HIV and it’s societal intricacies as the media reports it

Shibani Shankar Dash

24/05/2019

Any disease, along with it’s biological symptoms, brings with it a number of implications on the population that is directly perceptible to it. In the case of HIV AIDS, considering the fact that it spreads through human contact, exponentially increases the number of possible consequences in a country that is at the top of the charts(population-wise). We can also observe how far the administration in India has come towards this goal. This may be considered as an initial draft or starting point for a much deeper inquiry into the aforementioned subject.

HIV AIDS or Human immunodeficiency virus infection and acquired immune deficiency syndrome can be described as a series of conditions brought on by infection with the Human immunodeficiency virus (HIV).

There are three prominent ways to get infected, the first being transfusion of blood containing the Human immunodeficiency virus. This is caused when such individuals go about their daily lives, ignorant of the infection. This generally proves to be lethal for all the parties involved, if left as such.

Another way of getting infected is by unprotected sexual contact with a person who harbours the virus in question. This happens to be the most frequent mode of infection. However as we will see, the trend is different in India.

Finally, the virus may also be transferred from a mother to her child. This is the most common infection pathway when it comes to younger children.

We have tried to analyse the scenario in India. We collected news articles with the topic “Hiv” from the internet portal of the daily The Times of India, from the year 2018 by way of an R script (R Core Team (2019), Wickham (2019)). We then cleaned it up using the dplyr package (Wickham, François, et al. (2019)). Note that we only collected data from the news section while ignoring the articles in the lifestyle section.

We analysed this text using the some text mining approaches including bigram counts, tf-idf approach and pairwise correlations between words (Robinson and Silge (2018)). We then tried to visualise the content of the articles based on the statistics obtained from the forementioned approaches in plots using ggplot2 and a correlation graph using igraph and ggraph(Wickham, Chang, et al. (2019), Pedersen (2018)).

Results

We present the results of our analysis below.

Bigram count approach

We can see the plot obtained using this approach in Figure 1. We can immediately observe that the bigram with the highest frequency is ‘blood bank’. This suggests that news reports on HIV infections in blood banks were that much more common than the rest of the suspects. This gives us some evidence of blood transfusion being the most common infection pahway in India. There seem to be many other bigrams present just in this subset of all the bigrams that provide further proof for this claim. This data, however is still rather noisy and more cleaning is required for a stronger claim.

Tf-idf approach using bigrams

We used the bigram data from before and calculated the tf-idf values of the bigrams themselves as the individual words proved very noisy to deal with. We believe it would be worthwhile to clean the dataset further in future analyses. Here, we only subset on the condition if the first or the second part of the bigram contains some key words and get the top 50 with the highest tf-idf. See Figure 2 for the plot. We see that 'blood bank' trumps all other pairs, closely followed by 'sex workers' further suggesting that blood transfusion is a more common infection pathway than unprotected sex in India. Further down the line is 'drug addicts' suggesting drug abuse as another prominent infection pathway, as is in the rest of the world. However a more in-depth approach with extensive cleaning of the dataset is needed to draw any stronger conclusions.

Correlation graph

We again tokenized the dataset by words and calculated the correlations between words grouped by articles. We plotted a subset of this data in Figure 3. It was seen that a lot of words like 'blood', 'infected' and 'pregnant', 'woman' were grouped with Sattur, Virudhunagar and Madurai, all of which are cities in Tamil Nadu. This suggests some combination of societal factors which contribute to the infection rates in these cities. We can also see words like 'gay' and 'transgender' which suggest that among those infected some signficant proportion were of these categories. Also, notably, 'drug' and 'addicts' are also present which suggest a trend similar to the rest of world when it comes to infection through the illegal use of narcotics.

Online methods

The code for scraping, and plotting can be found at https://github.com/ssd71/toi-hiv-analysis

References

Pedersen, Thomas Lin. 2018. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks.* https://CRAN.R-project.org/package=ggraph.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, and Julia Silge. 2018. *Tidytext: Text Mining Using 'Dplyr', 'Ggplot2', and Other Tidy Tools.* https://CRAN.R-project.org/package=tidytext.

Wickham, Hadley. 2019. *Rvest: Easily Harvest (Scrape) Web Pages.* https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, and Kara Woo. 2019. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2019. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

