An analysis of Covid-19 vaccine between manufactures on Twitter

Yi Ren

Institute of statistics

University of Munich

06.03.2023

Table of content

[Introduction 4](#_Toc128853903)

[Structure 5](#_Toc128853904)

[Related work 6](#_Toc128853905)

[Materials and methods 6](#_Toc128853906)

[Data acquisition 7](#_Toc128853907)

[Methods 7](#_Toc128853908)

[- Sentiment analysis 8](#_Toc128853909)

[- Wordcloud 9](#_Toc128853910)

[- Network 10](#_Toc128853911)

[- Topic modelling 10](#_Toc128853912)

[- Time series 11](#_Toc128853913)

[Results 12](#_Toc128853914)

[- Sentiment analysis 12](#_Toc128853915)

[- Wordcloud 17](#_Toc128853916)

[- Network 19](#_Toc128853917)

[- Topic modelling 22](#_Toc128853918)

[- Time series 23](#_Toc128853919)

[Discussion and conclusions 27](#_Toc128853920)

Table of figures

[Figure 1 8](#_Toc128853880)

[Figure 2 Plot 1 12](#_Toc128853881)

[Figure 3 Plot 2 13](#_Toc128853882)

[Figure 4 Plot 3 14](#_Toc128853883)

[Figure 5 Plot 4 14](#_Toc128853884)

[Figure 6 Plot 5 15](#_Toc128853885)

[Figure 7 Plot 6 16](#_Toc128853886)

[Figure 8 Plot 7 16](#_Toc128853887)

[Figure 9 Plot 8 17](#_Toc128853888)

[Figure 10 Plot 9 Wordcloud for positive biontech tweets 17](#_Toc128853889)

[Figure 11 Plot 10 Wordcloud for negative biontech tweets 18](#_Toc128853890)

[Figure 12 Plot 11 Wordcloud for neutral biontech tweets 18](#_Toc128853891)

[Figure 13 Plot 12 Network of positive biontech tweets 19](#_Toc128853892)

[Figure 14 Plot 13 Network of negative biontech tweets 20](#_Toc128853893)

[Figure 15 Plot 14 Network of neutral biontech tweets 20](#_Toc128853894)

[Figure 16 Topics for biontech 22](#_Toc128853895)

[Figure 17 Topics for AstraZeneca 22](#_Toc128853896)

[Figure 18 TS for positive biontech 23](#_Toc128853897)

[Figure 19 TS for negative biontech 24](#_Toc128853898)

[Figure 20 TS for neutral biontech 24](#_Toc128853899)

[Figure 21 TS for positive AZ 25](#_Toc128853900)

[Figure 22 TS for negative AZ 25](#_Toc128853901)

[Figure 23 TS for neutral AZ 26](#_Toc128853902)

# Introduction

It has been two years since the breakout of the covid-19 pandemic in 2020, our lives seem have finally came back to normal conditions, but looking back at the past two years, we’ve been through a lot during that special situation: at the beginning period of pandemic, when newly developed high-tech vaccines were firstly announced and promoted to the market, there were many uncertainties about their safety and effectiveness from the public, this has led to vaccine hesitancy or even vaccine rejection. A lot of work and research related to that have been done to help more people get more trust in receiving vaccine. Now, the rate of vaccination coverage in most countries of the world has reached over 70%, what are the feelings of the public about vaccine after two years’ observation, what do people now care about this topic and what kind of expectations/worries do they have on different vaccine manufacturers. All these potential analyses could be an essential concern for public health authorities, journalists, government and manufacturers themselves when making political decisions, evaluating or promoting products, building indications for their own research or work. Hence, there’s a great value to do an analysis on covid-19 vaccine between manufacturers.

Survey is normally a usual way to do some national or global investigations, but instead of survey, here the information is collected on twitter. Main reasons for that are: 1) comparing with information from social media, survey is way too difficult to conduct, especially when the operator is not from a national institute. 2) the procedure of survey takes a long time in order to have the desired data, on the other hand, data acquisition from social media is much easier. 3) when the context refers to sensitive information, data from social media often have a higher accuracy. 4) it’s not feasible to collect the data through survey if the hidden circumstance is very dynamic, which is the case here for vaccine analysis. On the contrary, social media is an appropriate choice due to its characteristics.

Social media is increasingly used for discussing and sharing viewpoints about infectious disease outbreaks health topics[[1]](#footnote-1), Twitter, as one of the most active social media platforms, has more than 300 million global users who post millions of opinions on every possible topic every day, and therefore is a good resource of useful information.

# Structure

In this work, there are two main manufacturers of interest: Biontech and AstraZeneca. Datasets are collected separately for these two manufacturers. The data collected from twitter are very chaotic, including various attitudes and topics. To perform a meaningful analysis on data, a brief investigation on data is conducted. It’s decided that the data will be analyzed in two formats: one is with duplicated tweets for net-working and time series analyses the other is without duplicated tweets for the rest of analyses. Then according to the sentiments of each tweet, the tweets are divided into three groups that are used to do the comparison afterwards, they are namely: positive group, negative group and neutral group. The division is based on the scores for different sentiments and a subjective definition of positivity, negativity. For each group, the most frequently mentioned words are ranked and a wordcloud serves as a tool for visualization to give a better presentation for readers so that they can make an easier comparison between different groups. Furthermore, dataset with duplicated tweets is used to build a network between identified top 50 mentioned and therefore most influential users. It reveals besides the number and size of connected clumps consisted of top users, but also the interactions within them. At the end, it can be studied to understand the characteristics of the corresponding group, such as potential occupation, number of friends, number of followers and more. For both manufacturers, the discussing topics were searched separately. At the very end, a time series analysis that aims to give an overview of correlations between most mentioned words and their appearing dates is conducted for each group in both manufacturers. This could be seen as an indicator for the trends of popular themes. All the results from analyses mentioned above can be compared intra and inter both manufacturers and give readers a both broad and deep understanding of current situations about vaccine.

# Related work

Many different studies have been done before in analyzing people’s attitudes towards vaccination through twitter data, or use twitter data to identify sentiments [1], dominant opinions [2], themes [3], or do network between most influential users for a particular sentiment [4] , they all happened before 2022, there were also theses that have categorized activities from social media into vaccine hesitancy and anti-vaccine movements [5] or finding out that influential twitter accounts contributed a great part of vaccine-opposition messages in the US [6]. Other research with similar directions serves mainly for the purpose of improving vaccination rate. This work is a post-period comparison with combined methods which tries to give indications for all possible potential usages.

# Materials and methods

## Data acquisition

In R, function “search\_tweets” from “rtweet” package is used to collect related data from Twitter. Key words “biontech” and “astrazeneca” were given as inputs to retrieve tweets with these key terms from 08-01-2023 to 12-01-2023. The chosen language for tweets is set to English as it’s a globally most often used language and the data is supposed to collected internationally. For the same reason, there was no geographical limitation for information searching. Because of limited computational budget, the upper bound for searched tweets was set to 5000 and we had allowance for re-tweeted tweets to give more possibilities for the analysis later on. The datasets we got in designed way were with 43 variables: “created\_at”, “id”, “id\_str”, “full\_text”, “truncated”,” display\_text\_range”, “entities”,” metadata”, “source”, “geo”, “coordinates” and many other details. Among these, a new variable “date” is created through using information from variable “created\_at” so that only the date is extracted and restored. This makes a time series analysis for later more feasible.

For key term “biontech” the dataset pulled from Twitter consists of 4822 tweets and for “astrazeneca” 4890 tweets in total. After checking the datasets, it is found out that there were 3347 and 2661 duplicated tweets for each of them and the datasets were created by 4815 users and 4883 users separately. Considering the methods that will be applied later on, for dataset, two versions are used, namely dataset about biontech with duplicated tweets, dataset about biontech without duplicated tweets, dataset about astrazeneca with duplicated tweets and dataset about astrazeneca without duplicated tweets. Specifically, datasets without duplicated tweets were used for wordcloud and topic modelling since both methods are highly related to the absolute number of unique tweets, so if datasets with duplicated tweets were used, that would result in biased results.

# Methods

## Sentiment analysis

Sentiment Analysis is a text classification technique used to analyze natural language text[[2]](#footnote-2), it uses Natural Language Processing (NLP) to determine whether the sentiments expressed towards a subject are positive, negative or neutral [10]. There are mainly three approaches to realize sentiment analysis as shown in the figure 1 below: a machine learning approach, a lexicon-based approach and a hybrid approach.

