Health News in Twitter
Practical Application of Data Mining and Knowledge Discovery in R

Meryem Ben yahia [1]

Abstract

In this paper, I conducted a series of practical experiments on the Health News in Twitter Data (Karami, 2018) to apply and test the concepts covered in the Data Mining and Knowledge Discovery in R. The focus was finding interesting and valuable structures that are embedded in a large dataset.

1. Problem Understanding

An estimated 415 million people use X/Twitter globally in 2023, with over 95 million of them residing in the US (Statista Research Department, 2024). The number of users indicates the platform's significant power in disseminating public health information, updates on disease prevention, and acting as a support tool in crisis recovery (Merchant et al., 2011). The majority of people, including scientists and doctors, initially become aware of medical advancements through the media (Phillips et al., 1991).

In this paper, I aim to analyze health news tweets extracted from several health Twitter accounts in 2015 using Association Rule Mining. The accounts include cbchealth, everydayhealth, and others, providing a comprehensive dataset of health-related information. Many industries, including healthcare, employ data mining (DM) techniques including clustering, classification, and association rule mining (ARM) to extract knowledge (Rousidis et al., 2020).

The dataset used for this study can be found at the following link: Health News in Twitter (Karami, 2018). It was created by Dr. Amir Karami, Associate Professor of Quantitative Methods and Business Analytics at the University of Alabama and Scientist at the Center for Clinical and Translational Science (CCST). After carefully reviewing multiple health-related datasets, I discovered this particular dataset on the UC Irvine Machine Learning Repository. It emerged as the most relevant and accessible option following my initial exploration of the "Twitter Dataset for Mental Disorders Detection," provided by Miryam Elizabeth Villa-Pérez and Luis A. Trejo from Tecnologico de Monterrey (Villa-Pérez et al., 2023), which presented limitations due to Twitter's content redistribution policy.

2. Data Understanding

2.1. Health News Dataset Overview

The Health News in Twitter dataset, collected from major health news agencies via the Twitter API, contains over 58,000 tweets dating back to 2015. Each news outlet's tweets are stored separately over 16 text files annotated by ID, date, and text. The first step was to combine all of the sources and to observe the different types of features.

3. Data Preparation

Due to limited computational resources, all code was executed in the Colab Research Environment using an 'R' kernel.

3.1. Exploratory Data Analysis

The distribution of tweets across various health news sources is illustrated in Figure 1, which presents the total number of tweets associated with each source throughout the dataset period. This bar graph quantifies and compares the engagement levels and influence of each news outlet within the health discourse on Twitter.

[*]Equal contribution [1]Jean Monnet University, Saint-Étienne, France. Correspondence to: Meryem Ben yahia <meryem.ben.yahia@etu.univ-st-etienne.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

