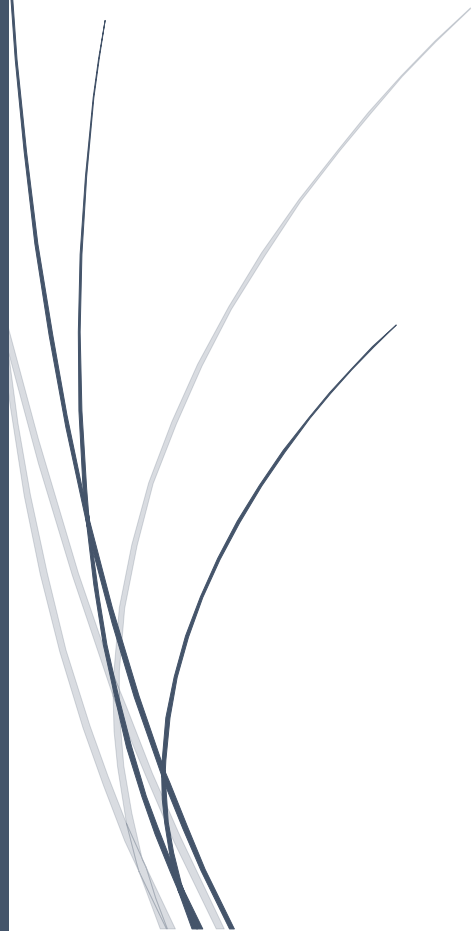
6-7-2018

The impact of #MeToo; a data visualization.

Bachelor Thesis

Xadya van Bruxvoort

Supervisors
Dr. Ir. M. van Keulen
Dr. A. Kamilaris



CREATIVE TECHNOLOGY 2018

UNIVERSITY OF TWENTE.

Abstract

#MeToo is a movement which started in 2006 and was further developed in 2017. This movement aims to raise awareness for the magnitude of sexual harassment. This movement was mainly present on Twitter last year, but has been covered on the news as well. Via tweets, users would state that sexual harassment had happened to them, by the use of a hashtag: #MeToo. During this project, there has been looked at how all these tweets and all this information could be summarized in a data visualization. Research has been conducted regarding data visualizations in general, the tools available for visualizing and on similar projects. By exploring similar projects, a survey was conducted of which the results were turned into ideas and designs. These designs were then made into visualizations using Microsoft Power BI. The first visualizations included an introduction about the topic, where the two most important persons, Alyssa Milano and Tarana Burke, were introduced. The topic was then further introduced by important news articles. After this, focus was put on sentiment of tweets and jokes being made about #MeToo. Then the magnitude of #MeToo and sexual harassment in general were explored via visualizations and finally, credits were given to authors of this project. A visualization of #MeToo has been created which guides users through the story of the movement and which included some shocking details. During testing, users became very quiet when these details were displayed, and users started discussing #MeToo after the test, which indicated that impact has been made. Overall users were very positive about it and only found some minor flaws that could easily be resolved. This project could be turned into a framework as long as a new topic complies with the following requirements: it should be a sensitive and underestimated topic, a topic that is happening on social media, and a topic for which news articles are available. Next to that, there should be looked at the story and other ethical implications for different topics. To gain better insight into #MeToo, further research has to be done to see what tweets and news articles have happened since early 2018. There should also be looked into similar hashtags or words, such as #TimesUp and 'me too'. This end product should be turned into a scrollable webpage, to make the product easier to use and lastly, there should be looked at the sentiment analysis. Now there were tweets which were positive regarding the movement, but negative in general which were marked as negative. Question is whether such a tweet is classified correctly or should actually be marked as positive, so another API should be used for this. In general, the end result led to positive reactions and open discussions, which indicates the main goals of the visualizations were achieved.

Index

Chapter 1 - Introduction

The main focus of this bachelor thesis is making a data visualization to make #MeToo more insightful for the target group. This will be done by analysing tweets of the second half of 2018 that include #MeToo and making visualizations of the results.

1.1 Background

Twelve years ago, Tarana Burke introduced the “Me Too” movement to raise awareness on sexual harassment in society. When Tarana Burke started the movement, it was because she was talking to a girl who was sexually abused by her mother’s boyfriend. Burke couldn’t take it anymore and sent the girl away, to another counsellor. After this she thought that she had to do something about it, so no girl had to experience what she had. The only thing to say that came to mind was ‘Me too.’. She felt some kind of shame about this situation and this made that she wanted to stand up, to make victim realize that they are not the only one.

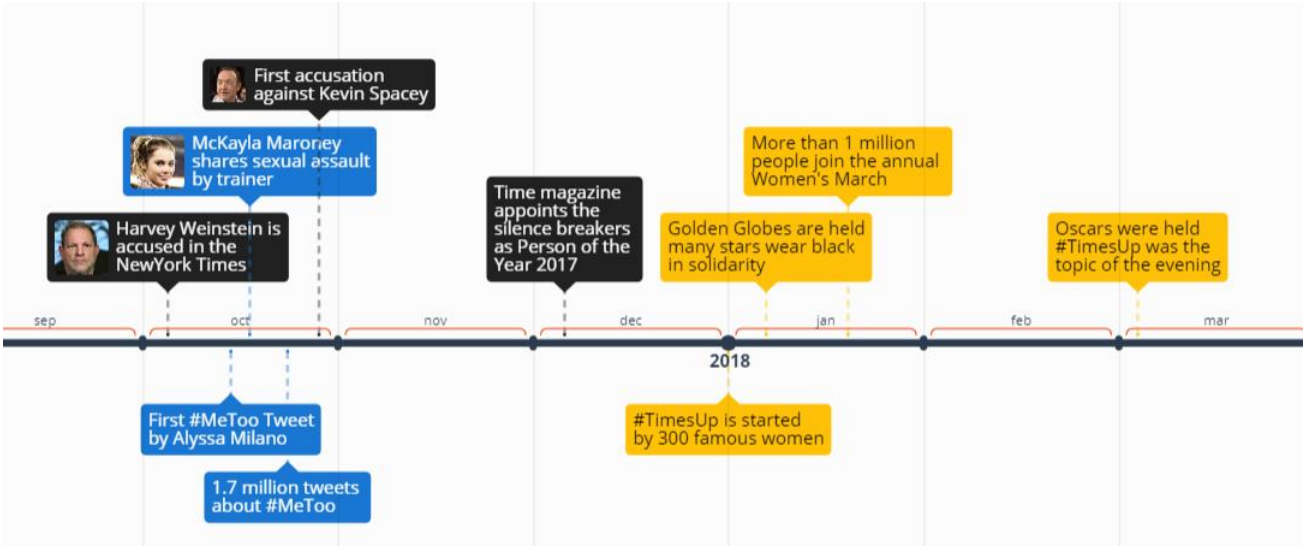


Figure 1 - An overview of the Me Too movement.

In October 2017, the movement gained a lot of attention. An overview of some big events around #MeToo can be found in figure 1. On October 5th Ashley Judd accused Harvey Weinstein of sexual assault. After this, many more celebrities came forward with similar stories to hers. On October 15th, Alyssa Milano tweeted: ‘Me too. Suggested by a friend: “If all the women who have been sexually assaulted wrote ‘Me too.’ As a status, we might give people a sense of the magnitude of the problem.’ The days after, many women and men came forward with their stories, for example McKayla Maroney and Anthony Rapp. With a record of 1.7 million tweets on October 24th, Me Too was a big happening in 2017. On the first of January 2018, 300 women who work in film, television and theatre started Time’s Up (2017); ‘a unified call for change from advertising leaders for women everywhere’. During the Golden Globes, many of these women brought an activist with them to the red carpet and many stars wore black in solidarity. The path continued with the annual Women’s March. More than 1 million people joined worldwide with slogans like: ‘Girls just want to have fun-damental rights’ and ‘I am an Object.’. This happening continued during the Oscars, where Time’s Up was the topic of the evening. Many people wore pins that said ‘Time’s Up’ and there were various speeches given about this topic.

What can be seen from this is that Me Too was a big happening last year and early 2018 as well. It was started as a call to show the world the magnitude of sexual harassment happening within society. After

this, Time's Up was founded by powerful women wanted no more waiting and no more silence, but mainly no more tolerance for discrimination, harassment or abuse.

1.2 Problem

The problem can be stated as the magnitude of the problem - sexual harassment - not being clear to society, as was also mentioned by Alyssa Milano. Aside from this, there are many more problems which Me Too wants to tackle. Many of these can be found at the Time's Up site (2017), which states sources of the facts as well. These problems are, for example, that one in three women between 18 and 34 have been sexually harassed at work and 71% of them did not report it. To focus on the main problem, when Burke started the movement, it was mainly to help others and to show them that they are not alone. Twelve years later the hashtag was initiated because of exactly the same reason. The reason Time's Up was started was also to draw attention to the existence of the problem, so this really seems the main problem at the moment.

What is also a problem at the moment is the fact that not many projects have focused on #MeToo and not many technical solutions are offered to tackle this problem. Because of this, people do not know the extent of sexual harassment.

1.3 Research questions

The main, societal, research question for this bachelor can then be defined as follows:

How can a data visualization help in making the magnitude of sexual harassment in the world clearer to people in the Netherlands?

By researching and answering this question, a data visualization will be developed to help people understand the magnitude of the problem. The target group are people between 18 and 35, which will be explained in chapter 2.8 and 3.5 later on in the report.

The technical research question that can be asked is defined as:

How can a framework be designed to help in tackling similar problems by means of a data visualization?

1.4 Challenges

There are a few challenges within the project. One of them being that the topic is a rather sensitive topic and can easily hurt or anger people. Therefore, a very sensitive approach should be used. Another challenge is the lack of data visualization tools available for big data analysis, so there has to be looked into a solution for this. The last problem concerns the privacy of data subjects since the data will be gathered from Twitter. Many strong opinions on this movement are being held and the main places to state these are often via social media, because of its accessibility and anonymity. There should also be taken a look into that.

1.5 Language

The report and main research will be in English. However, when talking to subjects when performing interviews or holding user tests, the language that is most convenient for both parties will be used, in order to make people feel most at ease. This will then be Dutch in case of the subject being Dutch, German when the subject is German and English if the subject is non-Dutch and non-German.

1.6 Goal

The goal of the project is to make people more aware of the magnitude of sexual harassment in the world by using a data visualization. A side goal of this is to make people think from a different perspective about this

topic. The focus will lie on the use of a static or interactive dataset, due to time restrictions. Research will be done on data visualization, other Me Too projects, project languages, data gathering and privacy. In the end, user input will be used to evaluate the project and to find some improvements for future work.

Another side goal is to offer a technical solution for this problem and similar problems about difficult topics by the means of researching how to convey this sensitive information in an efficient way.

1.7 Outline of the report

This thesis consists of 5 chapters. In chapter 2, there will be delved into visualizations in general, different tools available for visualizing and there will be looked at related work regarding similar projects. There will also be looked at the data that will be used and requirements for the project. Chapter 3 discusses visualizations about #MeToo that were found online and in this chapter a survey will be conducted to gather more information. This chapter concludes in a brainstorm on both design and ethical aspects. Chapter 4 described the realization of this project. In the beginning of this chapter, design will be made which will be turned into visualizations. In the meantime, the ethical side will be stressed and explored more. At the very end of this chapter, the result will be tested. Chapter 5 concludes this project. First it discusses both research questions and how to further improve this project and it end with acknowledgements.

Chapter 2 - Background Research

2.2 Data Visualizations

2.2.1 Introduction

Data is everywhere around us and businesses are getting increasingly dependent on these data. Whether they are the groceries one buys, the 'like' one gives to their friend or the bike counter that counts one passing by; data is all over the place. Data can be seen as big data when a dataset has no clear borders or even becomes infinite. These data can be presented in various data formats, most of them are not structural data flows (Gorodov & Gubarev, 2013).

In order to transfer data, often a data visualization is made, because graphical thinking is a very natural way of thinking for any human being. Since the phenomenon of big data is relatively new it is quite hard to represent and process big data in the same way as is done with smaller datasets (Sun, 2017).

The goal of this paper is therefore to find out what the best way of representing a big data set is. There are subquestions for finding the best colour and the best type of graph, in order to find out what attracts people most and what makes them want to look at a data visualization. After this, the question is what the differences are between a dynamic data and a static data visualization. This is because dynamic data keeps changing over time and might be more interesting to look at during a specific period of time compared to other times. This literature review ends with an elaboration of points for future research in data visualization.

2.2.2 Choosing a data visualization

colours

There are several recommendations when looking at the best colours for a data visualization, depending on the data and the message one wants to convey. The first recommendation about the number of colours is given by both Healey (1998) and Wang, Giesen, McDonnell, Zolliker and Mueller (2008) who point out that it is important to only use up to a maximum of seven colours, of which all must be linearly separable from their two neighbours, when deciding upon colours for a data visualization. Healey (1998) states that this distance should be above a certain threshold and Lujin Wang et al. (2008) add to this that these colours should differ in hue, saturation and brightness. How much they should differ or what the threshold is, is not further explained in these papers. Another recommendation about colour palettes is pointed out by Setlur and Stone (2016) who argue that automatically generated palettes are a good starting point for making data visualizations. However, a data visualization is faster to understand when using semantically meaningful concept-colour associations for data categories with a strong colour association according to Kelleher and Wagener (2011) and Setlur and Stone (2016). Another recommendation is made by Lujin Wang et al. (2008) and indicates that warm colours excite emotion, while cold colours create openness and distance.

The last recommendation is about the usage of light and striking colours is identified by Silva, Sousa Santos and Madeira (2011) and Lujin Wang et al. (2008) who both argue that there should be thought of what the main goal of a data visualization is and what should attract attention from the subject; this should be the most striking colour. Kelleher and Wagener (2011) suggest that the light colours should represent small values and the dark colours the large values when having a data visualization about quantitative data. However, they also state that the light colour should be used for the average and there should be two contrasting colours for the extremes if a data visualization is about averages.

There is not one answer to the question what colour is best to use for a data visualization, but there are some guidelines to improve visual design by using colours. Unfortunately, not much research has been done about big data visualization yet. Concluding from these articles it seems like these guidelines can still

hold for big data visualization, although this is not sure. Further research would be necessary to find out whether there are differences in the colour use when applying it to a big data set.

graphical representations

There are several guidelines for choosing the type of data visualization, depending on the type of data and what the message is one wants to convey. Spence and Lewandowsky (1991) found out that data visualization is not only a nice way to represent data in a clear and easily readable manner, but it actually makes the processing time for subjects shorter than having a table. Data can be represented by the use of many different graphs. The type of graph chosen for a certain representation is dependent on the purpose of the representation. The first guideline, as mentioned by Kozak (2010) and Strange (ad cited by (Kelleher & Wagener, 2011), is that line charts are a good representation of data when the dataset is meaningful over time, whereas bar charts are good for showing absolute values and comparing these with each other. Besides, pie charts, whether it is good to use them at all or not, are a good representation for comparing percentages with a whole or with each other. Spence and Lewandowsky (1991) compared the bar chart and the pie chart and found out that a bar chart is a better fit for easy tasks (comparing a vs. b), while the pie chart is better for more complicated tasks. Kozak (2010) argues that the pie chart should be avoided altogether and Kelleher and Wagener (2011) add that this is especially the case for large datasets. A second guideline is given for overlapping data, due to large amounts of data that need visualization. Kozak (2010) shows that, if this is the case, it might be a good idea to make the symbols either transparent or open, so density can be seen in a proper way. The beforementioned author adds to this that data should be organized by size, rather than by label in all cases and Spence and Lewandowsky (1991) agree with the latter.

However, Otjacques (as cited by (Lidong Wang, Wang, & Alexander, 2015) indicates that, while a lot of research has been done about small data representation, little has been done about big (dynamic) data representation. It is a very difficult job to represent big data in an adequate manner as mentioned by Lidong Wang, Wang and Alexander (2015), because traditional data visualization tools are often inadequate to handle big data. This is mainly because of the scalability and dynamics of these data. Some options to more easily represent data include removing outliers or clustering data, so there are less different categories. Gorodov and Gubarev (2013) argue that there are also some different types of graphs available for big data representation which each have their advantages and disadvantages per certain type of data, such as a treemap, a sunburst and a streamgraph. A good solution to this is given by Danyel Fisher who held a presentation for the Chalmers University of Technology (2014) in which he stated that taking random samples of a data set and using these for a data visualization instead of the entire dataset might be a solution for the tools not being ready for big data. What was also found was that there are quite some non-scientific sites that offer insight in different type of graphs for different users and different data types, such as datavizcatalogue and Tableau.

Like in the previous subquestion, there is not one answer to the question. It depends on the type of data one wants to represent and also about the size and dynamics of the data. From the abovementioned articles, it seems easier to represent small data than big data, mainly because of the amount of research that has been done and the tools available. However, there are some workarounds for graphs and tools to make it possible to represent (parts of) big data.

Data types

There are a couple of differences between dynamic data and static data. Dynamic data and static data differ mostly in the amount of research that has been done about it, and thus affecting the difficulty of visualizing it, and the amount of user engagement with a certain visualization. The definition of both, as mentioned by

Anselin (1999), Lidong Wang et al. (2015) and Cottam, Lumsdaine and Weaver (2012) is that dynamic data are data that change over time, either by filters applied by the user (inter-active data) or because the data is time-varying, whereas static data does not change at all.

The first difference is given by Lidong Wang et al. (2015) who indicate that analyses should be performed real-time or on frequent intervals when data is time-varying, in contrast to static data visualization which can be made once and stay untouched forever. Vande Moere (2004) suggests that users are often not concerned about the values, but more about trends when having time-varying data. Another difference is stated by Cottam et al. (2012) who point out that time-varying data can become fragile too, since it can take away the identity of a certain data visualization over time. Lidong Wang et al. (2015) and Anselin (1999) point out a second great difference, which is that users should be engaged in a visualization, because interactive data visualizations can bring the user to great insights. Thus, it might be good to look into the option of having a dynamic data visualization. What is especially great about interactive data visualizations is the brushing and linking between visualization approaches. However, a third big difference, according to Segel and Heer (2010) and Anselin (1999), is that static data visualization is better studied than dynamic data visualization, which makes visualizing dynamic data a very difficult job because of the few guidelines available. Lidong Wang et al. (2015) add to this that this is especially the case when these data are considered big data.

Concluding, the biggest difference between static and dynamic data seems to be the amount of research that has been conducted. This makes representing dynamic, mainly time-varying, data very difficult. Where the time-varying data visualization has some downsides, the four articles used do not mention any downsides about interactive data visualization and mainly very much support the use of it in order to get user engagement and to get great insights that otherwise would not have occurred.

2.2.3 Conclusion

The idea behind this literature research was to find out what the best way of representing big data was. Despite this being a very broad question, quite some guidelines were found for choosing a type of graph and a colour (palette). In short, one has to take into account the type of data when choosing upon a type of graph and a colour (palette). Next to this, there were some additions about choosing an open or transparent symbol when having a big data set and taking random samples from a big data set to represent only a small (random) part. For colour was stated that warm colours excite different feelings than cold colours. In addition to this, adopting visually distinctive colours is important, however a certain harmony in colour palettes should still be present. Furthermore, the most striking colour should be used for the part that needs a highlight. Lastly, the lightest colour should either be used for smallest value or for the average. There are two main differences between static and dynamic data visualization, which are the amount research conducted and the amount of user participation. The papers reviewed were utmost positive about inter-active visualization, mostly because of the new insights. Altogether, when choosing for the type of data visualization, the answer is not plain, but some handles are given which can help with the decision.

There is definitely room for more research. Something that would be very interesting for the future would, for example, be using virtual reality and augmented reality for a data visualization. In this way a third dimension would become available and such another element can be added to a graph. This could be researched in an equally themed literature review or a research on its own. The literature research gives an overview of research conducted about visualization, but not much of it is specialized on big data. For this, more research should be done about big data visualization and what the differences in perception are compared to data visualization as we know it. This could be done by performing user tests, for example. Lastly, this literature review gives an overview of some guidelines, however these need to be applied to

certain data. Some extra research might be needed for the specific data, but this paper can be used as a start.

2.3 Data visualization tools

The three market leaders in the area of business analytics are Tableau, Microsoft and Qlik (Schlegel, Sallam, Yuen, & Tapadinhas, 2013), as can be seen in figure 2. These three tools will be discussed below, and a comparison will be made in a table afterwards. There are other tools available, also within programming languages, but there has been chosen to make it within one of these frameworks, because of its attractiveness and ease of use for future projects.