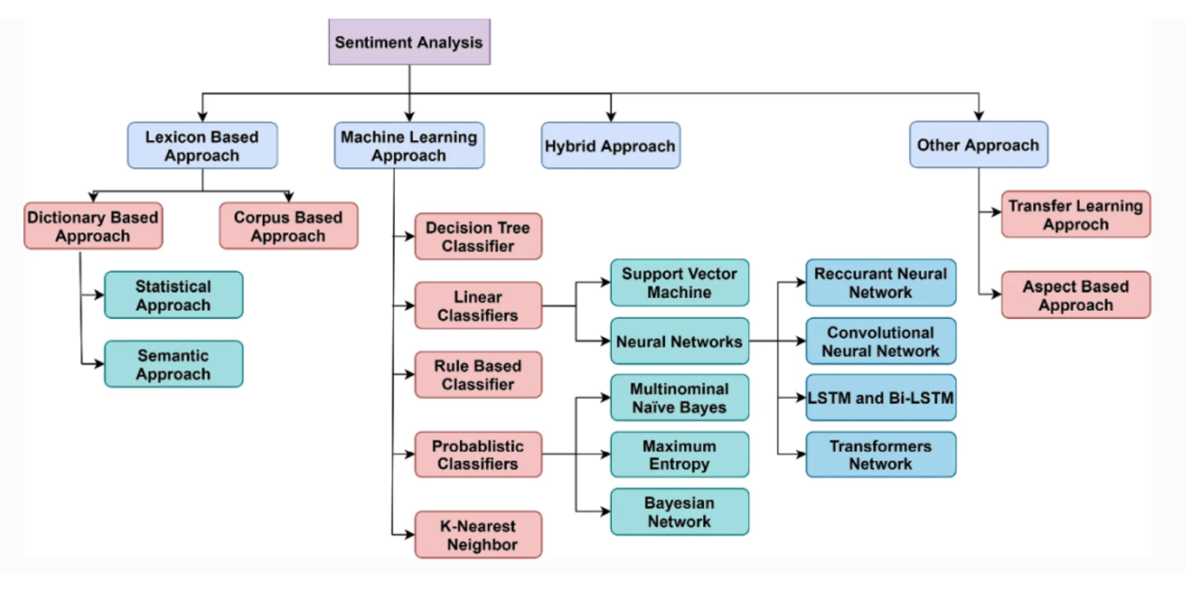


Figure 1

source:<https://link.springer.com/article/10.1007/s10462-022-10144-1>

To classify tweets from datasets into “positive”, “negative” and “neutral” three categories according to the sentiment they represent, Valence Aware Dictionary and sEntiment Reasoner (VADER) is used to assign sentiments to every tweet. VADER is a Python lexicon and rule-based sentiment analysis tool [9] and is designed to determine sentiments of social media posts based on individual words and sentences [11]. get\_nrc\_sentiment from syuzhet package is applied on duplicated and not duplicated datasets for both biontech and astrazeneca to get sentiment scores for each tweet from corresponding dataset. Different scores are given for “anger”, “anticipation”, “disgust”, “fear”, “joy”, “sadness”, “surprise”,” trust”, “negative” and “positive”. Whether a tweet has positive or negative sentiment is decided according to a rule based on author’s judgement: the scores of “joy”, “trust” and “positive” are summarized into one score, which can be seen as a total positive score. The scores of “anger”, “fear” and “negative” are added up as the final total negative score, “anticipation” and “surprise” were not considered for their ambiguity. accompanying with the original sentimental scores from get\_nrc\_sentiment function, the whole sentimental scores are added up to original datasets and the part, whose total positive scores are greater than total negative scores are set as dataset with positive sentiment, similarly, the part with greater final negative scores is seen as dataset with negative sentiment, if the scores for total positive and negative sentiments are the same, it will be treated as dataset with neutral sentiment.

Then, for every positive, negative and neutral dataset, the distribution of all aspects of sentiments is analyzed through counting the summarized score for them.

## Wordcloud

For the creation of wordcloud, datasets without duplicated tweets are used. Here we will take the positive dataset for biontech as example to do the explanation. After classification of the datasets, the first step is to do some data pre-processing and cleaning. Firstly, potential website address and usernames are removed from tweets’ text and so is the symbol &. Then, the cleaned texts are tokenized by words. After that, the stop words for English are uploaded through “stopwords” package. Stop words are the words in a stop list which are filtered out before or after processing of natural language data (text) because they are insignificant[[3]](#footnote-3) . the English stop words are continually deleted from tokenized words. After this process, words are ranked according to its frequency. The most frequently appeared but useless words for this work are manually picked out and listed for a deeper words-cleaning. Finally, wordcloud function from “quanteda” package was applied on the ultimate cleaning words to produce a wordcloud. Same work has been done for other sentiment groups from biontech and astrazeneca.

## Network

To begin with a network analysis, the till now used datasets of unique tweets’ context must be exchanged into the datasets with duplicated tweets’ text, because the interactions between twitter accounts is the essential meaning of a network analysis. Same as before, we take the dataset with positive sentiment from biontech as an example for explanation. The positive biontech dataset was firstly converted into a text corpus by using corpus function from “quanteda” package, then the texts were cleaned by removing punctuations, numbers, symbols and the same manually selected words from the list created at last step. English stop words were also filtered out for that corpus. After those steps, the texts were tokenized with 1grams. The work of network analysis is based on a document feature matrix, which is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document feature matrix, rows are corresponded to documents in the collection and columns correspond to features[[4]](#footnote-4). For this reason, the tokenized corpus is converted into a document feature matrix by using dfm function from “quanteda” package. Then the DFM (document feature matrix) was subset to a matrix that only contains usernames by using the pattern in regex “@\*”. On top of that, a feature co-occurrence matrix, which measure co-occurrence of features within documents, was created to serve for network analysis. In this work, only the top 50 most popular users were selected to build the network. This is realized through textplot\_network function from “quanteda. textplots” package.

## Topic modelling

Except for sentiments and network, important topics discussed for each of the two manufacturers is also of great interest. in this work, Latent Dirichlet Allocation (LDA) method is deployed to extract discussing topics for biontech and astrazeneca. LDA is a generative statistical model that explains a set of observations through unobserved groups, and each group explains why some parts of the data are similar [[5]](#footnote-5) . LDA assumes the following generative process for a corpus D consisting of M documents each of length :[[6]](#footnote-6)

1. Choose ~ Dir (α), where i∈ {1…M} and Dir (α) is a Dirichlet distribution with a symmetric parameter α which typically is sparse.
2. Choose ~ Dir (β), where k∈ {1…K} and β typically sparse
3. For each of the word position i, j, where i ∈ {1…M} and j ∈ {1…}

Choose a topic ~ Multinominal ()

Choose a word ~ Multinominal ()

LDA is applied on the DFM without empty documents and after several trials, the number of topics was set to 6 for biontech and 6 to astrazeneca. At the end, wished number of terms that could represent each topic was chosen and visualized since the LDA method won’t automatically give the name or summary of the topics, readers have to recognize them by observing those terms.

## Time series

What else could be interesting for potential target readers is a time series analysis that shows the time development of the most mentioned key words to track the movement of public’s interest. in this work, the datasets with duplicated tweets were used for this part. First of all, terms of interest were selected by the frequency listed in the DTM of each dataset, then, created DTM was turned into data frame that only includes columns of interested terms, after data transformation, a time series visualization is created through using ggplot. This is done for both manufacturers with different sentiments so that readers can have a comprehensive insight into the provided information.

# Results

## Sentiment analysis

Under help from VADER, the datasets of biontech and astrazeneca were categorized into positive, negative and neutral three sub datasets for both with and without duplicated tweets versions. A sentiment analysis is only applied on the datasets without duplicated tweets.

