 11), highlights seventeen words, including: “regions”, “food”, “climate”, “hunger”, “security”, “pandemic” and “crisis” amongst others. This suggests links between famine, the Climate Crisis and a pandemic (presumably the Covid-19 pandemic). Further exploration into this may be required as the words “food”, “security” and “risk” together may also indicate information linked to food-borne illnesses whose keywords did not make it to the word network. This cluster ultimately alludes to the risk of famine in a localised region due to climate change, which is likely to act as a driver for disease outbreak as

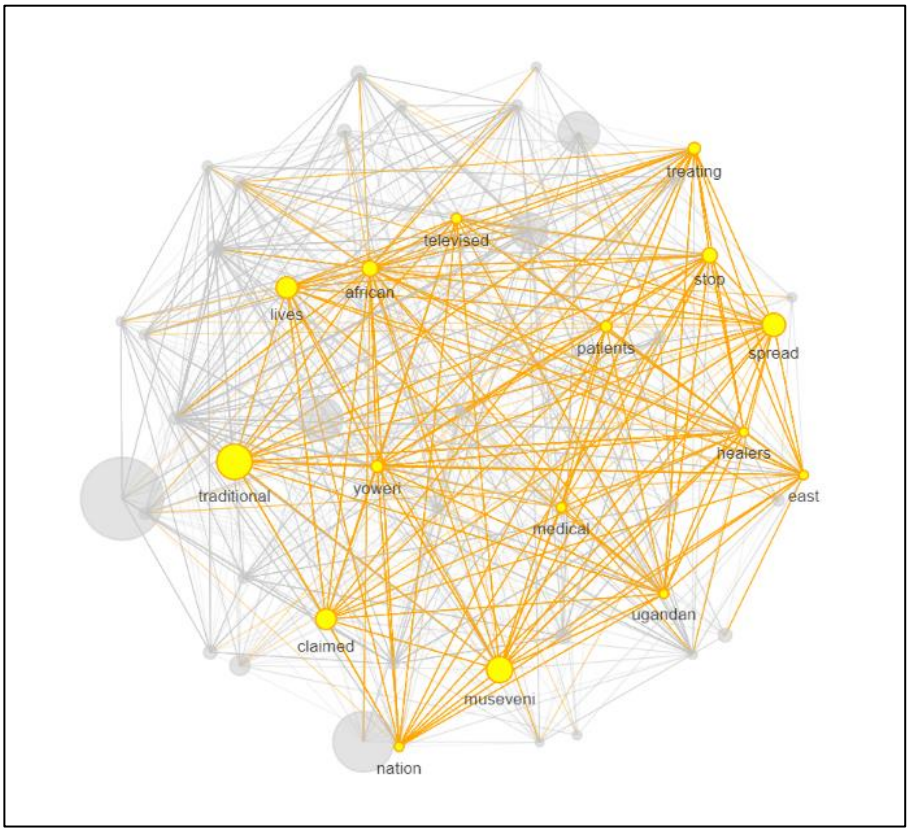

Figure 12

View of Clusters 3, highlighted in yellow, taken from Microsoft Edge Web Browser

populations there become vulnerable due to malnutrition (which is a non-communicable disease).

Cluster 3 (Figure 12), highlights seventeen words, including: “Ugandan”, “healer”, “Museveni”, “traditional”, “televised” and “stop”. One word not highlighted but connected through links is “Ebola”. This cluster is almost legible as a title to those already familiar with the name of the Ugandan president as “President Musevenni, in a televised message, told traditional healers to stop treating Ebola patients, as the outbreak spreads through the country and has claimed lives”. The topic of this cluster is clearly defined and is suggestive of the worsening of the ongoing Ebola epidemic in Africa due to the addition of complications potentially caused by the involvement of traditional healers.