Figure 2 - The magic quadrant of Schlegel et al.(2013)

2.3.1 Tableau[1]

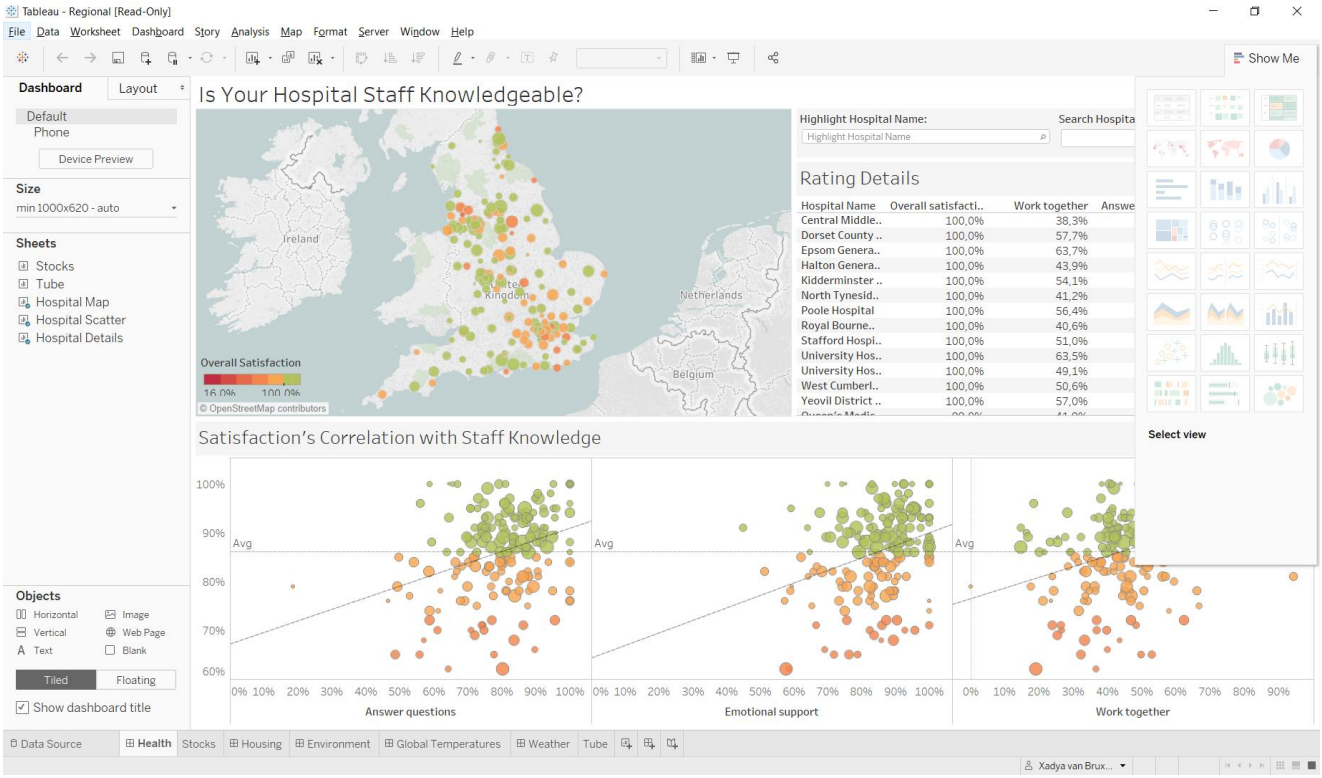


Figure 3 - Tableau

A small screenshot of Tableau can be seen in figure 3. This program is often seen as the Master of Data visualization tools. It had over 57.000 customer accounts in July 2017 and is one of the biggest tools. Tableau advertises themselves with being very good regardless of the type of data available. Advantages are that visualization made with Tableau can be published on a personal website, it is very intuitive to use, and it broadly used and becoming a BI standard. Tableau is especially good with maps and has also an option for word clouds (Savoska & Bocevska, 2016). A big disadvantage is that Tableau is not as distinctive anymore and it has a lack of complex data support (Schlegel et al., 2013). Tableau has some great colour palettes available that were tested by Setlur and Stone (2016) and proved to be at least a good starting point. Tableau comes with 1 GB storage in the free version, but can be upgraded (Ali, Gupta, Nayak, & Lenka, 2016).

> ***BI** stand for Business Intelligence. Refers to technologies, applications and practices for the collection, integration, analysis, and presentation of business information.*

[1] https://www.tableau.com/

2.3.2 Qlik[2]

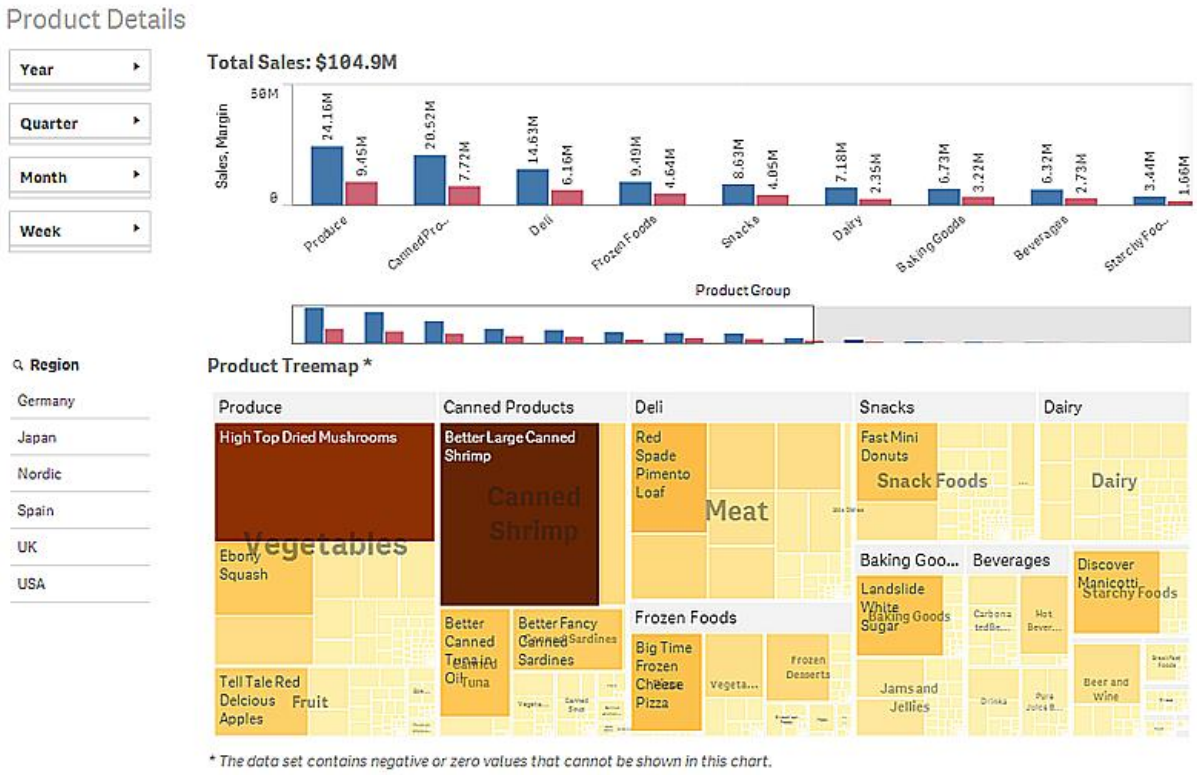


Figure 4 - Qlik

Figure 4 shows a screenshot of Qlik. This platform can be seen as the biggest competitor of Tableau with 40.000 customer accounts in 2018. Qlik has about 13 graphs available to choose from, which is significantly less than Tableau. Qlik's dashboard seems a bit less 'cluttered'. To make a map chart, an extra package should be bought. Qlik's advantages are its large partner network, the many trainings online available and its differentiation by its appealing dashboard and its ease of use. Disadvantages are its narrow use case; it is mostly used for parameterized reports and dashboards, and its technical support lags.

[2] https://www.qlik.com/us

2.3.3 Microsoft Power BI[3]

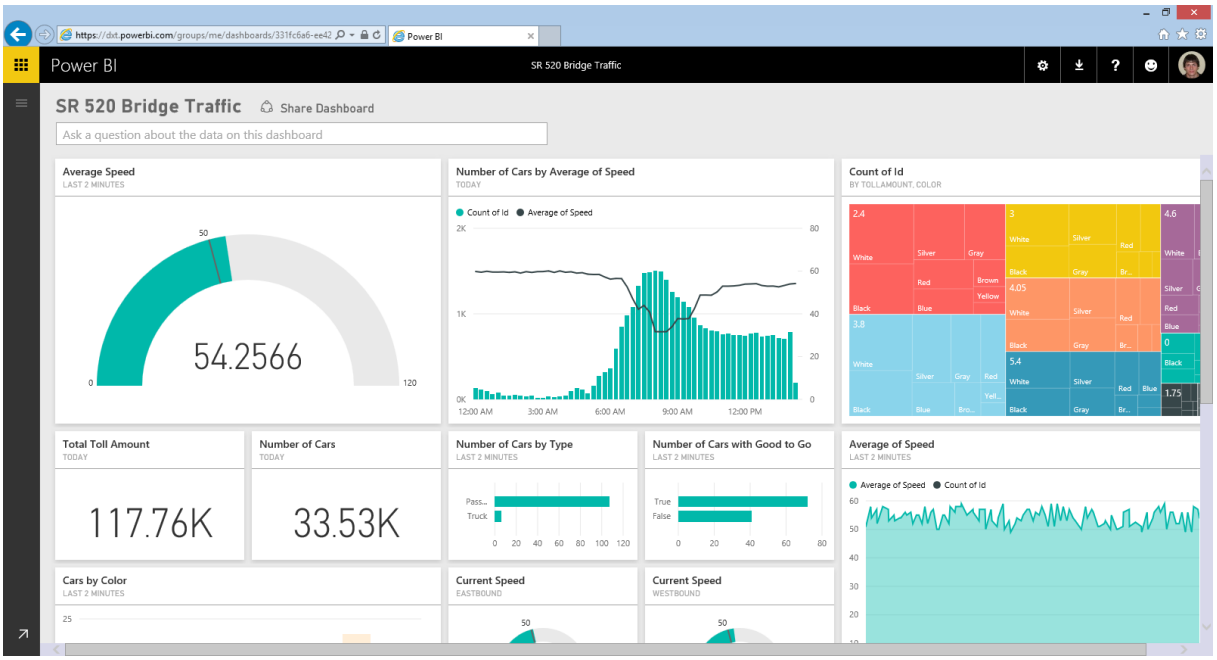


Figure 5 - Microsoft Power BI

As shown by the magic quadrant in figure 2 Microsoft Power BI is the most visionary tool of the three, mostly because of its monthly releases and some machine learning options. The advantages of Microsoft Power BI, according to Schlegel et al (2013) are its ease of use and its vision, while the weakest point is its immaturity which makes that the platform sometimes needs some help from other platforms such as Excel. This tool can also handle big data, just like the other two. This tool has about 24 graphs to choose from. An overview of the tool is given in figure 5.

2.3.4 Comparison

After introducing these three tools, they will be compared in the table below on a few demands that are important for the project. These demands are as follows:

Number of graphs – Marks the flexibility of a tool, the more graphs available the more can be visualized with it.

Ease of use – There is not really a preferred level for this as of right now, however it is good to keep in mind in order to restrain to the time frame available.

Sentiment analysis – It might be useful to perform a sentiment analysis on data available, so therefore there will be taken a look into it here already.

Inter-active dashboard – As known from the background research on data visualization was found that it might be a very good idea to make the visualization inter-active to obtain user engagement. Therefore, there should be looked at which tools are capable of doing that.

Maps – There seems to be a big difference in tools being able to do mapping visualizations. As for now it is not sure whether this is useful, it might be good to at least look at whether it would be possible at all.

Colours – What was also found out from the background research on data visualizations was that colours are important within a data visualization, so therefore there will also be looked at the palettes that are available.

[3] https://powerbi.microsoft.com/en-us/

Table 1 – Comparing the three tools

	Tableau	Qlik	Microsoft Power BI
Number of graphs	24	13	24
Ease of use	Intuitive, no scripting knowledge needed	Requires basic scripting	Easy to use with some technical knowledge
Sentiment analysis	Can be combined with Python to do sentiment analysis	Sentiment analysis APIs available in English	Sentiment analysis APIs available in English
Inter-active dashboards	Possible	Possible	Possible
Maps	Available	Not available, only when bought	Available
Colours	Palettes available	Only two palettes available	Palettes available

Concluding from here, it seems most logical to use Microsoft Power BI. Mainly because of its visualization abilities, the API and the potential this program has. Therefore, the rest of the report will be based on this tool.

2.4 Similar projects

Similar projects can be described as project that are either data visualizations about heavy-hearted topics or projects about #MeToo in general.

2.4.1 ISIS in the eyes of the Dutch

*A **retweet** is when a tweet of another user is reposted or forwarded. Often people do this for political statements with which they agree or facts that they want more people to know.*

Last year, Hendrikse, Habib and van Keulen (2017) performed research on the reaction of Dutch citizens on ISIS by analysing Twitter. There are some specific tips that can be gathered from their research. For example, they filtered out retweets, because many of their tweets (that were taken from a dataset with all tweets of 2015) were retweets and are not necessarily new information. Because the main focus of this research being the opinion of Dutch citizens, they also removed news headlines from their dataset. However, it could be that this information is insightful, but it can be useful to keep in mind. Furthermore, they used the MAchine Learning for LanguagE Toolkit was used to filter out all irrelevant tweets and to group the tweets with a common subject. Unfortunately, the visualization itself cannot be found online anymore, so nothing can be said about that.

2.4.2 ezyinsights

One of the better, more complete data visualizations was found by ezyinsights (2017) who compared the viral Me Too event of 2017, mostly based on the trends. On the webpage, they looked at the media coverage on Facebook and its spreading. In the article added to the visualizations, some side effects are mentioned, such as the effect on high profile individuals (like Harvey Weinstein and Kevin Spacey) and the impact on values and ethics. What is also mentioned is that something like this has not happened before and the comparison of #MeToo with other big happening like the Grenfell Tower fire and fake news. The main

question raised was how important #MeToo was as an event within the year 2017. This article does not provide the background research, but gives a nice insight in the data visualizations possible and made already, there should be looked at how useful this is apart from ideas and inspiration.

2.4.3 On the Impact of Twitter-based Health Campaigns: A Cross-Country Analysis of Movember

Dwi Prasetyo, Hauff, Nguyen, van den Broek, & Hiemstra (2015) researched the impact of the Movember campaign on donations done. Movember is a campagn where men do not shave during the entire month of November to raise awareness for men's health; especially for cancer and mental health. This is considered one of the few global events, even though it is localized, because each country runs its own campaign. In the paper, some data visualizations were made. One was a map about the number of tweets per country, a line chart was made about trends of tweets per day and a scatterplot to compare the number of visits/social tweets with the number of donations/social tweets. Next to that, to convey data, they also included some tables. What is interesting in this paper is the manner of analysing tweets, which was very accurately described. They used a machine learning algorithm to guess the country of origin of the user and the Naïve Bayes algorithm being used for deciding which topic the tweet was about, for example.

2.4.4 Researching technology for mining social media in times of crises

A project from Delft University of Technology and TNO was conducted about filtering relevant information from Twitter. To make this easier, they introduced Twitcident[4], a framework and Web-based system for filtering, searching and analysing information about real-world incidents or crises (Abel, Hauff, Houben, Stronkman, & Tao, 2012). The system is connected to the emergency services to get a clear overview of crises happening. Within Twitcident, users can easily apply filters to only see the most relevant tweets. Some basic graphs are also made, to let the user gain some more insight on, for example, the tweet timeline with a bar chart and a world cloud to gain insight on the content of tweets. A lot of focus is put on user-friendliness within this project.

2.5 Data gathering

Data gathering can be done by searching through all tweets of months October and November 2017. These data are available for use. There are, however, some ethical issues involved with using these data even though the usage of these data is legally approved. According to Townsend and Wallace (2016), some questions when performing research on social media data arise. The first one being whether data from social media sites can actually be considered private or public. Users agree to a privacy policy when signing up, but of the other hand, is it fair from researchers to justify their actions by simply stating that users agreed with it (Boyd & Crawford, 2012)? An answer to this question can be that there needs to be some reasonable expectation of privacy according to the British Psychological Society in (Townsend & Wallace, 2016). For example, data on a private forum for people who struggle with alcohol addiction would be considered private data, whereas an open Twitter discussion with the usage of a hashtag to familiarize your thoughts with the thoughts of other Twitter users would be considered public data.
In addition to this question, also the question about informed consent is raised and whether the user knows his data are being used. Where in academic research the user signs a consent form where they agree their data are being used for research, although this being much like a privacy policy, the user does not explicitly sign this (Webb et al., 2017).

[4] http://www.wis.ewi.tudelft.nl/twitcident/

Next to this, Anonymity is also key, but it is very hard to anonymize social media data for every individual Tweet in a big data set (Narayanan & Shmatikov, 2009). Where a data set in its whole is possible to anonymize, a single tweet is not, since a simple google search will often lead to the original Tweet.
Lastly, there is also a risk of harm, and this might be one of the bigger questions considering the topic of the data visualization being rather sensitive. To reduce the possibility of harming subjects, it is mainly important to ensure their privacy and anonymity. Townsend and Wallace (2016) however argue that this risk might not be present when the user is aiming for a broad readership by using hashtags, for example.
Townsend and Wallace (2016) also made a framework where guidelines are represented for the usage of social media data with the use of a workshop with key scholars and further research. These guidelines can come in handy for this research. Webb et al. (2017) suggest some solutions in order to have research integrity which should also be taken into account when researching.

In general, there should be thought about protecting the users of such platforms when performing research with their data. Sometimes not all data is necessary to have and display and sometimes these data can be anonymized. Some other tricks are available as well, for example a framework has been made. There should, eventually, be reflected on the way privacy was guaranteed within this graduation project and what could have been done better. For now, it is good to keep everything mentioned in this report in mind when performing the research in the next module.

2.6 Data representation

To decide what data to represent, it might be good to use a survey as a starting point and to send this to different people within the later defined target group. In here would be statements with which they can agree or disagree regarding #MeToo. With the use of this survey it would be easy to see what people think is true, while it is actually not, and these data would be interesting to represent, because then it can fulfil the side-goal. Of course, also the main goal should be fulfilled (representing the magnitude of #MeToo), so this should be kept in mind when holding the survey.

2.7 Programming languages

Before importing data to the tool that will be used for the visualization, the data should be filtered already. For this SQL will be used because of the ease and effectiveness. Furthermore, when the data is prepared R will be used within the Microsoft Power BI. For sentiment analysis, Microsoft has an API available which can be used within the tool itself. If there will be chosen to do something with named entity recognition, the Azure machine learning studio can be used in combination with, amongst others, Microsoft Power BI.

2.8 Target group

The target group should be very broad, but it should be people with some affinity with internet and data in general. To narrow it down, people with ages between 18 and 65 will be chosen. This is so the visualization is accessible for most computer users.

2.9 Requirements

The demands of this assignment can be listed as follows:

- The visualization should be easy to understand; *The user should not need too much explanation to use it, since a data visualization like this is normally freely available on the web as infographic and anyone can look at it or interact with it.*

- The visualization should reveal information that was not known before (if possible); *The data visualization should have an impact on the user and this is most easily done by showing users data that they did not know of before.*
- The visualization should be interactive; *According to 2.2 Data Visualizations, an interactive data visualization is highly advised, since this leads to more user engagement and can also lead to new insights. This is closely related to the second bullet point to reveal new information.*
- The project must protect the individual user from harm; *Next to some requirements regarding the data visualization itself, there should not be forgotten that there are personal data involved in this project. There will looked at the visualization from an ethical point of view to ensure that the people involved in the dataset are not subjected to harm.*
- The visualization must convey information; *The visualization must tell a story and therefore must have information that can be told to the user.*

2.10 Testing

In the end, the final product will be tested by means of user tests; where users of the target group get invited to see whether their idea of the product matched the idea behind the product. Here will be looked at how people interact with the product as well and their feedback will be considered to improve the visualization.

2.11 Conclusion

In this chapter, a background research on big data visualization in general has been conducted. From here can be seen what colours (not) to use, which graphs are useful for which visualization and the differences between static data visualizations and dynamic data visualization. The main points of focus of this part were that line charts are very useful for displaying trends and that the colours chosen should differ sufficiently from each other and up to seven colours can be used. Next to that, the most striking colour should be used for what should stand out the most and the lightest colour should either be used for the average or for the smallest value. Was what also concluded from here was that there should be looked into making an interactive data visualization.

After this there has been chosen which tools would be good to use. This would be Microsoft Power BI, because of its potential, flexibility and useful APIs available. Then there was looked at similar projects and a few projects were discussed. Some useful conclusions from this were that some of them developed a framework or a library which can be used. Next to this, there were some projects with data visualizations, so there can be looked at what has been done there to see what might be useful for this project. Lastly, some projects described how they analysed the data, which can also be used to find a good way to do it in this project.

Then there was looked at the more ethical side of the data gathering. This will be explored more later on, but for now some concerns were raised about user tweets for research, because the users were not informed about this and also because it is not clear whether the data on Twitter is private or public.

Finally, some ideas were raised about how to decide upon the visualizations and how to test the end product. To make this process a bit easier, a few demands were set up by combining the testing with the problem statement, goal and research performed earlier.

Chapter 3 - Ideation

3.1 Requirements

The visualization should reveal information that was not known before (if possible).