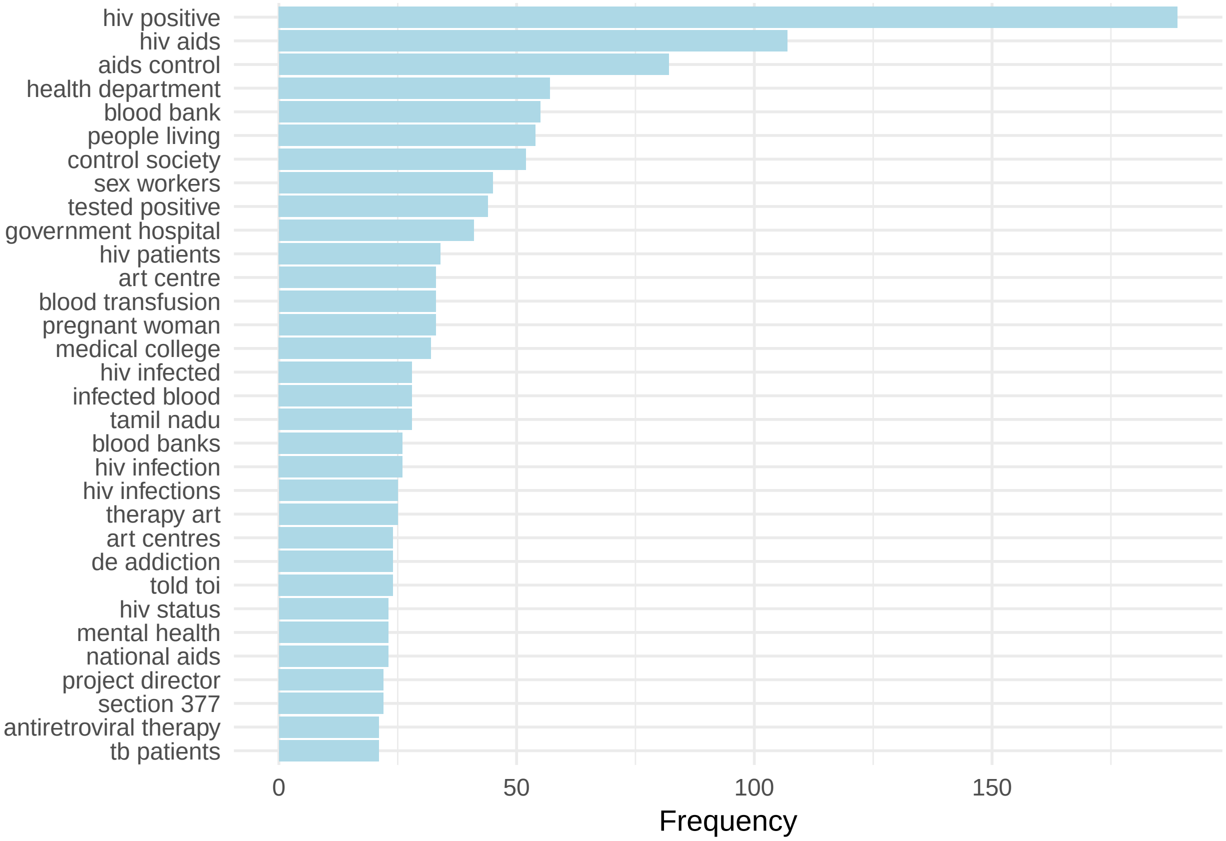

Figure 1: We tokenized using the bigram method to detect the most common pairs of words, and then eliminated any and all stop words from the corpus. We only print a part of the entire dataset where the frequency is greater than 20 partly because of the sheer vastness of the data and also because the frequency decreases ever so slowly from that point onwards