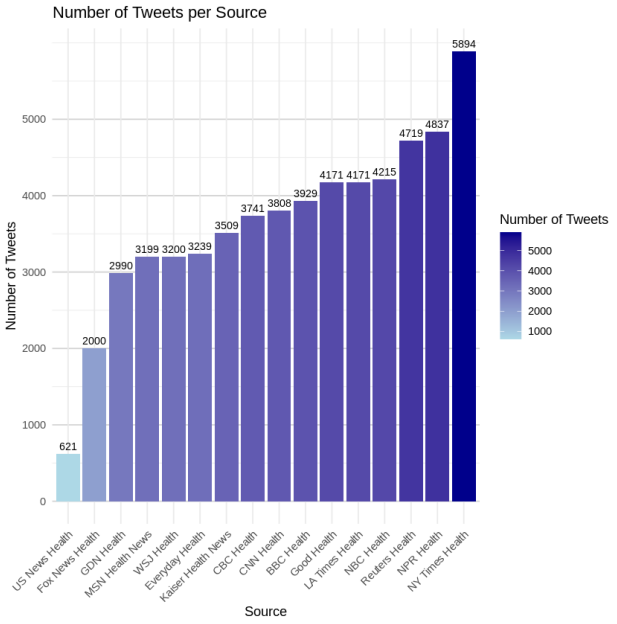

Figure 1. Number of Tweets per Source for the Entire Dataset

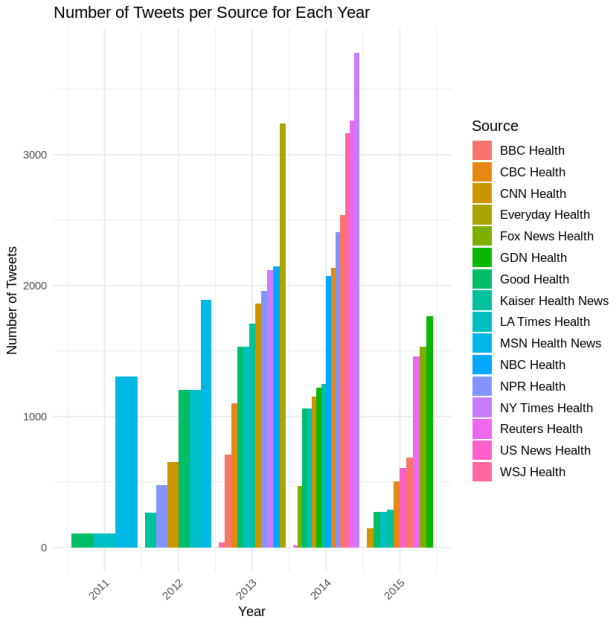

Figure 2. Number of Tweets per Source per Year

WSJ Health emerges as the leading source in terms of tweet volume, with a total of 5,894 tweets, indicating its prominent role in health news dissemination on social media. Reuters Health and NY Times Health also demonstrate significant engagement, with 4,837 and 4,719 tweets respectively, reflecting their extensive reach and active participation in health-related discussions.

Conversely, sources like US News Health, which recorded substantially fewer tweets (621), may reflect lesser focus or lower audience engagement in health topics compared to other major outlets.

The distribution of tweets per source over the years 2011 to 2015 is illustrated in Figure 2.

As shown in Figure 2, certain sources such as WSJ Health and NY Times Health demonstrate a significant increase in tweet volume, particularly in 2014 and 2015. In contrast, other sources like Fox News Health show relatively less growth in tweet volume.

Other observations to make were the volume of tweets per year, and their sources.

The distribution of tweets across different years, as illustrated in the pie chart (Figure 3), emphasizes 2014 as the year with the highest proportion of tweets, accounting for 43.9% of the total volume.

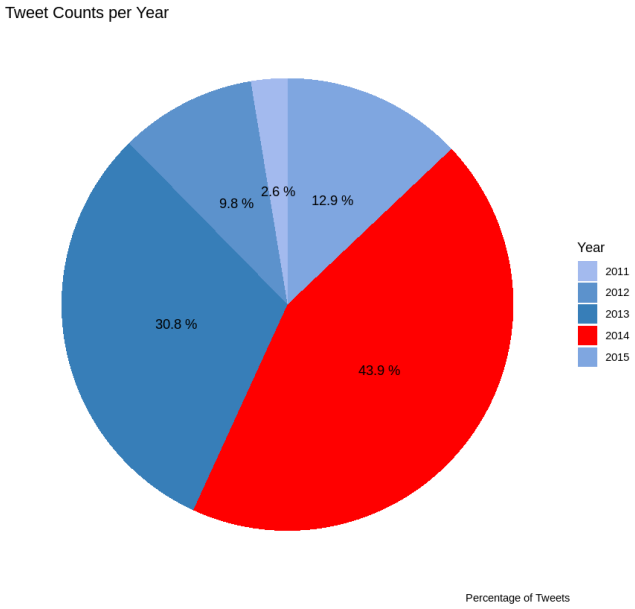

Figure 3. Pie Chart of Tweet Counts per Year

The year 2024 was also the one with the most distinct number of sources, as shown in 4.