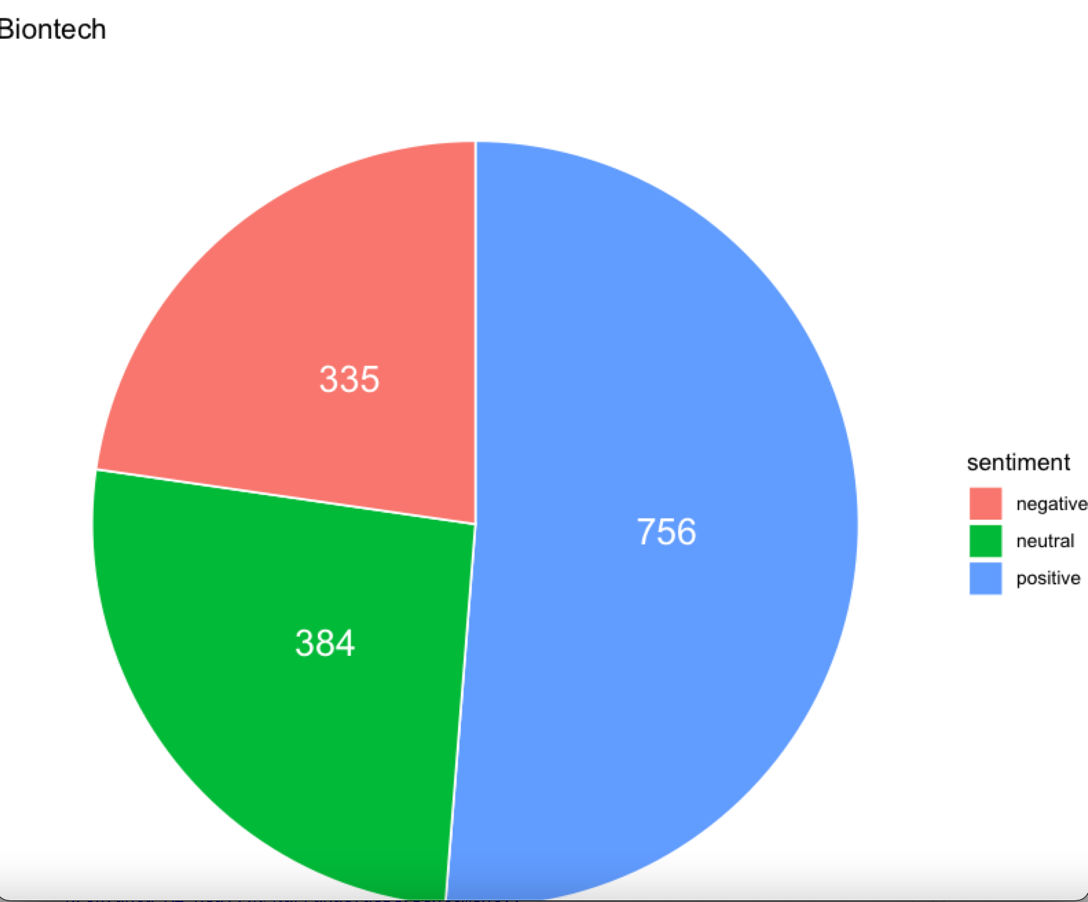


Figure 2 Plot 1

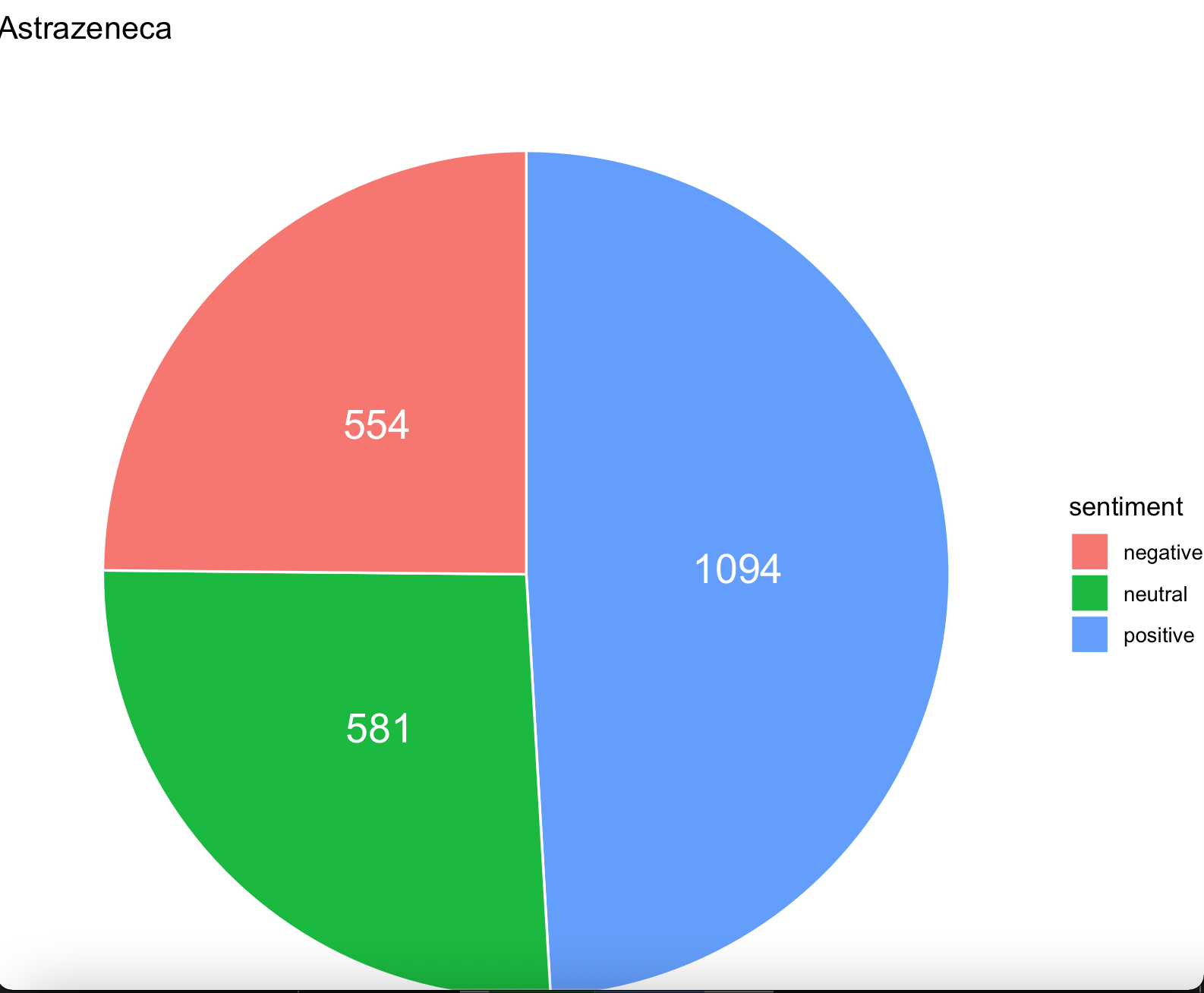


Figure 3 Plot 2

Plot 1 and plot 2 have shown the percentage of each sentiment for both manufacturers, As can see from the plots: the percentage of positive sentiment is over 50 percent for biontech with the absolute counts of 756, followed by neutral sentiment with 384 absolute counts and negative sentiment with 335 absolute counts. For Astrazeneca, the highest percentage is still from positive sentiment with 1094 counts, but is already less than half of all tweets, then comes neutral sentiment in the second place with 581 tweets, there were 554 tweets from astrazeneca with negative sentiment, which is pretty as many as tweets with neutral sentiment. Through these we can make a brief conclusion: Twitter users generally have a more positive impression for biontech than for astrazeneca even it seems like that more things are discussed about astrazeneca than about biontech.

