VINCENT: A visual analytics system for investigating the online vaccine debate

Anton Ninkov[1*], Kamran Sedig[1]

1. INSIGHT Lab, Western University, CANADA

Abstract

This paper reports and describes VINCENT, a visual analytics system that is designed to help public health stakeholders (i.e., users) make sense of data from websites involved in the online debate about vaccines. VINCENT allows users to explore visualizations of data from a group of 37 vaccine-focused websites. These websites differ in their position on vaccines, topics of focus about vaccines, geographic location, and sentiment towards the efficacy and morality of vaccines, specific and general ones. By integrating webometrics, natural language processing of website text, data visualization, and human-data interaction, VINCENT helps users explore complex data that would be difficult to understand, and, if at all possible, to analyze without the aid of computational tools.

The objectives of this paper are to explore A) the feasibility of developing a visual analytics system that integrates webometrics, natural language processing of website text, data visualization, and human-data interaction in a seamless manner; B) how a visual analytics system can help with the investigation of the online vaccine debate; and C) what needs to be taken into consideration when developing such a system. This paper demonstrates that visual analytics systems can integrate different computational techniques; that such systems can help with the exploration of online public health debates that are distributed across a set of websites; and that care should go into the design of the different components of such systems.

Keywords: Visual Analytics, Public Health, Vaccine Debate, Webometrics, Natural Language Processing, Data Visualization, Human-Data Interaction

Abbreviations: Visual Analytics System (VAS), Multi-Dimensional Scaling (MDS), Natural Language Processing (NLP), Natural Language Understanding (NLU), VINCENT (VIsual aNalytiCs systEm for investigating the online vacciNe debaTe)

*Correspondence: Anton Ninkov- aninkov@uwo.ca

DOI: 10.5210/ojphi.v11i2.10114

1. Introduction

As the use of the Internet expands, people engage in social discourse and debate in different areas of interest, generating a great deal of online data. One broad area of interest generating

such online information is public health. Public health data is often large, complex, and difficult, if at all possible, to analyze without the aid of computational tools. Public health informatics is a research area that focuses on "the systematic application of information, computer science, and technology to public health practice, research, and learning" [1]. Visual analytics systems (VASes) can be of great utility in public health informatics [2]. VASes are computational tools that combine data visualization, human-data interaction, and data analytics. They allow users to interactively control data visualizations to change how data is analyzed and presented to them. VASes make it possible for users to quickly make sense of online data that would otherwise be impossible or take more time and effort to accomplish.

In this paper, we report and describe a VAS designed to help public health stakeholders (users) make sense of data from websites involved in the online debate about vaccines. The VAS, VINCENT (VIsual aNalytiCs systEm for investigating the online vacciNe debaTe), allows users to explore visualizations of data from a group of 37 vaccine-focused websites (listed in Appendix 1). These websites range in their position on vaccines, topics of focus about vaccines, geographic location, and sentiment towards the efficacy and morality of vaccines, specific and general ones. While numerous VASes have been developed and studied previously, VINCENT is novel in that it integrates webometrics (i.e., co-link analysis), natural language processing (i.e., text-based emotion analysis), data visualization, and human-data interaction.

The research questions this paper examines are as follows:

1. Is it feasible to integrate webometrics, natural language processing of website text, data visualization, and human-data interaction in a seamless manner to develop a VAS?
2. Can such a VAS help with the investigation of the online vaccine debate?
3. What are some of the considerations that need to go into developing such a system?

The remainder of this paper is organized as follows. Section 2 provides a conceptual and terminological background--i.e. vaccine debate, visual analytics systems, webometrics, and natural language processing. Section 3 describes the development of VINCENT and includes an in-depth discussion of the various components of the VAS. Section 4 provides a summary and conclusions.

2. Background

This section provides a conceptual and terminological background for this paper. We will first describe the issue that VINCENT aims to clarify--i.e. the vaccine debate. Next, we will review visual analytics. Finally, we will discuss the data analytics methods (webometrics and natural language processing) that are used in this research.

2.1 Vaccine Debate

In light of increased recent news coverage of outbreaks of diseases such as measles and whooping cough, the anti-vaccination movement appears to be a new and emerging phenomenon

[3-5]. The World Health Organization has listed the rise of the anti-vaccination campaign as a top ten health emergency in 2019 [6]. However, anti-vaccination views and sentiments are not a recent development. Since Edward Jenner's discovery of the smallpox vaccine, vaccination has garnered much attention both positive and negative. From the beginning, some have felt that the practice of vaccination is ineffective, violates personal freedoms, and is "unchristian" [7]. However, the Centers for Disease Control reports that vaccines have had a positive impact on global health and are "one of the greatest achievements of biomedical science and public health" [8].

Despite the medical community's unified support of immunization, there are many reasons for the persistence of anti-vaccine views. There is some suggestion that increasingly polarized political views (especially in the United States) have generated an environment in which the rejection of scientific facts has become more prevalent and accepted [9]. This erosion of trust in scientific findings among segments of the population may also contribute to this increased polarization. Additionally, the rise in accessibility to, and widespread use of, the Internet has played a role in amplifying the voice of the anti-vaccination movement [10,11]. [11] states, "The connective power of the Internet brings together those previously considered on the fringe. Members of marginalized groups (e.g. Holocaust deniers, 9/11 'Truthers', AIDS deniers) can easily and uncritically interact with like-minded individuals online… Anti vaccine groups have harnessed postmodern ideologies and by combining them with Web 2.0 and social media, are able to effectively spread their messages". Hence, the Internet plays an important role in the anti-vaccination movement, helping spread their message and promoting their views on vaccination dangers.

The polarity of the vaccine debate is creating a clear divide and this has been revealed through both qualitative classification of inlinks [12] and quantitative co-link analysis [13]. The divide is having harmful effects on the health of the general population. "Providers and policymakers must begin to recognize the jagged, context-dependent, equifinal nature of how parents sort through vaccination-related information or account for their vaccination decisions in order to reverse declining vaccination rates" [14]. Some of the themes of the discussion that have developed in this polarized debate include those related to autism and vaccines, evil government conspiracies, and technological developments [15]. A more automated approach that would allow an analysis of such online discussions and information could help illuminate this public health problem.

2.2 Visual Analytics Systems (VASes)

In today's environment of big data, people are often victims of information overload. They can get lost in and overwhelmed by the voluminous data and its meaning that they encounter [16]. By combining human insight with powerful data analytics and integrated data visualizations and human-data interaction, VASes can help alleviate this problem. VASes can enable potential stakeholders to make sense of data. "Just like the microscope, invented many centuries ago, allowed people to view and measure matter like never before, (visual) analytics is the modern equivalent to the microscope" [17].

VASes are composed of three integrated components: an analytics engine, data visualizations, and human-data interactions [18,19]. The analytics engine pre-processes and stores data (e.g., data cleaning & fusion), transforms it (e.g., normalization), and analyzes it (e.g., multi-dimensional scaling, emotion analysis) [20]. Examples of data analytics techniques that can be integrated into the analytics engine are webometrics and natural language processing (NLP). Data visualizations in a VAS can be visual representations of the information derived from the analytics engine. Visualizations extend the capabilities of individuals to complete tasks by allowing them to analyze data in ways that would be difficult or impossible to do otherwise [19,21]. For instance, a scatterplot can be used to visually represent coordinates of entities, and this, in turn, helps the user determine quickly the proximity between data points. Human-data interaction is used in VASes to allow the user to control the data they see and the way the data is processed. Interaction in VASes supports users through distributing the workload between the user and the system during their exploration and analysis of the data [18,22,23]. Some examples of the numerous human-data interactions that can be incorporated into VASes include filtering, scoping, and drilling of data [24], with each interaction supporting different epistemic actions on information by the user.

One of the theories that can help with the conceptualization of VASes is general systems theory. Systems theory views a system as composed of entities, properties, and relationships [25]. VASes are complex, multi-level systems, consisting of systems within systems [18]. These multi-level systems consist of super-systems, systems made up of other systems, and sub-systems, together making up a super-system [25]. With this understanding of systems theory, we can see how VASes work. When building and examining VASes, the interactions of the user with the system can have an impact on any of these levels. At the highest level, super-system interactions will change the overall display of the VAS. At lower levels, the interaction sub-system will change specific components of the system. These interactions, regardless of level, are important to the functioning of the VAS and necessary for making sense of the data being presented.