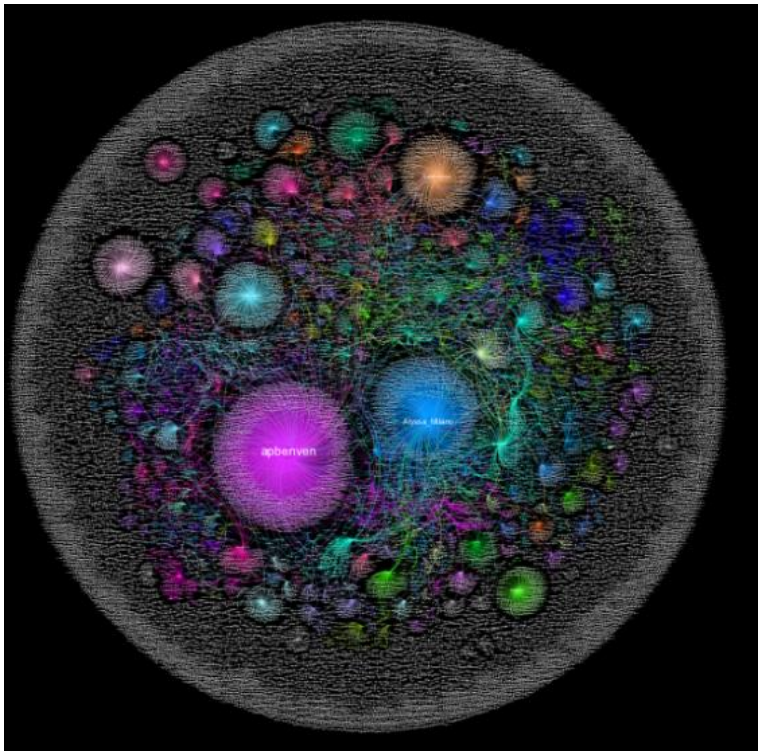
It is first needed to get familiarized with the data. In order to do this, first some already existing data visualization will be explored to see what these are about. After this, these visualizations will be remade in Power BI with the real dataset, to see if they match. The existing visualizations will be analysed to find out their broader topics and these topics will be used as input to make the survey. This is interesting, because the data could show something that is completely different than what people think the data would say and this would then be very interesting to show in the final data visualization, because that will make an impact on people.

3.2 Existing visualizations

These sources are not scientific, but they are publicly available and read by many people. What is interesting is to see which assumptions these infographics imply and to determine the topics these visualizations are about in order to gather questions for the survey.

3.2.1 Medium.com[5]

Medium.com has gathered many infographics from different sources.



First, they show a network graph which shows the largest nodes within the Twitter network between 16 October and 18 October 2017. What can be seen is that the largest nodes are apbenven and Alyssa_Milano whose tweets can be seen below in figure 7 an 8. These tweets were the most retweeted and liked during this timeframe.

Figure 6 - Network graph of most popular users regarding #MeToo

[5] https://medium.com/@erin_gallagher/metoo-hashtag-network-visualization-960dd5a97cdf

Figure 7 - apbenven's tweet

This visualization is not entirely accurate, since it only shows data of two days, just after the initial start of the hashtag. If this would have been done over multiple weeks or months, the visualization would probably be more scattered. It seems like the first topic of interest are the most popular tweets.

A similar visualization is made for #YoTambien, the Spanish version of #MeToo and can be seen in figure 9.

Figure 8 - Alyssa_Milano's tweet

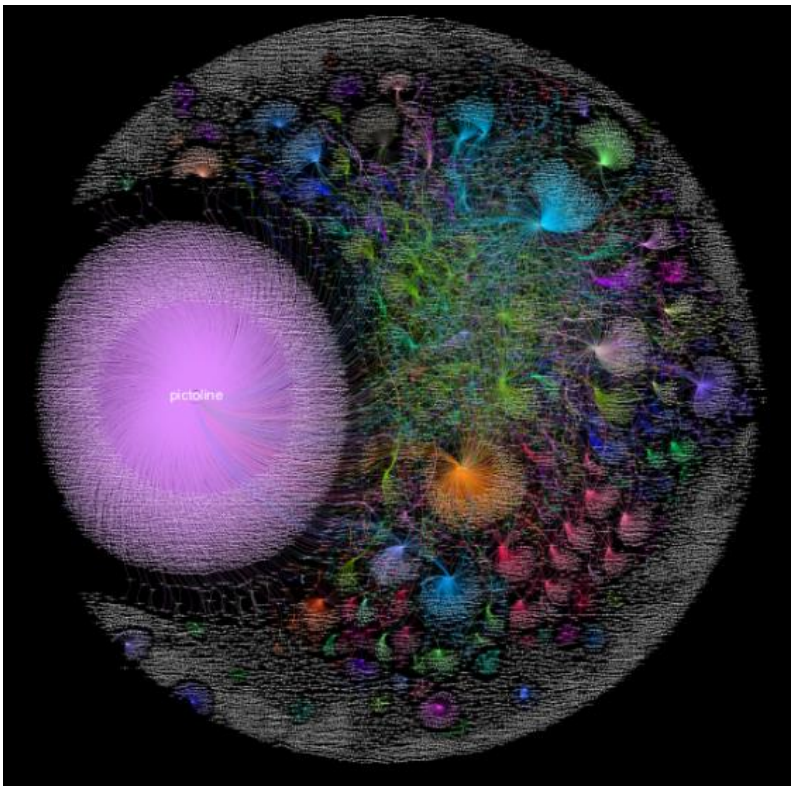


Figure 9 - Wordcloud of #YoTambien

Here only one account is very popular: pictoline. These two network graphs are very different from each other because of the amount of (big) nodes. It has the same pitfall as the English graph, namely, the data is only from two days.

Then a visualization about the spreading of the hashtag around the world is shown on medium.com. What can be seen from here is that the hashtag was most popular in the United States and Europe. What is maybe more interesting is to see that, for example, Russia, Australia and most of Africa do not use the hashtag very often. This might be because of a language barrier or because the topic is not that big in those countries. The second topic of interest can then be stated as the countries in which the hashtag is or was used. The fifth visualization on this site also addresses this topic. This visualization differs from the second in color and dynamics, but the conclusion is the same; America and Europe are most active and Russia, Australia and Africa are least active. What is noteworthy for both visualizations is that in Asia, not much activity is shown, however in India the hashtag is quite popular.

The third visualization is one from the Telegraph about the gender of the people tweeting #MeToo, their origin and the amount of tweets, replies and reactions. This graphs are simple bar charts and can be seen below.

Figure 10 - Global breakdown of people tweeting #MeToo

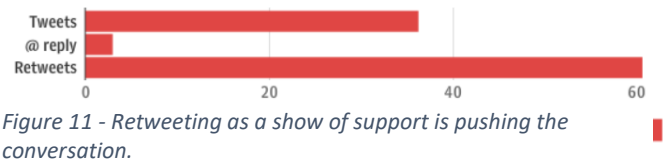


Figure 11 - Retweeting as a show of support is pushing the conversation.

These visualizations are not really telling a story, but mostly giving a brief overview of the hashtag. Gender and Tweets/replies/Retweets are new topics in these visualizations.

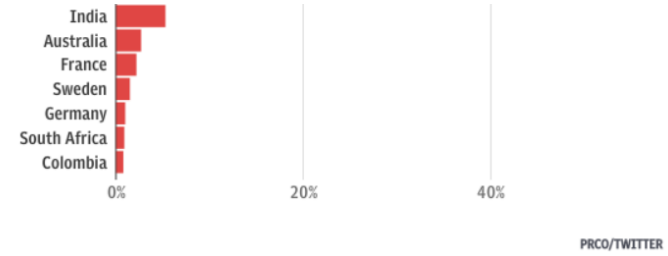


Figure 12 - Where is the #MeToo hashtag most used?

The fourth visualization is a word cloud with associated hashtag with #MeToo. What can be seen is that it are mostly translations of 'me too' in Spanish, French and Chinese. Next to that are some different phrases of me too, such as no more, sexual harassment and I hear you. This visualization can be seen on the right and the main topic of this visualization is associations.

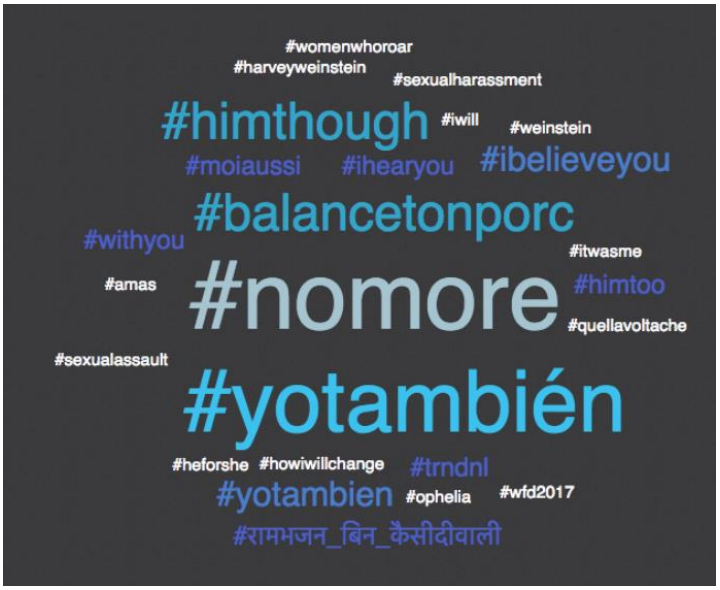


Figure 13 - Word cloud of other hashtags associated with #MeToo.

Concluding from this, there are five topics addressed in these visualizations on medium.com:

- Popular accounts
- Popular countries
- Gender
- Tweets/replies/Retweets
- Associated hashtags

What is mostly lacking in these visualizations is the storytelling aspect. These graphs are secluded from each other and tell a small part of the story, but since the styles of the visualizations are so different, they cannot be combined into one story. This can thus be added as an extra requirement for this project.

3.2.2 Google

Next to medium.com, Google also paid attention to the topic Me Too.

Figure 14- Me Too rising, a data visualization of Google about #MeToo via Google trends

Google made a noteworthy data visualization based on Twitter trends, as can be seen in figure 14. They made a 3d map of the top searching cities in the world. The user can interact with it by moving the time slider on the bottom. The user can also click on a city to see all searches from that city or to create alerts when new searches come up. In this visualization the user can quickly see what was a hot topic in certain cities. What is again lacking is the storytelling aspect, after clicking a few times around cities nearby and interesting cities, the user is done with the visualization without getting any idea about the magnitude of it or further thinking it through. There is no confrontation, except for when the user is clicking through the different hot news topic to further inform them. What is very interesting though, is that they display data based on city instead of on country, this might be something to consider.

3.3 Data

In order to further become familiar to the dataset that will be used during this project, some visualizations mentioned before (in chapter 3.2) will be remade to see whether their results are comparable to the findings mentioned previously. Note that the dataset used for this part includes the first 2200 #MeToo tweets after 15 October 2017.

3.3.1 Associated hashtags

Figure 15 – Wordclouds of hashtags associated with #MeToo. Left #MeToo is excluded, right it is included.

What can be seen is that Weinstein is in this remade data visualization as well, just like WomenWhoRoar, a famous Twitter account. The rest of this visualization does not seem to have similarities to the one of medium.com, but this can be because of the different time frame.

3.3.2 Popular countries

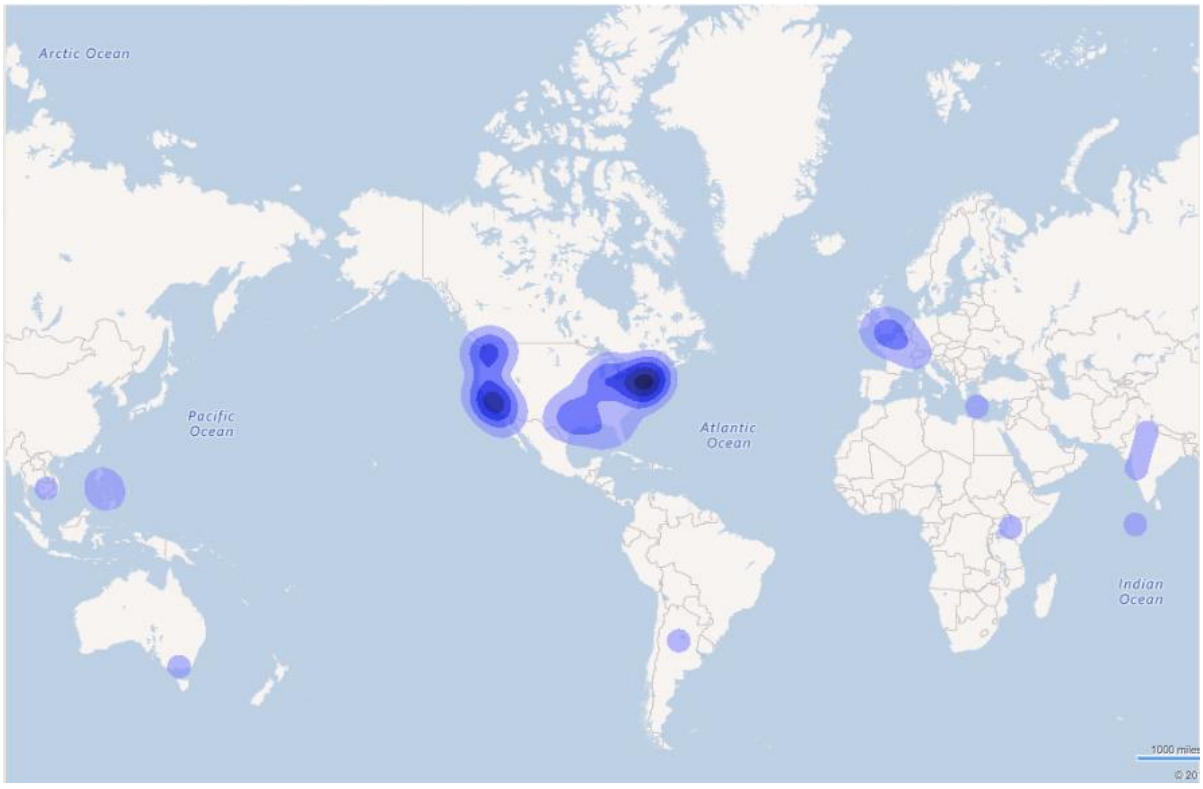


Figure 16 - Contour map of location #MeToo tweets

In figure 16, a map of the locations of 47 of the 2200 #MeToo tweets can be seen. Twitter has an opt-in option to use location of the user. In only 47 tweets (about 2.1%), the user allowed Twitter to use their location, but this will scale when the total amount of tweets scales as well. This map is quite similar to the map visualization from medium.com; the United States and Europe are most popular. There are no tweets in

Russia and only one in Africa and Australia. What can be seen again is that Asia is not popular regarding tweeting #MeToo, except for India.

3.3.3 Gender

Gender is 'unfortunately' no longer saved by Twitter because it is considered to be special personal data. There are strict regulations regarding saving and using these, so they are not included in the dataset. There are, however, APIs available that can guess the gender of a user based on their name, but this estimation of gender is fairly unsure, especially for non-English names. A probability of 1 can be assigned to Peter being a man, but when looking at other names this probability is not that high (Nguyen et al., 2014). On top of that, users do not have to include their real name; their 'names' can be anything from AnActualBlackCat to RustyBertrand. Namsor[6] can be a good API if gender turns out to be an important factor from the survey, but otherwise it seems like a good idea to not include this gender in the visualization.

3.3.4 Tweets/replies/Retweets

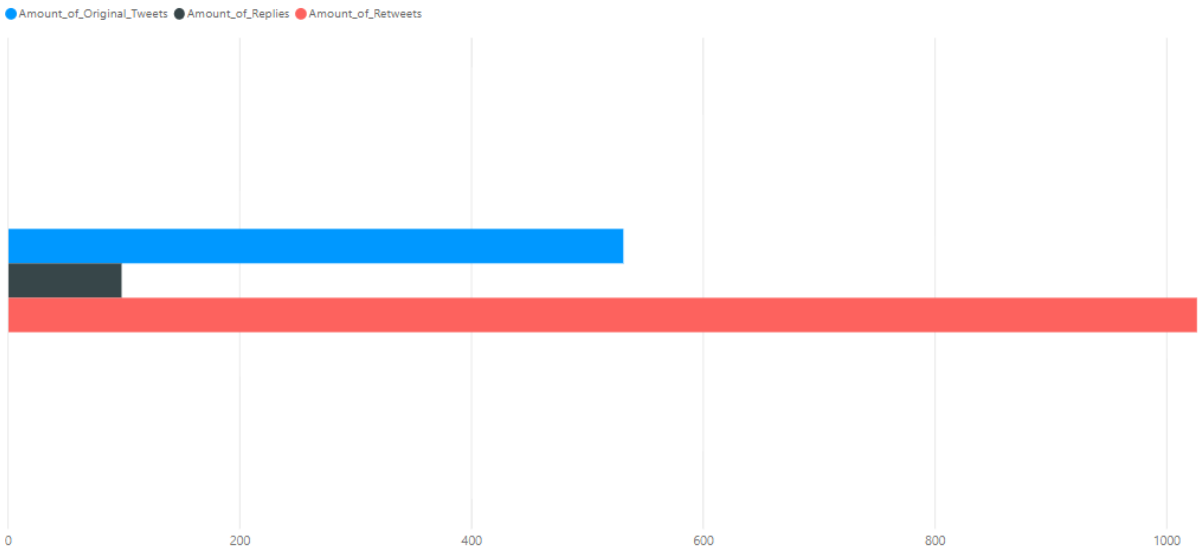


Figure 17 - Tweets/replies/Retweets in a stacked bar chart

After this, the number of original Tweets, replies and Retweets were compared to the visualization from medium.com, as can be seen in figure 17. This graph is practically identical to the one displayed in figure 11.

[6] https://github.com/namsor/namsor-powerbi-connector/releases

3.3.5 Popular accounts

Figure 18 - Popular user accounts associated with #MeToo

Alyssa Milano is again one of the most popular accounts, but apbenven is not included in figure 18. This could be because the dataset used in figure 18 is over a different period than the one in the network graphs shown before. What could also be is that in figure 18 was only looked at tweets and Retweets, but maybe there are a lot of replies too.

What can be seen from this part is that there are five factors that might be interesting, however one of them, gender, is not very feasible to analyse since Twitter does not save these data about users. Using an API can be a solution, but it is questionable how accurate this would be. It will, however be taken into account with survey, because maybe there are other solutions if people find gender to be a very interesting factor.

- Popular accounts
- Popular countries or cities
- Tweets/replies/Retweets
- Associated hashtags or words
- Gender

Two small additions were made to the factors to include them in the survey. The first one is to include associated words, since these might be interesting as well aside from the associated hashtags. Just like the addition to popular countries, because it might become closer to home when using cities instead, like Google did.

3.4 Data preparation

The files used for this project included 25667 tweets (about 1% of tweets) from the 15th of October until the 31st of December 2017 that included #MeToo. The data were gathered from archive.org[7], a website where users can upload all types of files, including datasets. User Jason Scott uploaded a Spritzer Twitter grab of every month in 2017 on this website, which were in total about 340 million tweets. Spritzer is a free version of the Firehose Twitter grab and grabs about 1% of the total twitter stream randomly. The data was first put in logical order; a folder a day which included 23 folders each of about 60MB zipped (representing one hour) which all included 59 files of about 10MB zipped (representing one minute). These tweets first needed to be prepared in order to perform data analyses and visualizations. To be able to do this, R and RStudio[8] were used, because of its availability and ease of use. These files were then loaded in R, untarred/unzipped, filtered on #MeToo in text and then added to a new subset. In the end the subset was exported and will be used for further analysis. In practice, only filtering on #MeToo (and not on 'Me Too.') can exclude some inexperienced or casual users (Moe & Larsson, 2012). Therefore, the dataset has a bias towards experienced users.

3.5 Survey

A survey was made to gain better insight in what people thought about #MeToo. The factors identified in 3.3 Data were included in this survey. The survey itself can be found in appendix A. As can be seen, a score between 1 and 5 could be given to all closed questions, where 1 meant strongly disagree and 5 meant strongly agree. What can be seen is that there are quite some questions which are the same but asked in a different manner. This is mostly to make sure that people react on the statement and not on the manner of asking the question and also to see whether people fill in the survey seriously instead of randomly. The survey itself was handed to 23 subjects on a paper version, because the response rate is higher in this way. This could have influenced the survey however, because people may not like to write much, or they might make up their own options (which is maybe a good thing with this topic). The full results can be found in appendix B, the summary can be found below.

3.5.1 Popular accounts

In the survey, people were not very interested about what celebrities think about #MeToo, both questions received a score lower than 3 (the average). Next to this, when asked which of the accounts the subjects knew about, not many of the names were recognized. Most people knew Kevin Spacey (20) and Harvey Weinstein (15), but no one recognized Tarana Burke, the real founder of Me Too. Some people also mentioned Alyssa Milano as the founder when asked to explain Me Too.

3.5.2 Popular countries or cities

Most subjects do think that #MeToo is more popular in America and in western countries in general. Respondents are interested in what different countries in the world think about #MeToo, this seems to be because they think there are countries that react more negatively than the Netherlands. Next to this, people are also interested in the opinion of different cities in the Netherlands, although being a bit less than the interest in difference of opinion between countries. Interest in the spreading of the hashtag scored below the average.