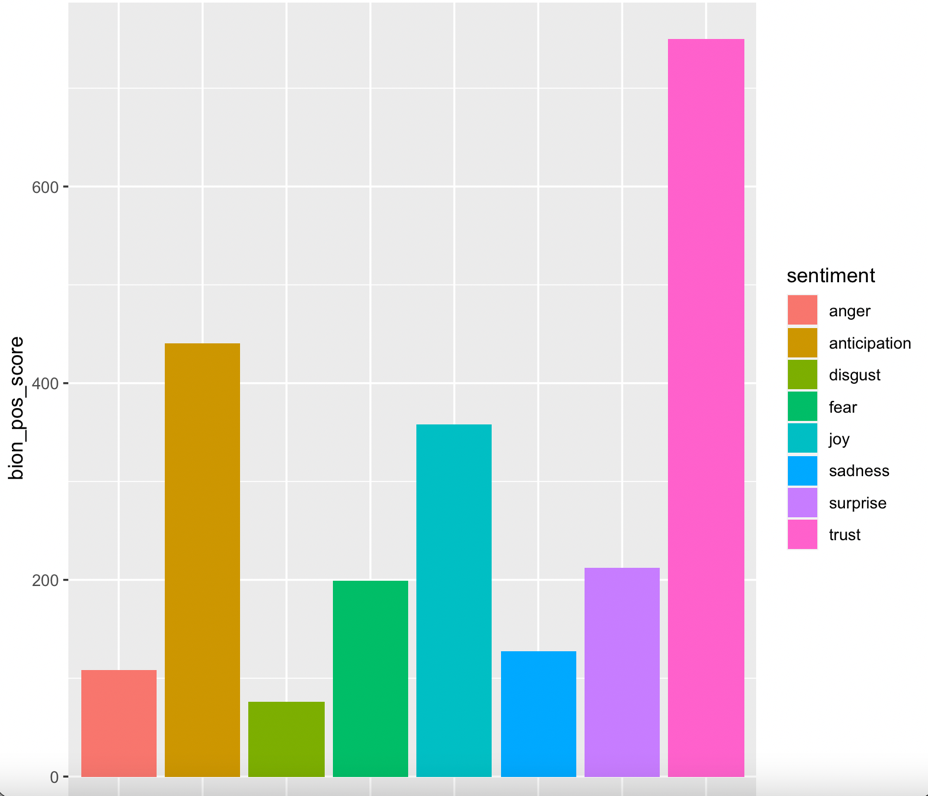


Figure 4 Plot 3

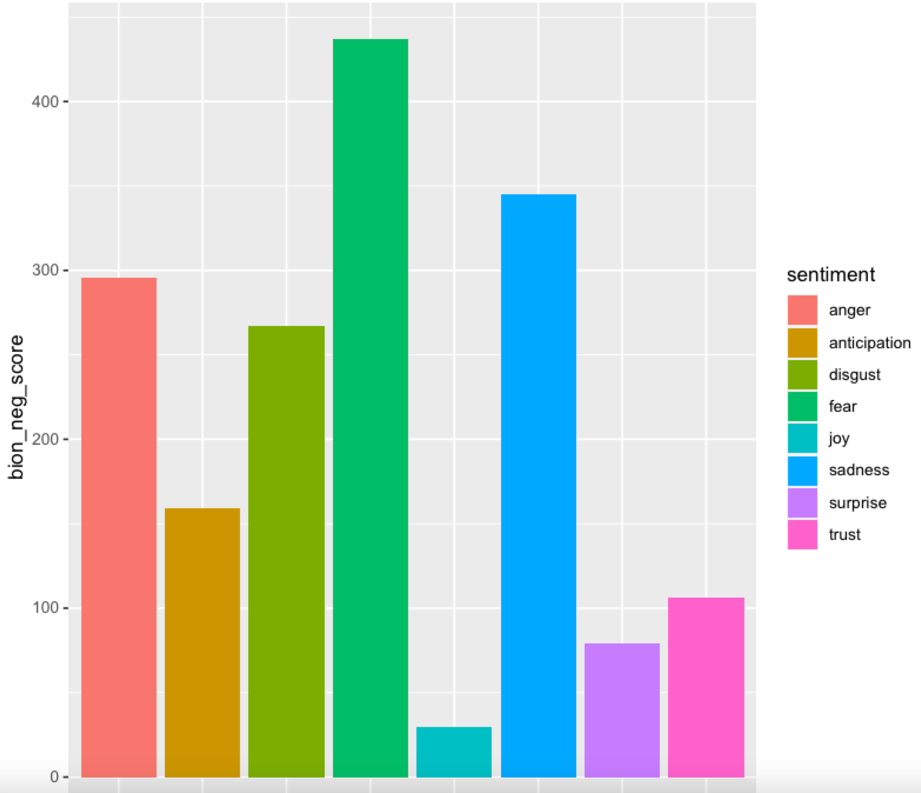


Figure 5 Plot 4

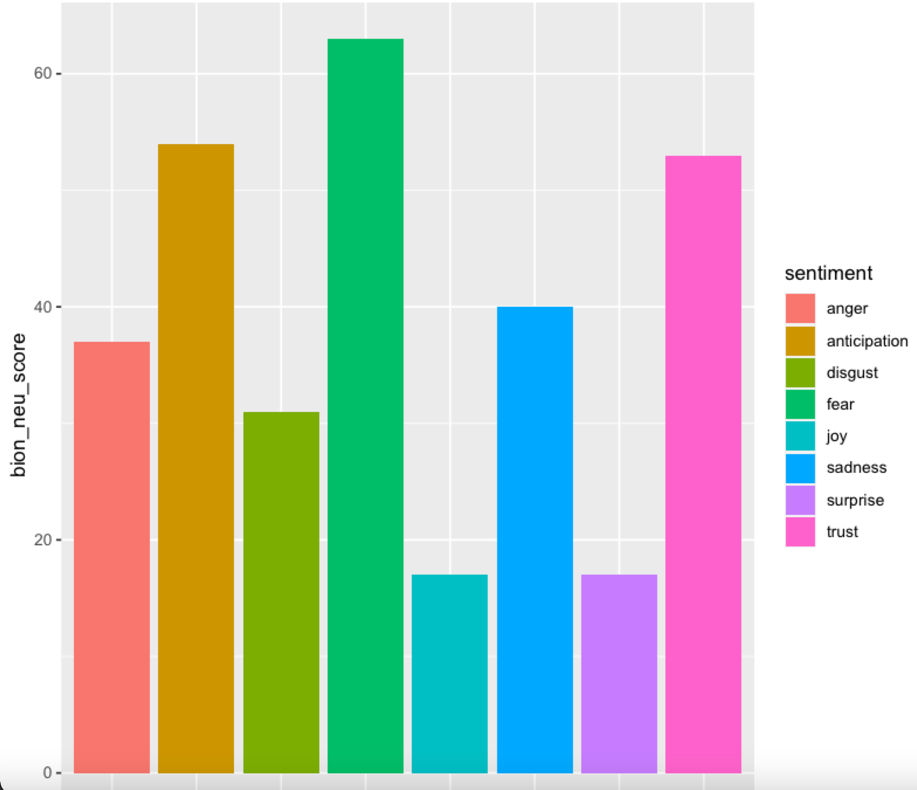


Figure 6 Plot 5

In order to have a deeper insight into each sentiment, the distribution of 8 sentiments from VADER are shown in plot 3. We can see, biontech positive sentiments consists mainly of “trust”, “anticipation” and “joy”, others have noticeable lower scores. While for negative tweets from biontech, the scores for “anger”, “disgust”, “fear” and “sadness” are much higher than other sentiments, as shown in plot 4. For neutral tweets from biontech, the total score of “anger” and “fear” seems a bit higher than the total score from “trust” and “joy”, but not much if we pay attention to the value of y axis in plot 5.

The situation for astrazeneca is very similar: for positive sentiment, “joy” and “trust” have the dominance, while for negative tweets “anger” and “fear” have the most scores. The scores are quite balanced distributed for neutral sentiment as shown in the following plots.