There are several resources available to assist in developing VASes. Two of the most widely used VAS resources include the open source D3.js JavaScript library [26] and Tableau software [27]. The advantage of D3.js is the almost limitless customization capabilities it offers, as it is bound only by programming constraints, and the fact that it is open source. However, the time, effort, and programming skills required by developers to create systems is greater for D3.js than other solutions, as there are fewer templates and starting points to work with. Tableau, on the other hand, is a proprietary data visualization software that provides users with the ability to develop interactive data visualizations with only minimal coding effort. One feature that makes Tableau particularly appealing is that there are several templates available to users to build their own interactive visualizations. As well, Tableau allows users to create dashboards easily, which place multiple interactive visualizations together in one system that automatically connects data together. While both D3.js and Tableau can be useful solutions for developing visual analytics, Tableau has been used in this research because of its ability to create a functioning and useful visual analytics system while at the same time reducing the programming workload.

VASes incorporate one or more data analysis techniques including (but not limited to) supervised learning (i.e. decision trees or SVM), or cluster analysis [16]. Previous VAS research

has incorporated similar data analysis techniques used in VINCENT. For example, researchers have investigated how incorporating multi-dimensional scaling of co-occurrence data (discussed in section 2.3) in VASes help users investigate entities and identify clusters in a variety of data sets [28,29]. As well, researchers have utilized emotion analysis (discussed in section 2.4) in VASes that help users investigate online text from both social media and the general web regarding a variety of topics [30-32]. Both these data analysis techniques have been implemented in VAS research independently of each other, however there have been no published studies examining the integration of the two techniques in a single VAS, as proposed in VINCENT.

2.3 Webometrics

Webometrics is the "quantitative study of web-related phenomena" [33]. With the ever-increasing adoption of the Internet, the various metrics used for analyzing its data, such as hyperlinks, become important to investigate. Two types of webometrics research methods exist: evaluative and relational [34,35].

Evaluative webometrics can include examining webpages for properties such as (but not limited to) the number of external inlinks they receive (links directed to a website from another website) and the website location [12,35,36]. Examining the number of inlinks a website receives has been shown to be an indicator of performance in a variety of measures for organizations [33,37-39]. Additionally, geographic location has demonstrated to be a valuable resource in conducting evaluative webometrics research [40,41].

Relational webometrics focuses on "providing an overview of the relationships between different actors" [35]. Co-occurrence measurements to indicate similarity are important for relational analysis in webometrics [35,36,42]. The concept behind this method is that the more entities share occurrences, the more likely they are to be similar in some way [34]. This method can apply to webometrics in the study of co-links to help analyze similarity in terms of shared online presence between websites [41,43-45]. To represent and examine co-link data, numerous studies have been conducted with multi-dimensional scaling (MDS)--studies using MDS to analyze business [45], university [46], government [47], and political domains [48,49]. All these studies found that using MDS to analyze co-links generated worthwhile insights into the data.

2.4 Natural Language Processing (NLP)

NLP is a vast area of research that focuses on using computational methods to understand and produce human language content [50]. NLP encompasses a wide range of research topics, two of which are text-based emotion detection and word frequency [51].

Text-based emotion detection has been examined previously in NLP research [51-54]. One resource, in particular, that has been developed that makes it possible for researchers to automatically conduct this type of analysis is IBM's Natural Language Understanding (NLU) API [55]. The NLU API (formerly referred to as the AlchemyAPI) has been widely used by many researchers to study topics, sentiments, and emotions in text [56-59]. The NLU API allows researchers to either input text directly or pull text from URLs of webpages and return a number of different NLP analyses, one of which is emotion analysis. Furthermore, the NLU API can not

only detect emotion on the entirety of a text/webpage, but can also return emotion scores for specified target words/phrases [55].

The study of word frequency in text has been examined and used in NLP research [60-62]. One of the main concerns for word frequency analysis is how to manage meaningless or unimportant words. In English, like any language, there are many words that are repeated frequently that are not necessarily the key point of interest to a reader. Some of the more obvious examples of these words are “the”, “and”, and “of”. Other types of undesirable words can exist depending on the domain of interest (e.g., dates or numbers). To deal with this issue, the technique of filtering for a list of stop words has been used, and preliminary lists of these words have been created that allow researchers to automatically exclude words that are not of interest [63]. To display word frequency data, word clouds have been used successfully [60-62]. Word clouds display identified words in varying sizes, with larger words being the more frequent. Word clouds are useful because they allow users to quickly see the most prevalent words of a text document and enable them to make quick assessments about what the overall text of a document/website may be discussing.

3. System Design

The design of VINCENT, displayed in Figure 1, consists of three primary components: the analytics engine, data visualizations, and human-data interactions. In this section, we will discuss these components of the system and explain how the data was collected and managed. VINCENT was developed in Tableau, version 10.5.

Figure 1: VINCENT: A Visual Analytic System

3.1 Analytics Engine

The analytics engine of VINCENT utilizes webometrics and NLP as its data analysis methods. In this section, we will discuss how, using these methods, data was collected, transformed, and processed. For webometrics, this included leveraging inlink data and geographic location data. For NLP, this included leveraging word frequencies and emotion detection analysis.

The list of 37 vaccine websites (Appendix 1) in VINCENT was created based on a list produced for a study on co-link analysis of vaccine websites which included a total of 62 websites [13]. Websites from that study could be included in VINCENT if they had a central focus on the vaccine debate and a minimum of 200 inlinking domains. The reduction from the original list was primarily due to the elimination of website that were more minor, websites that had increased their scope beyond just vaccination, and websites that had gone obsolete or merged with another website to form a new website. This list should not be viewed as comprehensive of all vaccine websites, but rather as a sample of some of the more major English-based ones from both sides of the polarized debate.

3.1.2 Webometrics

Inlink data was collected from each website using MOZ's Link Explorer tool (https://www.moz.com/link-explorer). Diverging from some of the previous webometrics research using inlink data, which mostly investigated inlinks coming from pages (41,43–45) and sites (13), VINCENT uses inlink data about the inlinking domains. Changes in September 2018 to the data provided by MOZ required us to adapt and examine the feasibility of using domain-level inlink data. After comparing domain-level inlink data to data collected for a previous study (13), we determined that the domain-level inlink data was a suitable replacement and would be used in the analytics engine of VINCENT.

The shared online presence between the set of websites (Appendix 1) was analyzed using MDS. Following similar data analysis techniques to that of previous MDS research [13], the inlink data collected on each website was used to create a similarity matrix, which is based on the number of co-links each website shared with one another. Using a computer program originally developed for a previous study [13], this co-link data was generated from the collected raw data. Using the output co-link matrix, the data was input into SPSS version 25 and an MDS analysis was conducted. The results of this analysis provided a scatter plot in which each data point was plotted according to the number of co-links they shared, or in other words their shared online presence. Websites that shared more inlinks (and therefore more online presence) were more similar and plotted closer together, while those with fewer inlinks were plotted further away from each other. The goodness of fit between the output scatter plot and the co-link matrix had a stress value of less than 0.05, which suggests a good fit between the two.

Data was also collected regarding the geographic location of the websites. This data was collected through two primary means. The first way of collecting location data was through the sites themselves. Many of the websites had identifying information about the managing owner or organization. The data usually came from an "about us" or "contact us" page and required manual labor to find. For those that did not indicate on their website a location, ICANN WHOIS

registration data was collected. For each of the various collected locations, latitude and longitude coordinates were generated to plot each website on the map of website locations.