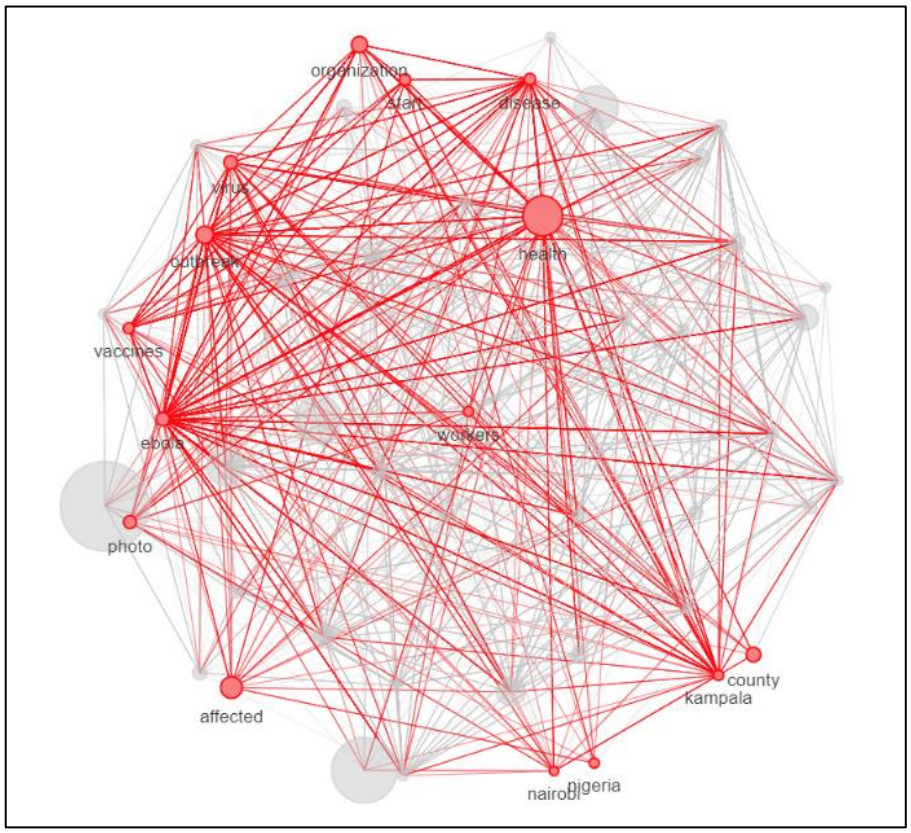

Figure 13

View of Clusters 4, highlighted in red, taken from Microsoft Edge Web Browser

Cluster 4 (Figure 13) highlights fourteen words while significantly linking to eight more. Key highlights include “vaccines”, “Ebola”, “disease”, “outbreak”, “Nairobi”, “Nigeria”, “Kampala”. These can be understood as a vaccine breakthrough in the fight against the ongoing Ebola epidemic also mentioned in Group 2. Based on the place names given, one cannot immediately sure where the events are taking place as the regions are far from each other, therefore further reading would be required.