[7] https://archive.org/
[8] https://www.rstudio.com/

3.5.3 Tweets/replies/Retweets

The subjects seemed to think that most tweets are retweets, the second-most are replies and the least are the number of original tweets. Mostly the retweets vs. the original tweets are expected to have the greatest difference (3.91). Only two respondents use Twitter themselves.

3.5.4 Associated hashtags or words

Although this being like the popular accounts, some extra questions were asked to see if people wanted to know about more 'nearby' issues of #MeToo. People were quite curious to know what their friends think about it, but some people knew this already from talking to them about this topic. People did not seem to be sure whether it happens in their industry. They gave a score of 3.04 when asked whether they think it does not happen a lot in their industry, while they gave it a 2.87 when asked whether they think it does happen a lot in their industry.

3.5.5 Gender

The score for thinking there is a difference in opinions between men and women was the highest of the entire survey (4.22). So almost everyone thought there was a difference between the genders and they felt that women are generally more positive about the hashtag. People were also interested to know more about male versions of the hashtag (such as #MenToo).

3.5.6 General

In general, people knew quite well what #MeToo is, although sometimes it was written down a bit bluntly. Almost no one wrote down that Me Too is about the magnitude of sexual harassment, but almost everyone knew that it is about sexual harassment. People did not seem to think #MeToo is just a hype. One person mentioned that they do not want to know more about #MeToo at all, while most people were quite interested. There was one comment about the questions being too generic. What was also interesting is that some people mentioned that they know #MeToo mostly from jokes. Lastly what drew attention was that one person actually wrote #MeToo as a comment.

Some critical notes are that there is an age gap above the age of 35 year, where not so many respondents were found. Many were asked but did not want to be involved because of the sensitive topic. Next to this, there were some (older) people who filled in the survey but scored very different on the opposite negative questions. For example, scoring men being more negative a 4 and scoring men being more positive also a 4, these were filtered out because it was not sure whether the survey was filled in seriously or not. Next to this, most of the replies were from people between 18 and 25. This was not exactly planned, but people were very enthusiastic about the topic and filling out the survey. In the end, this age range is probably most interested in seeing a data visualization about this kind of topic. To ensure a fair research, the target group of this project will be people between 18 and 35, since this group filled in the survey mainly and this ensures research can be conducted about this group.

What can be concluded from this part is that there are some strong opinions about the difference between men and women, but this is very hard to test unfortunately, since Twitter does not store these data. It might be an idea to replace this factor with tweets that include hashtags like #MenToo. Looking at the other four factors, it seems like people do not know exactly the ratio between tweets, retweets and replies, because there are more original tweets than they expected there to be. Next to that, the subjects thought that the hashtag was more popular in America and in western countries in general, while we have seen that the hashtag is indeed popular in America, but also in Europe and India, so this might surprise them.

People are not very interested in knowing more about their favourite celebrities regarding #MeToo, but it might be an idea to put in the persons they did not recognize (i.e. Tarana Burke). People are interested in knowing what their friends think about it and it seems like this is partly because they think it does not happen around them. In general people do not seem to know exactly that the hashtag is about the magnitude of sexual harassment, because of the open question at the beginning. Lastly, it might be an idea to include the jokes that are being made about #MeToo in the actual visualization.

3.6 Brainstorming

Concluding from the last part, there are five factors that came out of the survey that are worth conceptualizing with:

1. Nearby issues regarding #MeToo
2. Gender (maybe including #MenToo)
3. #MeToo also happening in Europe/India
4. Magnitude of sexual harassment
5. Jokes about #MeToo

What should be kept in mind is that the visualizations must tell a story and should thus somehow be related to one another. It would be very nice to include a shock factor, if possible.
Often during conversations or when handing in the survey people said: 'but many of the tweets are still fake.'. While that might be true, it falls into the same category as jokes about #MeToo and thinking that it does not happen right next to you. Therefore, it seems logical to let the purpose of the visualization be to let people face the facts, also close to home. To do this maybe some jokes can be displayed and then made into real data (for example, '#MeToo is only for whiny stuck up girls', then show #MenToo). This way points 1, 3 and 5 can be combined, with maybe a part of point 4 in it as well.

Another idea can be to first show #MeToo in America and after this bringing it closer to home. This way points 1, 3 and 4 can be tackled. Then people can first experience a 'well that doesn't happen here, luckily' moment and then experience that it actually does happen here.

A third idea is to let people enter their demographics first (e.g. male, 23, Enschede, IT) to then show them how #MeToo is in their industry and maybe let them compare it to other industries and environments.

The last idea, made because people were very enthusiastic to fill in the survey, is to let people tell their opinion about #MeToo and compare it to their demographics or to the Netherlands in general.

3.6.1 Jokes

Many jokes are made about MeToo, both in casual conversations as on Twitter (see figure 19). An interesting example of these jokes was made by Bill Cosby. Bill told Laura McCrystal, Inquirer reporter: 'Please don't put me on MeToo, I just shook your hand like a man.'[9]. Now Bill is accused of over 50 counts of sexual misconduct.

Figure 19 - Joke about #MeToo

9 https://twitter.com/LMcCrystal/status/951255936726859777?ref_src=twsrc%5Etfw&ref_url=https%3A%2F%2Fwww.teenvogue.com%2Fstory%2Fbill-cosby-me-too-joke&tfw_site=teenvogue

So, it is clear that not everyone takes #MeToo seriously. What could thus be a very interesting thing to do is to show these jokes in contrast to the real numbers available. The user experience would then first start with a bit of a giggle about the jokes, they are kind of funny after all, and then go on to seriousness, because it happens so often. What is something to take into account with this is to not publicly name and shame the people who make the jokes, but on the other hand, the people posting tweets that include #MeToo did it because they wanted their joke to be out there. Still their privacy should be taken into account somehow, there can either be looked at public statements, comedians or at jokes that are made more often, since then they can't be traced to the original user anymore or was meant to be out there anyways. Another thing that could help is to rephrase the joke a bit so people cannot search for it. It should always be taken into account that the visualization itself should not include personal data, since everyone can see these data and it may potentially cause harm to the subject. If this idea gets a follow-up, there should be looked into the ethical side of it.

3.6.2 America first, Europe second

Since many of the respondents seem to think that #MeToo is more something that is popular in America and not in Europe, it can be a nice idea to play with this a bit. This can be done by mapping the origin of user accounts or, if users have enabled this, based on the location tweeted on. What is hard about this is that people in Europe often tweet in their own language, which is not always English, while many analysing tools are only available for free for English text. The sentiment analysis, which is maybe one of the most interesting analyses to perform on this dataset, is only in English and Spanish. If there are sufficient tweets from the Netherlands that are in English, this is an option. What could be done is to compare it to the United Kingdom, but that does not seem relevant for the research questions. Another suggestion for this is to search the tweets for names entities that are countries and group them by continent to see what (mostly) American users think about other countries. However, that also does not hit close to home unfortunately.

So, if it is not possible to do something with the actual content of the tweets, there probably should be focused on the number of tweets and their location. From the survey was concluded that people were not very interested in the spreading, so there should be looked at the magnitude of the topic. The number of tweets might thus be a very interesting thing to use in the visualization.

3.6.3 Demographics

To let #MeToo really hit close it might be a solution to ask people for their demographics and then show them the opinion of their peers. For example, someone could fill in that they work in IT and the data would filter accordingly resulting in only showing the tweets about IT (i.e. '#MeToo, a.k.a. How to ambiguously lump yourself in with victims of rape after you were cat called or the IT guy asked you out for coffee.') or writer ('#MeToo a writer/an actor/two producers. If you believe that it happened to me = good. Better = help ensure it doesn't happen to anyone else.'). A quick search through the sample data set does not show many results, but it shows some.

Unfortunately, Twitter does not store many demographic data other than (guessed) location. Other demographics are possible to guess, such as age and gender, but the correctness of these tools is relatively low (around 75%) (Nguyen et al., 2014). It is the question, again, whether it is good to include 'guessing' gender when having such a sensitive topic to talk about.

3.6.4 Input field for opinions

The idea is that people can input their own opinion in a text field along with some demographics. This can then be compared with their peers and also with the big dataset to see similarities and differences. Microsoft Power Bi does not support this feature yet, even though it is already a suggestion for over two years now, they recently (March 2018) reacted that this feature is something they will look into in the far future. It can be done with a simple web application, but the problem with that, again, is the sentiment analysis. The API is free up to 5 million tweets a month, if the data needs to be updated frequently the limit will be reached very soon if the visualization proves to be popular. What can be done is to choose to only keep the top 100 opinions per hour, such that the monthly total of analysed rows is 72.000 opinions, far below the 5 million maximum. This would mean that an extra web application needs to be made in order to gather the opinion of people.

Concluding from the part above, the first and fourth idea seems most appealing, because they are most related to the survey and both quite possible to make without too many assumptions. More thought will be put in the chosen ideas in the next part of this report.

3.7 Elaboration on brainstorm

To choose between the two initial ideas mentioned above, a good look needs to be taken at what both ideas consist of, how they relate to chapter one and how they relate to the survey.
The first idea focuses on showing what other people think and how they sometimes mock #MeToo. This can therefore be an eyeopener to people, since they may also have their doubts about it. It is good to show people both sides of the story, so they can decide for themselves what to think about it. People do not have a way to insert their own opinion, but they could maybe tweet or comment on it to still show their opinion, without changing the visualization.
The second idea focuses more on what the user thinks and comparing that to the world and other users. Therefore, this becomes more like a real-time data visualization where users feel heard.
When looking at chapter one, a few conclusions about the main data visualization were made. First of all, it was mentioned that a real-time data visualization was not a very good idea, since it can make the visualization very vulnerable. It is not clear which data will be displayed and maybe the data right now are completely different from the data a month ago. This can cause the visualization to get a different (unwanted) meaning. Next to this, the visualization can also change when it is kept running for a few years without regular updates. It is fairly certain that the Tweets about #MeToo in the year 2025 will differ from the Tweets right now.
Regarding the survey, both ideas could fit well, since both jokes and opinions were mentioned by participants of the survey. What is remarkable is that both ideas have not yet been executed with this topic, as can be seen in the beginning of this chapter.

Considering these remarks, the first idea would be a better fit, mostly because the data visualization is at least very stable and because a story can be told without it possible being changed by different user input. A nice addition to this might be some attention to #MenToo as well.

3.8 Ethical aspects

There is a thin line whether tweets are public or private data, as briefly mentioned in chapter one already. There are no clear rules about this either at the moment of writing, merely guidelines. The question arises

whether people know their tweets are legally public data even though it is stated very clearly in Twitter's privacy policy[10]. There is also a difference in people tweeting with a hashtag included or without. The people who include it often aim for a broad readership or want to associate their thoughts on a subject with others' thoughts (Townsend & Wallace, 2016). The people making jokes about #MeToo are often including the hashtag in it. According to the sixth case study (page 15) of Townsend and Wallace (2016) the quotes can be used without informing the subject. However, it would be a good idea to remove ID handles from everyday users who are not a celebrity, to protect these users from potential harm. Tweets are also being compared to letters to newspapers and they are being published and used as well. Next to this, Twitter is by default set to public and people are able to change this to private, so people are in control of their data. When users change it to private, their tweets can then no longer be accessed without following the subject. If a researcher would do this, this would imply a potential closeness (Moreno, Goniu, Moreno, & Diekema, 2013). All in all, there can be concluded that these tweets have a public character. However, caution must be taken when handling such vulnerable data. When displaying text with user ID, there should be looked into the subject being over 18 and not being a vulnerable adult. Lastly, all tweets include an 'is sensitive' Boolean in the database, so it might also be a good idea to not include these tweets for qualitative analysis, if present.

Aside from the public/private argument, another serious issue with using Twitter data is that it is written, and therefore can easily be taken out of context, as mentioned by Bath, Kennedy, Whiteman et al. (2016). It might be that a certain tweet is sarcastic/ironic and then incorrect assumption might be taken. What can also be the case is that the hashtag was used for another purpose than what it grew into. It might be that people using it, mainly in the beginning, used it in a more innocent way. For quantitative research this might not be a very big issue, but for qualitative research there should be manually looked into the tweets and whether they were indeed written with a sarcastic tone.

What was also brought up by Bath, Kennedy, Whiteman et al. (2016) is that subjects should not become victims of a witch hunt of followers. Media can override ethical considerations with public interest. This could be prevented by the same method as mentioned in the first paragraph; to only use celebrities' tweets with ID handles and other people's tweets without ID handles. Another important step is to check whether the tweet was deleted afterwards or not before publishing a contextual analysis of it.

Concluding, there is not one answer to what is ethical versus what is legal, and research often provides more questions than answers. Tweets can be used for quantitative analysis without informing the subjects, but for displaying tweets in itself there should be looked at the person who posted it. If this person is a celebrity, the tweet can be posted with ID handles. If this person is under the age of 16 or a vulnerable adult, the tweet cannot be posted. If neither of these apply, the tweet can be posted without the ID handles. There should also be looked at whether the tweet is sarcastic or ironic before screenshotting it. On top of this, before posting the exact tweet, there should be checked whether the tweet was deleted afterwards or was marked sensitive. This topic will be further explored in 4.3.

3.9 Revised requirements

The requirements can now be altered according to these new findings and, since it has become clearer, this will be done with the MoSCoW method. This method prioritizes requirements according to whether they must, should, could or will not be in the end product.

[10] https://twitter.com/en/privacy

The visualization must:

- Be interactive; *to comply with 2.2 Data Visualizations*
- Tell a story; *To comply with 3.2*
- Convey information; *The visualization must tell a story and therefore must have information that can be told to the user.*
- Protect the subjects from potential harm; *As briefly touched in 3.8, this will be further explored in 4.3*
- Be easy to understand; *So, anyone can use it*
- Tell about the magnitude of sexual harassment; *To comply with the original goal of Me Too*

The visualization should:

- Include a couple of the interesting topics concluded from the survey; *To comply with 3.5 Survey*
- include the 'right' graphs and colours; *To comply with 2.2*

The visualization could:

- Reveal information that was not known before; *To make an impact*
- Be Easy to use; *So, all users from the target group can interact with it*

The visualization will not:

- Let the user give input by themselves; *As explained in 3.7*
- Be real time; *As explained in 2.2*
- Include ID handles of non-famous persons, children or vulnerable adults; *As briefly touched in 3.8, this will be further explored in 4.3*
- Include analysis on gender or age; *As explained in 3.3.3 and 3.6.3*

Chapter 4 - Realization

In this chapter, the idea will be worked out and made into a design. Later on, this will be turned into a full data visualization using Microsoft Power BI.

4.1 Jokes

First, several jokes about #MeToo have to be found. As mentioned in the previous chapter, these jokes must be made by famous persons to protect others from potential harm. Jokes about #MeToo can be organised in two categories; positive jokes and negative jokes. There will be looked at both categories. A list of possible jokes can be found below. When using this project as a framework, there can be looked at jokes about other topics as well or some more general jokes.

4.1.1 Positive jokes

Figure 20 – Opening ceremony at the Film Independent Spirit Award 2018 https://youtu.be/OW0Wt30EeCc?t=2m43s

In figure 20 a screenshot of the film independent spirit award can be seen. Nick Kroll (right) and John Mulaney (left) opened this ceremony with a monologue. Here they started of this piece with 'What a year..' and they proceed with sharp remarks regarding the allegations on Weinstein, Spacey and Allen. They mostly joke about the apologies these men made. For example, Spacey saying: 'I don't remember doing this, but I know I was drunk when I did it.'.

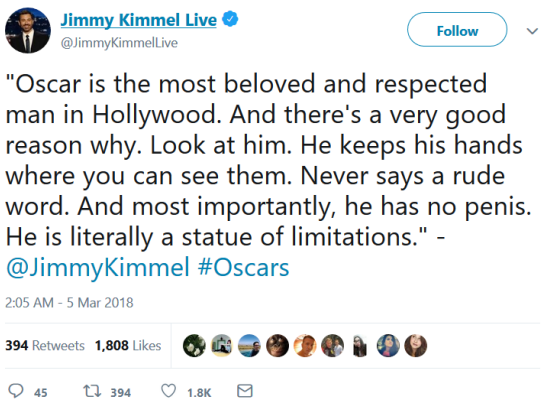


Figure 21 - Indirect joke about #MeToo

Jimmy Kimmel, an American television host, comedian, writer and producer, also made some indirect jokes regarding #MeToo. This tweet is a quote from the opening act of the Oscars in March 2018. During his speech he addresses the gender inequality, not only #MeToo, but also the number of women directing movies versus the number of men for example.

During the Golden Globes 2018, comedian Seth Meyers also made a sharp remark regarding #MeToo in Hollywood.

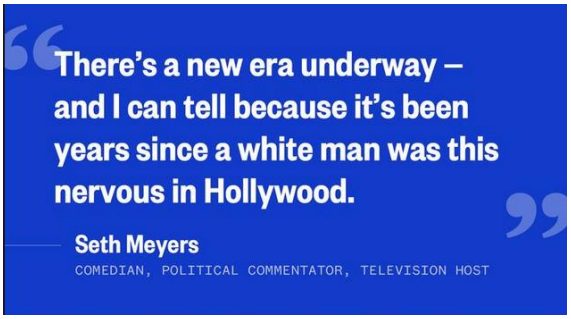


Figure 23 - Joke by Seth Meyers at the Golden Globes 2018

Figure 22 - The Mash Report's on the confusion around sexual harassment

Lastly, The Mash Report including Rachel Parris and Nish Kumar also provides a good job in positively joking about #MeToo. Rachel Parris says: 'Even talking to a woman is now completely off limits, isn't it?', to which Nish Kumar responds with: 'No'. 'No indeed, but it is fun to pretend to be confused about it', Rachel then answers. A screenshot of this show can be found in figure 22.

4.1.2 Negative jokes

Furthermore, there are some heavily criticized 'jokes' as well. Such as the poem of Sean Penn, as can be seen in figure 24. He received criticism from both men and women about the poem he published in his book 'Bob Honey Who Just Do Stuff'. He clearly does not agree with the movement and mocks it comparing #MeToo to a toddler's crusade.

Though warrior women
Bravely walk the walk,
Derivatives of disproportion
Draw heinous hypocrites
To their flock.
[....]
Where did all the laughs go?
Are you out there, Louis C.K.?
Once crucial conversations
Kept us on our toes;
Was it really in our interest
To trample Charlie Rose?
And what's with this 'Me Too'?
This infantilizing term of the day...
Is this a toddler's crusade?
Reducing rape, slut-shaming, and suffrage to reckless child's play?
A platform for accusation impunity?
Due process has lost its sheen?
But, fuck it, what me worry?
I'm a hero,
To Time Magazine!

Figure 24 - Poem of Sean Penn

Funny how some of ugliest looking women are claiming to have been daily eve teased n molested!

Figure 25- Negative Tweet

There are many negative jokes made by people whom are not celebrities. While it is the question whether they should be named, their quotes could be used. Figure 19 already displayed one, but figure 25 and 27 have some examples as well.

I'd love it if a group of women harassed me. Does High Spirits offer that as well?

Figure 27 - Negative Facebook status

Lastly, some people made online greeting cards with a negative remark regarding the movement, as can be seen in figure 26.

Figure 26 - Greeting card about sexual harassment

4.2 Designs

The visualizations can both be made into a webpage and a report. Every report can be published in HTML and then embedded in a site. A site would allow users to have easier access to the data, because they do not need to download Power BI as well as to simply scroll through instead of clicking on new tabs when being done viewing the previous one. The publishing onto an HTML page is merely technical additions to an already existing dashboard and because of time limitations, no focus will be put on this during this project.

The designs will be briefly explained here. An overview of all dashboards and their order can be found in figure 28. A more extensive description can be found in the next paragraph, where the dashboards will be made.