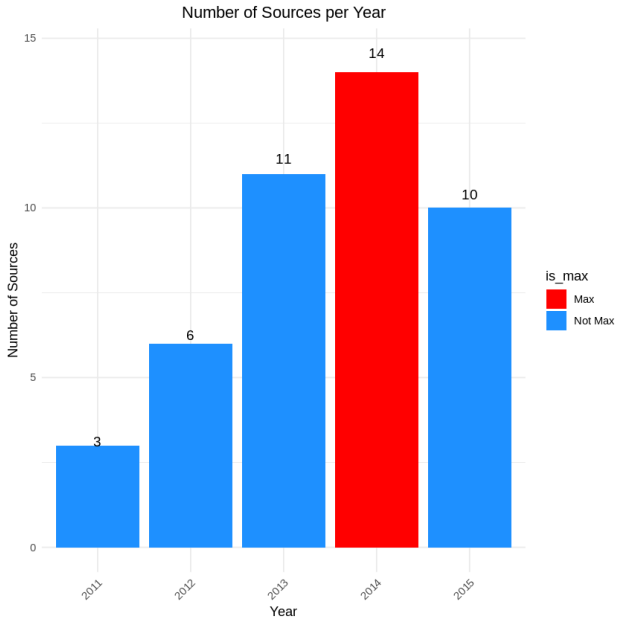

Figure 4. Counts of Sources per Year

To obtain the most possible balanced and diversified sub-sample of data, I focused exclusively on entries for the year 2014. Additionally, I chose to exclude sources with the lowest tweet counts, such as US News Health, which only had 13 tweets in 2014.

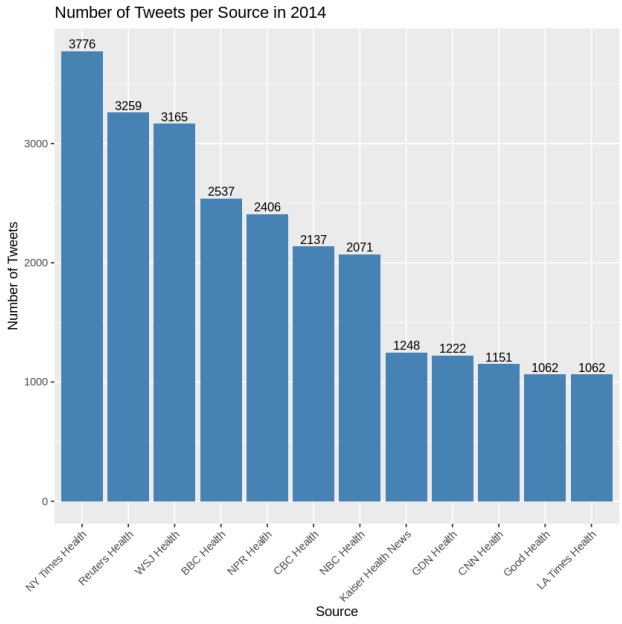

Figure 5. Number of Tweets per Source in 2014

As shown in 5, the final dataset has 23,996 tweets and 10 different sources with at least 1,000 tweets each.