3.1.2 NLP

Word frequency data was collected using the following process. First, each website was analyzed and crawled using InSite5, a software package developed by InSpyder (https://www.inspyder.com/products/InSite). With this software, we were able to obtain a CSV export file containing a list of all the words contained on each website, along with the frequency of those word occurrences. After collecting all the raw data about each website, the word frequency lists were filtered to meet the requirements of our analysis. In other words, we wanted only unique words related to the vaccine debate to be displayed. In this effort, we manually created a stop words list to remove irrelevant or common words. The list was built, first, using the Natural Language Toolkit list of stop words for English [63]. This list of stop words contains some of the most common English words (e.g., “I”, “you”, “too”). From this starting point, the list was expanded to include words that needed to be removed including, but not limited to, letters (e.g., “A”, “B”), dates (e.g., “January”, “Wednesday”), self-reference names (e.g., “NVIC”, “Voices for Vaccines”), people’s names (e.g., “Tom”, “Katie”), Internet words (e.g., “blog”, “post”), and common vaccine debate words (e.g., “vaccines”, “vaccination”). In total, the stop words list, used to refine the word frequency data, consists of 1231 words.

After finalizing each of the website’s individual word frequency list, combined word frequency lists were created for 3 sets of websites: all websites, anti-vaccine websites, and pro-vaccine websites. For each word, the sum of the word frequency was normalized by sum of the total number of words in that set. This generated a proportional count of each word’s presence on the website for each website’s top 25 words. This was a more accurate reflection of the presence of the word on the site rather than simply counting the word frequency totals as some sites had more total words than others. With these proportional word frequencies generated, a list of top 25 words for the 3 sets of websites was also created: all websites, anti-vaccine websites, and pro-vaccine websites.

Text-based emotion detection in the website was conducted with the use of IBM’s NLU tool. This tool provides NLP automation through the use of their API and, specifically, can do targeted phrase emotion detection. A user can input text or a URL of a webpage of interest and specify target phrases of interest. The NLU API will return scores for the level of emotion detected for those phrases. Five different emotions (joy, fear, anger, sadness, and disgust) are provided for analysis, which is an overrepresentation of negative emotions [64]. For this system, we did not want to bias our data by over-representing negative emotions. Consequently, the data was cleaned up by merging the 4 negative emotions into one and the labels were changed to reflect a binary of positive emotion (joy) and negative emotion (fear, anger, sadness, and disgust). The vaccines of interest that were examined included: flu, MMR, measles, chicken pox, whooping cough, HPV, polio, hepatitis B, and meningitis. The text was processed using the NLU API’s targeted emotion analysis tool. For each of the vaccines, we manually sampled 2 webpages that contained meaningful discussion about the specified vaccine. Several alternate ways of referencing the vaccines were all targeted. For example, with the MMR, targeted phrases included “MMR”, “MMR Vaccines”, and “MMR Vaccination”, among others. The data from

each of these different phrases for a vaccine were then merged to reflect the total emotion detected about the specified vaccine.

3.2 Data Visualizations

VINCENT is comprised of four main visualization components: an online presence map, a word cloud, a map of website locations, and an emotion bar chart. Each of these visualizations represents an important aspect of the websites' information and involves some type of webometrics or NLP data analytics. In this section, each of these visualizations will be discussed, looking at the decisions that were made to represent the data.

3.2.1 Online Presence Map

The online presence map, displayed in Figure 2, is a representation of the hyperlink data analyzed from each website. The generated MDS scatter plot map of the websites displays each website in proximity to each other based on their shared online presence. Websites that are plotted closer together share more online presence, while those plotted further away share fewer. Based on this map, polarity between the anti- and pro-vaccine websites was evident, similar to findings in previous related research [13]. All anti-vaccine websites ended up on the left side of the map, while all pro-vaccine websites are located on the right side with a space in the middle dividing the two. To display the existence of this polarity, a line dividing the two groups of websites was added to the map with labels for the anti-vaccine and pro-vaccine sides.

Online presence for each website was encoded as a circle representing each of the websites. In this representation, each of the circles was sized based on their total number of inlinking domains. The larger a circle, the more inlinks and, therefore, the larger online presence it has. For reference, the site with the most inlinking domains (9,986) is immunize.org, while the site with the fewest inlinking domains (248) is Vaccine Injury Help Center.

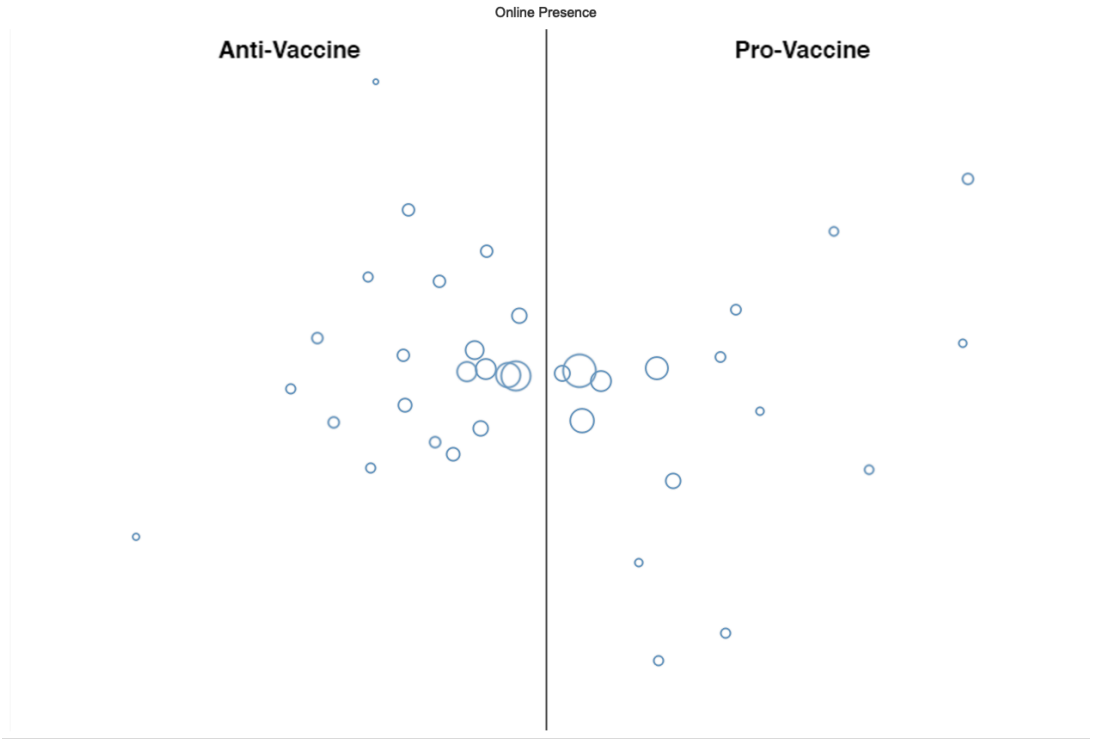

Figure 2: MDS Similarity Map

3.2.2 Word Cloud

The word cloud, displayed in Figure 3, is a representation of the 25 most common unique words that are related to the vaccine debate from each website or group of websites. Words are sized based on the frequency with which they appeared on the website or group of websites. The bigger a word is on the word cloud, the more frequently it is used on the website, while the smaller a word is, the less frequently it is used. Each word was colored differently to assist with differentiating words from each other.