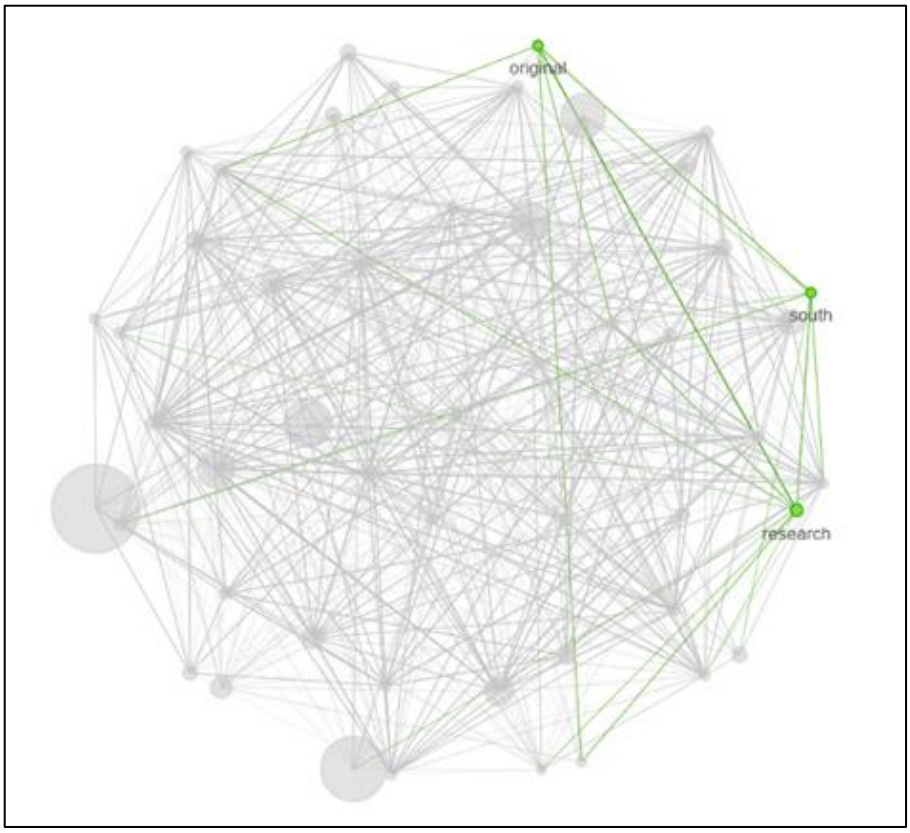

Figure 14:

View of Clusters 5 highlighted in green, taken from Microsoft Edge Web Browser

Cluster 5, highlighted in green, is the smallest of all clusters with only three highlighted words: “south”,” research” and “original”. This cluster does not contain a lot of clues as to its components, and therefore reading its component articles is required to understand the link between the words.

Clusters 3 and 4 demonstrate the way two clusters can have the same topic at heart, but focus on different areas within it, allowing them to be grouped separately.

The results offer an easy overview of the 420 articles collected. A human mind still has to work out what the actual topics are about and how they relate to one another in real life, but the ideas are summarised in an easily interpretable form.

6 Discussion

An objective evaluation of digital media for driver surveillance will ultimately require an unbiased dataset that links media reports of disease drivers to actual outbreaks. That would enable the determination of the specificity and sensitivity of the system. Notably, it will be difficult to develop such a system because of the circular nature of these data (i.e., the system would be evaluated based on the system used to collect the data). In the absence of independently collected "gold standard" data, one cannot determine the degree to which unknown unknowns are missed.[3]

The purpose of the project is to simplify the job of experts looking to spot signs off early outbreaks by clustering results to form helpful visual aids. With the help of network analysis and clustering algorithms, using computers to group topics together. This could also be a list form, but less user friendly.

Challenges overcome

Searches

- Defining the scope of the results of interested.
- Using specific spelling to in search terms and for RSS links.
- How to optimize search parameters for the thirty-seven diseases
- Utilising the EMM filters: there were over fifty available options, choosing too many or too broadly lead to unspecific results and many duplicate articles which put a lot of pressure on the workflow. The four most relevant ones were eventually settled on.
- Despite the specific search queries and filters applied, a major news event is still able to leach into the results, which is especially frustrating due to the limiting fifty word cap.

Language

- Selecting the languages of the input data: originally, we were going to use English only, but due to the international nature of the Lake Victoria Basin and the promising nature of the translating features of KNIME, the "all" language filter was selected

        - This led to having to decide the main languages to filter for in the translation meta-node and devise an effective translating mechanism
    - Translation nodes
        - Addition of an “emptiness” checker before the start of each language tract. Without this mechanism, when no articles of a particular language appear in the search results, the translation program shuts down, stalling the entire workflow.
        - Working around the maximum limit of words imposed by the translation component-node; Solved by:
            - the translation meta node: splitting the languages to the most common languages (English, French, Swahili, simplified Chinese, and the “others”), and translating them in separate tracks.
            - Capping the amount of input data by keeping the time interval of interest under seven days.
        - Auto translation works fine for smaller texts, but if it reaches a certain threshold, it does not do the job. So, the separate language pipeline could be used in these cases.
        - Need time and opportunity to compare accuracy and efficiency of the different flows