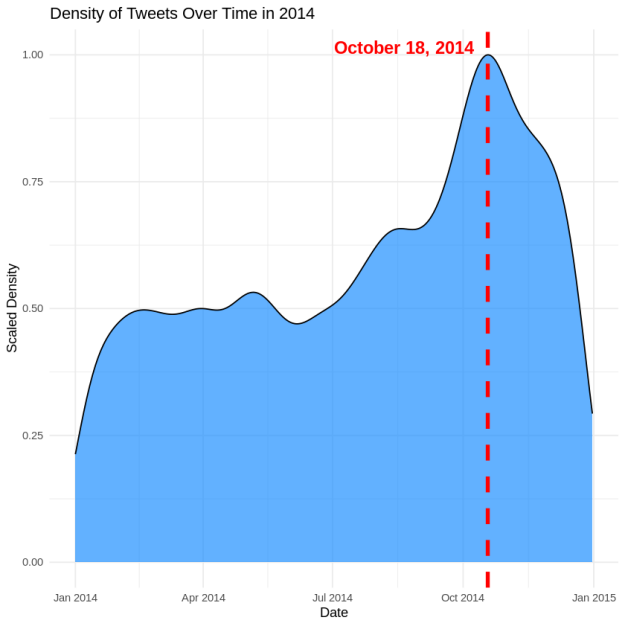

Figure 6. Number of Tweets per Source in 2014

Further exploration shows use the trends and the density of tweets in 2014. In Figure 6, the overall trend in the graph shows an initial increase in activity in the early months, a slight decline mid-year, and then a gradual build-up leading to the peak in October. The plot shows a significant spike in tweet density around October 18, suggesting an event or a series of events that garnered substantial attention or engagement related to health.

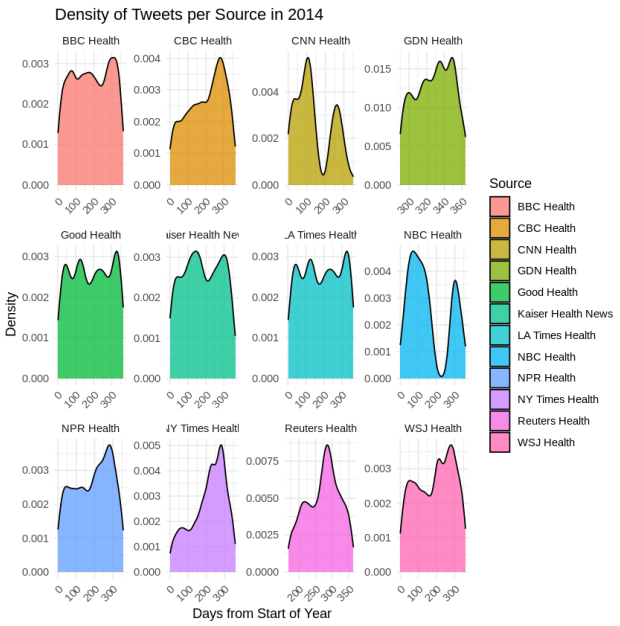

Figure 7. Density Plots of Different Sources in 2014

For the density of each source during 2014, we can observe different trends. In Figure 7, each subplot represents a distinct news outlet, demonstrating the temporal dynamics of tweets related to health topics. Notably, sources such as CNN Health exhibit significant peaks towards the end of October, aligning with the overall heightened activity observed in the general tweet density analysis for 2014. In contrast, outlets like BBC Health, CBC Health, and Reuters Health show peaks earlier in the year, which may reflect their responsiveness to specific events or editorial preferences.

3.2. Pre-processing

Data preprocessing is a step that typically involves data transformations prior to analysis. The function used was clean_text.

$$\text{clean\_text}(\text{text}) \rightarrow \text{cleaned\_text}$$

For each step within the function:

1. gsub(”s”, ””, text): Removes occurrences of special characters followed by ’s, which might result from encoding issues.
2. gsub(”bRTb”, ””, text): Eliminates the string ”RT”, commonly found in retweets, since retweets are not considered relevant.
3. gsub(”@w+”, ””, text): Deletes Twitter usernames (strings starting with ’@’) from the text.
4. gsub(”https?://S+s?”, ””, text): Erases URLs (strings starting with ”http” or ”https”) from the text.
5. rm_twitter_url(text): further removes any remaining Twitter-specific URL patterns.
6. gsub(”[Â-Za-z]”, ” ”, text): Replaces any non-alphabetic characters with a space, effectively removing punctuation and special characters.
7. tolower(text): Converts all text to lowercase to ensure consistency.
8. removeWords(text, stopwords(”english”)): Eliminates common English stopwords (such as ”the”, ”and”, ”is”, etc.), which typically do not contribute to the meaning of the text.
9. gsub(”bw1b”, ””, text): Removes single-character words, which are often irrelevant or artifacts of previous cleaning steps.
10. stripWhitespace(text): Removes any extra whitespace characters.