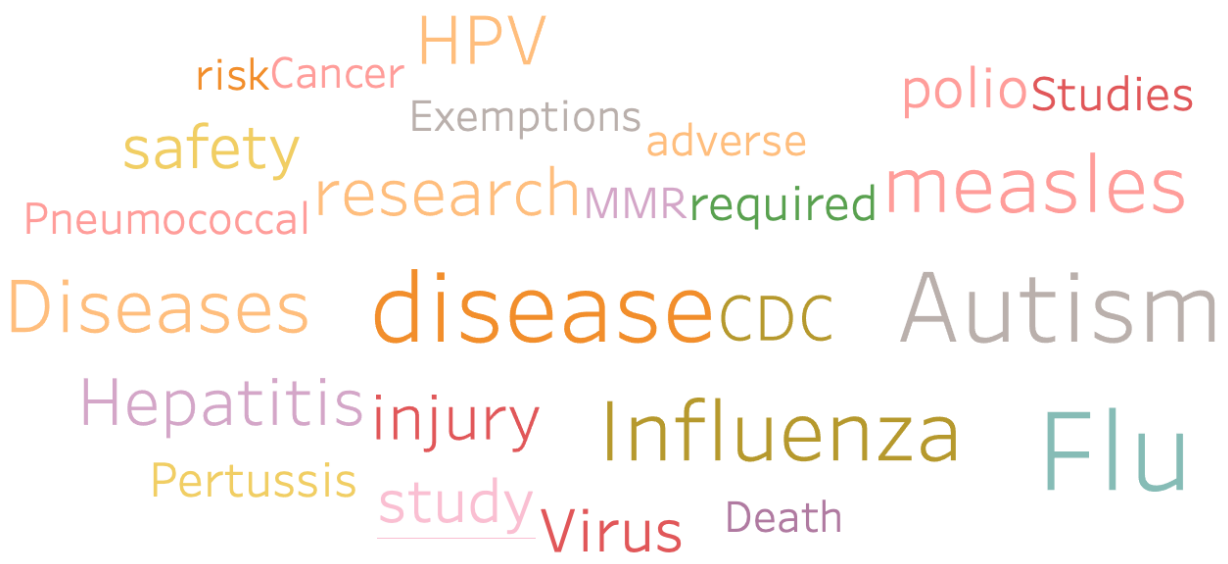

Figure 3: Word Cloud for all Websites

3.2.3 Map of Website Locations

The map of website locations, displayed in Figure 4, shows a representation of the locations of each website on a world map. Website location is an important piece of data as it allows users of the system to explore the geographic diversity of the websites and identify where clusters of websites may exist. Similar to the online presence map, the website locations use circles to encode each website. Different from the online presence map, the circles were all sized equally to help the user see the location of each website, and to avoid confusion with excessive overlapping and occlusion of the circles.

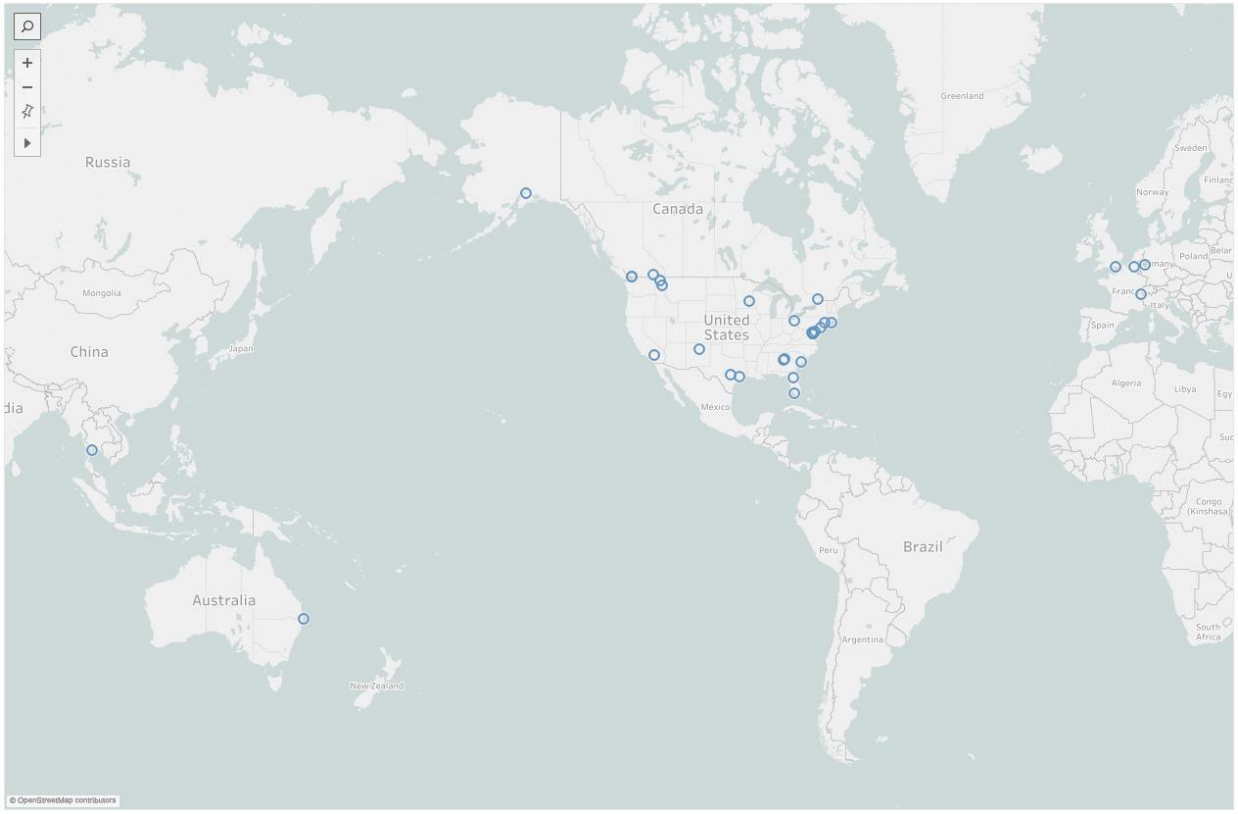

Figure 4: Map of Website Locations

3.2.4 Emotion Bar Chart

The emotion bar chart, displayed in Figure 5, represents positive and negative emotions for a selection of each website's text about a set of vaccines. The two bar charts represent the negative (red) and positive (green) emotions detected by the API. Each bar is composed of individual rectangles that refer to individual websites in the set studied. The width of each of these individual rectangles represents the degree of detected emotion on that specific website. The wider the rectangle, the more that emotion is detected. The entire bar is made up of all the smaller rectangles (websites). This bar then represents the overall detected emotion in the text of the complete website set. The negative and positive bar charts will change in response to the data set that is selected. This will be discussed in more detail in Section 3.3.2.

Figure 5: Emotion Bar Chart

3.3 Human-Data Interactions

To support users to gain insight into the data and explore the online vaccine debate, many interactions are built into VINCENT. These interactions take place on a global level as well as in the sub-systems of VINCENT. In this section we will explore these interactions and discuss how they will assist users to explore the data.

3.3.1 Global System Interactions

There are several interactions that users can perform on VINCENT that occur at the global system level. These interactions not only affect displayed data at individual, sub-system levels of VINCENT, but also change displayed data at the level of the whole system. Global system interactions in VINCENT include website selection and filtering of websites.

The website selection interaction allows users to focus on a single website. Using this interaction (see Figure 6), users can highlight a single website's data throughout the system in order to determine quickly the website's position on vaccination, online presence, location in the world, and emotion about specific vaccines. Consider the following use case. A user is interested in learning more about the website "SaneVax". They would select this website (Figure 6) from the existing options. VINCENT would then highlight the data points associated with this website, as displayed in Figure 7. For this selected website, the user can immediately find that the website's position is anti-vaccine, that it has strong online presence, that it is located in North Western part of North America, that it has more negative emotions regarding vaccines than positive, and that it discusses many issues related to HPV (i.e., Cervarix, Gardasil, Cancer, Silgard, HPV).