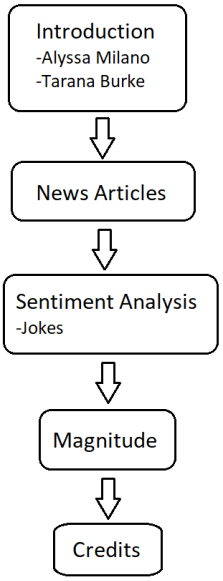


Figure 28 – Overview of order of designs

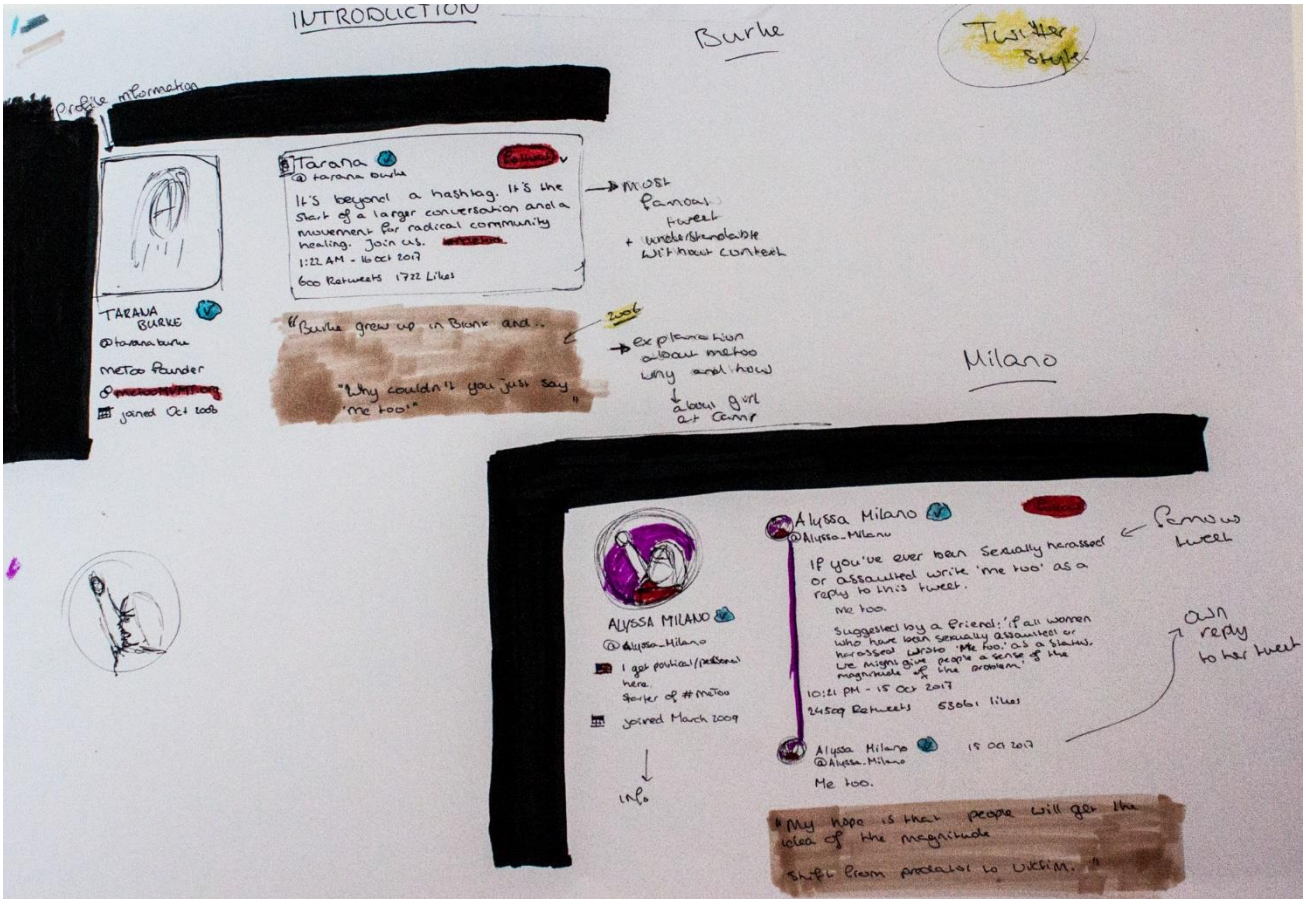


Figure 29 – Design: Introduction Tarana Burke & Alyssa Milano

At first, an introduction to the topic is being given by introducing the two people who started #MeToo: Tarana Burke and Alyssa Milano. This design can be found in figure 29. They are introduced by their twitter profile and an important tweet. For Alyssa, this is the tweet with which she introduced the hashtag and for Tarana, this a tweet where she states her support for the hashtag.

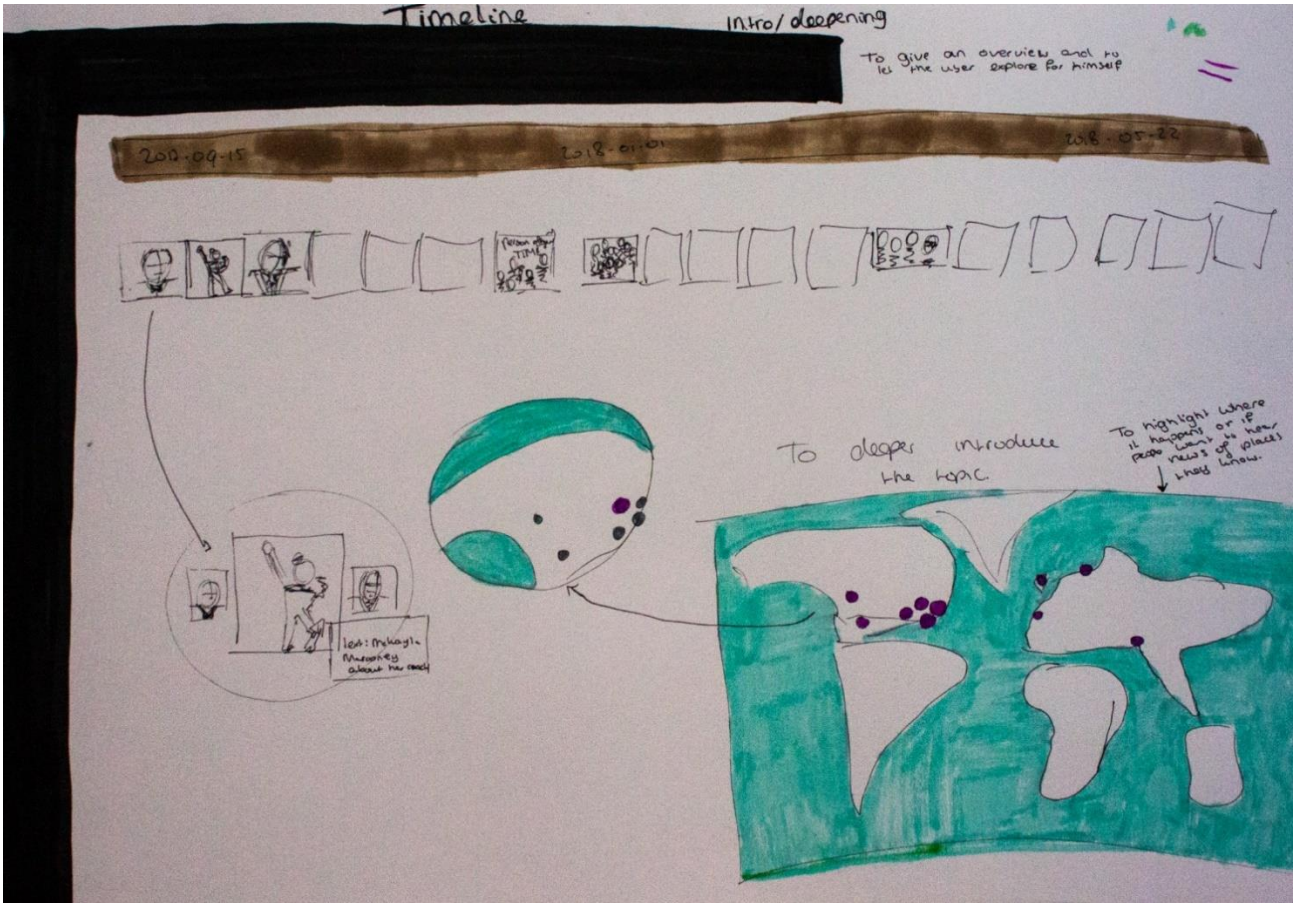


Figure 30 – Design: Introduction #MeToo via news

After this, as can be seen in figure 30, the user can explore more about #MeToo by scrolling through some important news. The news is organized on a timeline and can be filtered on location, so the user can also see what news happened at their hometown. This is done so people can see for themselves what happened and can choose whether they want to know more about the topic before going on. Later on, they can also go back from the magnitude tab if they see something interesting that they want to compare to the news. The introduction phase is then done and next will be deepening.

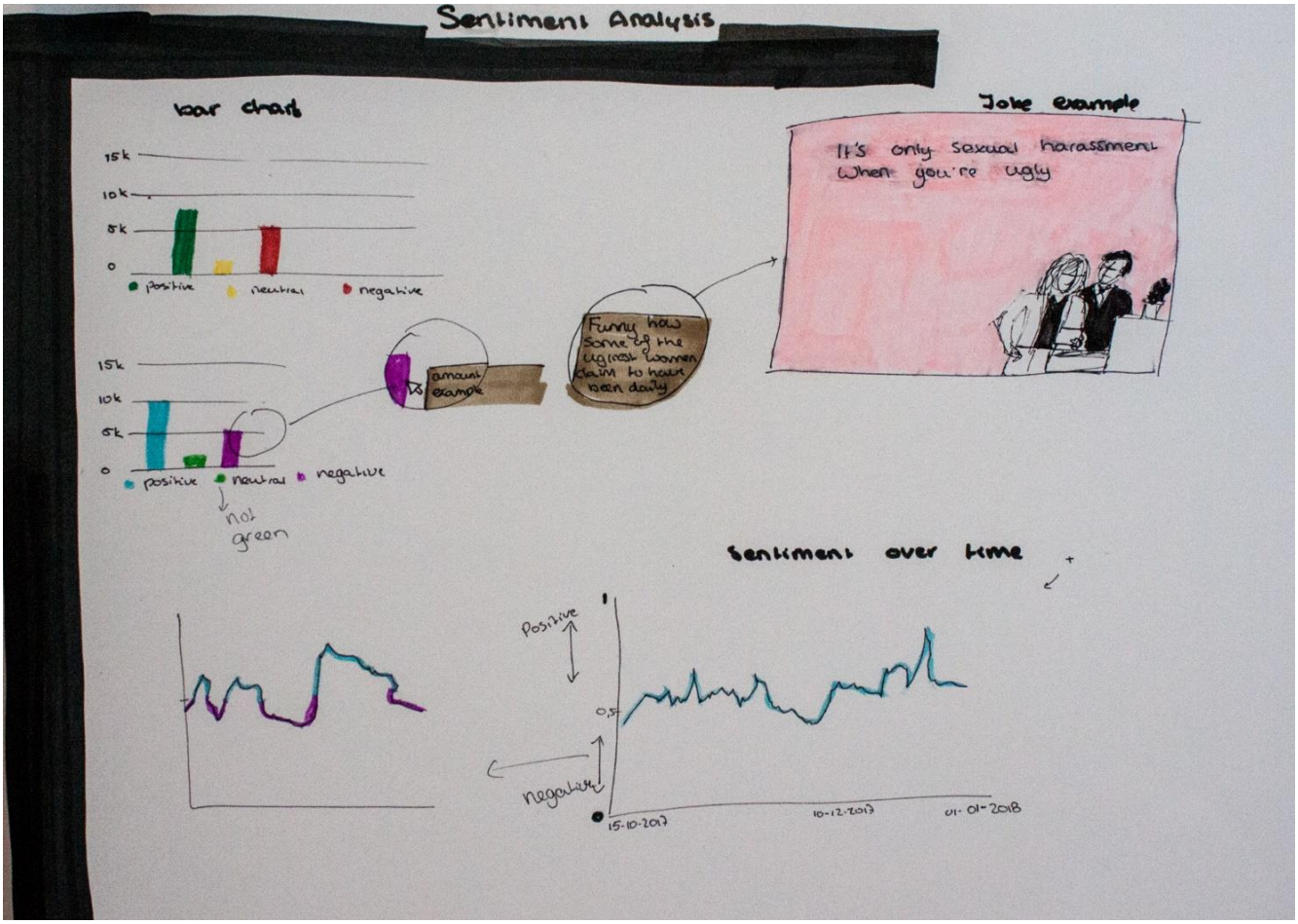


Figure 31 – Design: Deepening on the topic: sentiment analysis

The deepening will be done by a sentiment analysis, as shown in figure 31. This gives a brief introduction to the different opinions about #MeToo. This will be supported by jokes made about #MeToo, so both sides can be shown, and people can relate to that, as was found out with the survey. The sentiment over the entire timeframe will also be shown by a line graph. This was mostly done because of chapter 1 which states that line graphs can be very useful when dealing with large datasets, since mostly trends are interesting in such case. The ratio between the sentiment categories will be displayed in a bar chart, because of the different opinions regarding pie charts, as discussed in the background research.

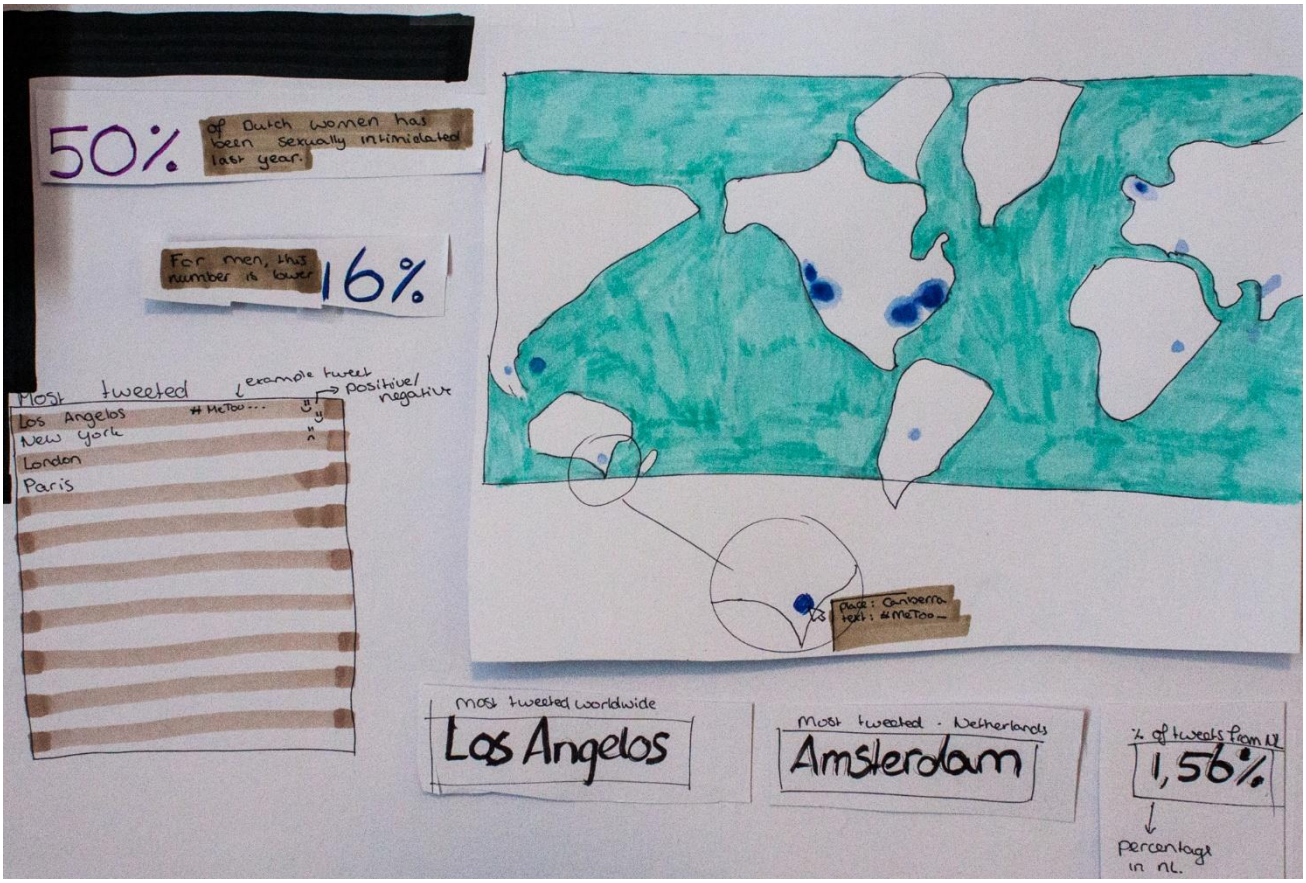


Figure 32 – Design: Further deepening: locations & in the Netherlands

The next dashboard, displayed in figure 32, is to give an idea about the magnitude of #MeToo. It displays a map where the locations of the people who tweeted are presented. The size then represents the number of tweets from that location and a table next to it will show the top cities. No exact numbers are shown here since location is opt-in on Twitter, so only 430 people (which is about 1,67%) allowed their location to be saved with their tweet. Giving these numbers would confuse many users who might think that only a few people tweeted at all. Next to these graphs, some numbers will be given about the number of men and women being sexually harassed in the Netherlands to let people face the numbers and also because people wanted to know more about #MeToo nearby, according to the survey. The top city in the world tweeting about #MeToo and the top city in the Netherlands tweeting about #MeToo are shown so people get an idea where it happens mostly. Lastly the percentage of tweets from the Netherlands is calculated. This is done mostly because, according to the survey, people wanted to know more about #MeToo nearby.

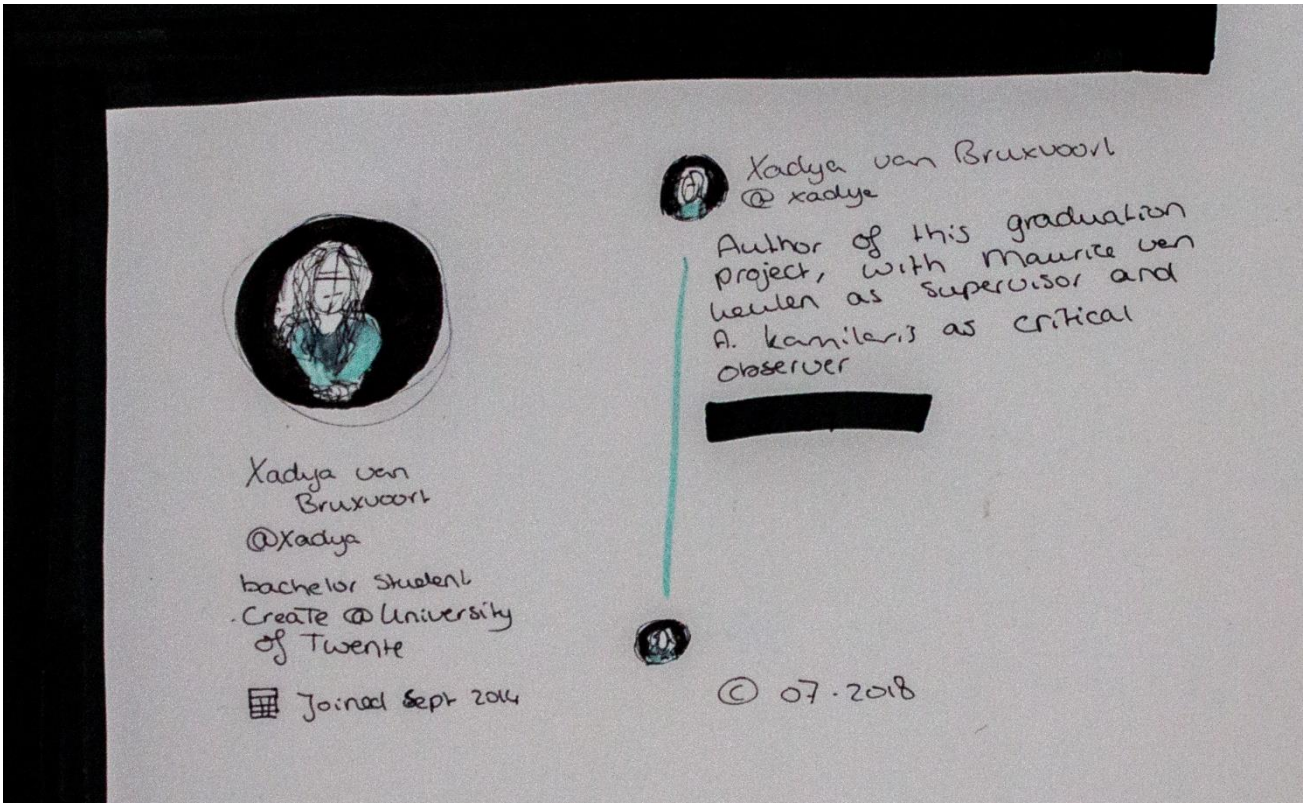


Figure 33 - End and introduction of the author

The last dashboard, in figure 33, is about the author of this graduation project, introducing her by, again, a Twitter profile. Here, the contributors to this project, Maurice van Keulen and Andreas Kamilaris are introduced as well. Finally, some extra details about the project are displayed here as well.

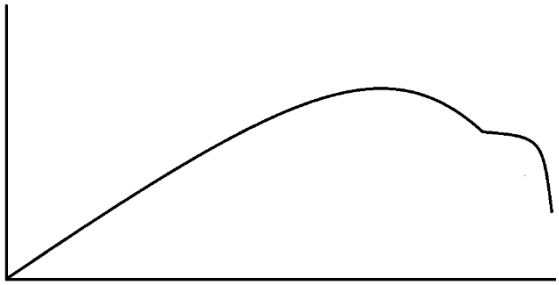


Figure 34 - Storyline of visualizations

With this order, the idea is to first let people get to know more about the topic, by introducing two important people and the news section, which they can explore themselves. After this, the deepening on these data comes, which is information that people probably do not know yet; this is the climax, as displayed by the peak in figure 34. Then the deepening goes on with information on location of Twitter users regarding #MeToo, which is information that they might know. The confronting numbers here is data that some of them might already know, but they maybe do not want to admit; this is the second climax, the second peak in figure 34. The ending is simple credits and the same style as the beginning. The colour scheme is consisting throughout the visualizations as well, by using the colours Alyssa Milano and Tarana Burke use on their Twitter profile, used in the visualizations again.