Content	processed_content
<chr>	<chr>
RT @cspanwj: VIDEO: @jrovner, @KHNews Senior Correspondent, on #Medicare and the #ACA http://cs.pn/1B1PCuh http://pbs.twimg.com/media/B6HQ67nCAAAp57I.jpg	video senior correspondent medicare aca
How A State□s Choice On Medicaid Expansion Affects Hospitals, @SarahVarney4 reports in @NewsHour: http://khne.ws/1A57NOj	state choice medicaid expansion affects hospitals reports
Some doctors who treat #Medicaid patients are going to get a big pay cut. How will it affect patients? http://khne.ws/1wD8dvD	doctors treat medicaid patients going get big pay cut will affect patients
Ebola Doctor Makes Tough Choice To Save The Lives Of Two Colleagues http://khne.ws/1zhSfnF	ebola doctor makes tough choice save lives two colleagues
DETAILS: FDA Proposes Easing Lifetime Ban On Blood Donations By Gay Men http://khne.ws/1GXjlGL	details fda proposes easing lifetime ban blood donations gay men
Millions Have Already Enrolled In 2015 Health Policies, Deadline Still 7 Weeks Off, @jrovner reports: http://khne.ws/1AXR3Is	millions already enrolled health policies deadline still weeks reports

Figure 8. Table showing text before and After Pre-processing

By tokenizing the pre-processed contents and creating a table for counts, I was able to generate a wordcloud.

Central to the word cloud are terms such as ”health,” ”Ebola,” and ”cancer,” which evidently were major topics of discussion. The prominence of ”Ebola” suggests a specific time period during which there was a significant public and media focus on the Ebola outbreak.

Other terms such as ”risk,” ”hospital,” ”FDA,” and ”treatment” indicate discussions centered around healthcare services, regulatory actions, and treatment options, pointing to a community deeply engaged in healthcare practices and policies. The presence of words like ”insurance,” ”Medicaid,” and ”Obamacare” reflects the ongoing discourse on healthcare reform and policy changes in the United States during the timeframe of the data.

Smaller but still significant words like ”study,” ”research,” and ”data” underline the importance of scientific research

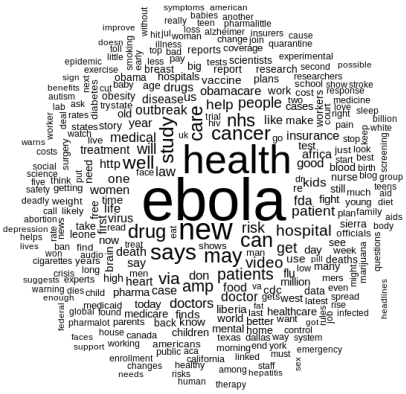

Figure 9. Wordcloud of Most Frequent Words in Health News Sources in 2014

and data in health-related conversations. Additionally, words like "outbreak," "virus," and "disease" suggest a focus on public health crises and infectious diseases.

4. Modelling

4.1. Topic Extraction

4.1.1. LDA

Latent Dirichlet Allocation (LDA) is a generative probabilistic model utilized extensively in natural language processing for the purpose of topic modeling. In LDA, documents are viewed as distributions over topics, and topics are distributions over words. The model assumes that each document is a mixture of various topics, and each word within the document is attributable to one of those topics. Through statistical inference techniques, LDA aims to uncover these latent topic distributions from a corpus of documents, providing insight into the underlying thematic structure of the data (Science, 2024).

The approach involves a systematic procedure for topic modeling using Latent Dirichlet Allocation (LDA). Initially, raw text data is preprocessed to create a corpus of documents. A Document-Term Matrix (DTM) is then generated from this corpus, applying TF-IDF weighting to highlight important terms. After converting the DTM into a sparse matrix format for efficiency, a Term-Document Matrix (TDM) is derived. An LDA model is initialized with a set number of