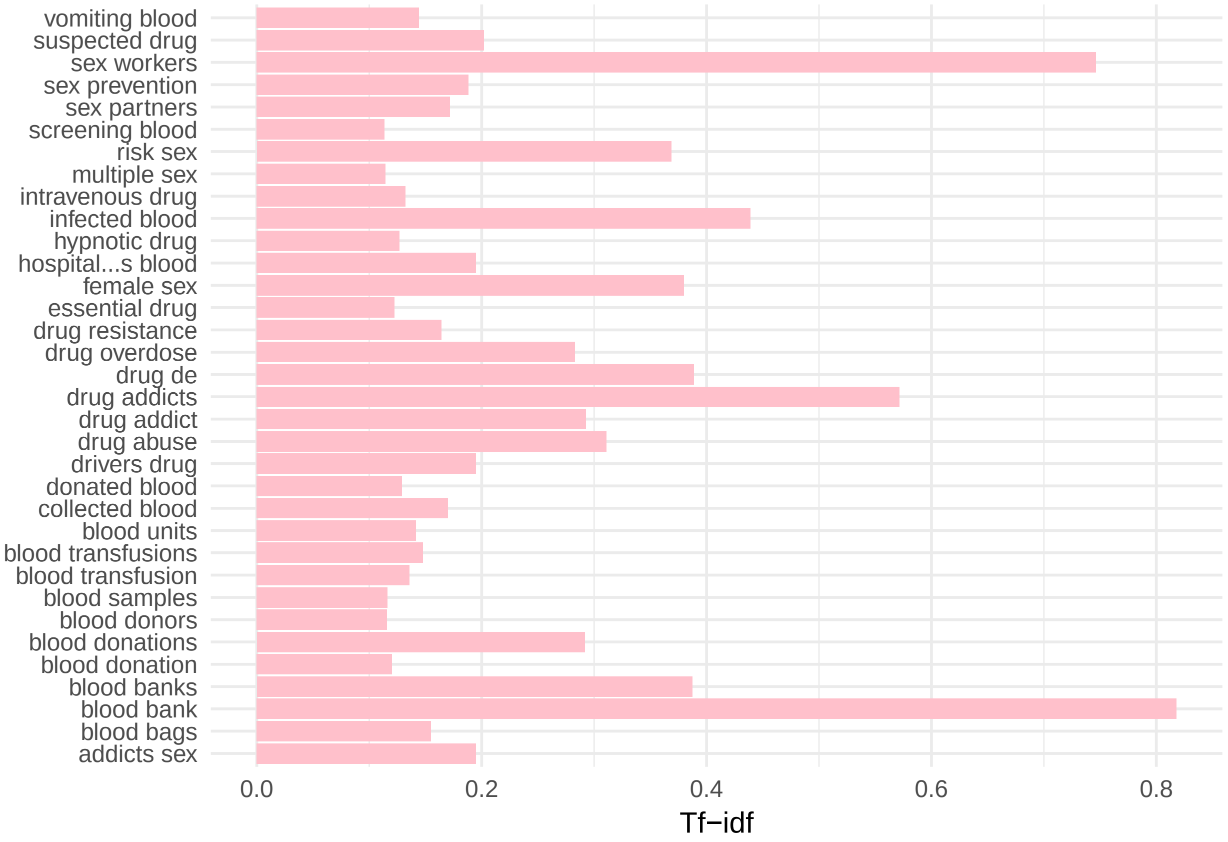

Figure 2: We used the tokenized text to produce a tf-idf variable for each bigram, sorted in descending order of tf-idf value, which we then subset on the condition that each contains either of the words “blood”, “sex” “drug” in order to focus on those aspects of it.

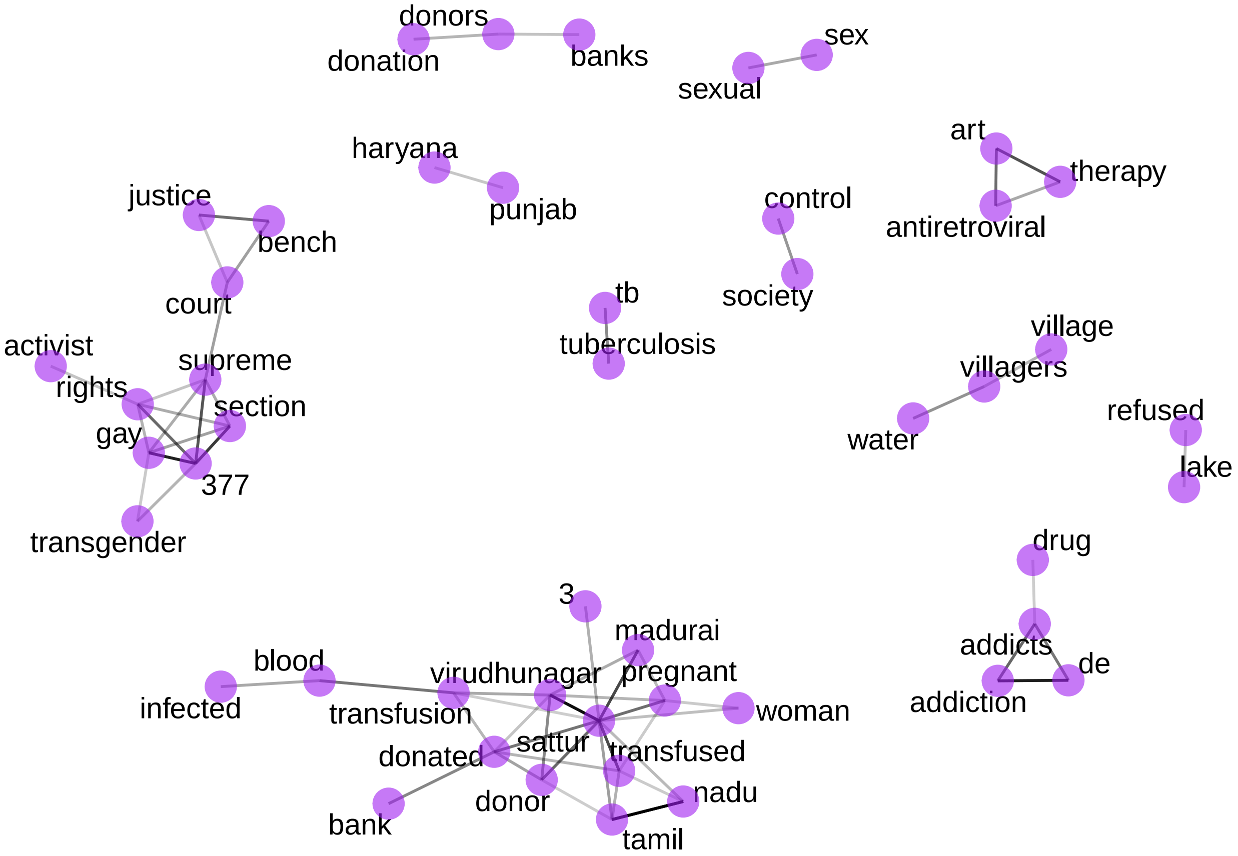

Figure 3: We first tokenized the text from the scraped articles(by each word), preserving the article from whence it came. We then calculated the correlations between the words by article and plotted a subset of these correlations in a graph, considering only pairs whose correlation is greater than 0.48