Twitter Sentiment Analysis – Smarter and Responsible AI

Aaron Xiao, Kiran Sidhu, Paul Flemming, Rozi Hagos, Tara Singh

Executive Summary

We were engaged by the Big Data and Analytics Conference organizers in Toronto to conduct an analysis on topics related to AI that were relevant to local users for their inaugural event. We used search terms specifically related to smarter and responsible AI covering the period from January 2020 to October 2020. We recommend using the themes of "global goals" and "ethics" as central topics to market and promote the event as well as increase revenue and help make future events successful.

Business Objectives and Problem Formulation

When hosting an event, attracting sponsors, vendors and attendees is central to the event's success. Therefore, the organizers want to ensure there is a market for their event in their chosen location. Their main business objective is to see how it will be received in the local community. Another important goal was to perform a sentiment analysis, find topic themes to help build the agenda, draw attendees, assist with marketing efforts, and select the keynote speakers. We obtained tweets on the topic of AI from Twitter to find out if the sentiment is related to smarter and responsible AI. We were able to extract a dataset using search terms related to AI.

Methodology

A few key search terms were used to gather information related to "Smarter and Responsible AI". Several graphs and models were created to perform the sentiment analysis which provide insight into descriptive words, top words, group clustering and sentiment related to the search terms. Further insight into the results will be provided.

Preliminary Analysis, Data Manipulation, Descriptive Analysis

The base Twitter API does not return results older than 7 days, so we used the Sandbox API which allows access to all Twitter data. The limitations we had were a quota of requests/tweets we could pull each month, and 128-character limit to our query. This meant we were restricted in the number of tweets we could pull at this time.

The words used in our query were "#AI", "#artificialintelligence" and "#smarterAI". All non-English tweets were filtered out and were limited to a specific geolocation. The center of the geolocation was located at

the following coordinates [-79.386672 43.773886] which was in a central location in Toronto (near the intersection of highways 401 and 404) and had a radius of 25 miles. Getting data from the last 7 days would have provided thousands of tweets, it is more difficult to obtain historical data. We widened the timeline from January to October and honed in on keywords that were more relevant to the topic. This would provide a more in-depth perspective.

With our dataset complete, we cleaned the data for analysis by removing "http", converted all text to lowercase, removed all punctuation, and “stop” words. The total data set contained 838 tweets and 90 columns to analyze. The focus of the work was on the text column which contained the content of the tweets.

Model Building and Evaluation

A line graph was created to display the Tweets by hour (Figure 1). A word cloud was used to visualize the most popular words and frequency in which they appeared in the text column (Figure 2).

Bar graphs were used to find unique and top words connected to our search terms (Figure 3)

To identify groupings and relationships between words, a network analysis graph (Q Graph) was run to find the associations (Figure 4). Cluster Dendrogram (Figure 5) is another approach used to demonstrate hierarchical clustering. An EGA Graph (Figure 6) was run to produces a "visual guide—network plot— that not only indicates the number of dimensions to retain, but also which items cluster together and their level of association” (Reference 2).

The bigram shows the grouping of two words used together in a sentence. The bigram shows what words follow each other or come before each other (Figures 7, 8).

We performed a Bing sentiment analysis and created separate bar charts for negative and positive sentiments (Figures 9,10). NRC lexicon sentiment analysis was also conducted (Figures 11, 12). This

lexicon “Categorizes words in a binary fashion (“yes”/“no”) into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust." (Reference 1)

Insights and Summarizing Results

In the Tweet by hour line graph we found that the peak time of day for tweets was around 15:00 UTC/GMT (See Figure 4). Also, between 5:00 and 10:00 UTC/GMT was non-peak time for tweeting activity on our topic.

In the word cloud, the largest words are the words that appear more frequently and are related to our search terms. Other related words are also featured in the cloud with smaller fonts. The largest words are "tech", "learning", "innovation", "data", "fintech", "globalgoals" and "deeplearning". We also noticed that "COVID-19" appeared which we thought was interesting (See Figure 2).

In the Bigram model "ai" is mainly grouped with "ml", "globalgoals" and "fintech" etc. (See Figure 8). The bigram graph also shows words that would typically come before or after the word "ai". All of the words that are listed are aligned with all of the results that we have previously seen. The only new terms that we see popping up are the words "will" and "new" (See Figure 8).

In the Unique Word Count Chart, we found that the top unique word was "ai" (we will disregard because it is the search term used). The actual top 5 unique word were "machinelearning", "tech", "learning", "data" and "innovation" (See Figure 3). Some of the interesting and important words that popped up on the unique word count when we conducted the twitter search were: "deeplearning", "fintech", "sustainable" and "globalgoals" (See Figure 3). We noticed that "sustainable" and "innovation" also appeared in the Contribution to Twitter Sentiment bar graph (Figure 9) as a related term.