After making this in Power BI, testing will be done to find flaws to improve.

4.3 Ethical aspects

As explored a bit before, now the designs are ready the privacy and ethical part should be reconsidered. Here will be looked deeper into already introduced topics from previous chapters.

4.3.1 Private versus public data

Tweets are personal data and therefore should be handled carefully. When creating a Twitter profile, the user agrees to a set of terms and conditions which state the following: *'By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license authorizes us to make your Content available to the rest of the world and to let others do the same.'* [1]
While Twitter has legally covered the usage of these data in their terms and conditions and every user has to agree upon these (thus given consent), ethically nothing is said about it by the company itself.

On the other hand, Boyd & Crawford (2012) argue that just because data is accessible does not make it ethical. They explain this case with an example of a research conducted from 2006 until 2008 by a Harvard-based research group. This group gathered 1,710 college-based Facebook users to study how their interests and friendships changed over time (Lewis, Kaufman, & Christakis, 2008). They used mailing lists that included students' full name and searched for these names on Facebook. Afterwards, these Facebook data were shared with the world, so other researchers could use them as well. Zimmer (2010) used this example to perform a case study and concluded that the methodology used during this project lacked understanding of the privacy implications of research in social networking spaces. The result of this was that, according to Zimmer(2010), the privacy of the subjects of this study was threatened. This case raised many questions and resulted mainly in a dilemma: It might be unreasonable for researchers to ask for consent from all subjects, but it might also be problematic for researchers to justify their actions as ethical simply because the data are accessible.

When looking at the extremes, the Guardian published an article[11] about Raytheon, a US defence contractor, who developed the Rapid Information Overlay Technology software. With this software, freely accessible data from social networks and data related to a certain IP address could be combined such that a profile of a user could be made. This has not been sold to any clients, but it has been shared with the US government and industry.

It is clear that not one solution can be found to the question where the border between public and private data lies and that there are different views on what is acceptable. Since many different views are prevalent, it is very hard to find the 'right' one, as many topics within the field of ethics are. It is not only hard because of the different opinions, but also because what is right changes over time. This is clearly explained with an example. Take homosexually, this used to be a shame, because people were not designed to be homosexual as was thought by the broader public. Nowadays our society made a big change regarding this topic, since many countries allow homosexual weddings. We now think it is absurd for people to think about homosexuals as we did in the past, but this does not mean that people back then were morally wrong. They very much believed in homosexuality being wrong and it was also very normal to believe so. On the other hand, the author of this paper argued in a previous paper (van Bruxvoort, 2018, Privacy[12]) that privacy has not changed very much over the last years, what mainly changed was the scale on which privacy is happening. Where it used to be the info about a neighbour whose curtains were not closed completely, nowadays breach of privacy can include data about millions of people, such as during the Cambridge Analytica case in 2018[13].

[11] https://www.theguardian.com/world/2013/feb/10/software-tracks-social-media-defence
[12] Can be found on Blackboard under reflection 1 assignment
[13] https://www.theguardian.com/news/2018/mar/26/the-cambridge-analytica-files-the-story-so-far

4.3.2 Protection from potential harm

Another important question is how to protect subjects from potential harm. This question is already partly covered by the previous question; to look at private versus public data before using it, but is also important to protect subjects from potential harm, such as embarrassment, reputational damage or prosecution (Townsend & Wallace, 2016). It is hard for researchers to correctly oversee the damage certain research can do in the future. What is a difficult topic is to publish direct quotes of subjects, since these can often be retraced to the subject via search engines which on their turn might harm the subject. Townsend & Wallace (2016) argue that people posting on Twitter with a hashtag aim for a broader public and therefore risk of harm might not be present. McKee (2013) adds that it would be a good idea to paraphrase quotes to protect a subject, even though this can introduce bias since it interferes with the integrity of the data. She later states that paraphrasing does not have to happen when using Twitter data since it is apparent that the initiators expected, and in many cases, wished theses tweets to be quoted. This because they were speaking publicly to advocate a particular position. While that seems a bit blunt to say, I think it should at least be kept in mind while reflecting on the visualizations, to prevent users from getting 'witch hunted'.

An addition, while also looking at the recently updated GDPR[14], is to ensure data minimization; all data that are not necessarily needed for the research should be deleted. This is more of a precaution, since datasets with tweets may not be republished anyways.

Aside from these points, another serious issue with using Twitter data is that it is written, and therefore can easily be taken out of context, as mentioned by Bath, Kennedy, Whiteman et al. (2016). It might be that a certain tweet is sarcastic or ironic and then incorrect assumptions might be taken. What can also be the case is that the hashtag was used for another purpose than what it grew into. It might be that people using it, mainly in the beginning, used it in a more innocent way.

In general, it seems like there are some guidelines that state that directly quoting individuals might not be a good idea, but these do not seem to apply for tweets because of the different origin. Tweets including hashtag are put out there to voice an opinion, a statement, so they are public in its essence. Still I think it is good to minimize the use of such quotes and only use them directly when it is necessary.

4.3.3 Consent

Where regular research often includes a part where informed consent must be given by the subject, online research does not have such a field. Online research uses naturally occurring data rather than generated data. As mentioned before, all users agree to the terms and conditions when creating a profile. They also agree to their data possibly being used for research and therefore somehow give consent. The question arises whether people do this consciously or if they 'just tick the box and continue'. A survey performed by Deloitte concluded that 91% of consumers do not read the terms and conditions before installing apps, registering Wi-Fi hotspots, accepting updates and signing on to online services such as video streaming (Deloitte, 2017). This number lies even higher for consumers between the age of 18 and 35 (which is the target group of the graduation project research), namely 97%.

As often mentioned before, Twitter is often considered to be public, however Woodfield et al. (2013) mention that a tweet might be intentioned to only be shared with followers, which might consist from only a small group of friends and family. They also mention a counterargument, which is again the fact that all users must agree to the terms and conditions of Twitter before using the service.

4.3.4 Anonymization

Another guideline given by Townsend & Wallace (2016) states that it might be a good idea to anonymize the data. This is mainly because it could harm subjects. However, Webb et al. (2017) mention that Twitter's User

[14] https://www.eugdpr.org/

Development policy requires any reproductions of tweets to be done in full. They do add, however, that anonymization can be acquired by correcting spelling errors or removing or substituting word such that it does not affect the integrity of the analysis but provide a mean to shroud the identity of the poster. This will result in a few results when copy pasting the text in the Twitter search engine. Through this, contriving the original poster becomes a bit harder, but it is still fairly easy to find out. This means that anonymization can only really be achieved by leaving certain data out and not by changing or censoring data.

4.3.5 Discussion

While we will try to find a solution to these questions regarding this project, we thus also have to keep the different point of views and the current view on privacy with the expanding magnitude in the back of our minds.

The most common solution, as for example mentioned by Townsend & Wallace (2016), is that there needs to looked at the expectation of privacy regarding certain data. For example, data on a private forum for people who struggle with alcohol addiction would be considered private data, whereas an open Twitter discussion with the usage of a hashtag to familiarize your thought with the thoughts of other Twitter users would be considered public data. To determine whether some data can be seen as private or public, one thus has to look at the environment of the post and the settings of the source. Next to this, researchers should keep reflecting on the importance of using certain data and stop assuming that ethics boards will do the important work (Boyd & Crawford, 2012).

To determine how to use the data during this graduation, there can be looked at one of the case studies performed by Townsend & Wallace (2016). The second case study described in this paper is about collecting Twitter data about marijuana. The usage of marijuana is illegal in the country the research is performed in and there might be people under 18 in the dataset. This situation is somehow similar to the #MeToo discussion; quite a sensitive topic about a, sometimes, illegal act. The solution they describe is to see the data as public data, because it is posted on Twitter, where the default setting of tweets is public. Furthermore, the usage of hashtags implies that the users want to contribute to a community and therefore expect a greater number of people to see the data. Because the data is about a sensitive topic, ID handles (name, username etc.) will be removed. So, in short, it is acceptable to use Twitter data which include hashtags, since these can be considered public data because of the default Twitter setting and the usage of a hashtag.

It is arguable whether direct quotes can be used, because on one hand this case looks a bit like the, beforementioned, second case study of Townsend & Wallace (2016), which argues that direct quote may be used but only with informed consent of the subject. On the other hand, this case is fairly similar to their fifth case study as well, which argues that people using a certain hashtag use it because they want to let others know what their opinion is. Because of these people wanting to convince others, their quotes may be used directly. Because it is, again, quite a dilemma, there has to be looked for a middle way.

4.3.6 Regarding this project

There are two parts of the visualization that may harm the privacy of subjects. One are the tweets from a certain location, which are shown as plain text, another one is the gif with jokes about #MeToo.

To prepare the part with tweets to be ethically ‘correct’, the username, name and all other ID handles will be removed, except for the location. The location is opt-in and relatively hidden in the settings menu, so we can argue that people who put it on knew what they were doing. Therefore, we can, arguably again, consider this as informed consent, but we should not use this as an opportunity to show everything; these people should still be protected. Their tweets will be shown as plain text, because these people used #MeToo in their tweet, which indicates that they want attention for their tweet and also because they could have put their tweets on private in the same settings menu as the location option, so it seems like they knew

what they were doing. The content of the tweets will not be altered (censored) since this is not allowed by Twitter and also to not alter the data and ensure integrity. There should be made sure that the only tweets used for displaying the text are tweets dated after the hashtag became popular (15th of October 2017). Tweets that seem to not go about sexual harassment will be kept in the dataset to, again, secure data integrity.

The gif displaying jokes about #MeToo first included tweets of users with text such as: 'I'd love to be harassed, is there a group of women offering this?'. At first, they were in the visualization with their name, username and date but that seemed unethical since they might experience harm by it. This was then changed to only quoting these tweets, but that still did not feel right because people could still look them up and these people might not have known what the implications of their actions were when posting. So eventually it has been decided to only directly quote celebrities, since these people often aim to reach as wide an audience as possible (Townsend & Wallace, 2016).

Concluded, the visualization now only has people who are public figures, who are quoted directly and people who opted in for using their exact location to be sent with each tweet. Other direct quotes are now removed, even though they were posted in a public setting with the usage of hashtags to aim for a broader audience, to protect subjects. The rest of the visualization do not include (direct) personal data, but rather quantitative analysis. When looking at this project as a framework, these guidelines should be reconsidered for every topic and choice of graphs before publishing news visualizations.

4.4 Realization in Microsoft Power BI

After completing the designs, the visualizations were realized in Power BI and will be explained below.

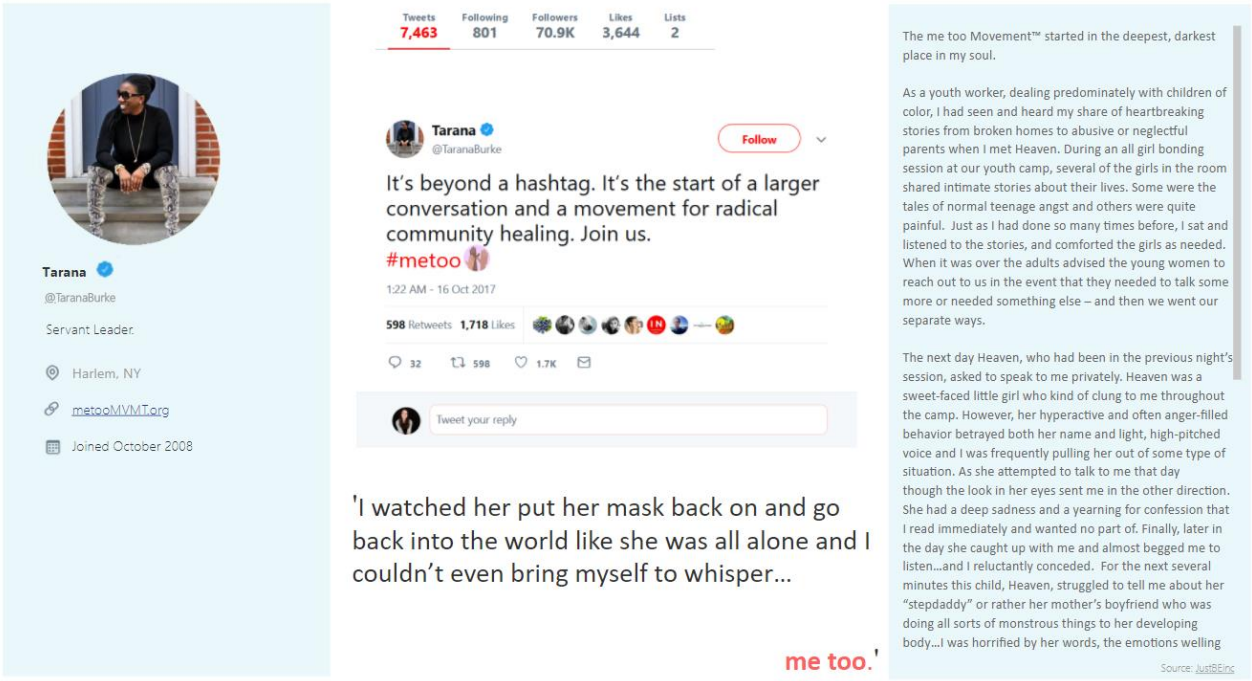


Figure 35 - Introduction of Tarana Burke dashboard

As mentioned in the previous chapter, the visualizations start with introducing the two greatest names regarding #MeToo and the first big name is, as also mentioned before in chapter 1, Tarana Burke; The founder of the Me Too Movement. The page is set up like a Twitter page. This means that on the left, the user profile is present, which is filled with the information Tarana provided on her Twitter. On top, her current Twitter statistics are shown. Then in the middle, one of her most famous tweets, which was placed at the beginning of #MeToo, is shown. On the left, her story about how she came up with Me Too as a phrase is presented and the most important sentence of that story is placed at the bottom. There is no user interaction on this page, except for maybe the scrolling on the right. This page is merely an introduction to the topic which some people might already know as well.

Figure 36 - Introduction of Alyssa Milano dashboard

The second introduction dashboard is for Alyssa Milano, who invented the hashtag. It has the same setup as the page of Tarana. The information on the left is of Alyssa's Twitter. On top are again the Twitter statistics of Alyssa. On the right her story is displayed as published on Time, with an explanation who she is and what she did. Lastly, in the middle her world-famous tweet is displayed, the one with which she started #MeToo.

One issue with this page is that the layout does not look exactly like Twitter's. This is mainly because Power BI's page format is different than Twitter's and scrollable dashboards are not a thing yet. This could be solved by publishing this on a webpage. Then both the URLs and the rest of the information can be displayed in a more Twitter-like style (including the 'banner').

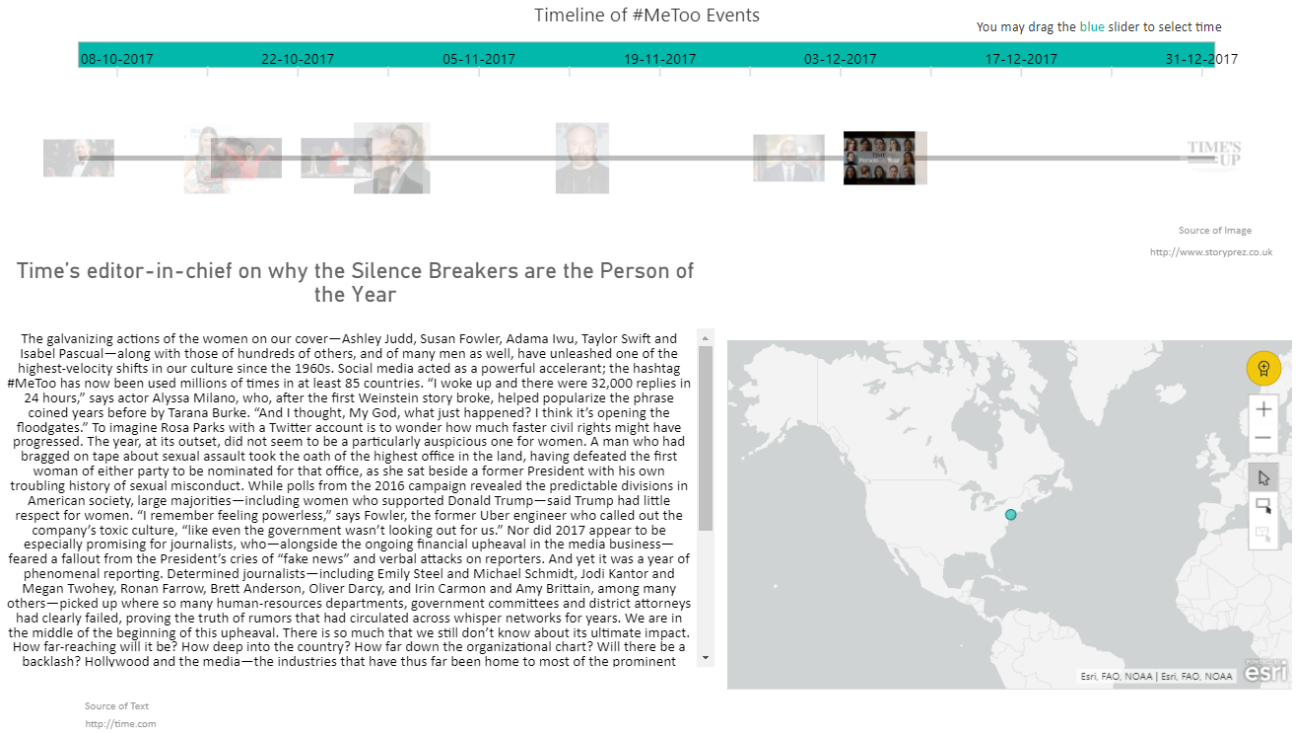


Figure 37 - News dashboard

In figure 37, the news dashboard can be seen. Here users can familiarize themselves with big news that has happened over the time period of October 2017 – December 2018. On this dashboard, people can choose for themselves what they exactly want to know about the topic. On top is a timeline, which shows the news articles over the time frame. These news articles can be selected by clicking. The time period can also be changed by dragging the sides of the blue bar of the timeline. When selected, the title and text of the article are displayed on the bottom left. On the bottom right, the location of the news can be seen. Users can also select the article by clicking on the map, since the survey concluded that people like to know about news from nearby.

For displaying the news article, a custom visual that was not available in the Power BI marketplace had to be downloaded called long text viewer[15]. Without this custom visual, the text would not be scrollable, thus not be able to be shown completely.

Improvements for this page lie mostly in the timeline. This is a custom visual from Power BI, which means that it was not originally made by Power BI, but by the community. For example, the dates on the blue bar run out of the bar on the right side. It is also not possible to decide how to show the dates, it can be displayed by months, but then these months will occur multiple times (see figure 38).

Figure 38 - Timeline bar when date format is changed to Month-Year

One last, a slightly bigger issue, is that images can only be imported and displayed when they come from a HTTP source. HTTPS is supported, but some bug in the timeline code prevented it from showing up.

[15] https://file.ac/bwAcEdqzh8M/LongTextViewer.0.3.0.pbiviz

At the time of writing, GitHub issues and Trello cards had been made to solve this issue, but they have not been solved yet. The workaround now is to only use images from an HTTP source, but these are less safe and maybe not the 'perfect' images for the news article.

Another improvement is that the text box right now is not able to have white lines. That is why, in this visualization, no paragraphs are present in the news article text. This make readability a bit worse than it could be with paragraphs.