KNIME

- It is a large program, that can easily glitch/ freeze or shut down itself or other programs running on a laptop if it is not strong enough.
    - I used a 128GB; 4GB memory Microsoft Surface Go laptop, with a Pentium Gold processor and ran into the above issues. Meanwhile my Supervisor’s Macintosh could run it with other programs operating in the background without issue.
- R & python plugins
    - They are lost each time KNIME updates to a new version.
    - They cap the word network at fifty words maximum. Computers with better computing power may be able to go to seventy, but that was not possible on my device.

    - Their proper function is crucial to the workflow's performance; however, their glitches are hard to spot and often only noticed when they eventually stall the workflow all the message says is error in the program
    - Date and time restrictions
        - Searching for longer periods leads to the entire workflow shutting down. This is thought to be because of the translation meta-node, but other nodes may also have limits that we have not yet encountered.
    - Different network analysis plugins could be tried out to compare results
- Text mining
    - Pre-processing:
        - Determining what words are "clutter" when limited to only 50 key words was challenging, and took weeks to get focused
    - Despite the specific search queries and filters applied, a major news event is able to leach into our results, which is especially frustrating as we are limited to a 50 node cap; to combat this, we filtered out days of the week, months, numbers, county names as well as names of influential figures and groups, assuming that if the news is relevant enough, then that part of the network is deducible without using the precious space

Lessons Learned:

In doing this project and overcoming its challenges, many lessons were learned, and room identified for improvement.

The accessing old news media was the first and most difficult challenge to overcome. Using an RSS feed creator opened-up many resources that would have otherwise been impossible to use. And while EMM was a vital source of articles, the issue of duplicates often arose. Further work could be done to reduce the number of duplicate articles entering the workflow to free up computational power for other processes.

The computational limitation of both the KNIME program and the devices that ran it were often a limiting factor when it came to the quantity of data that could be analyzed. Because the various individuals who ran the program on their personal computers had different experiences with the workflow, issues could be identified as either being linked to program error or computational power. The project aimed to skew to the weakest link, enabling computers with smaller processing capacity to still run the workflow and achieve results in a reasonable time. This included the addition of the fifty-word limit, as more powerful machines were able to produce networks with up to eighty words, while weaker ones got error messages or crashed completely.

Additionally, in the process of modifying the original DEMETER workflow, opportunities were found to experiment with different methods and materials. The translation meta-node currently offers two methods of translating the input texts, and further investigations would be required to establish which process is the more efficient and accurate. Furthermore, additional inquiries could be conducted into which of the many open-source text mining and data analysis programs work best. This thesis used “R” and “¨Python” as these were the most convenient at the time, but additional comparisons would be beneficial to optimize the process.

Further tests could be conducted on the workings of the text mining step and the accuracy of the co-occurrence counters in place. Experimenting with various levels of co-occurrence counting (e.g.: sing co-occurrences within per fifty words instead of per sentence) to see which gets the more pertinent results, along with refining the pre-processing procedure, would be incredibly valuable future steps for the optimization this method.

Overall, the team is satisfied with the current status of the workflow and has hopes for its future as a tool in the fight to prevent outbreaks of infectious disease.