Visually, we can identify 3 groups in the network analysis graph (Figure 4). For the grouping in the top right section of the diagram, the nodes include" globalgoals", "sustainable", "sdgs" (which stands for sustainable development goals), "tech", "fintech" and "innovation". On the bottom right section, the nodes include "data", "new", "amp", "will" and "cloud". We noticed that "cloud" was standing alone and not closely related to any of the other nodes. The section on the left has several words and terms that are closely related to our search terms. The grouping includes: "artificial", "intelligence", "ml", "can", "datascience", "deeplearning", "machinelearning", "artificalintelligence", "via" and "learning".

In the Cluster Dendrogram which features hierarchical clustering, reinforces what we noticed earlier, there are two major branches that include "tech", "innovation", "fintech", "globalgoals", "sdgs" and

"sustainable". The other major branch was related to "artificialintelligence", "can", "data", "artificial", "intelligence", "datascience", "ml", "amp", "new" and "cloud" (See Figure 5).

In the EGA graph, the central node is "globalgoals". The red nodes on the top are centred around "deeplearning". On the bottom right, the green nodes are centred around "machinelearning" and the blue nodes on the bottom left are tightly interconnected with "sdgs" as the centre of that node (See Figure 6).

In the Bing Contribution to Twitter Sentiment the top three negative terms were "cloud", "risk" and "drones". The top three positive terms were "innovation", "sustainable" and "intelligence" (See Figure 9). At a glance positive sentiment count also outweighs the negative sentiment. It was noted that ethical appears to have a positive sentiment (Figure 9). The NRC Twitter Sentiment Graph shows similar findings where the positive sentiment far outweighs the negative. The positive emotions are "joy", "positive", "surprise", "anticipation" and "trust". The negative emotions are "anger", "disgust", "fear" and "sadness". (See Figures 11, 12). "Trust", "anticipation" and "joy" appeared most often in the tweet sentiments.

Recommendations

If you want to use Twitter itself as a marketing tool, we suggest tweeting at these popular times for tweets (15:00 UTC/GMT) and not tweeting at non-peak times (5:00 and 10:00). In addition, using the following key words that came up in our descriptive analysis ("tech", "learning", "innovation", "data", "fintech", "globalgoals", "machinelearning", "tech", "learning", "data", "innovation", "fintech", "sustainable" and "deeplearning"), you can leverage these terms to create compelling hashtags and create more conversations around your topics which should increase interest and awareness.

With the same process we can identify the key trends to benefit the advertising company and the event organizers who can use the keywords found in the analysis to develop the themes and agenda for your event. You can then find speakers that specialize in these topics and use their presence to create your targeted marketing campaign. This will attract your big sponsors, vendors, and attendees. This is important because the conference earns revenue through sponsorship fees, exhibition fees and entrance fees.

The two major branches we found in hierarchical clustering were: Branch 1 ("tech", "innovation", "fintech", "globalgoals", "sdgs" and "sustainable") and Branch 2 ("artificialintelligence", "can", "data", "artificial", "intelligence", "datascience", "ml", "amp", "new" and "cloud"). "Globalgoals" and "fintech" also showed up together in the bigram analysis. From this we can recommend that you can develop two distinct streams for your seminars, recruiting, and training using these branches as a guide. In the EGA graph, "globalgoals" was the central node, connected to all the other nodes. As such, it would be a good central theme for your event.

Since this is a new event, the overall idea we can see in our sentiment analysis is that the local market has positive feelings towards AI, this means a conference is a viable business opportunity. We also can be sure there will not be any negative media attention, publicity, protests etc. In addition, because the feelings are so positive, this will create "excitement" about your topic and help spread the word for future events.

Conclusion

Although our analysis of Twitter data has given us central themes and keywords for your event and let us know that AI is perceived as a positive technology, it's what's missing in the results/analysis that is a concern. Ethics did not come up as an important keyword in any of our findings. This shows that there is a gap in the conversation and the marketplace. Since this topic is important to your organization, we would recommend inviting a keynote speaker to address this topic and make it one of the main themes of your event along with global goals.