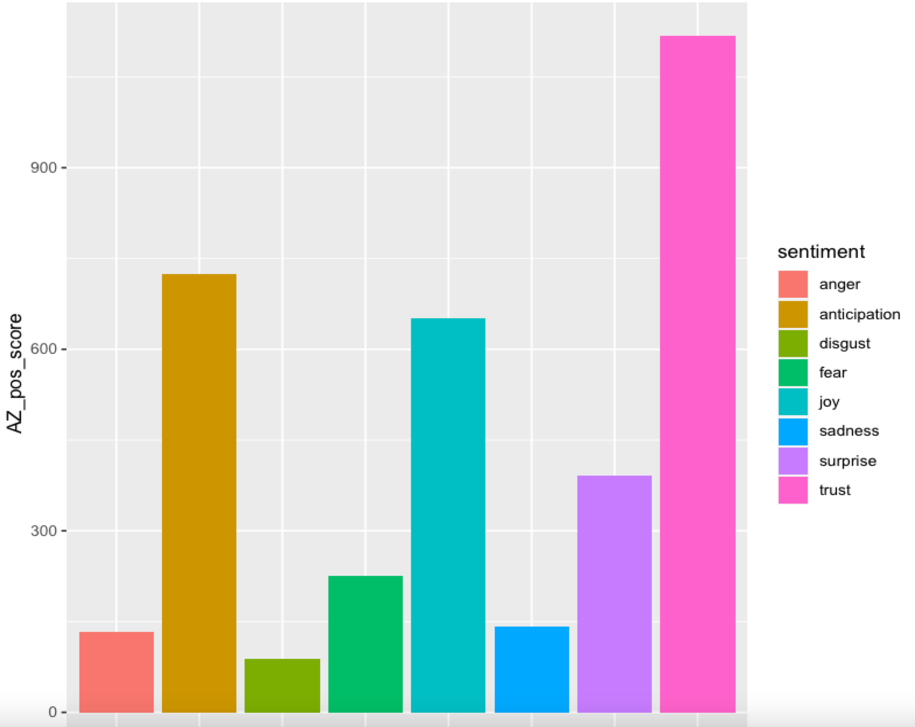


Figure 7 Plot 6

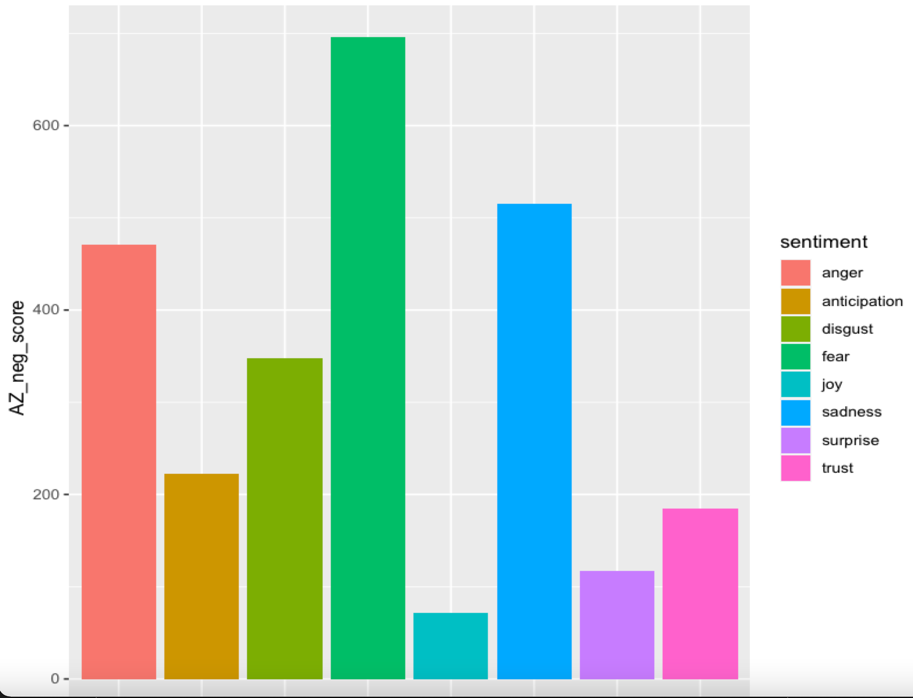


Figure 8 Plot 7

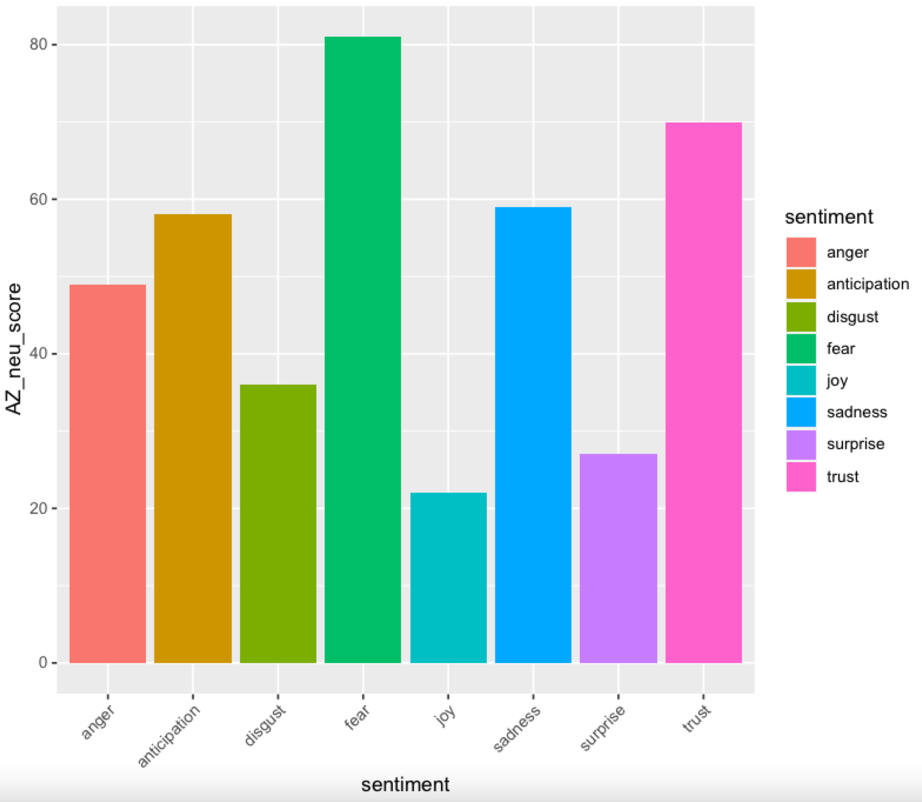


Figure 9 Plot 8

## Wordcloud

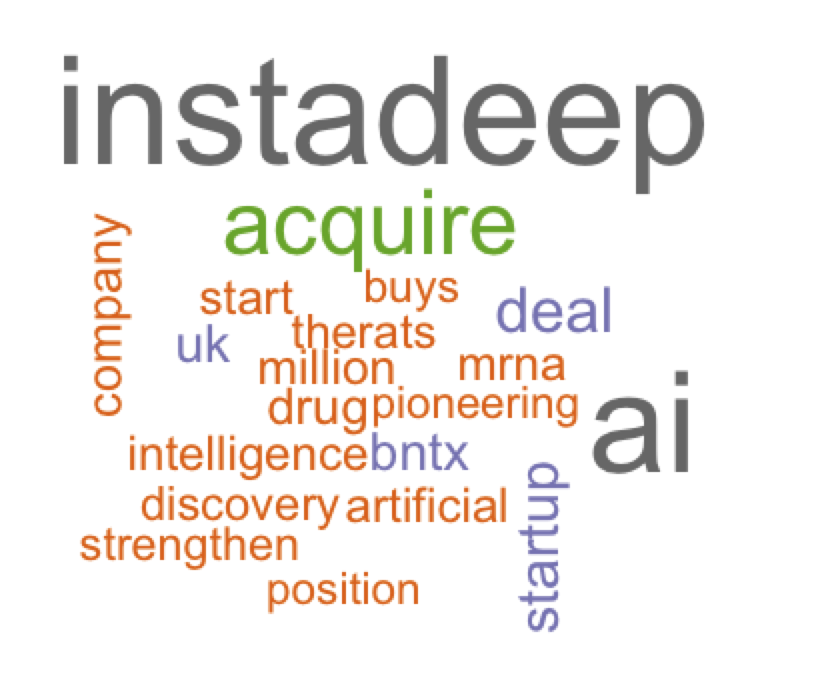


Figure 10 Plot 9 Wordcloud for positive biontech tweets

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Figure 11 Plot 10 Wordcloud for negative biontech tweets



Figure 12 Plot 11 Wordcloud for neutral biontech tweets

Wordcloud is based on an analysis of unique tweets about positive, negative and neutral sentiment respectively to visually show readers the corresponding important key words according to the frequency of their occurrences in the texts. In this work, after data cleaning, tokenization and removing of meaningless words, we can conclude from the resulting wordclouds shown as above that for positive tweets of biontech there were more things about new technical (AI) startup-instadeep and its potential contributes to the pharma company biontech talked about, in the meanwhile, we can see that there were more doubts and worries about mrna and its original purpose from the negative tweets of biontech. For neutral tweets, it seems like there were just some objective statements about some facts.

The wordcloud which resulted from the positive tweets about astrazeneca tells the acquisition of the us biopharmaceutical company by astrazeneca and its investments on some research. From the wordcloud of negative tweets about astrazeneca, besides the acquisition of new company, we can also see some words like “vaccinedeaths” and “myocarditis” are mentioned quite often. The neutral wordcloud seems like a combination of both negative and positive tweets.

Through wordclouds readers can not only have an overview about the 20 most mentioned words for each sentiment, they also have confirmed the correctness of data classification according to VADER scores.

## Network

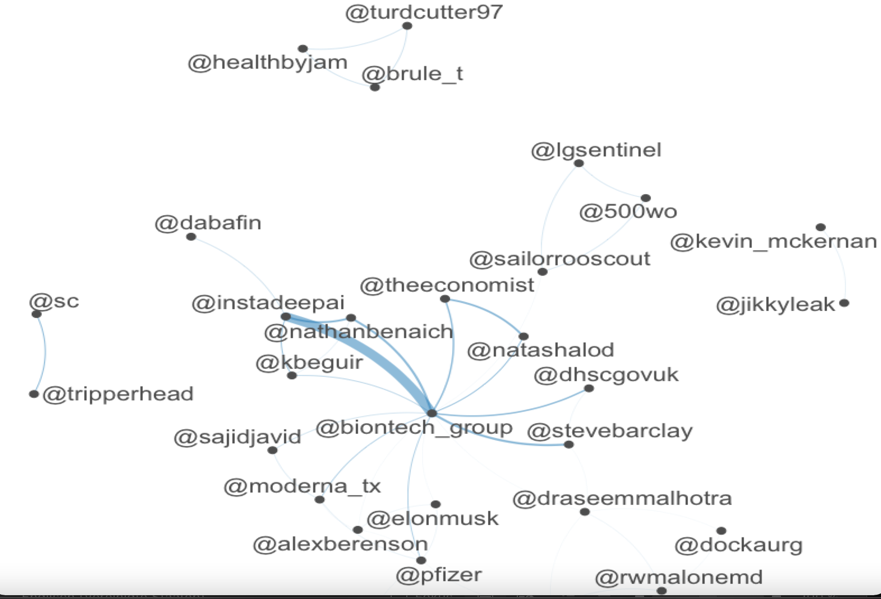


Figure 13 Plot 12 Network of positive biontech tweets

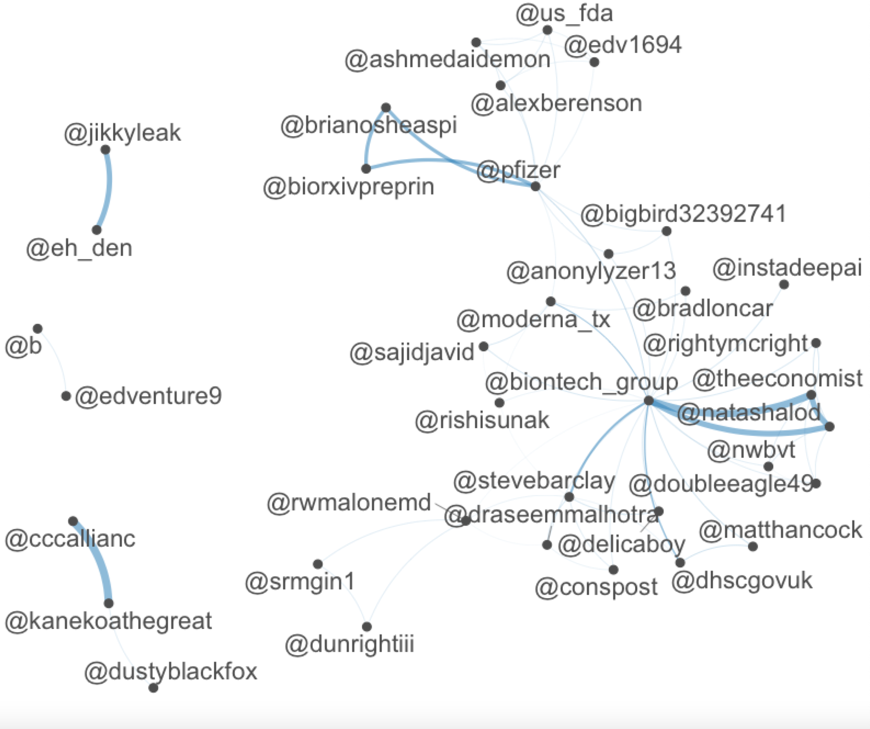


Figure 14 Plot 13 Network of negative biontech tweets

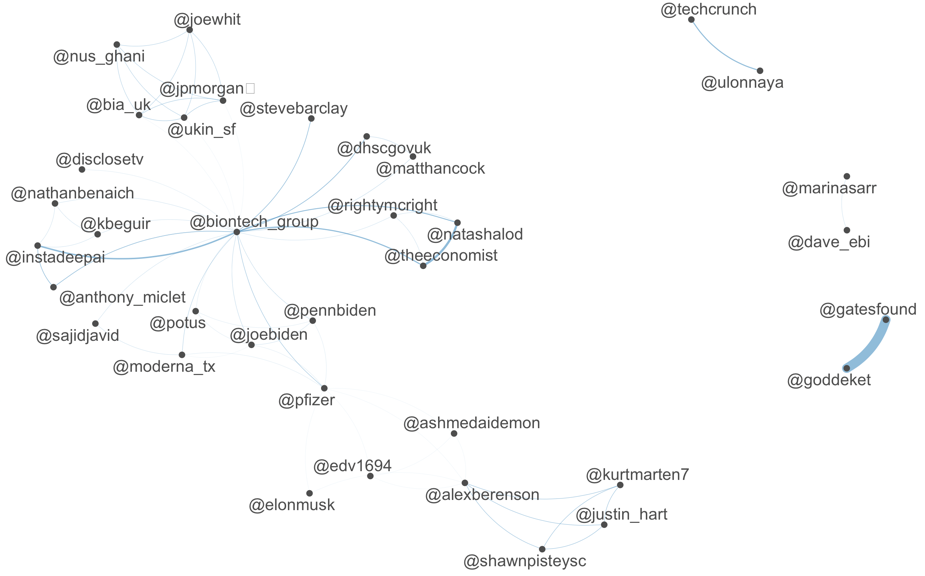


Figure 15 Plot 14 Network of neutral biontech tweets

through the resulted network plots of different sentiments shown above, we can read that for positive messages about biontech there was mainly one big network group consisting of pharma company biontech, the AI company instadeep, the influential media newspaper-the economist and some other private accounts. Among them all, the connection between pharma company biontech and AI company instadeep is strongest, which corresponds to the fact shown by the positive wordcloud. Besides, there were also other tiny groups with two to three members within them that communicated with each other about positive tweets of biontech. For negative biontech tweets, the situation is similar, there was one big group with diverse components, but the members with most information exchanges were between pharma company biontech, a private account @natashalod and the journal-the economist, this phenomenon meets the analysis from corresponding wordcloud. In contrast to positive tweets, the connections within other small groups are significantly stronger. For neutral tweets of biontech, even though there was a big network group that includes almost all possible kinds of users, the strongest connection didn’t happen between them. The exchange of neutral information appeared between a non-operating private foundation and a private account at most instead, which is not difficult to see through if we take the explanation from wordcloud as background into account.