7 References

1. Jones K (2008) Global trends in emerging infectious diseases. Nature 451:990–993. https://doi.org/10.1038/nature06536
2. Brownstein JS, Freifeld CC, Madoff LC (2009) Digital Disease Detection — Harnessing the Web for Public Health Surveillance. New England Journal of Medicine 360:2153–2157. https://doi.org/10.1056/NEJMP0900702
3. Olson SH, Benedum CM, Mekaru SR, Preston ND, Mazet JAK, Joly DO, Brownstein JS (2015) Drivers of Emerging Infectious Disease Events as a Framework for Digital Detection - Volume 21, Number 8—August 2015 - Emerging Infectious Diseases journal - CDC. Emerging Infectious Diseases, CDC 21:1285–1292. https://doi.org/10.3201/EID2108.141156
4. O'Shea J (2017) Digital disease detection: A systematic review of event-based internet biosurveillance systems. Int J Med Inform 101:15. https://doi.org/10.1016/J.IJMEDINF.2017.01.019
5. Maudlin I, Eisler MC, Welburn SC (2009) Neglected and endemic zoonoses. Philosophical Transactions of the Royal Society B: Biological Sciences 364:2777–2787. https://doi.org/10.1098/rstb.2009.0067
6. Beyrer C, Villar JC, Suwanvanichkij V, Singh S, Baral SD, Mills EJ (2007) Neglected diseases, civil conflicts, and the right to health. Lancet 370:619–627
7. Aagaard-Hansen J, Nombela N, Alvar J (2010) Population movement: A key factor in the epidemiology of neglected tropical diseases. Tropical Medicine and International Health 15:1281–1288. https://doi.org/10.1111/J.1365-3156.2010.02629.X
8. Feldmann H, Czub M, Jones S, Dick D, Garbutt M, Grolla A, Artsob H (2002) Emerging and re-emerging infectious diseases. Med Microbiol Immunol 191:63–74. https://doi.org/10.1007/S00430-002-0122-5
9. Allen T, Murray KA, Zambrana-Torrelio C, Morse SS, Rondinini C, di Marco M, Breit N, Olival KJ, Daszak P (2017) Global hotspots and correlates of emerging zoonotic diseases. Nature Communications 2017 8:1 8:1–10. https://doi.org/10.1038/s41467-017-00923-8
10. Semenza JC, Lindgren E, Balkanyi L, Espinosa L, Almqvist MS, Penttinen P, Rocklöv J (2016) Determinants and Drivers of Infectious Disease Threat Events in

Europe - Volume 22, Number 4—April 2016 - Emerging Infectious Diseases journal - CDC. Emerg Infect Dis 22:581–589. https://doi.org/10.3201/EID2204.151073

11. Origin NRC (US) C on ASGC for S and R to ED of Z, Keusch GT, Pappaioanou M, Gonzalez MC, Scott KA, Tsai P (2009) Drivers of Zoonotic Diseases
12. Wepner B, Giesecke S (2018) Drivers, trends and scenarios for the future of health in Europe. Impressions from the FRESHER project. European Journal of Futures Research 6:1–10. https://doi.org/10.1007/S40309-017-0118-4/TABLES/1
13. Gething PW, Smith DL, Patil AP, Tatem AJ, Snow RW, Hay SI (2010) Climate change and the global malaria recession. Nature 465:342–345. https://doi.org/10.1038/NATURE09098
14. Morse SS, Mazet JAK, Woolhouse M, Parrish CR, Carroll D, Karesh WB, Zambrana-Torrelio C, Lipkin WI, Daszak P (2012) Prediction and prevention of the next pandemic zoonosis. The Lancet 380:1956–1965
15. Brice J, Soldi R, Alarcon-lopez P, Guitian J, Drewe J, Baeza Breinbauer D, Torres-cortés F, Wheeler K (2021) The relation between different zoonotic pandemics and the livestock sector. Publication for the committee on the Environment, Public Health, and Food Safety, Policy Department of Economic, Scientific and Quality of Life Policies
16. Meijer N, Filter M, Clark B, Józwiak Á, Comber R, Mylord T, Kerekes K, Willems D, van Asselt E, Frewer L, Czyz M, Fischer A, Marvin H (2018) Project DEMETER: Concept Note for an Emerging Risks Knowledge Exchange Platform (ERKEP) Framework. EFSA Supporting Publications 15:. https://doi.org/10.2903/sp.efsa.2018.en-1524
17. Tetra Tech, Land Trees and Sustainability (LTS) Africa (2016) Lake Victoria Basin Ecosystem Profile Assessment Report
18. Geheb K, Kalloch S, Medard M, Nyapendi AT, Lwenya C, Kyangwa M (2008) Nile perch and the hungry of Lake Victoria: Gender, status and food in an East African fishery. Food Policy 33:85–98. https://doi.org/10.1016/J.FOODPOL.2007.06.001
19. Beuving JJ (2010) Playing Pool Along the Shores of Lake Victoria: Fishermen, Careers and Capital Accumulation in the Ugandan Nile Perch Business. Africa 80:224–248. https://doi.org/10.3366/AFR.2010.0203
20. Fèvre EM, de Glanville WA, Thomas LF, Cook EAJ, Kariuki S, Wamae CN (2017) An integrated study of human and animal infectious disease in the Lake Victoria