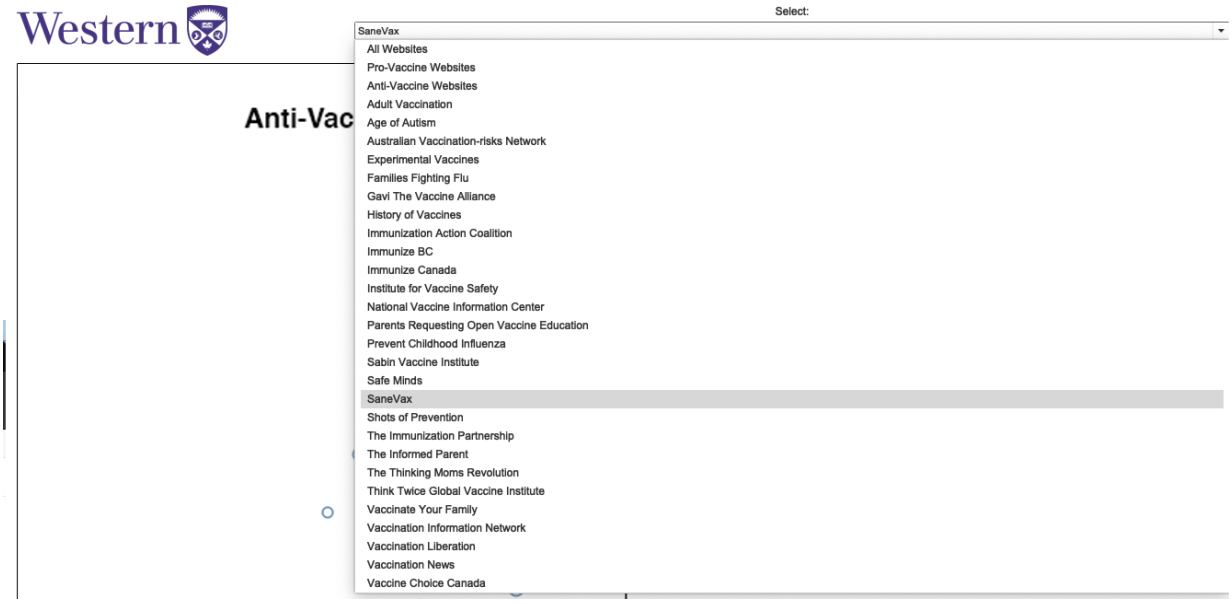

Figure 6: Website Selection Interaction

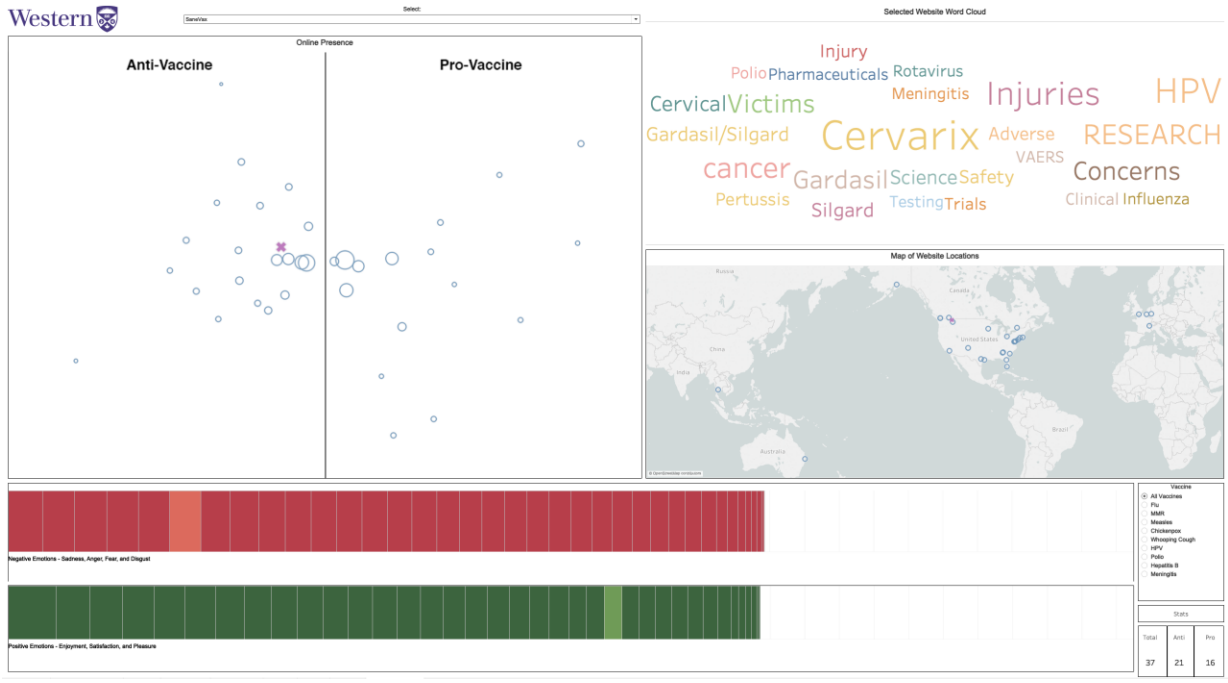

Figure 7: VINCENT after Website Selection Interaction

In addition to the website selection interaction, users have the ability to filter the data to focus on a selected group of websites. Users can highlight and select websites using any of the 3 visualizations, thereby filtering and isolating the data points of a subset of websites. This can be done using the online presence map, map of website locations, or emotion bar chart. Consider a sample use case. A user is curious to learn more about the websites located in North Eastern part of North America. The user goes to the map of website locations and picks websites located in that geographic region. In reaction, the data points on the online presence map and the data of the emotion bar charts are filtered to show only these data points, as displayed in Figure 8. Simultaneously, the stat tracker on the bottom right changes to give the user a numeric count of how many websites they are utilizing now, and how many of each vaccine position is included. The user will quickly see that they have selected 15 websites (10 pro-vaccine and 5 anti-vaccine websites), that the websites are wide ranging in shared online presence, and that they have approximately equal degree of positive and negative emotions associated with the vaccines.

Figure 8: Global Filtered Selection (North Eastern North America)

3.3.2 Sub-System Interactions

There are a number of interactions that can be performed at the sub-systems level of VINCENT. These interactions are focused on isolated elements of the system. They include such interactions as filtering the emotion bar chart to display selected vaccines, hovering display elements to expand an information box, and navigating the map of website locations.

The vaccine selection interaction allows users to filter the displayed data on the emotion bar chart. Upon opening VINCENT, the emotion bar chart displays the overall vaccine emotion data. When a user selects a specific vaccine, the bar chart changes to display only the emotion data that is collected about that specific vaccine. Consider a sample use case. A user is curious about the emotions of the entire set of websites regarding the MMR vaccine. The user would select this vaccine (see right-hand panel in Figure 9), and the bar charts change to display the data. The user can immediately see that there is a greater level of negative emotion on the set of websites than positive emotion regarding the MMR vaccine.

Figure 9: Filtered Vaccine Selection (MMR)

Users also have the option to hover over the online presence map, map of website locations, or emotion bar charts to expand an information box (this is referred to in Tableau as a tooltip) about each specific data point. When a user hovers off the data point, the information box disappears. Again, a sample use case is illustrative of this. A user is interested in identifying which of the pro-vaccine websites have the greatest online presence. To do this, the user would examine the online presence map, determine which token on the pro-vaccine side of the map is the largest, and hover the mouse icon over the token to reveal the information (see Figure 10). In this case, it would be "Immunization Action Coalition". Similarly, if the user were interested in knowing more about a website at a specific location or emotion score, they would hover over those data points to reveal that information.