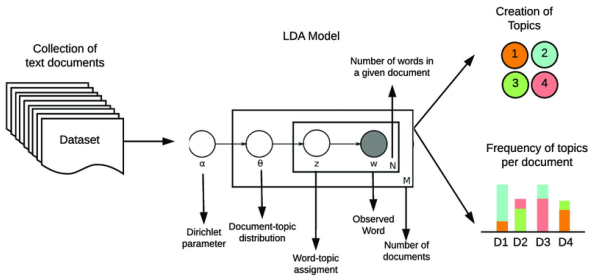

Figure 10. Schematic-of-LDA-algorithm. Source: (Science, 2024).

topics and trained on the TDM to infer topic distributions for each document. Top words for each topic are extracted from the model output, and a vocabulary is constructed from the corpus tokens.

```
INFO  [21:27:20.538] early stopping at 75 iteration
INFO  [21:27:28.501] early stopping at 50 iteration
          X1         X2    X3       X4         X5     X6         X7         X8
1     health cigarettes  well     know        nhs cancer      video      video
2  obamacare       food  good   doctor     health  study       help     weight
3  insurance      audio   amp     life healthcare   drug marijuana      water
4        law    smoking   can   things    doctors   risk        can        fat
5       care    tobacco   age     well       free    may       live  bacteria
6      court       well   old medicine       care  drugs    surgery       lose
7   coverage      brain   new  parents       drug  heart       baby        run
8   abortion    smokers   day   living   medicare  blood    alcohol        way
9        new      risks today   likely       data  women cigarette challenge
10     plans    warning   eat     take        fda breast christmas       ill
          X9      X10
1     ebola    ebola
2  outbreak  patient
3     cases hospital
4    africa    nurse
5       flu patients
6   liberia    video
7     virus   dallas
8      mers      cdc
9      case    tests
10     says      new
```

Figure 11. LDA Model Results

Figure 11 shows the results of the Latent Dirichlet Allocation (LDA) model, identifying ten distinct topics by clustering the distribution of words within the documents. Utilizing `LDA$new`, the model was configured with ten topics, and hyperparameters were set at 0.1 for document-topic prior and 0.01 for topic-word prior, to control the sparsity of the distributions. The model underwent fitting on a term count matrix (`tcm`) through `lda_model$fit_transform()`, iterating up to 1000 times or until achieving a convergence tolerance of 0.01. This iterative process was monitored for convergence every 25 iterations. Subsequently, the ten most prevalent words for each topic were extracted using `lda_model$get_top_words(n = 10)`, and the results were organized into a DataFrame, `topic_terms_df`, which was printed to provide a clear representation of the topics.

The output reveals diverse health-related themes encapsulated in the ten topics identified by the LDA model. For instance:

1. **Topic 1** includes words like "health," "cigarettes," "cancer," indicating discussions around health impacts of smoking.

2. **Topic 2** emphasizes on "ebola," "outbreak," "Africa," pointing to discussions centered around the Ebola outbreak.

3. **Topic 3** is rich with terms like "insurance," "obamacare," "healthcare," suggesting a focus on health insurance and policy.

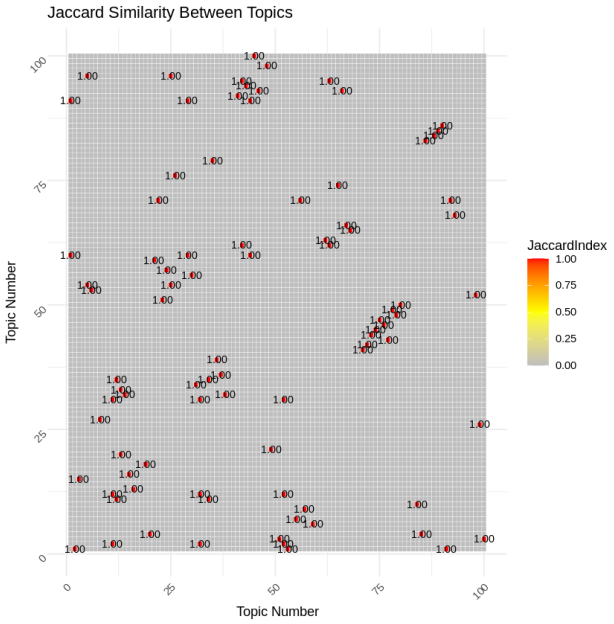

Figure 12. Jaccard Similarity between Topics in 2014

4.1.2. Jaccard Similarity

The Jaccard index, also referred to as the Jaccard similarity coefficient, serves as a vital statistical measure in assessing the similarity and dissimilarity between two sample sets. It quantifies the overlap between the sets by dividing the size of their intersection by the size of their union (Wikipedia, 2024). It's a measure of similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Where:

$J(A,B)$: Jaccard index between sets A and B

A : Set A

B : Set B

Using the Jaccard similarity with Latent Dirichlet Allocation (LDA) in natural language processing helps to gauge how similar or different the topics extracted by LDA are. It can reflect the topic stability.

Figure 12 shows the Jaccard Similarity between topics as a map. A significant number of topic comparisons show a Jaccard similarity index of 1.0, as indicated by the predominance of red points with a value label of "1.00." This suggests that many topics have a high degree of overlap in terms of the words they contain. The latter indicates that the majority of the news sources in the data have approached the same topics in relation to same areas.

4.1.3. Coherence Similarity

Topic Coherence Analysis measures the semantic similarity between high scoring words in each topic, offering a quantitative approach to evaluating topic model quality. The coherence score, typically derived using measures such as pointwise mutual information (PMI), cosine similarity, or others, assesses how frequently words appear together, providing insights into the interpretability and distinctiveness of identified topics (Sarkar, 2019).

Figure 13 show coherence scores peak at specific numbers of topics suggesting optimality at these configurations. However, there is still a high variability, which indicates not all topics are equally meaningful.

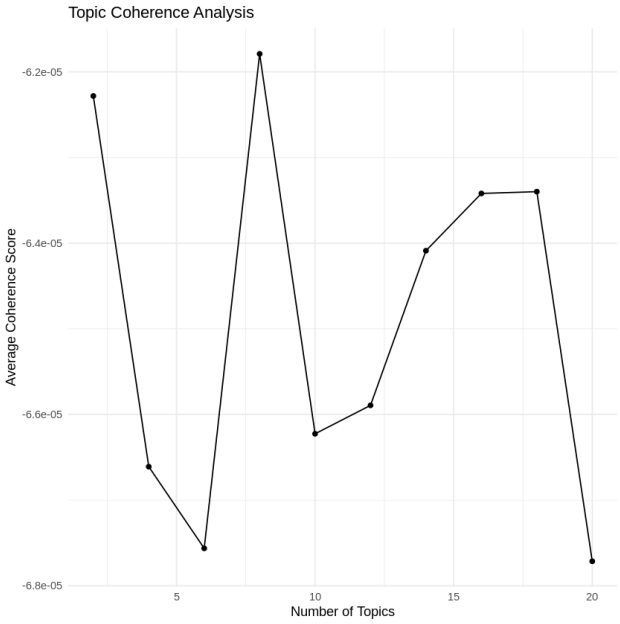

Figure 13. Topic Coherence Analysis

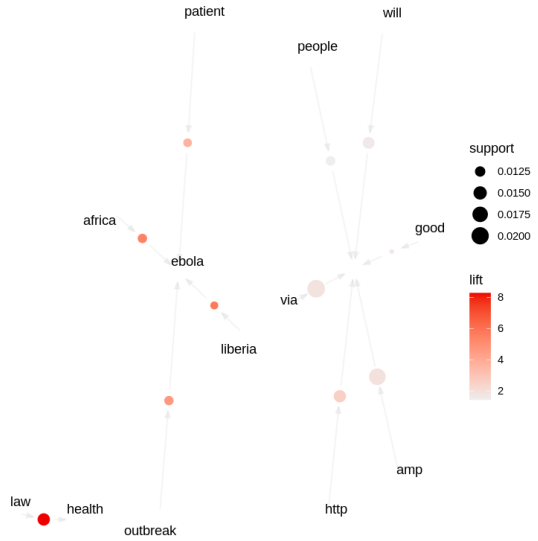

Figure 14. Association Rule Graph

5. Association Rule Mining

Association rule mining is a data mining technique used to uncover interesting relationships, patterns, and associations among sets of items in large databases.