crescent small-holder crop-livestock production system, Kenya. BMC Infectious Diseases 2017 17:1 17:1–14. https://doi.org/10.1186/S12879-017-2559-6

21. Prabhu M (2022) When it rains: how the global climate crisis is already threatening public health on the shores of Lake Victoria. In: Gavi, The Vaccine Alliance. https://www.gavi.org/vaccineswork/when-it-rains-how-global-climate-crisis-already-threatening-public-health-shores-lake-victoria. Accessed 5 Feb 2022
22. Marshall B, Ezekiel C, Gichuki J, Mkumbo O, Sitoki L, Wanda F (2009) Global warming is reducing thermal stability and mitigating the effects of eutrophication in Lake Victoria (East Africa). Nature Precedings 2009 1–1. https://doi.org/10.1038/npre.2009.3726.1
23. Berthold MR, Cebron N, Dill F, di Fatta G, Gabriel TR, Georg F, Meinl T, Ohl P, Sieb C, Wiswedel B (2006) KNIME: the Konstanz Information Miner. In: Workshop on Multi-Agent Systems and Simulation (MAS&S), 4th Annual Industrial Simulation Conference (ISC). Palermo, Italy, pp 58–61
24. European Commission (2018) Europe Media Monitor (EMM). Ispra
25. Munyua P, Bitek A, Osoro E, Pieracci EG, Muema J, Mwatondo A, Kungu M, Nanyingi M, Gharpure R, Njenga K, Thumbi SM (2016) Prioritization of Zoonotic Diseases in Kenya, 2015. PLoS One. https://doi.org/10.1371/journal.pone.0161576
26. Sekamatte M, Krishnasamy V, Bulage L, Kihembo C, Nantima N, Monje F, Ndumu D, Sentumbwe J, Mbolanyi B, Aruho R, Kaboyo W, Mutonga D, Basler C, Paige S, Behravesh CB (2018) Multisectoral prioritization of zoonotic diseases in Uganda, 2017: A One Health perspective. PLoS One 13:e0196799. https://doi.org/10.1371/JOURNAL.PONE.0196799
27. Hao Y, Baik J, Fred S, Choi M (2022) Comparative analysis of two drought indices in the calculation of drought recovery time and implications on drought assessment: East Africa's Lake Victoria Basin. Stochastic Environmental Research and Risk Assessment 36:1943–1958. https://doi.org/10.1007/S00477-021-02137-3

8 Appendix

Appendix A Table of Figures:

Appendix B Search terms and their links

1. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&country=BI%2CKE%2CRW%2CTZ%2CUG&dateto=end_dateT23%3A59%3A59Z&atLeast=aanthra+Trypanosoma+Brucella+%28rift+valley+fever%29+echinococcosis+%28Non+TTyph+Salmonellosis%29+%28Q+fever%29+mycobacterium+cysticercosis+dengue+malmala+leptospirosis+Schistosomiasis+%28yellow+fever%29+rickets+Rickettsiosis+TaeniT+Sarcopsis+Cryptosporidiosis+Leishmaniasis+Ebola+Marburg+%28Crimean-Congo+haemorrhagic+fever%29+%28antimicrobial+resistance%29+%28drug+resistancr%29%28antimicrobial+resistance%29+%28multidrug+resistance%29+%28antibiotic+rresistanc%29+Dermatophilosis+Cryptococcosis+Listeriosis+Aspergillosis+MERS+SASA+Plague+Chikungunya+%28West+Nile+Virus%29&datefrom=start_dateT00%3A00%3A00Z

2. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&country=BI%2CKE%2CRW%2CTZ%2CUG&dateto=end_dateT23%3A59%3A59Z&datefrom=start_dateT00%3A00%3A00Z&category=AnimalHealth
3. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&country=BI%2CKE%2CRW%2CTZ%2CUG&dateto=end_dateT23%3A59%3A59Z&datefrom=start_dateT00%3A00%3A00Z&category=CommunicableDiseases
4. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&country=BI%2CKE%2CRW%2CTZ%2CUG&dateto=end_dateT23%3A59%3A59Z&datefrom=start_dateT00%3A00%3A00Z&category=FoodSafety
5. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&country=BI%2CKE%2CRW%2CTZ%2CUG&dateto=end_dateT23%3A59%3A59Z&datefrom=start_dateT00%3A00%3A00Z&category=FoodSecurityFoodAid
6. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&dateto=2021-08-29T23%3A59%3A59Z&atLeast=Kenya+Burundi+Tanzania+Ruanda+Uganda&datefrom=2021-08-23T00%3A00%3A00Z&category=AnimalHealth
7. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&dateto=end_dateT23%3A59%3A59Z&atLeast=Kenya+Burundi+Tanzania+Ruanda+Uganda&datefrom=start_dateT00%3A00%3A00Z&category=CommunicableDiseases
8. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&dateto=end_dateT23%3A59%3A59Z&atLeast=Kenya+Burundi+Tanzania+Ruanda+Uganda&datefrom=start_dateT00%3A00%3A00Z&category=FoodSafety
9. https://emm.newsbrief.eu/rss/rss?language=all&type=search&mode=advanced&dateto=end_dateT23%3A59%3A59Z&atLeast=Kenya+Burundi+Tanzania+Ruanda+Uganda&datefrom=start_dateT00%3A00%3A00Z&category=FoodSecurityFoodAid
10. https://smartfarmerkenya.com/feed/
11. https://www.africanfarming.net/component/obrss/african-farming-rss-feed
12. https://theorganicfarmer.org/feed/
13. https://theorganicfarmer.org/?s=Animal+health
14. https://eastafrican-agrinews.com/feed/

15. https://ea-agribusiness.com/feed
16. https://www.theeastafrican.co.ke/tea/science-health
17. https://swara.co.ke/category/news/latest-news/east-africa-news-updates/feed/
18. https://www.ajol.info/index.php/ajol/Gsearch/google?q=animal%20health
19. https://ajlmonline.org/index.php/ajlm/gateway/plugin/WebFeedGatewayPlugin/atom
20. https://africa-health.com/feed/

HuVetA

ELECTRONIC LICENSE AGREEMENT AND COPYRIGHT DECLARATION*

Name: Zorka Mwendwa Rakonczay .....……

Contact information (e-mail):.zrakoncz99@outlook.com.....

Title of document (to be uploaded):

Automated News Screening for Emerging Infectious Disease Risk Identification in the....

Lake Victoria Basin.....……..

Publication data of document: November 15, 2022...……...

Number of files submitted: …two.……...…....

By accepting the present agreement the author or copyright owner grants non-exclusive license to HuVetA over the above mentioned document (including its abstract) to be converted to copy protected PDF format without changing its content, in order to archive, reproduce, and make accessible under the conditions specified below.

The author agrees that HuVetA may store more than one copy (accessible only to HuVetA administrators) of the licensed document exclusively for purposes of secure storage and backup, if necessary.

You state that the submission is your original work, and that you have the right to grant the rights contained in this license. You also state that your submission does not, to the best of your knowledge, infringe upon anyone's copyright. If the document has parts which you are not the copyright owner of, you have to indicate that you have obtained unrestricted permission from the copyright owner to grant the rights required by this Agreement, and that any such third-party owned material is clearly identified and acknowledged within the text of the licensed document.