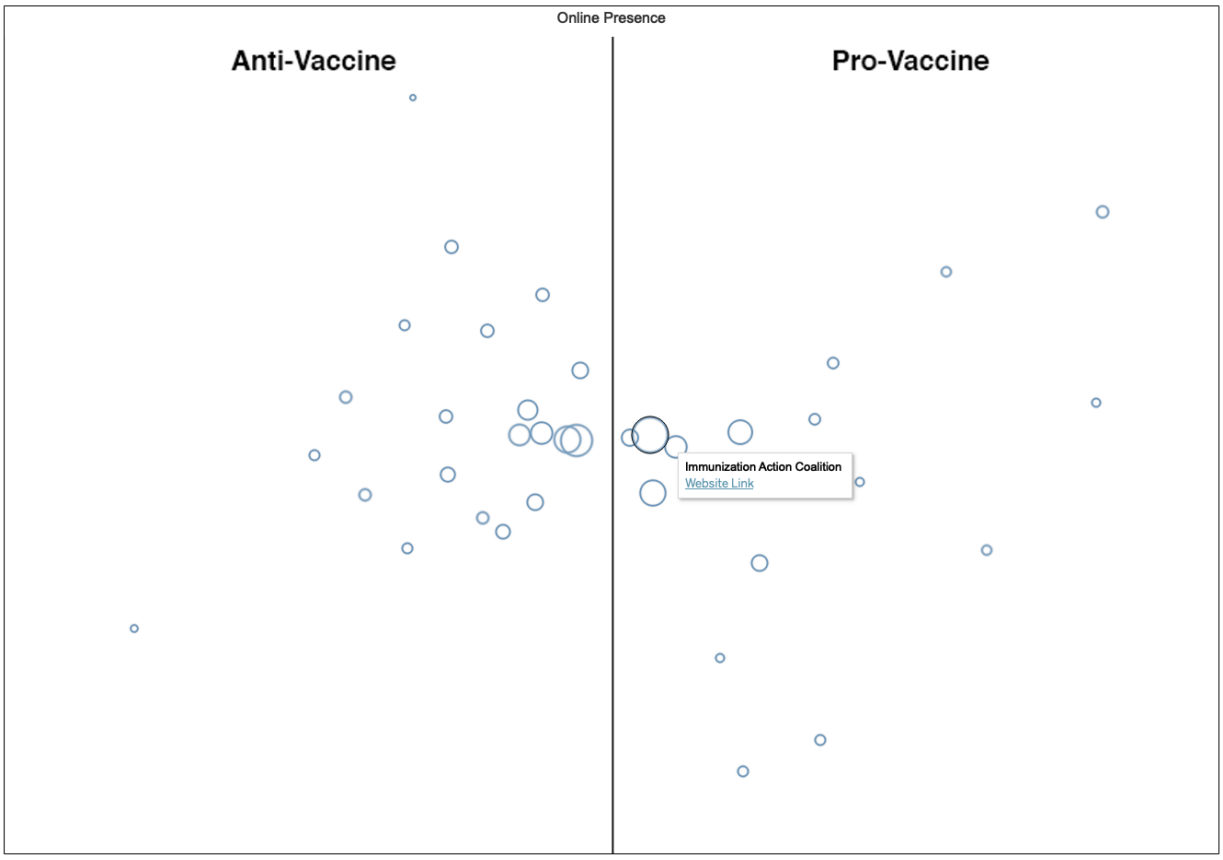

Figure 10: Hover to Expand Information Box

Finally, on the map of website locations, users have the ability to navigate through the set of websites. On the map of website locations, users can zoom in and out of the map to focus on specific areas. As well, users can click on and drag over the map to move the area of focus. Consider a sample use case. A user is interested in looking at websites in Europe to get a better sense of where exactly they are located. By zooming in on the map and going to Europe (as seen Figure 11), they can clearly identify four websites located there in England, Germany, Belgium, and Switzerland.

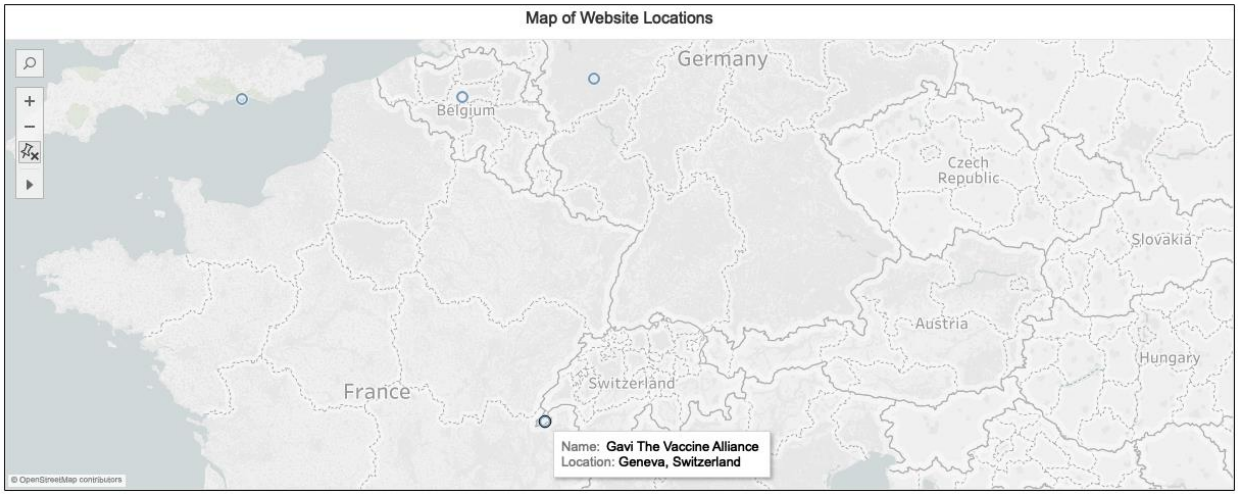

Figure 11: Navigate Map of Website Locations

4. Summary and Conclusions

In this paper, we have reported the development of VINCENT, a VAS to help with the investigation of data from websites involved in the online debate on vaccination. VINCENT was created using Tableau, version 10.5. VINCENT incorporates three main sub-systems, each comprised of other sub-systems. An analytics engine made up of webometric (co-link analysis) and NLP (text-based emotion detection) data analysis components; visualization, made up of several different data visualizations; and interaction, made up of a set of different human-data interactions. The development of VINCENT demonstrates that it is feasible to integrate webometrics, natural language processing of website text, data visualization, and human-data interaction into a VAS. VINCENT is novel in its incorporation and integration of the data analysis techniques used (i.e. co-link analysis and text-based emotion analysis) with data visualization and human-data interaction, which had never been previously attempted. VINCENT supports user exploration of data derived from a set of 37 vaccine websites and enables the user to investigate and develop an overall perspective on the vaccine debate. By looking at data from individual websites and groups of websites, a user can identify the breakdown of pro- and anti-vaccine websites, the emotions contained within these websites about specific vaccines, the locations of these websites, and the frequency of vaccine words that appear in these websites. Furthermore, by integrating the data from these different websites, users can associate the various types of data and uncover patterns that would be otherwise difficult to identify.

Several considerations should go into creating VASes such as VINCENT. First, deciding which tool to use to create the VAS is important. There are advantages and disadvantages to using more programming intensive solutions (such as D3.js) versus more rigid, yet easier to use, toolkit-based solutions (such as Tableau). As well, identifying the appropriate data sources is a challenge that is unique to each project. Online data sources are constantly changing; therefore, it is important for researchers to keep abreast of the current available data. Depending on the resources available to the developer, alternate methods and sources for acquiring proprietary data could improve the value of the system. Next, determining which visualizations are most

appropriate for each type of studied dataset is important. For example, the emotion bar charts, presented here, went through several iterations. At first, tree maps were tested but were found to be inadequate at representing certain aspects of the data. Researchers who develop similar VASes need to consider all facets of their data and desired interactions and test various iterations of their system. Finally, incorporating meaningful interactions into the VAS is important. It is necessary to analyze the tasks that users would need to perform, and then determine what combinations of interactions would facilitate the performance of these tasks. In the case of VINCENT, such tasks included comparing websites, identifying groups of websites, and identifying trends in the entire set of websites.

VINCENT was developed to help users make sense of the data from vaccine websites and, ultimately, the online vaccine debate. However, there are many other areas, both within and outside of public health, for which a system such as this could also prove useful. In public health, a similar VAS would be useful for surveillance of other online health debates, such as debates on the efficacy of alternate health claims or debates regarding different medications and drugs. Outside of public health a system similar to VINCENT could prove beneficial in the areas of business, academia, or politics.

One example of such an area that would be well served by a similar VAS is the online discussion about cannabis use. There are diverging positions regarding the risks and benefits of cannabis, and a system similar to VINCENT could enable users to further investigate the debate and make sense of the data from existing websites. With such a system, users would be able to quickly identify the positions of different websites (i.e., pro- or anti-cannabis, medical or recreational focus on cannabis, and so on), obtain a geographic breakdown of website locations, determine the focus of each website, and identify the detected emotions about various concepts related to cannabis (i.e., “essential oils” or “epilepsy”). Performing tasks such as these could help researchers acquire valuable insight into the online debate on cannabis and determine what (if any) actions could be taken (or policies adopted) to improve public health in this area.

4.1 Limitations

There were two key limitations to the development of VINCENT. The first set of limitations was related to the data and analysis tools that we used. Social media data could have generated very rich and revealing data for investigation, but these types of data are proprietary and not freely accessible to conduct research of this scale. Moz Link Explorer provided only enough data on inlinks for an adequate co-link analysis at the domain inlink level; getting data for the page- or site-level analysis was not feasible due to the associated cost. As the trend in the area of webometrics is towards collected data becoming increasingly proprietary, researchers need to consider alternative ways of making do with the limited data availability. Additionally, resources like the NLU API are limited in their ability to analyze the websites emotions. Tools like NLU API are essentially only in the infancy of their development. In the future, tools for emotion detection and NLP will certainly improve and be able to achieve a broader range of analysis and better results than are currently possible.