Unlike positive tweets of biontech, the ones about astrazeneca in network plot were talked primarily by two groups: one is with less members but the communication between four of all members was much more frequent than the rest. In the other group, the range of connections between members was much broader crossing pharma companies, professional persons, government, organizations and private accounts, meaning there were more people from different backgrounds talking about some positive context about astrazeneca. As for negative tweets of astrazeneca, there was a clear classification of four groups and the most frequent connections were within the same group between pharma companies, government, some organizations, some private accounts and political persons like Boris Johnson or Elon musk. For the neutral tweets, there was a complex big network with three main communicators: the pharma company astrazeneca, the account from UK government and a private user. Generally, it’s to notice that the networks from astrazeneca are more complicated than the ones from biontech, which indicates that the theme about astrazeneca could be more controversial.

## Topic modelling

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Figure 16 Topics for biontech

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Figure 17 Topics for AstraZeneca

It’s not apparent to see what each topic is about right away from the results shown above. So, after some web-searching by typing in those key words, the context has become clear for each topic immediately: the first topic for biontech with those seven key words was about the collaboration of BioNTech and AI company instadeep and the so called self-amplifying mrna technology. The second topic can be best described through the title of a news- “cancer vaccine trials could start in the autumn-UK signs deal with BioNTech”[[7]](#footnote-7). The third topic seems like to be some discussions between media and BioNTech about the mrna technology. The fourth topic is about the CEO of BioNTech has purchased British AI startup Instadeep for millions of Pounds. The fifth topic is about BioNTech’s pioneering position in the field of the AI-powered Drug discovery. The last topic seems overlaps with the fourth topic, but with different emphasis.

While for astrazeneca, even though the number of topics was set to six, according to the result represented from below, we can conclude it as four topics: the first one is about the content of a news- “Astrazeneca boots heart, kidney business with $1.8 bln Cincor deal”[[8]](#footnote-8) and the second one is likely to be some relationship between AstraZeneca, mrna and the UK government. The third topic is quite easy to see, which is the acquisition of the Pharma Cincor. The last topic is something about the side effects of astrazeneca, for example, blood clot.

We also noticed that even though the topics for astrazeneca seem to be less than the ones of BioNTech, but according to the analysis from network, its complexity of the connections is higher than BioNTech, which can be due to the degree of negativity and we can see this trend also here from the topics of astrazeneca.