The copyright owner defines the scope of access to the document stored in HuVetA as follows (**mark the appropriate box with an X)**:

☒ I grant unlimited online access,

☐ I grant access only through the intranet (IP range) of the University of Veterinary Medicine,

☐ I grant access only on one dedicated computer at the Ferenc Hutÿra Library,

☐ I grant unlimited online access only to the bibliographic data and abstract of the document.

Please, define the **in-house accessibility of the document** by marking the below box with an **X**:

☒ I grant in-house access (namely, reading the hard copy version of the document) at the Library.

If the preparation of the document to be uploaded was supported or sponsored by a firm or an organization, you also declare that you are entitled to sign the present Agreement concerning the document.

The operators of HuVetA do not assume any legal liability or responsibility towards the author/copyright holder/organizations in case somebody uses the material legally uploaded to HuVetA in a way that is unlawful.

Date: Budapest, ...15….day ……11……..month……2022……..year

Author/copyright owner signature

HuVetA Magyar Állatorvos-tudományi Archívum – Hungarian Veterinary Archive *is an online veterinary repository operated by the Ferenc Hutÿra Library, Archives and Museum. It is an electronic knowledge base which aims to collect, organize, store documents regarding Hungarian veterinary science and history, and make them searchable and accessible in line with current legal requirements and regulations.*

HuVetA relies on the latest technology in order to provide easy searchability (by search engines, as well) and access to the full text document, whenever possible.
Based on the above, HuVetA aims to:

- *increase awareness of Hungarian veterinary science not only in Hungary, but also internationally;*
- *increase citation numbers of publications authored by Hungarian veterinarians, thus improve the impact factor of Hungarian veterinary journals;*
- *present the knowledge base of the University of Veterinary Medicine Budapest and its partners in a focussed way in order to improve the prestige of the Hungarian veterinary profession, and the competitiveness of the organizations in question;*
- *facilitate professional relations and collaboration;*
- *support open access.*

I hereby confirm that I am familiar with the content of the thesis entitled

Automated News Screening for Emerging Infectious Disease Risk Identification in the Lake Victoria Basin

written by Zorka Mwendwa Rakonczay

which I deem suitable for submission and defence.

Date: Budapest, 14/11/2022

Ákos Józwiak

Supervisor name and signature

Digital Food Chain Education,
Research, Development and Innovation Institute

Thesis progress report for veterinary students

Name of student: Zorka Mwendwa Rakonczay

Neptun code of the student: OAA5WE

Name and title of the supervisor: Ákos Józwiak, deputy director

Department: Digital Food Chain Education, Research, Development, and Innovation Institute

Thesis title: Automated News Screening for Emerging Infectious Disease Risk Identification in the Lake Victoria Basin

Consultation – 1st semester

	Timing			Topic / Remarks of the supervisor	Signature of the supervisor
	year	month	day		
1.	2022	03	07	Scope, objectives, storyline, introduction to data analysis	
2.	2022	03	21	Literature review principles, data analysis pipeline first draft	
3.	2022	04	20	Data analysis pipeline: extending the input sources	
4.	2022	05	02	Data analysis pipeline: translation nodes	
5.	2022	05	16	Data analysis pipeline: network analysis and visualisation	

Grade achieved at the end of the first semester: ……5……..

Consultation – 2nd semester

	Timing			Topic / Remarks of the supervisor	Signature of the supervisor
	year	month	day		
1.	2022	07	26	Thesis first draft: literature review	
2.	2022	08	22	Thesis next draft: materials and methods	
3.	2022	09	14	Data analysis pipeline: finalisation and first analysis	
4.	2022	09	30	Thesis next draft: results and discussion	

5.	2022	10	10	Thesis finalisation	

Grade achieved at the end of the second semester: …….. 5

The thesis meets the requirements of the Study and Examination Rules of the University and the Guide to Thesis Writing.

I accept the thesis and found suitable to defence,

……...
signature of the supervisor

Signature of the student: ……

Signature of the secretary of the department: ……..……

Date of handing the thesis in…15.11.2022……....