The second set of limitations was related to the interaction capabilities afforded by Tableau as a toolkit. For example, it was not possible in Tableau to allow the filtering interaction to also filter

the word cloud selection. Ideally, a user would want to be able to see word clouds of the top 25 words of any subset of websites selected in the other visualizations. However, given the manner by which Tableau allows for the structure of data, and the data management solutions it works with, this was not possible to achieve. The work-around we used for this was to create the website selection interaction that allowed individuals to filter for a specific website throughout VINCENT.

4.2 Future Research

In a follow up paper, we plan to conduct user testing of VINCENT to evaluate whether there is observable benefit to using VINCENT, and, if so, to what extent and in what ways. The findings of this research will lead to the development of best practices for creating similar VASes. They will also help with the identification of potential benefits of VINCENT-like systems that can support exploration of similar public health issues.

References

1. O'Carroll P. Introduction to Public Health Informatics. In: O'Carroll P, Yasnoff WA, Ward ME, Ripp LH, Martin EL, Editors. Public Health Informatics and Information Systems. New York, NY, USA: Springer-Verlag; 2003. p. 3–15.

2. Ola O, Sedig K. 2014. The challenge of big data in public health: an opportunity for visual analytics. *Online J Public Health Inform*. 5(3):e223, 1–21.

3. Otterman S. New York Confronts Its Worst Measles Outbreak in Decades. New York Times [Internet]. 2019 Jan 17; Available from: https://www.nytimes.com/2019/01/17/nyregion/measles-outbreak-jews-nyc.html

4. Oliviero H. Whooping cough is making a comeback. Here's why. The Toronto Star [Internet]. 2018 Sep 4; Available from: https://www.thestar.com/life/2018/09/04/whooping-cough-is-making-a-comeback-heres-why.html

5. Abbott B. Washington State Becomes Latest Hot Spot in Measles Outbreak. The Wall Street Journal [Internet]. 2019 Jan 23; Available from: https://www.wsj.com/articles/washington-state-becomes-latest-hot-spot-in-measles-outbreak-11548281172

6. Who.int. Ten health issues WHO will tackle this year [Internet]. 2019 [cited 2019 Feb 12]. Available from: https://www.who.int/emergencies/ten-threats-to-global-health-in-2019

7. Durbach N. 2000. 'They might as well brand us': working-class resistance to compulsory vaccination in Victorian England. *Soc Hist Med*. 13(1), 45-63. PubMed https://doi.org/10.1093/shm/13.1.45

8. Fox S, Rainie L. The online health care revolution. Pew Internet & American life project. Available from: https//www.pewinternet.org/2000/11/26/the-online-health-care-revolution/. 2000

9. Lewandowsky S, Oberauer K. 2016. Motivated rejection of science. *Curr Dir Psychol Sci*. 25(4), 217-22. https://doi.org/10.1177/0963721416654436

10. Kata A. 2010. A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine*. 28(7), 1709-16. PubMed https://doi.org/10.1016/j.vaccine.2009.12.022

11. Kata A. 2012. Anti-vaccine activists, Web 2.0, and the postmodern paradigm - An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*. 30(25), 3778-89. PubMed https://doi.org/10.1016/j.vaccine.2011.11.112

12. Ninkov A, Vaughan L. 2017. A webometric analysis of the online vaccination debate. *J Assoc Inf Sci Technol*. 68(5), 1285-94. https://doi.org/10.1002/asi.23758

13. Vaughan L, Ninkov A. 2018. A new approach to web co-link analysis. *J Assoc Inf Sci Technol*. 69(6), 820-31. https://doi.org/10.1002/asi.24000

14. Brunson EK, Sobo EJ. 2017. Framing Childhood Vaccination in the United States: Getting Past Polarization in the Public Discourse. *Hum Organ*. 76(1), 38-47. https://doi.org/10.17730/0018-7259.76.1.38

15. Mitra T, Counts S, Pennebaker JW. Understanding Anti-Vaccination Attitudes in Social Media. In: Proceedings of the Tenth International Conference on Web and Social Media; Cologne, Germany. 2016. p. 269–78.

16. Keim D, Andrienko G, Fekete JD, Görg C, Kohlhammer J, et al. Visual analytics: Definition, process, and challenges. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2008. p. 154–75.

17. Börner K. Atlas of Knowledge: Anyone Can Map. Boston, USA: MIT Press; 2015.

18. Sedig K, Parsons P. 2016. Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework. *Vol. 4, Synthesis Lectures on Visualization*. 1–185 Available from: http://www.morganclaypool.com/doi/abs/10.2200/S00685ED1V01Y201512VIS005

19. Sedig K, Parsons P, Babanski A. 2012. Towards a Characterization of Interactivity in Visual Analytics. *J Multimed Process Technol Spec Issue Theory Appl Vis Anal.* 3(1), 12-28.

20. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Burlington (MA), USA:Morgan Kauffman Publishers; 2011.

21. Shneiderman B, Plaisant C, Hesse BW. 2013. Improving healthcare with interactive visualization. *Computer*. 46(5), 58-66.

22. Salomon, G. No distribution without individuals' cognition: A dynamic interactional view. In: Saloman, G., editor. Distributed cognitions: Psychological and educational considerations. Cambridge, U.K.: Cambridge University Press; 1997. p. 111–38.

23. Liu Z, Nersessian N, Stasko J. 2008. Distributed cognition as a theoretical framework for information visualization. *IEEE Trans Vis Comput Graph*. 14(6). PubMed

24. Sedig K, Parsons P. 2013. Interaction design for complex cognitive activities with visual representations: A pattern-based approach. *AIS Trans Human-Computer Interact.* 5(2), 84-113. https://doi.org/10.17705/1thci.00055

25. Skyttner L. General systems theory: problems, perspectives, practice. Singapore: World Scientific Publishing Co.; 2005.

26. Bostock M, Ogievetsky V, Heer J. 2011. D3data-driven documents. *IEEE Trans Vis Comput Graph*. 17(12), 2301-09. PubMed https://doi.org/10.1109/TVCG.2011.185

27. Nair L, Shetty S, Shetty S. 2016. Interactive visual analytics on Big Data: Tableau vs D3.js. J e-Learning. *Knowl Soc*. 12(4), 139-50.

28. Hund M, Böhm D, Sturm W, Sedlmair M, Schreck T, et al. 2016. Visual analytics for concept exploration in subspaces of patient groups [Internet]. *Brain Inform*. 3(4), 233-47. doi:https://doi.org/10.1007/s40708-016-0043-5. PubMed

29. Cao N, Gotz D, Sun J, Qu H. 2011. DICON: Interactive Visual Analysis of Multidimensional Clusters. *IEEE Trans Vis Comput Graph*. 17(12), 2581-90. PubMed https://doi.org/10.1109/TVCG.2011.188

30. Cho I, Wesslen R, Volkova S, Ribarsky W, Dou W. CrystalBall: A Visual Analytic System for Future Event Discovery and Analysis from Social Media Data. In: 2017 IEEE Conference on Visual Analytics Science and Technology (VAST); Phoenix, USA. 2017. p. 25–35.

31. Pathak N, Henry MJ, Volkova S. Understanding Social Media's Take on Climate Change through Large-Scale Analysis of Targeted Opinions and Emotions. In: 2017 AAAI Spring Symposium Series; Palo Alto, California. 2017. p. 45-52.

32. Beigi G, Hu X, Maciejewski R, Liu H. An overview of sentiment analysis in social media and its applications in disaster relief. In: Sentiment analysis and ontology engineering. New York, USA: Springer; 2016. p. 313–40.