Appendix

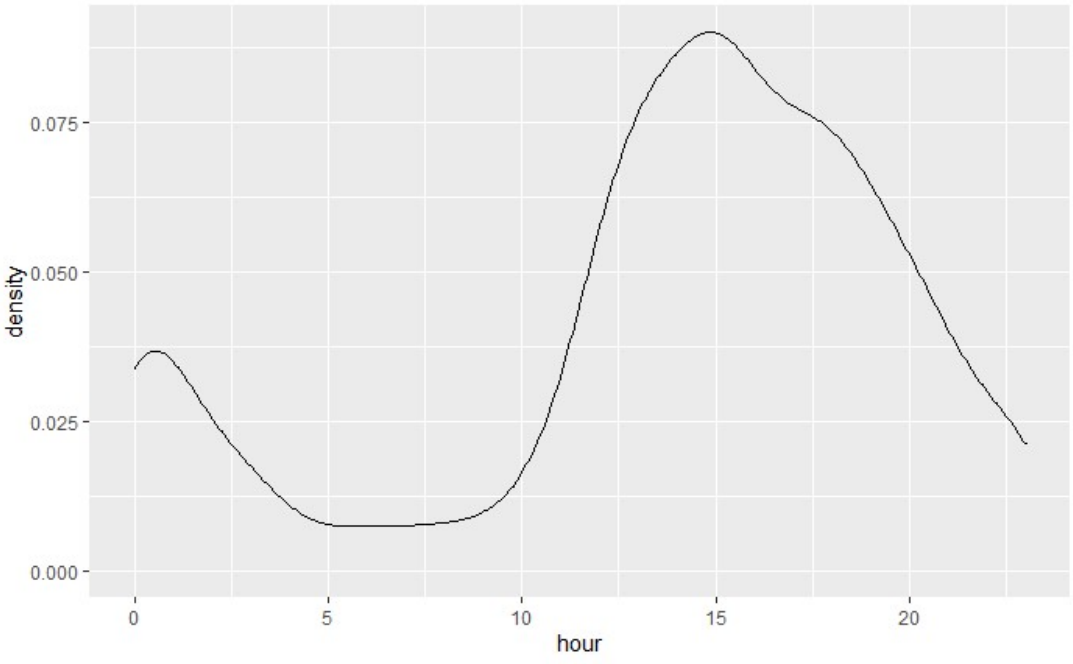

Figure 1: Tweets by hour

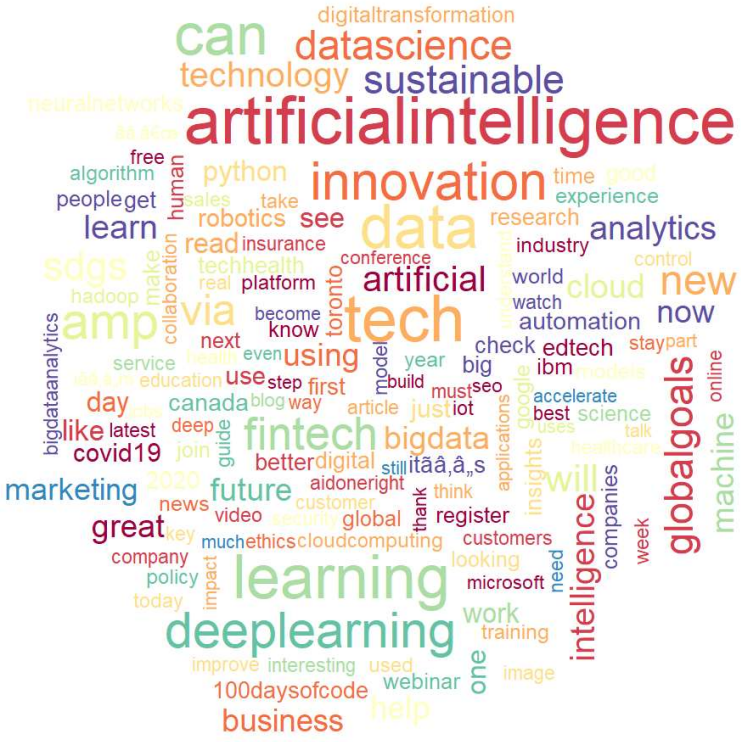

Figure 2: AI Wordcloud

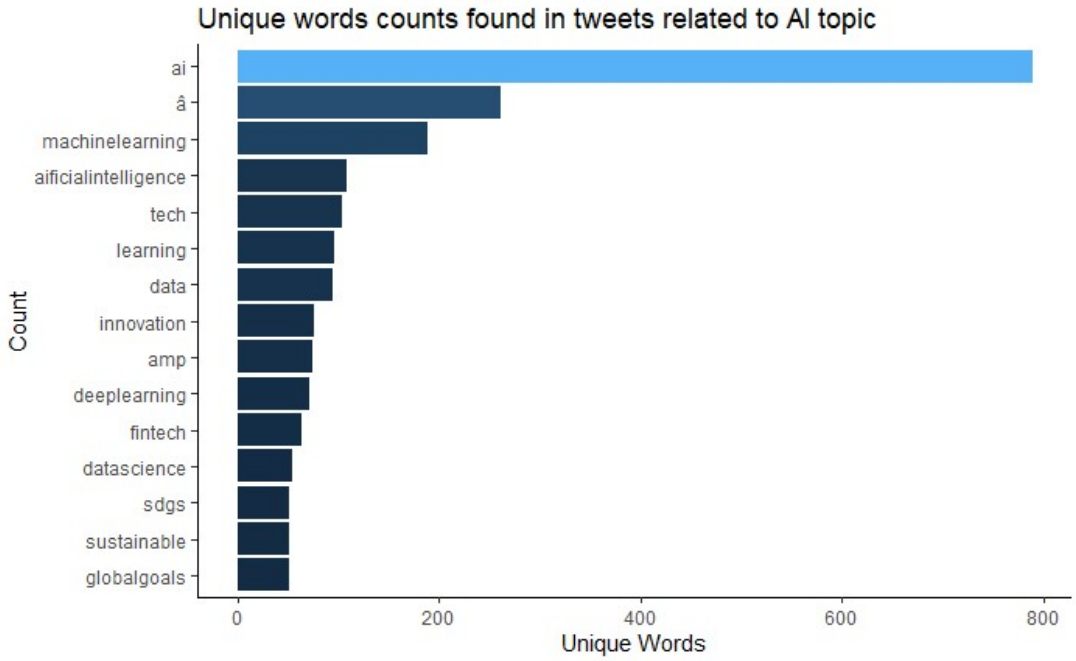