I treated each tweet as a transaction, and its words are considered as items. The Eclat algorithm was applied to the data to identify frequent itemsets with a minimum support threshold of 0.01, ensuring that only those terms appearing in at least 1% of the transactions were considered. Subsequent to itemset generation, association rules were induced with a confidence threshold of 0.5, implying that the likelihood of occurrence of consequent items given antecedent items in the rules was at least 50%. To assess the strength and significance of these rules, interest measures including lift and leverage were computed. Lift values, which compare the observed frequency of A and B occurring together with the frequency expected if A and B were independent, provide a measure of the strength of an association.

In Figure 14 displays the association rule graph, focusing on the relationships and co-occurrences of specific terms, particularly within the context of the Ebola outbreak. The terms ”Ebola,” ”Africa,” ”Liberia,” and ”outbreak” are prominently connected, which reflects their significant role in the discussions captured by the dataset. The strong connection between ”Ebola” and geographic identifiers like ”Africa” and ”Liberia” with a darker edge color suggests a high lift value, indicating these terms appear together far more often than would be expected if they were independent. Terms like ”health,” ”law,” and ”patient” also appear but are connected with lighter edges to the main cluster, indicating weaker associations.

6. Conclusion

In this project, I employed a comprehensive suite of data mining techniques to analyze health-related tweets from 2014, revealing significant insights into public discourse on health issues as represented on Twitter. The analysis pinpointed substantial topics related to critical health events, notably the Ebola outbreak, which dominated discussions particularly towards the end of the year. There was consistency across what all methods revealed in terms of associations and connections in the dataset for 2014 across all health news sources.

These findings not only tie back to the initial problem understanding—which emphasized Twitter’s significant role in disseminating public health information and acting as a

support tool in crisis recovery—but also showcase the potential of data mining techniques in extracting meaningful insights from social media data.

Moving forward, future works could expand on these findings by exploring more diverse data, sentiment analysis and applying newer analytical techniques.

7. Acknowledgement

I would like to express my sincere gratitude to Professor Fabrice Muhlenbach for providing me with the opportunity to work on this intriguing project. This project has been a substantial opportunity for my academic growth.

References

Karami, A. Health News in Twitter. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5BW2Q.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Merchant, R. M., Elmer, S., and Lurie, N. Integrating social media into emergency-preparedness efforts. *N Engl J Med*, 365:289–291, 2011. doi: 10.1056/NEJMp1103591.

Phillips, D. P., Kanter, E. J., Bednarczyk, B., and Tastad, P. L. Importance of the lay press in the transmission of medical knowledge to the scientific community. *New England Journal of Medicine*, 325:1180–1183, 1991.

Rousidis, D., Koukaras, P., and Tjortjis, C. Social media prediction: a literature review. *Multimedia Tools and Applications*, 79(9–10):6279–6311, 2020. doi: 10.1007/s11042-019-08291-9. URL http://dx.doi.org/10.1007/s11042-019-08291-9.

Sarkar, D. D. Understanding topic coherence measures, 2019. URL https://towardsdatascience.com/understanding-topic-coherence-measures-4aa41339634c. Accessed: 2024-05-06.

Science, T. D. Latent dirichlet allocation (lda), 2024. URL https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2. [Online; accessed 5-May-2024].

Statista Research Department. X (formerly Twitter) - statistics & facts, Mar 18 2024. URL https://www.statista.com/topics/2462/x-twitter/.

Villa-Pérez, M. E., Trejo, L. A., Moin, M. B., and Stroulia, E. Extracting mental health indicators from english and spanish social media: A machine learning approach. *IEEE Access*, 11:128135–128152, 2023. doi: 10.1109/ACCESS.2023.3332289.

Wikipedia. Jaccard index — wikipedia, the free encyclopedia, 2024. URL https://en.wikipedia.org/wiki/Jaccard_index. [Online; accessed 5-May-2024].