33. Thelwall M, Vaughan L, Björneborn L. 2005. *Webometrics. ARIST*. 39(1), 81-135.

34. Thelwall M. 2008. Bibliometrics to webometrics [Internet]. *J Inf Sci*. 34(4), 605-21. http://journals.sagepub.com/doi/10.1177/0165551507087238. https://doi.org/10.1177/0165551507087238

35. Stuart D. Web metrics for library and information professionals. London, U.K.: Facet publishing; 2014.

36. Thelwall M. Link Analysis: An Information Science Approach [Internet]. Bringley, U.K.: Emerald Group Publishing Limited; 2004. 88–91 p. Available from: http://linkanalysis.wlv.ac.uk/index.html

37. Thelwall M. 2001. Extracting macroscopic information from web links. *J Am Soc Inf Sci Technol*. 52(13), 1157-68. https://doi.org/10.1002/asi.1182

38. Thelwall M, Zuccala A. 2008. A university-centred European Union link analysis. *Scientometrics*. 75(3), 407-20. https://doi.org/10.1007/s11192-007-1831-8

39. Vaughan L, Wu G. 2004. Links to commercial websites as a source of business information. *Scientometrics*. 60(3), 487-96. https://doi.org/10.1023/B:SCIE.0000034389.14825.bc

40. Holmberg K, Thelwall M. 2009. Local government web sites in Finland: A geographic and webometric analysis [Internet]. *Scientometrics*. 79(1), 157-69. doi:https://doi.org/10.1007/s11192-009-0410-6.

41. Ortega JL, Aguillo IF. 2008. Visualization of the Nordic academic web: Link analysis using social network tools. *Inf Process Manage*. 44(4), 1624-33. https://doi.org/10.1016/j.ipm.2007.09.010

42. Thelwall M. 2009. Introduction to webometrics: Quantitative web research for the social sciences. *Synth Lect Inf Concepts Retr Serv*. 1(1), 1-116. https://doi.org/10.2200/S00176ED1V01Y200903ICR004

43. Vaughan L, You J. 2008. Content assisted web co-link analysis for competitive intelligence. *Scientometrics*. 77(3), 433-44. https://doi.org/10.1007/s11192-007-1999-y

44. Vaughan L, You J. 2010. Word co-occurrences on Webpages as a measure of the relatedness of organizations: A new Webometrics concept. *J Informetrics*. 4(4), 483-91. https://doi.org/10.1016/j.joi.2010.04.005

45. Vaughan L, You J. Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. In: Scientometrics. 2006. p. 611–28.

46. Holmberg K. Webometric network analysis: Mapping cooperation and geopolitical connections between local government administration on the web. Turku, Finland: Åbo Akademis förlag-Åbo Akademi University Press; 2009.

47. Holmberg K. Webometric network analysis: Mapping cooperation and geopolitical connections between local government administration on the web. Åbo Akademis förlag-Åbo Akademi University Press; 2009.

48. Kim JH, Barnett GA, Park HW. 2010. A hyperlink and issue network analysis of the United States Senate: A rediscovery of the web as a relational and topical medium. *J Assoc Inf Sci Technol*. 61(8), 1598-611.

49. Romero-Frías E, Vaughan L. 2010. European political trends viewed through patterns of Web linking. *J Assoc Inf Sci Technol*. 61(10), 2109-21. https://doi.org/10.1002/asi.21375

50. Hirschberg J, Manning CD. 2015. Advances in natural language processing. *Sciencenat.* 349(6245), 261-66. PubMed https://doi.org/10.1126/science.aaa8685

51. Liu B. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press; 2015.

52. Rubin VL, Stanton JM, Liddy ED. Discerning emotions in texts. In: The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT). Stanford, USA. 2004.

53. Ptaszynski M, Masui F, Rzepka R, Araki K. 2014. Emotive or Non-emotive: That is The Question. In: Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; Baltimore, USA. 2014: p. 59-65. https://doi.org/10.3115/v1/W14-2610

54. Tokuhisa R, Inui K, Matsumoto Y. Emotion classification using massive examples extracted from the web. In: Proceedings of the 22nd International Conference on Computational Linguistics; Manchester, U.K.; 2008. p. 881–8.

55. Vergara S, El-Khouly M, El Tantawi M, Marla S, Lak S. Building Cognitive Applications with IBM Watson Services: Volume 7 Natural Language Understanding. In: Tech rep. IBM Corporation; 2017. p. 98.

56. Palomino M, Taylor T, Göker A, Isaacs J, Warber S. 2016. The Online Dissemination of Nature–Health Concepts: Lessons from Sentiment Analysis of Social Media Relating to "Nature-Deficit Disorder.". *Int J Environ Res Public Health*. 13(1), 142. PubMed https://doi.org/10.3390/ijerph13010142

57. Meehan K, Lunney T, Curran K, McCaughey A. Context-aware intelligent recommendation system for tourism. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2013. 2013. p. 328–31.

58. Rizzo G, Troncy R. Nerd: A framework for evaluating named entity recognition tools in the Web of data. In: International Semantic Web Conference, Demo Session; Bonn, Germany. 2011; p. 1–4.

59. Saif H., He Y., Alani H. Semantic Sentiment Analysis of Twitter. In: Cudré-Mauroux P, et al., Editors. International Semantic Web Conference, Lecture Notes in Computer Science; Boston, USA. 2012; p. 508 – 524.

60. Katsuki T, Mackey TK, Cuomo R. 2015. Establishing a link between prescription drug abuse and illicit online pharmacies: analysis of Twitter data. *J Med Internet Res*. 17(12), e280. PubMed https://doi.org/10.2196/jmir.5144

61. McAuley J, Leskovec J, Jurafsky D. Learning attitudes and attributes from multi-aspect reviews. In: Data Mining (ICDM), 2012 IEEE 12th International Conference on Data Mining; Brussels, Belgium. 2012. p. 1020–5.

62. Healey C, Ramaswamy S. Visualizing twitter sentiment. Sentim Viz [Internet]. 2011; Available from: https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

63. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. Sebastopol, USA: O'Reilly Media, Inc.; 2009.

64. Grimes S. Sentiment, emotion, attitude, and personality, via Natural Language Processing [Internet]. IBM. 2016 [cited 2019 Jan 20]. Available from: https://www.ibm.com/blogs/watson/2016/07/sentiment-emotion-attitude-personality-via-natural-language-processing/

Appendix 1 - Set of Websites

Name	Domain
Adult Vaccination	http://www.adultvaccination.org/
Age of Autism	http://www.ageofautism.com/
Australian Vaccination-risks Network	http://avn.org.au/
Experimental Vaccines	http://experimentalvaccines.org/
Families Fighting Flu	http://www.familiesfightingflu.org/
Gavi The Vaccine Alliance	http://www.gavi.org/
History of Vaccines	http://www.historyofvaccines.org/
Immunization Action Coalition	http://www.immunize.org/
Immunize BC	http://www.immunizebc.ca/
Immunize Canada	http://immunize.ca
Institute for Vaccine Safety	http://www.vaccinesafety.edu/
National Vaccine Information Center	http://www.nvic.org/
Parents Requesting Open Vaccine Education	http://vaccineinfo.net/
Prevent Childhood Influenza	http://www.preventchildhoodinfluenza.org/
Sabin Vaccine Institute	http://www.sabin.org/
Safe Minds	http://www.safeminds.org/
SaneVax	http://sanevax.org/

Shots of Prevention	http://shotofprevention.com/
The Immunization Partnership	http://www.immunizeusa.org/
The Informed Parent	http://www.informedparent.co.uk/
The Thinking Moms Revolution	http://thinkingmomsrevolution.com/
Think Twice Global Vaccine Institute	http://thinktwice.com/
Vaccinate Your Family	https://www.vaccinateyourfamily.org/
Vaccination Information Network	http://www.vaccinationinformationnetwork.com/
Vaccination Liberation	http://vaclib.org/
Vaccination News	http://www.vaccinationnews.org/
Vaccine Choice Canada	http://vaccinechoicecanada.com
Vaccine Injury Help Center	http://www.vaccineinjuryhelpcenter.com/
Vaccine Injury Info	http://www.vaccineinjury.info/
Vaccine Liberation Army	http://vaccineliberationarmy.com/
Vaccine Resistance Movement	http://vaccineresistancemovement.org/
Vaccine Truth	http://vaccinetruth.org/
Vaccines Today	http://www.vaccinestoday.eu/
Vaccines.gov	http://www.vaccines.gov/
Vaxxter	http://vaxxter.com
Voices for Vaccines	http://www.voicesforvaccines.org/
World Association for Vaccine Education	http://novaccine.com/