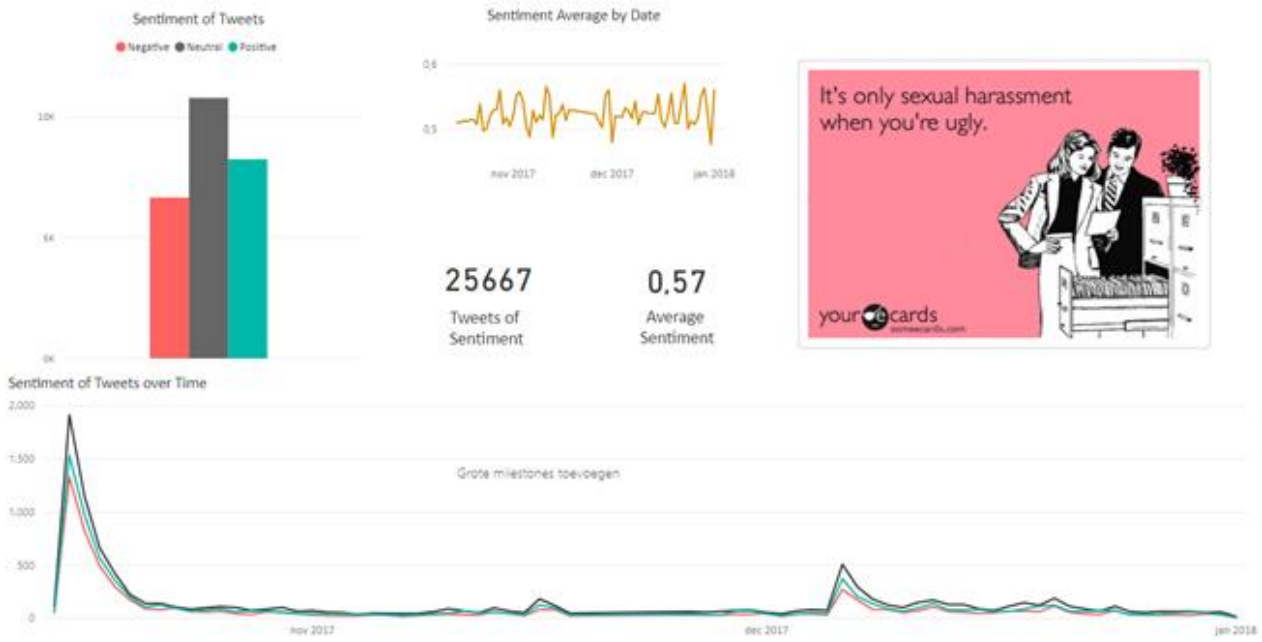


Figure 39 - Sentiment dashboard

The fourth dashboard that has been made is based on sentiment of tweets. The Microsoft Azure API on text analysis was used for this dashboard. This API can recognize sentiment based on text and assigns a sentiment score to every tweet in the dataset with an accuracy of 98%. This score differs from 0 to 1, where 0 is negative and 1 is positive. First these scores were used to make a bar chart. The bar chart shows the number of tweets that are negative (score from 0 to 0.3), neutral (score from 0.3 to 0.7) and positive (score from 0.7 to 1). Next to it, the total tweets in the dataset is displayed and a line chart with the average sentiment over time. On the right a gif with jokes made about #MeToo is looped. The gif was put in there because the survey resulted in people recognizing #MeToo mainly from jokes they or their friends make about it. Finally, at the bottom, a line chart with the number of tweets with a certain sentiment over time is shown.

The bar chart and the line chart are clickable, and users can select a sentiment category. All graphs, except for the gifs will then only display the results for that sentiment type. For example, if negative is chosen in the bar chart, the others will be greyed out in the bar chart, the line chart will only display the red line, the number of tweets will chart to the number of negative tweets and the yellow line chart will show the average negative sentiment over time (which is about 0.16).

An improvement here, before testing, is mainly the gif. It was planned to let there be an image which displays negative jokes only when selecting negative, but this was not possible in Power BI. The only thing possible in Power BI was to let one image be displayed with a sentiment type, but this meant that there could only be three images in total. Therefore, a gif has been chosen, because there can be as many images present as wished upon, but the downside is that this does not filter on sentiment type. It is actually not really possible to show gifs at all; this is only because of a bug in the program. The gif is essentially a line chart with no data where the plot area is changed into the gif. It would be very nice to make a table with negative, positive and neutral images which will randomly select one with the same sentiment as selected, once that is possible in Power BI.

Figure 40 - Magnitude dashboard

The fifth dashboard focuses on the magnitude of #MeToo. Top left are some statistics about sexual harassment as researched by Reuters Kenniscentrum voor Seksualiteit in the Netherlands. These are displayed because people, according to the survey, wanted to know about #MeToo happening nearby and people thinking that sexual harassment is not a very big problem in the Netherlands. Top right, a map is shown with locations of people tweeting with #MeToo and on the bottom left the corresponding tweets are shown. People can select locations on the map or tweets/locations in the table and both visualizations will filter accordingly. This way, people can see the tweets in locations of their interest. Lastly, on the bottom right, the top locations with tweets in the Netherlands and in the world are shown, just like how many tweets are from the Netherlands. This is shown in percentages because there are not many people who opted in for Twitter sending their location with each tweet, so displaying the number of tweets would show a lower number than the truth. With displaying percentages, we assume that the sprinkler used for collecting tweets randomly added tweets to the set, so this percentage is about the same as the percentage would be when everyone would include their location.

The only 'issue' with this page is that not so many people included their location, so there are only about 430 tweets displayed in about 320 locations. Compared with a dataset of 25667 tweets, this is about 1.67%.

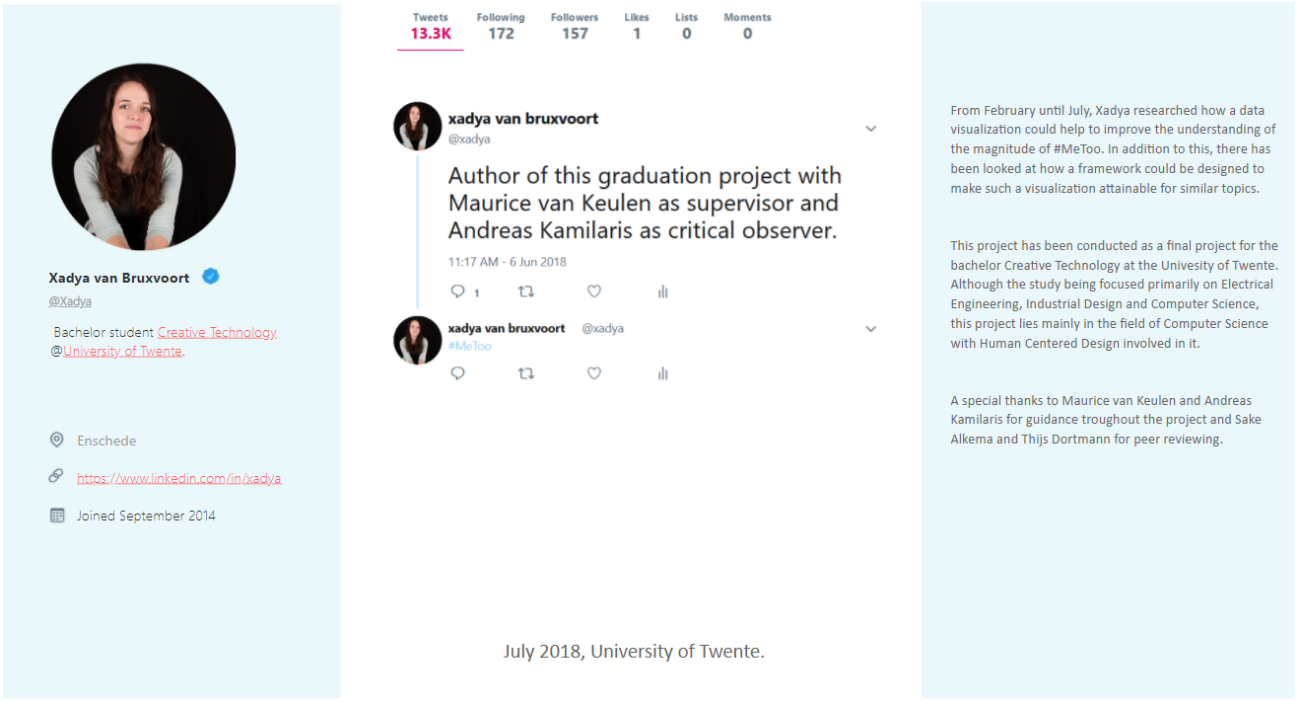


Figure 41 - Credits dashboard

The last page is for crediting the author, supervisors and people who gave extensive feedback. This page is in the same style as the first two pages and has a small description about the project on the right.

4.5 Testing

After realizing the visualizations in Power BI, it was time to test the visualizations. This was done by showing 7 people the visualization and asking them to think out loud and give feedback whenever they felt they missed something or when they did not understand something.

4.5.1 User test

From the first user test, quite some things were found out. First of all, the conversion from Power BI desktop to Power BI browser was not a smooth as hoped for. There were some minor bugs found by the testers, such as the images not displaying correctly, some text boxes that were too small (and therefore displayed a scroll bar on the side) and some tiles that were generally unclear. There were, however, a few bigger issues.

- The first issue was that it was not clear to people who Tarana Burke was and some people were not patient enough to read through the entire text. Other people mentioned that it was not clear at the beginning of the text that the story is told by Tarana Burke and they suggested an introduction first. This can be fixed by indeed introducing her a bit more and, maybe in bold, adding a small (tl;dr) title on top of the text box.
- Secondly, there were some issues with the image timeline in the news dashboard. This would disappear when hovered over and most people did not know how to filter. It was suggested by one person to add a small arrow near the scrollable bar to let people know where they should drag (since this is very precise).
- On the third place, some people did not know what was meant by ‘sentiment’, so a small explanation will be added to the page. This should also include the thresholds used to define

positive, neutral and negative categories. Similar to this, there should be added that 1 means positive and 0 means negative at the yellow line graph.

- While everyone laughed at the gif displaying jokes, some people mentioned that they expected the jokes to change depending on the selected sentiment. Although this was the original idea, as was mentioned before, this was not yet possible in Power BI at the time of making the visualization. To make the idea behind the gif clearer, they suggested to add a title or subtitle that explained some more.
- At this same page, it was also suggested to make the line graph at the bottom into a logarithmic scale, instead of a linear one. This is to let the user see more easily what the changes are to also determine really small peaks, instead of seeing one large peak and two smaller ones.
- Some extra addition to the line graph were suggested, such as comparing the three lines with the total tweets and adding news to the peaks of the line graph. There will be looked into these.
- The last bigger change was suggested for the magnitude dashboard, where someone mentioned that it was not clear to let the map and table filter with each other while they were situated diagonally from each other, instead of horizontally or vertically. There was added that it might be nice to have all static number on one side, while dynamics are grouped as well. It might be good to situate all statics, i.e. percentages and top tweeted cities, on the left, while having dynamics on the right. This will be changed in a new version.

There were also some general observations during the user tests. Users were afraid of clicking at first, until they reached the news section. Here the text displays: 'please select a news article', so users would eventually click on an article. After this dashboard, most users clicked on practically every aspect of a dashboard. This dashboard is really needed to get users comfortable with the visualizations. Next to that, most users became very quiet when they reached the magnitude dashboard, since the red numbers hit them quite hard. Almost all of them asked the question whether these number could be true. The most enthusiastic reactions were received about the news dashboard and the magnitude dashboard, but overall there were no big issues found during user testing. Most remarks were small and easily solvable and in general people were very positive.
Some critical notes on the user test are that it was very hard to find people who wanted to say something about the visualization, instead of about the topic itself. Some users talked a lot about what their opinion about #MeToo was and it felt like they did not really want to know more about it, since they already had an opinion ready for themselves.

4.5.2 Improvements

After testing, improvements were made according to the results of 4.5.1 and will be discussed below per page again.

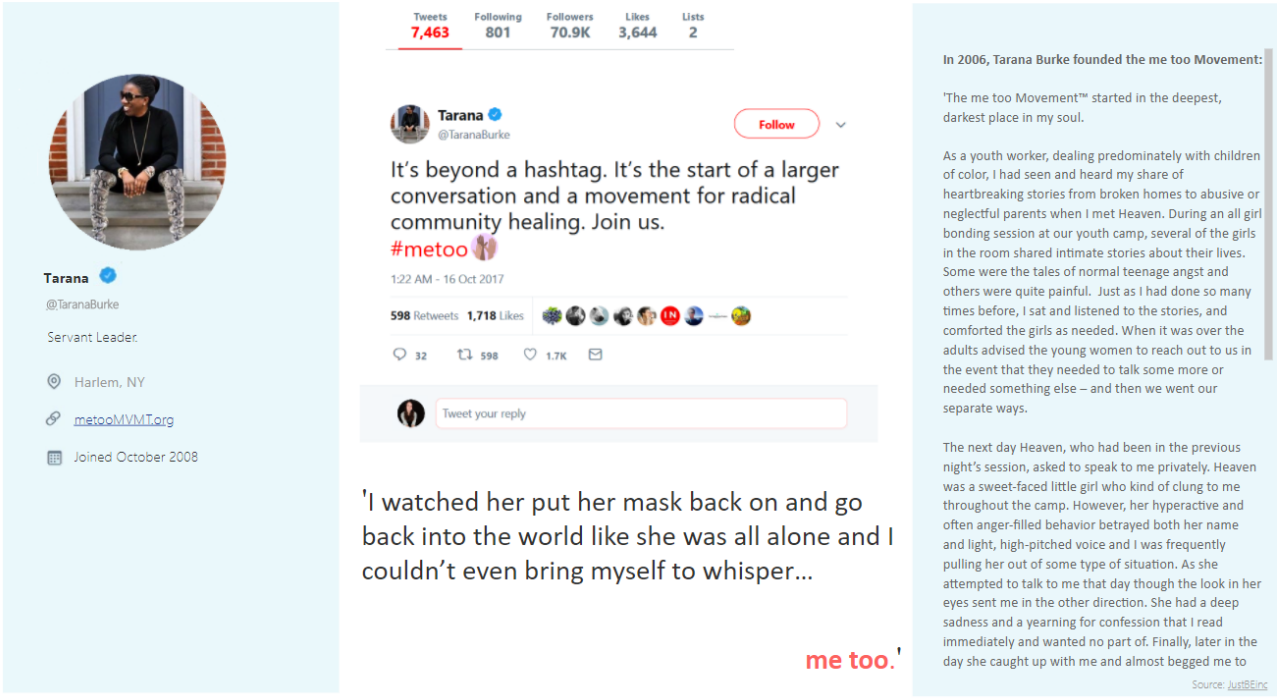


Figure 42 - improved introduction page of Tarana Burke

Not much had changed about Tarana's page, mainly the small, bold, header which introduced Tarana Burke as founder of the Me Too Movement. Next to this, some small changes were made regarding the size and placement of the text box (to make it equal to the placement with the page of Alyssa) and to change the font. The font, at first was meant to be the same as Twitter's font (Segoe), but people found this font ugly or unreadable. Therefore, this has been changed to Calibri; the same font as used in the rest of the visualization.

Figure 43 - Improved introduction page of Alyssa Milano

For Alyssa's page, the same minor changes were made: changing the font and placement of the text box. Next to this, this page has remained the same.

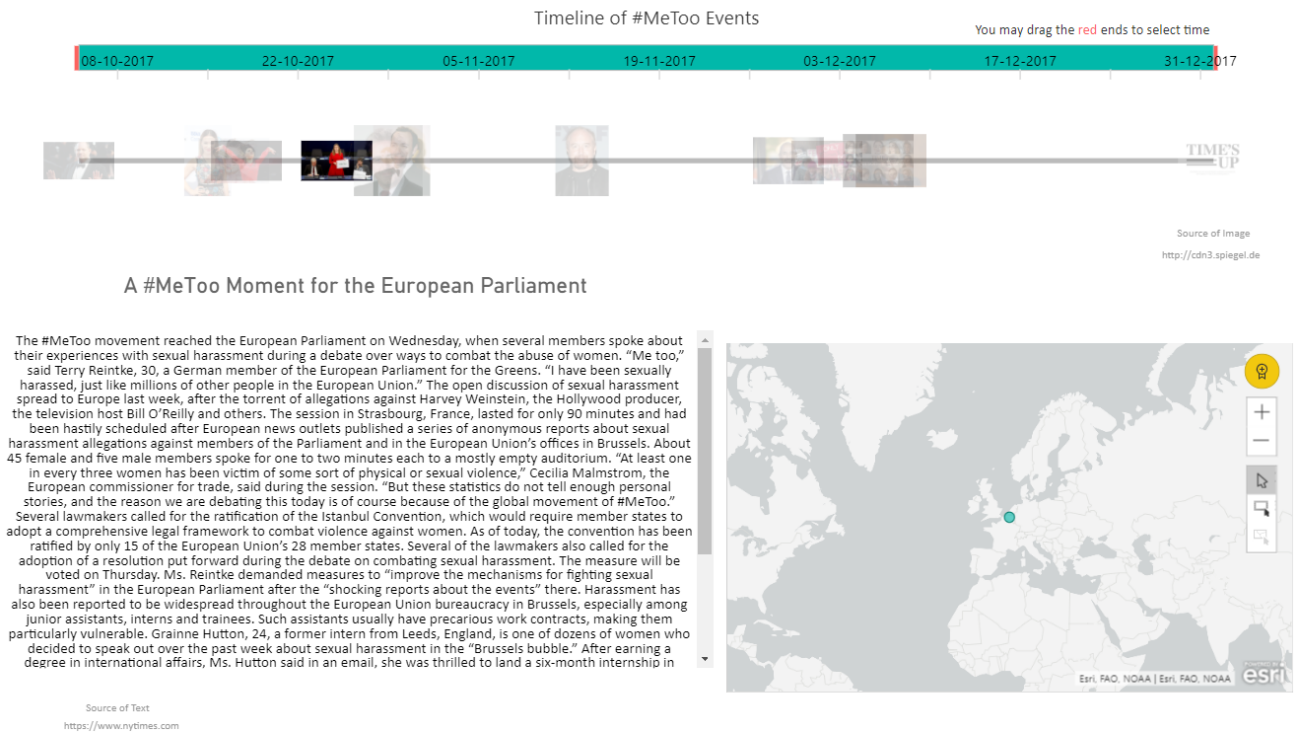


Figure 44 - Improved News Dashboard

To the news dashboard, minor changes were made. The timeline now has red ends, to make the filter option clearer. Other than that, no changes were made, except adding another news article. The overview of this dashboard can be found in figure 44.

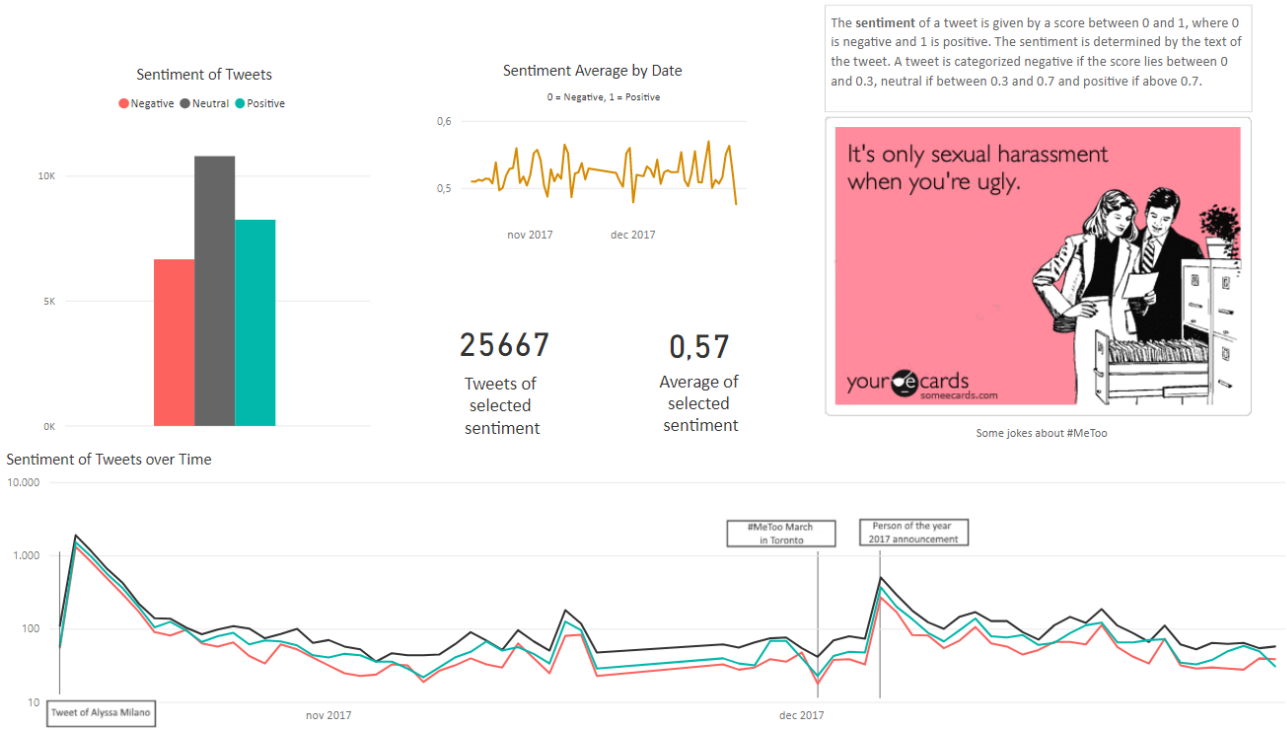


Figure 45 - Revised sentiment dashboard

The sentiment dashboard is probable the dashboard where most changes were made. The entire layout changed to make the topic of time clearer. Now users see that as one of the first parts on this dashboard (since it is on top and in the brightest colour, see 2.2) and therefore the concept of time is clear earlier in the process. Next to that, a brief explanation of sentiment was added to this page, just like the legend for the yellow graph. With these changes, users should be able to understand more easily what sentiment exactly means in this context. Next to this, the – very often requested – news articles were added to the peaks on the timeline at the bottom. This is so users can more easily place the peaks in context with big news. There were not many peaks and some peaks has absolutely no news connected to them, but it was added for three cases nonetheless. There were also some minor changes, such as a small caption was added to the gif and a small textual change was made to the numbers in the middle of the page so is it clearer that the number are based on what is selected. Lastly, January has been excluded because only a few tweets were available of January first and it was misleading.