Figure 3: Unique Word Counts found in tweets related to AI

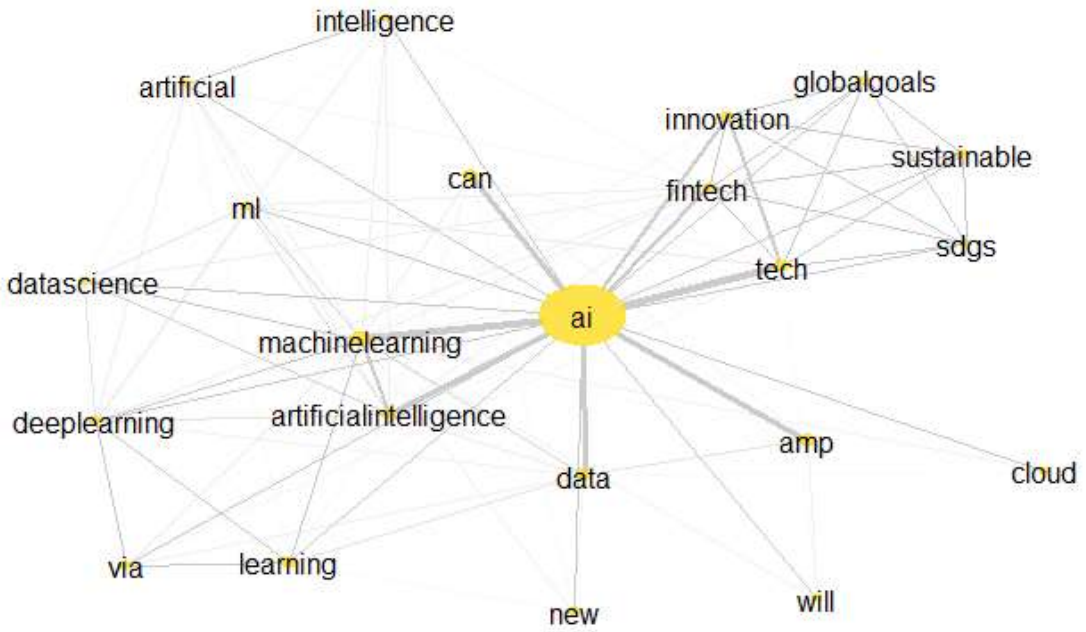

Figure 4: Q-Graph

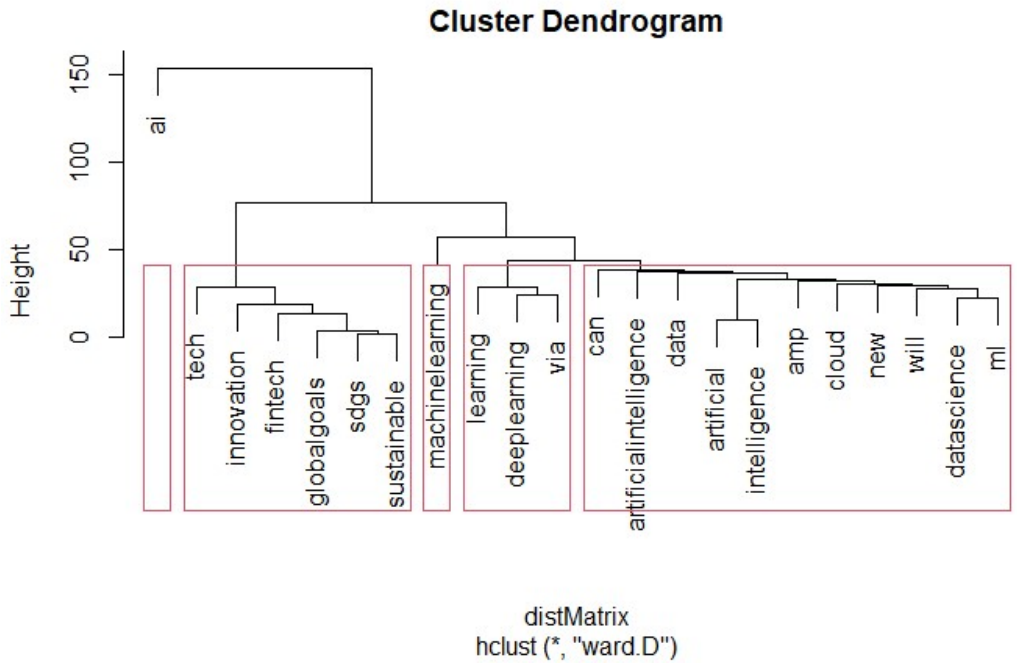

Figure 5: Clustering

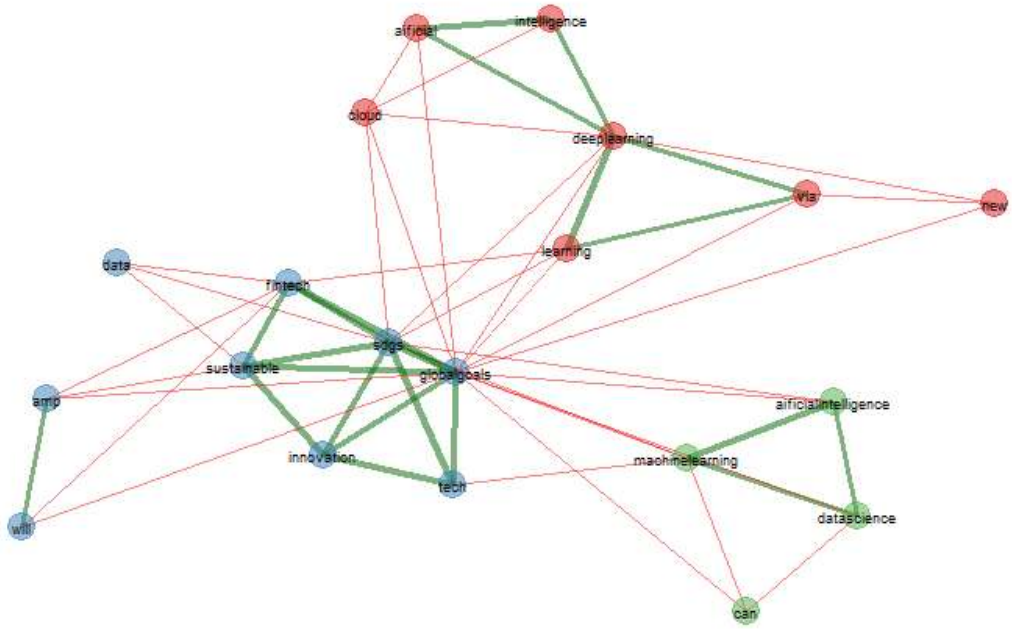

Figure 6: EGA Graph