## Time series

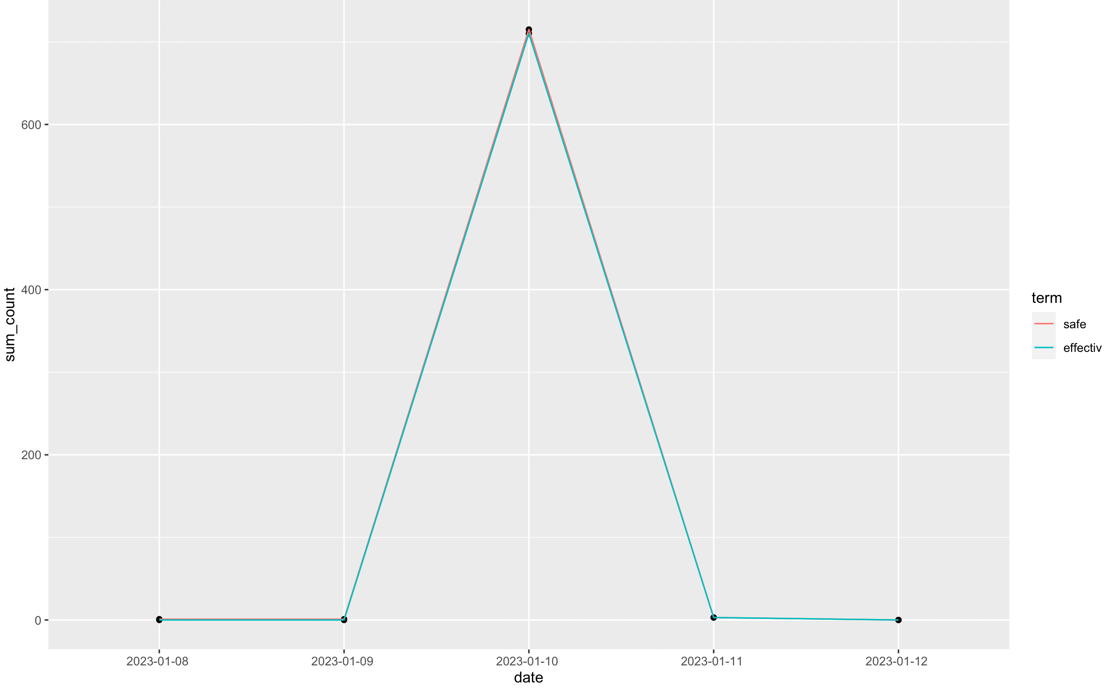


Figure 18 TS for positive biontech

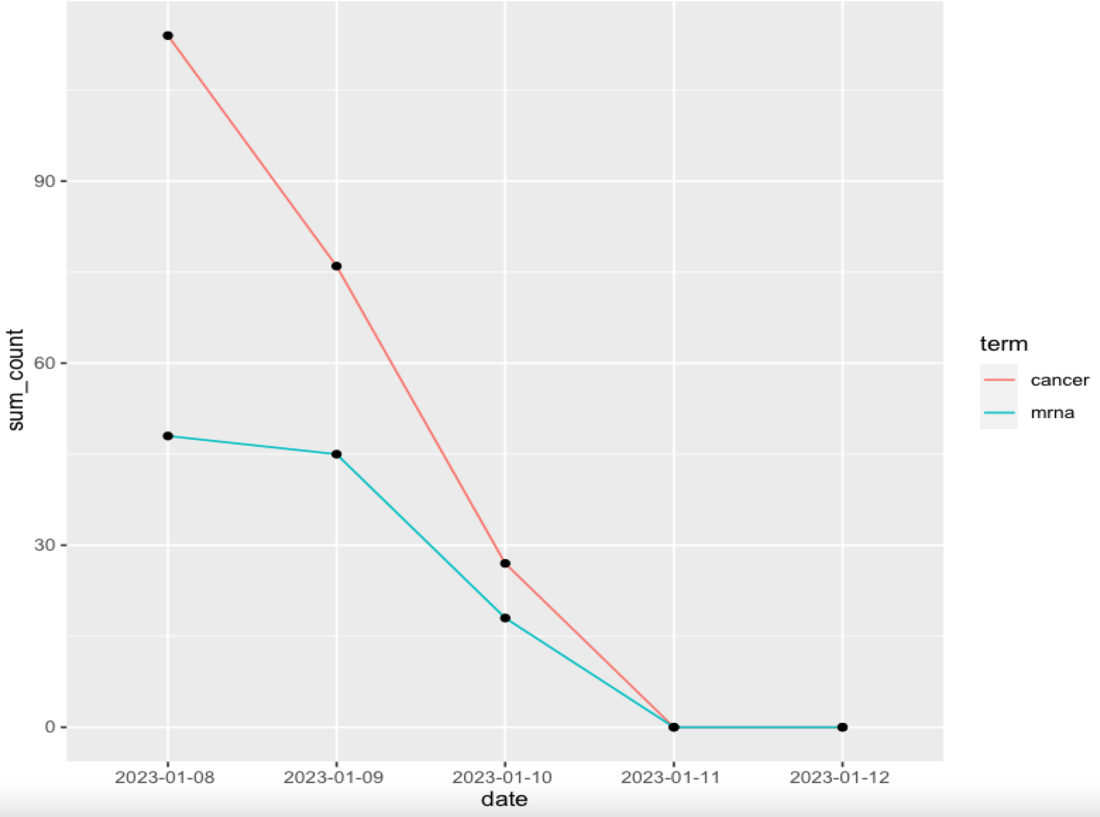


Figure 19 TS for negative biontech

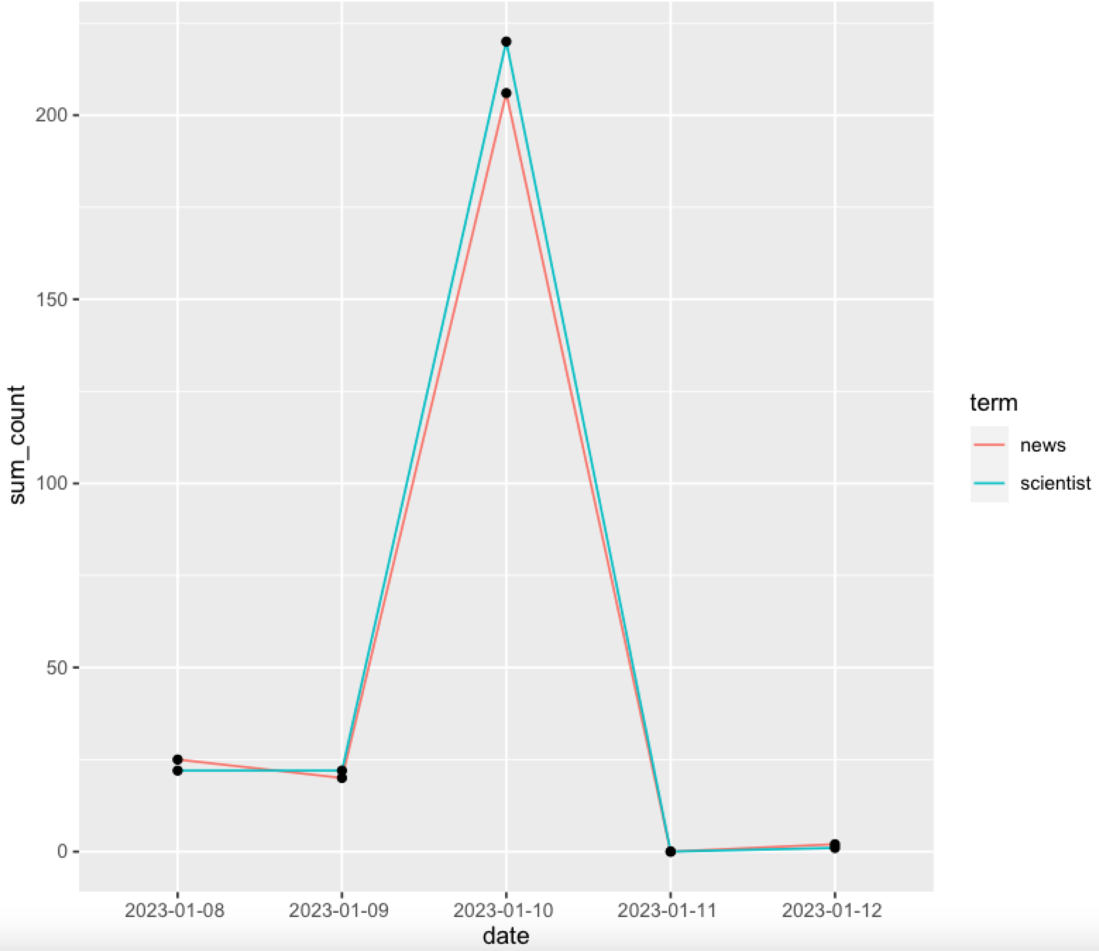


Figure 20 TS for neutral biontech

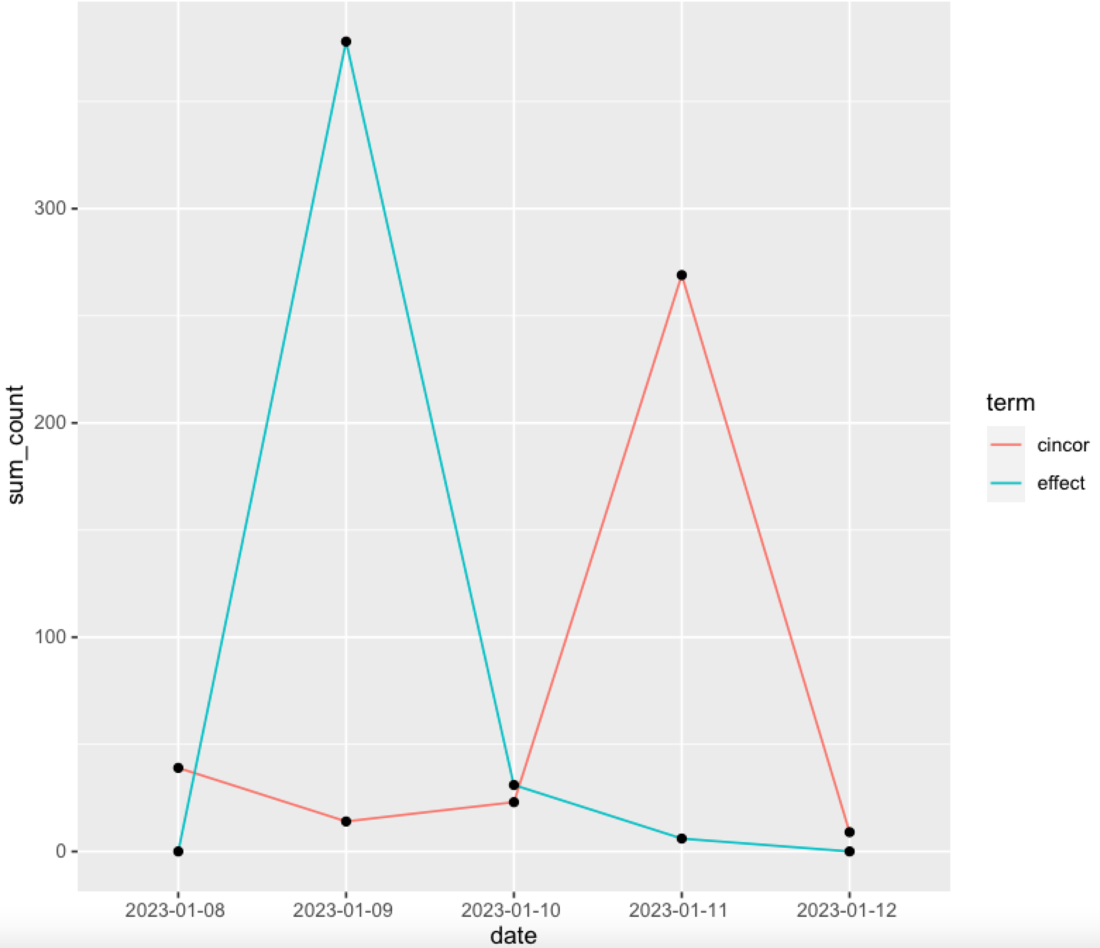


Figure 21 TS for positive AZ

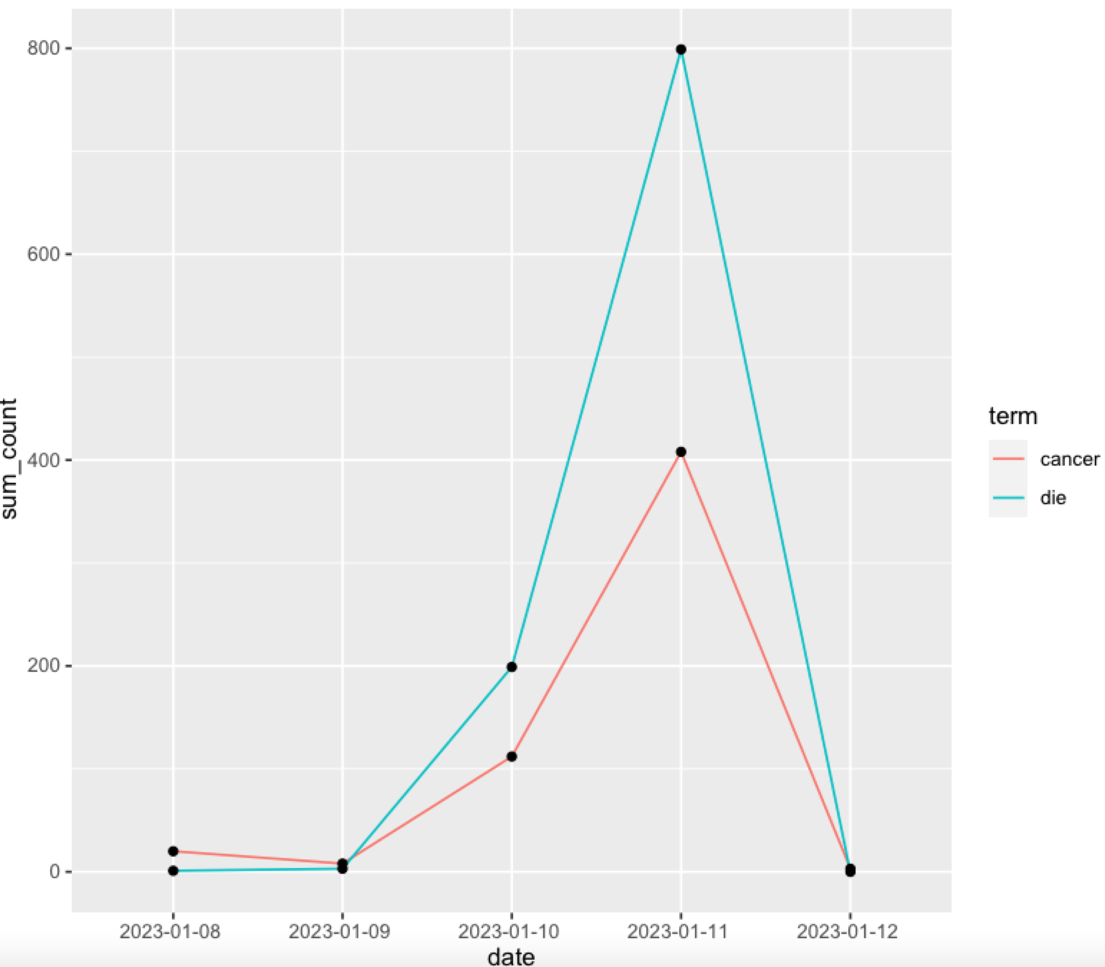


Figure 22 TS for negative AZ

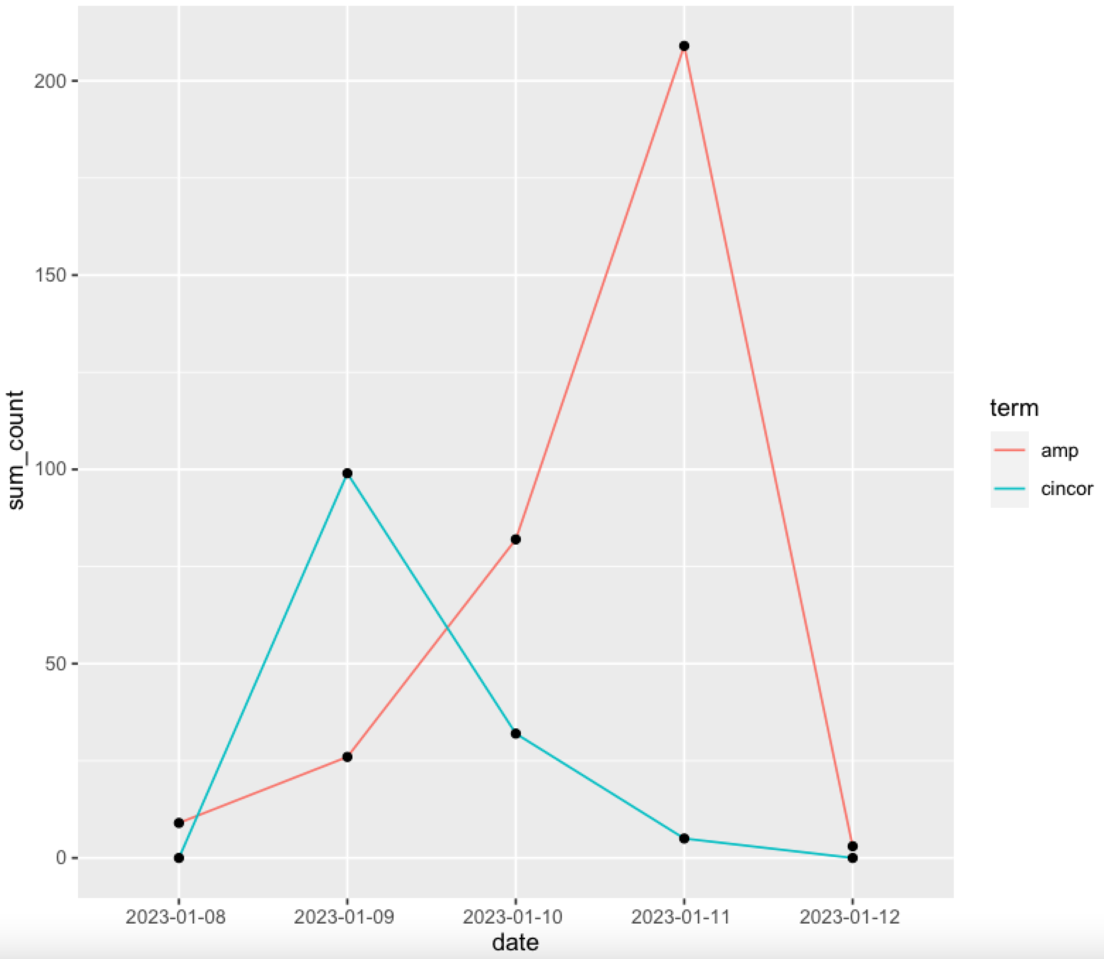


Figure 23 TS for neutral AZ

Through the time series represented above, we can infer that: for positive biontech tweets, the words “safe” and “effective” have almost the same trend, on the date of 10-01-2023 both of them were mentioned at most for more than 700 times, on the rest of the dates except 11-01-2023, nobody has used either of them and for 11-01-2023, there were also only very few people who mentioned those words. For negative biontech tweets, there was an apparent comparison between “cancer” and “mrna”. First if all, both of were no more mentioned after 11-01-2023, secondly, from 08-01-2023 till 11-01-2023, “cancer” appeared more frequently than “mrna” through the time, the last point is that both of them have a descending trend with the highest count on 08-01-2023 for “cancer” over 100 and for “mrna” over 45. For neutral biontech tweets, even though the words “news” and “scientist” have similar trends again, there were some differences at some time points: for “scientist” the count is quite stable between 08-01-2023 and 09-01-2023, while for “news”, on 08-01-2023 it was mentioned a little more than “scientist”, but then am 09-01-2023, it was below “scientist”. On 10-01-2023 both of these words have reached their highest point for more than 200 and it’s easy to see that on that day “scientist” was again higher than “news”, then from 11-01-2023, the count dropped again to almost 0. If we compare the three plots of biontech together, we can notice that the count for positive tweets were highest, then followed the neutral tweets, the lowest counts go to tweets with negative sentiment.

For positive AZ tweets, the highest point for “effect” is on 09-01-2023 with more than 350 counts, on rest of the dates the count was under 50 all the time, for “cincor” on the other hand, the highest point appeared on 11-01-2023 with more than 250 counts. For rest of the dates, the count was also under 50, but never 0. For negative AZ tweets from 09-01-2323 till 12-01-2023, “die” had always the higher count than “cancer” and the highest point reached 800 for “die” on 11-01-2023 while for “cancer” only the half on that day. Then for both words the count dropped to 0 on 12-01-2023. For neutral AZ tweets, word “cincor” was mostly mentioned n 09-01-2023, then it gradually falls to 0 on 12-01-2023. Meanwhile for “amp”, it started with about 10 on 08-01-2023, then climbed over 200 on 11-01-2023 rapidly, but on 12-01-2023 with 0 mention again. For AZ tweets, the negative information was dominant overall, then followed positive tweets with most frequently mentioned word has only the half count of that from negative tweets.

# Discussion and conclusions

This work tried to classify tweets from Twitter according to their sentiment scores and do various types of analyses based on that, hoping to provide different potential target groups more than enough intra and inter classificational analyses. Inspired from this work, there could be some other analyses using similar structure in the future if there’s the need.