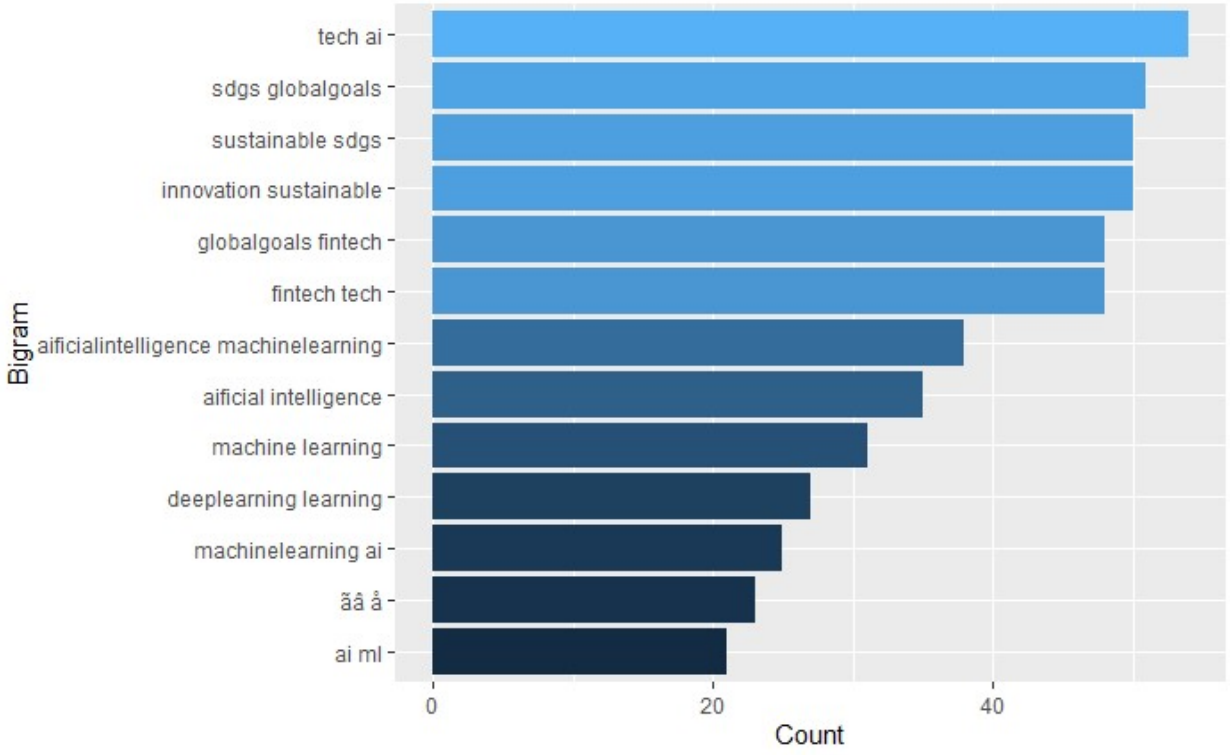

Figure 7: Bigram words

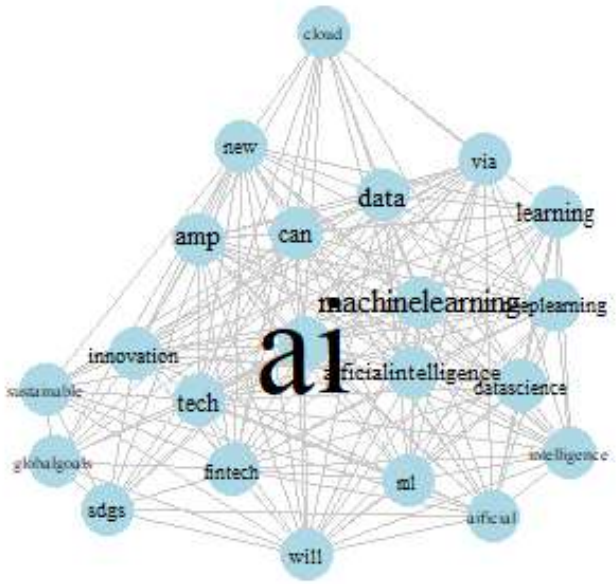

Figure 8: Bigram

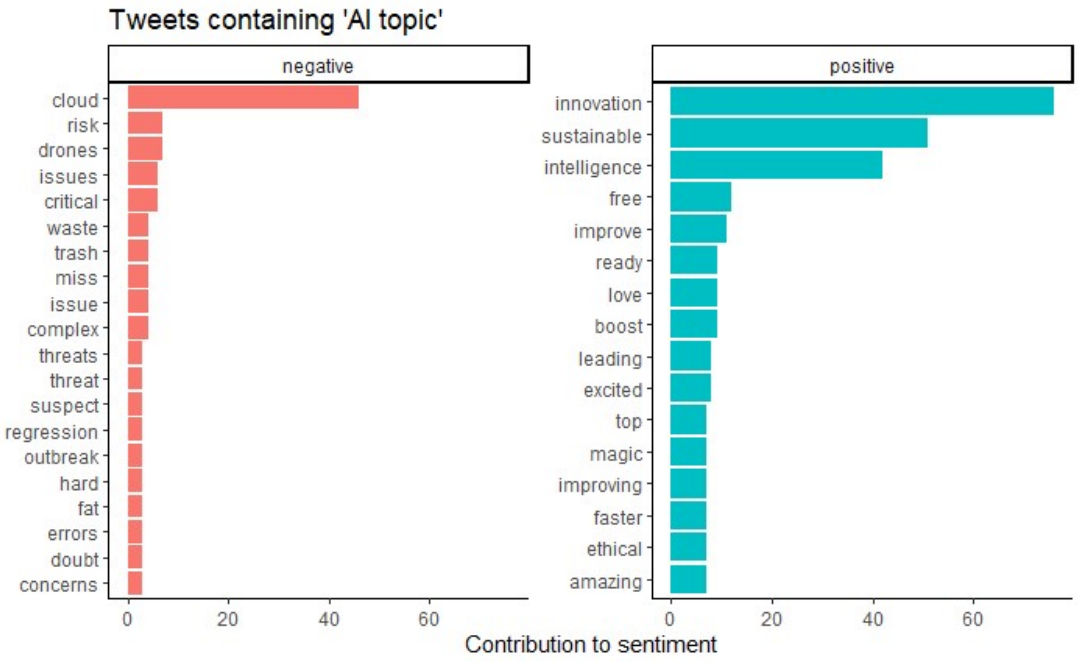

Figure 9: Contribution to Twitter sentiment