Despite the efforts that have been done to achieve the succuss for this work, there are still plenty of pitfalls that were unavoidable during analyzing:

1. Because of the limited authorization for data acquisition and compute budget, there were only about 5000 tweets pulled from Twitter. If the amount is suitable for an analysis aiming at the whole countries using biontech or astrazeneca remains to be discussed.
2. There could actually have more broad analyses using geographical or other interesting attributes, but because of authorization problem and also because the data at hand accidently has less information, many deeper analyses were not feasible.
3. The analysis for this work is based on self-defined definition for positive/negative sentiment, which is not objective enough and can vary according to different researchers.
4. For wordcloud and topic modelling, unigram was chosen as the basic of analysis, whether bigram or others is better is worth discussing. For example, we can see from topic modelling that “bill” and “gates” were presented separately even though they refer to the same person.
5. As already mentioned earlier in this work, the number of topics for topic modelling is decided based on manual trials and may be not accurate enough.
6. For time series, there were only data for the last seven days available, the results resulted from those datasets were therefore not as meaningful as expected. But as long as the access to more allowance exists, the trust of this time series analysis can be expanded.

from the results

References

[1] Piedrahita-Valdés H, Piedrahita-Castillo D, Bermejo-Higuera J, Guillem-Saiz P, Bermejo-Higuera JR, Guillem-Saiz J, et al. Vaccine hesitancy on social media: sentiment analysis from June 2011 to April 2019. Vaccines 2021];9(1):28.

[2] Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res 2016];18(8):e232.

[3] Nuzhath T, Tasnim S, Sanjwal RK, Trisha NF, Rahman M, Mahmud SF, et al. COVID-19 vaccination hesitancy, misinformation and conspiracy theories on social media: A content analysis of Twitter data. 2020 doi:1031235/osfio/vc9jb.

[4] Kang GJ, Ewing-Nelson SR, Mackey L, Schlitt JT, Marathe A, Abbas KM, et al. Semantic network analysis of vaccine sentiment in online social media. Vaccine 2017];35 (29):3621–38.

[5] Tangherlini TR, Roychowdhury V, Glenn B, Crespi CM, Bandari R, Wadia A, et al. “Mommy blogs” and the vaccination exemption narrative: results from a machine-learning approach for story aggregation on parenting social media sites. JMIR Public Health Surveill 2016];2(2):e166.

[6] Bonnevie E, Gallegos-Jeffrey A, Goldbarg J, Byrd B, Smyser J. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. J Commun Healthc 2020a];1–8.

1. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Quart 2014];92(1):7–33. [↑](#footnote-ref-1)
2. https://link.springer.com/article/10.1007/s12652-022-03805-0 [↑](#footnote-ref-2)
3. Rajaraman, A.; Ullman, J. D. (2011).[*"Data Mining"*](http://i.stanford.edu/~ullman/mmds/ch1.pdf)*(PDF)*.Mining of Massive Datasets. pp. 1–17. [↑](#footnote-ref-3)
4. https://en.wikipedia.org/wiki/Document-term\_matrix [↑](#footnote-ref-4)
5. https://en.wikipedia.org/wiki/Latent\_Dirichlet\_allocation [↑](#footnote-ref-5)
6. <https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation>, Generative process [↑](#footnote-ref-6)
7. https://www.gavi.org/vaccineswork/cancer-vaccine-trials-could-start-autumn-uk-signs-deal-biontech [↑](#footnote-ref-7)
8. https://www.reuters.com/business/healthcare-pharmaceuticals/astrazeneca-buy-us-based-cincor-pharma-18-bln-deal-2023-01-09/ [↑](#footnote-ref-8)