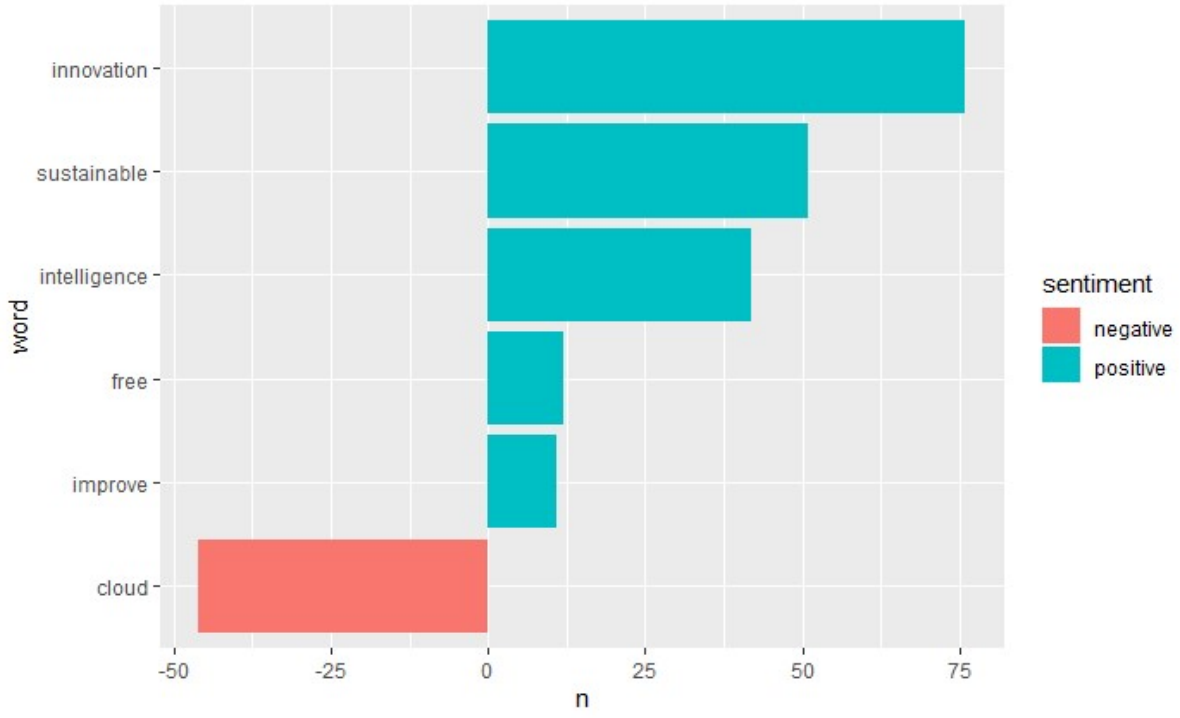

Figure 10: Positive and negative sentiment

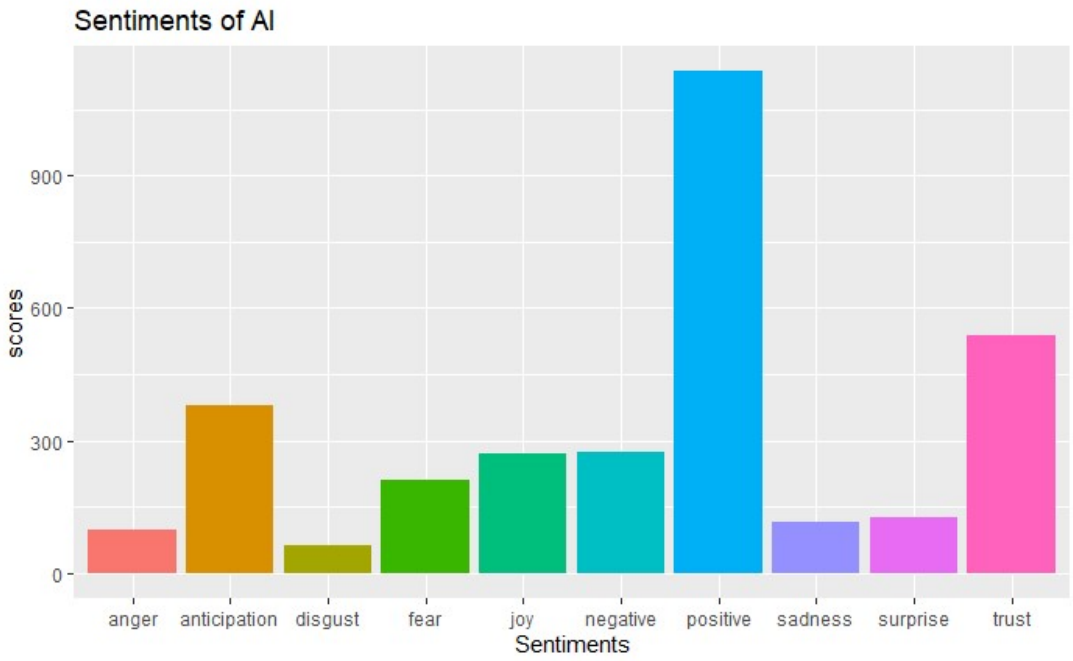

Figure 11: Twitter sentiment

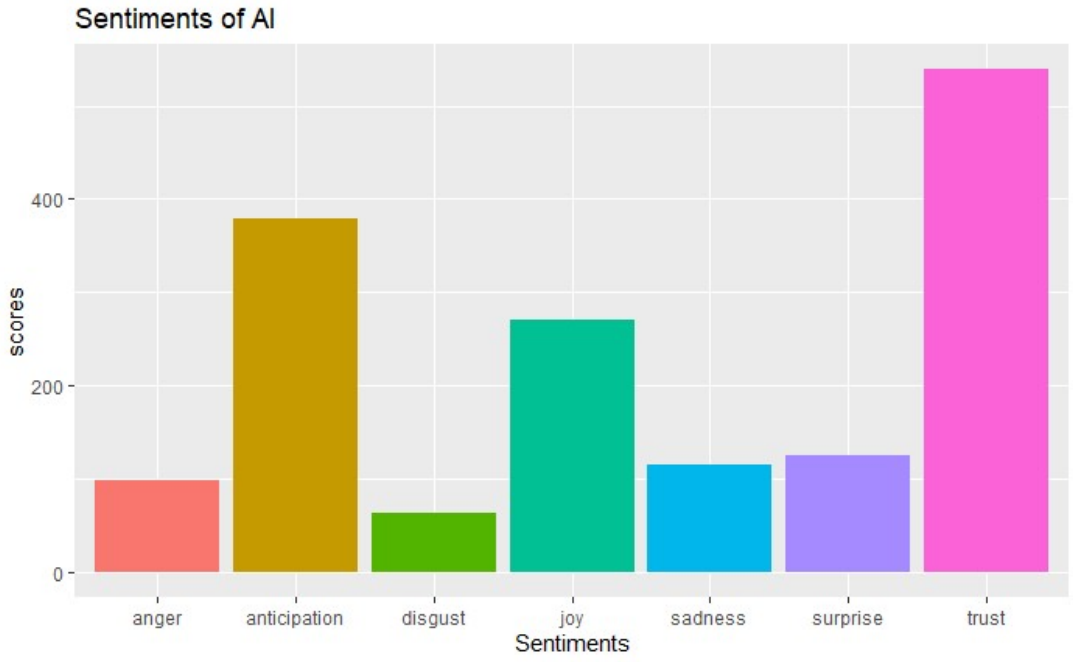

Figure 12: Twitter sentiment

References

(1) https://uc-r.github.io/sentiment_analysis

(2) Investigating the performance of Exploratory Graph analysis and traditional techniques to identify the number of latent factors: a simulation and tutorial. Hudson Golino, Dingjing Shi, Alexander P. Christensen, Luis Eduardo Garrido, Maria Dolores Nieto, Ritu Sadana, Jotheeswaran Amuthavalli Thiyagarajan & Agustín Martínez Molina.