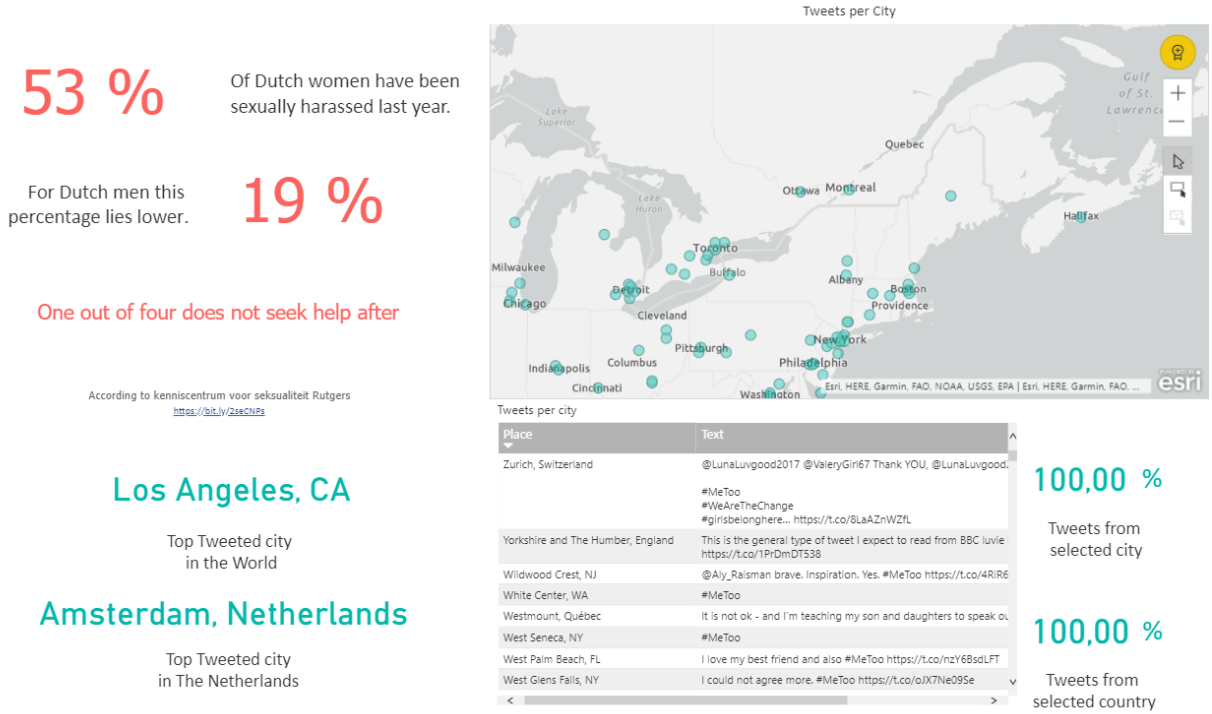


Figure 46 - Revised magnitude dashboard

What was changed to the magnitude dashboard was mainly the layout. At first the table and map were placed diagonally from each other, but with this layout it was not clear that they belonged to each other and filtered with each other. Now they are both placed on the right side and this became clearer. Next to that, there were some static statistics about the number of tweets from the Netherlands et cetera. In turned out that people wanted to know statistics about more countries than only the Netherlands, so this has been changed to statistics based on what is selected. The static statistics are still there, on the bottom left. They are in the same colour as the map and table to indicate that they are connected. The general statistics are displayed in red so it is the most striking colour and so it is clear that these numbers are not related to the blue data.

No changes were made to the credits dashboard, the final version can be found in figure 41.

4.6 Conclusion

In general, people were very enthusiastic during the user test. There was, of course, a bias since only people were asked who wanted to know something about the subject. This was done because otherwise people would give feedback on their opinion about #MeToo and not on the visualization itself. It is a question whether or not the second (not asked) group of people would have clicked on a data visualization with this topic when found on the internet on the first place. However, the people that did the user test, spent about 15 minutes with the visualization and most of them started discussing #MeToo with the author or with others after testing. This was a good sign, since this indicates that people were affected by the visualization. No graphs were found confusing during user testing, which is a very good sign. There were, as also mentioned in 4.4, some bugs with the software, but these will be probably resolved in following updates. This is mainly because Power BI as it is right now is very up and coming, especially compared to other tools.

In general, there can be said that this visualization made an impact on people and that people were very enthusiastic towards the visualization.

Chapter 5 - Implications and Conclusion

In this chapter there will looked at a generalization for this project. Next to this, a conclusion will be drawn regarding this project, recommendations for future work will be made and acknowledgements will be given.

5.1 Generalization

This project could be further turned into a framework for making comparable visualizations about similar subjects. The procedure of this project can be followed with some adjustments. These subjects should have some factors in common with #MeToo in order for them to work with this framework. First of all, the subject needs to be sensitive and something that people underestimate happening, otherwise the red numbers on the magnitude dashboard will not work. Secondly, there should be tweeted about the subject. This could be replaced by other social media like Facebook, but it should at least be on a type of social media. If Twitter is used, about 2000 tweets must be collected in order to have about 100 locations on the magnitude dashboard. Thirdly, there should be news articles about the subject available, both for the news dashboard and for the sentiment dashboard. The introduction and credits dashboards can be altered according to the subject in order to fit.

Next to these guidelines, the visualizations as described in the realization phase are tested. Three dashboards were made: news articles, sentiment analysis and magnitude. These dashboards could however be divided differently and could include different graphs. Per situation there should be looked at what makes an impact and how to guide users through a story. There can be concluded that introduction dashboards are really useful for making the user feel comfortable with the software. Secondly, it is good to have a climax where people are shocked by the result. This was achieved by first making users laugh about jokes and then showing the reality. This worked very well, since users were shocked, so it is advised to keep a shock factor when using this framework. What should not be forgotten are the ethical considerations as described in 4.3.

5.2 Conclusion

During this thesis, there has been looked into making a data visualization to make the magnitude of #MeToo clearer to the general public. The end product is a data visualization in Power BI. The product itself can be used for making an impact on users with the topic being #MeToo or on very least to make people think about #MeToo and sexual harassment more.

At first, there has been looked at what already existed, and it was concluded that there were not so many visualization projects available yet. This is mainly since the topic is very new (October 2017). This meant that there were not so many handles at the beginning of this project. In order to generate handles, there had been looked at some infographics available and the topics these infographics were about were used as input for making survey questions. This survey was then held, and the results were turned into general ideas and topics people found interesting. Then there was looked at the feasibility of these and decided upon one idea. This was then made into a visualization using Power BI, with the guidelines provided in chapter 1 in the back of mind. User tests have been conducted with the result and the visualizations have been adjusted accordingly.

Since people were so enthusiastic, or better to say people became quieter near the end, and also since people started discussing afterwards, the impact has been made. People know more about the topic after using the product and mostly they combine new info with things they already knew about #MeToo. The magnitude became clearer but can become even clearer as will be discussed in future work.

The end product could be further turned into a framework for making visualizations about similar subjects. These subjects should have at least three factors in common with #MeToo in order for them to work with this framework. These factors are: a sensitive and underestimated topic, a topic that is happening on social media, and a topic for which are news articles available. Next to that, there should be looked at the story and other ethical implications for different subjects.

5.3 Recommendations for future work

A few recommendations can be made for further improving this work. First these recommendations regarding the data will be made and later regarding the entire project.
For gathering the data, a sprinkler API was used. This API collects, as mentioned before, a random set of 1% of total tweets. However, nobody knows how this random 1% is selected and there is very little documentation available. Therefore, it might be an idea to look into this and maybe collect all tweets or to collect tweets with use of a garden hose (which collects 10% of all tweets).

Another nice addition to this project would be to also include data from 2018. A lot has happened in 2018, such as the Oscars[16], Cannes film festival[17] and the Eurovision Song Contest[18]. It would be good to include these together with tweets of this time period to show people that #MeToo is not a (quickly dying) hype but happening over a longer timespan now. This would also be good, so people recognize news articles because they heard of these articles recently in the media.

There should be looked into not only filtering tweets on #MeToo, but also on #metoo, #TimesUp, Me Too and metoo for example. It might be that there are some interesting results from also filtering these tweets. Next to that, a mentioned before, it might be that this project has a bias towards more experienced users since they are able to use a hashtag, whereas more inexperienced users maybe do not know how to use hashtags. This would be solved by also filtering on the other keywords.

Regarding the project itself, there are three subjects in which could be looked into. The first one is that the dashboard as it is now, would be easier to use if turned into a scrollable webpage. This because a webpage would be easier to find on the web and easier to link to. Next to that, it might feel a little definitive to click on a new dashboard, scrolling to the next dashboard might be less of a threshold.

The second subject is regarding the sentiment analysis; as of now it places tweets into a category based on the content. #MeToo is about a negative topic, but the movement itself is positive. A tweet like 'It is breaking my heart seeing sexual harassment happening so often. This shit has to stop. #MeToo' Gets a score of 0.28, being negative. The tweet itself is negative, but it is a positive thing towards the movement that the tweet is being posted. The question arises whether it really is negative then and what the meaning of negative is in this case. This could be solved by maybe using another API or by developing one that is able to recognize false positives and false negatives.

Lastly, when working on this project, many people bluntly shouted their opinions about #MeToo at the author. She was prepared but did not expect this to happen as often and bluntly as it did. If wanting to further investigate this topic or use this framework, one should be prepared for this.

5.4 Acknowledgements

First of all, a special thanks is more than in place for Maurice van Keulen, supervisor of this project. Where sometimes he would leave me with more questions than answers, he would also provide a logical way of thinking about problems I encountered during this project.

[16] https://www.bbc.com/news/entertainment-arts-43253715
[17] https://www.vanityfair.com/hollywood/2018/05/cannes-2018-metoo-debate
[18] https://www.bbc.com/news/av/entertainment-arts-44073911/netta-meet-eurovision-2018-s-metoo-voice

Secondly, Andrea Kamilaris, critical observer, provided a great help in sanity checking me and my insane planning. Without this I could not have started in such a peaceful manner as I could now.

Thirdly, the Power BI community deserves a big thank you as well. They helped me when I was really stuck and some people on this platform made time to share their expertise on the query language with me, for which I am very grateful.

Lastly, I would like to thank Sake Alkema and Thijs Dortmann for taking the time and effort to give feedback on many parts; from colour choices up until spellchecking the entire report, thank you.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., & Tao, K. (2012). Semantics + filtering + search = twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media - HT '12* (p. 285). https://doi.org/10.1145/2309996.2310043

Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016). Big data visualization: Tools and challenges. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 656–660). https://doi.org/10.1109/IC3I.2016.7918044

Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. *Geographic Information Systems: Principles, Techniques, Management and Applications*, 253–266.

Bath, P., Kennedy, H., Whiteman, N., & Bates, J. (2016). The Ethics of Internet-based and Social Media Research focusing on the ethical challenges of internet-based research, with a particular focus on research using social media, (July). Retrieved from https://www.sheffield.ac.uk/polopoly_fs/1.644904!/file/Report_Ethics_of_Social_Media_Research_Jul16.pdf

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Cottam, J. A., Lumsdaine, A., & Weaver, C. (2012). Watch this: A taxonomy for dynamic data visualization. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 193–202). Seattle: IEEE. https://doi.org/10.1109/VAST.2012.6400552

Deloitte. (2017). *2017 Global Mobile Consumer Survey: US edition*. *Deloitte*. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-tmt-2017-global-mobile-consumer-survey-executive-summary.pdf

Ezyinsights. (2017). #MeToo – The Viral Event of 2017 – EzyInsights. Retrieved March 29, 2018, from https://ezyinsights.com/metoo-viral-event-2017-1/

Fisher, D. (2014). Visualization of Big Data. Retrieved from https://www.chalmers.se/en/areas-of-advance/ict/events/Documents/Danyel Fisher_Visualization of Big Data.pdf

Gorodov, E. Y., & Gubarev, V. V. (2013). Analytical Review of Data Visualization Methods in Application to Big Data. *Journal of Electrical and Computer Engineering*, *2013*, 1–7. https://doi.org/10.1155/2013/969458

Healey, C. G. (1998). On the Use of Perceptual Cues and Data Mining for Effective Visualization of Scientific Datasets. *Proceedings Graphics Interface*, 177–184. Retrieved from https://www.csc2.ncsu.edu/faculty/healey/download/gi.98.pdf

Hendrikse, B., Habib, M. B., & Van Keulen, M. (2017). ISIS in the Eyes of the Dutch. Retrieved from https://ris.utwente.nl/ws/portalfiles/portal/14020206

Kelleher, C., & Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, *26*(6), 822–827. https://doi.org/10.1016/J.ENVSOFT.2010.12.006

Kozak, M. (2010). Basic principles of graphing data. *Sci. Agric. (Piracicaba, Braz.)*, *67*(4), 483–494. https://doi.org/10.1590/S0103-90162010000400017

Lewis, K., Kaufman, J., & Christakis, N. (2008). The taste for privacy: An analysis of college student privacy settings in an online social network. *Journal of Computer-Mediated Communication*, *14*(1), 79–100. https://doi.org/10.1111/j.1083-6101.2008.01432.x

McKee, R. (2013). Ethical issues in using social media for health and health care research. *Health Policy*.

https://doi.org/10.1016/j.healthpol.2013.02.006

Moe, H., & Larsson, A. O. (2012). Methodological and Ethical Challenges Associated with Large-scale Analyses of Online Political Communication Methodological Challenges Associated with Data Gathering. *Nordicom Review*, *33*(1), 117–124. Retrieved from https://s3.amazonaws.com/academia.edu.documents/33000820/359_moe_larsson.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1526390195&Signature=Ev2udx%2BuPynRQ65TF5WzEPzMSuQ%3D&response-content-disposition=inline%3B filename%3DMethodological_and_ethical_challenges_wi.pdf

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of Social Media Research: Common Concerns and Practical Considerations. *Cyberpsychology, Behavior, and Social Networking*, *16*(9), 708–713. https://doi.org/10.1089/cyber.2012.0334

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing social networks. In *Proceedings - IEEE Symposium on Security and Privacy* (pp. 173–187). https://doi.org/10.1109/SP.2009.22

Nguyen, D., Trieschnigg, D., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., & de Jong, F. (2014). Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. *Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014, Dublin, Ireland*, 1950–1961. Retrieved from http://www.dongnguyen.nl/publications/nguyen-coling2014.pdf

Prasetyo, N. D., Hauff, C., Nguyen, D., Broek, T. van den, & Hiemstra, D. (2015). On the Impact of Twitter-based Health Campaigns : A Cross-Country Analysis of Movember. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi)* (pp. 55–63). Retrieved from http://wwwhome.cs.utwente.nl/~hiemstra/papers/louhi2015.pdf

Savoska, S., & Bocevska, A. (2016). Data Visualization in Business Intelligent &Analysis – Analysis of First Positioned Tools According to Gartner's Magic Quadrant in Ability to Execute. https://doi.org/10.20544/AIIT2016.29

Schlegel, K., Sallam, R., Yuen, D., & Tapadinhas, J. (2013). Magic Quadrant for Business Intelligence and Analytics Platforms. *Gartner Research*, *2*(February), 1–27. https://doi.org/G00239854

Segel, E., & Heer, J. (2010). Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

Setlur, V., & Stone, M. C. (2016). A Linguistic Approach to Categorical Color Assignment for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 698–707. https://doi.org/10.1109/TVCG.2015.2467471

Silva, S., Sousa Santos, B., & Madeira, J. (2011). Using color in visualization: A survey. *Computers and Graphics (Pergamon)*, *35*(2), 320–333. https://doi.org/10.1016/j.cag.2010.11.015

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, *5*(1), 61–77. https://doi.org/10.1002/acp.2350050106

Sun, B. (2017). *A literature study of bottlenecks in 2D and 3D Big Data visualization*. Blekinge Institute of Technology. Retrieved from http://www.diva-portal.org/smash/get/diva2:1155657/FULLTEXT02

Time's Up. (2017). Time's Up Now. Retrieved March 27, 2018, from https://www.timesupnow.com/

Townsend, L., & Wallace, C. (2016). Social Media Research: A Guide to Ethics. *University of Aberdeen*, 1–16. Retrieved from https://www.gla.ac.uk/media/media_487729_en.pdf

Vande Moere, A. (2004). Time-Varying Data Visualization Using Information Flocking Boids. In *IEEE Symposium on Information Visualization* (pp. 97–104). Austin: IEEE. https://doi.org/10.1109/INFVIS.2004.65

Wang, L., Giesen, J., McDonnell, K. T., Zolliker, P., & Mueller, K. (2008). Color design for illustrative visualization. In *IEEE Transactions on Visualization and Computer Graphics* (Vol. 14, pp. 1739–1746). https://doi.org/10.1109/TVCG.2008.118

Wang, L., Wang, G., & Alexander, C. A. (2015). Big Data and Visualization: Methods, Challenges and Technology Progress. *Digital Technologies*, *1*(1), 33–38. https://doi.org/10.12691/dt-1-1-7

Webb, H., Jirotka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., … Burnap, P. (2017). The Ethical Challenges of Publishing Twitter Data for Research Dissemination. In *Proceedings of the 2017 ACM on Web Science Conference - WebSci '17* (pp. 339–348). https://doi.org/10.1145/3091478.3091489

Woodfield, K., Morrell, G., Metzler, K., Blank, G., Salmons, J., Finnegan, J., & Lucraft, M. (2013). Blurring the Boundaries? New social media, new social research: Developing a network to explore the issues faced by researchers negotiating the new research landscape of online social media platforms. *NCRM Networks for Methodological Innovation Report*, 1–23. Retrieved from http://eprints.ncrm.ac.uk/3168/1/blurring_boundaries.pdf

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325. https://doi.org/10.1007/s10676-010-9227-5

Appendix

A - Survey

#MeToo

#MeToo became a phenomenon in the media last year.

This survey is for preparation of a bachelor assignment of Creative Technology at the University of Twente. The results of this survey will be used for the background research of the making of a data visualization about #MeToo.

These results will be kept anonymous at all times and will only be used in the report, guiding the data visualization. If you feel uncomfortable with this, please do not fill in the survey.

Have you heard about #MeToo or the Me Too movement?

- ○ Yes
- ○ No

Can you describe #MeToo?

What is your age?

- ○ 18-25
- ○ 25-35
- ○ 35-45
- ○ 45-55
- ○ 55-65
- ○ 65+

What is your gender?

- ○ Male
- ○ Female
- ○ Prefer not to answer
- ○ Other

Where have you seen #MeToo?

- ☐ In the news
- ☐ Instagram
- ☐ Facebook
- ☐ Snapchat

☐ Twitter ☐ None
☐ Other, please specify:

Which of the following social media are you on?

☐ Facebook ☐ Twitter
☐ Instagram ☐ Snapchat

I use hashtags

○ Never ○ Once a year
○ Once a month ○ weekly
○ Daily

Which of the following have you heard about regarding #MeToo?

☐ TimesUp ☐ The silence breakers
☐ Kevin Spacey ☐ WomenWhoRoar
☐ MoiAussi ☐ Harvey Weinstein
☐ Tarana Burke ☐ YoTambién
☐ McKayla Marooney ☐ Alyssa Milano
☐ None of these

Retweet – When a tweet of another user is reposted or forwarded. Often people do this for political statements with which they agree or facts that they want more people to know.

Please answer how much you agree with the following statements	Importance				
	Strongly disagree	**Disagree**	**Neither disagree nor agree**	**Agree**	**Strongly agree**
I think #MeToo is **not** happening a lot in my industry (e.g. study/job)	1	2	3	4	5
#MeToo is less popular in western countries	1	2	3	4	5
I think there are more original tweets than retweets about #MeToo	1	2	3	4	5
I think there are more replies than retweets about #MeToo	1	2	3	4	5
I think women are more negative regarding #MeToo	1	2	3	4	5

I would like to know what my favourite famous persons (e.g. actors, singers etc.) think about #MeToo	1	2	3	4	5
I think there are more original tweets than replies about #MeToo	1	2	3	4	5
I think #MeToo is just a hype	1	2	3	4	5
I do not want to know more about #MeToo	1	2	3	4	5

Please answer how much you agree with the following statements	Importance				
	Strongly disagree	**Disagree**	**Neither disagree nor agree**	**Agree**	**Strongly agree**
I think #MeToo is more popular in Europe than in America	1	2	3	4	5
I think women are more positive regarding #MeToo	1	2	3	4	5
I would like to know what people in the Netherlands think about #MeToo	1	2	3	4	5
#MeToo is more popular in western countries	1	2	3	4	5
I would like to know what people in different countries around the world think about #MeToo	1	2	3	4	5
I think #MeToo is happening a lot in my industry (e.g. study/job)	1	2	3	4	5
I would like to know more about the spreading of the hashtag	1	2	3	4	5
I think there are more retweets than replies about #MeToo	1	2	3	4	5
I think there is a difference in feelings about #MeToo between men and women	1	2	3	4	5

Please answer how much you agree with the following statements	Importance				
	Strongly disagree	**Disagree**	**Neither disagree nor agree**	**Agree**	**Strongly agree**
I think there are certain countries where people react more negative on #MeToo than in the Netherlands.	1	2	3	4	5
I would like to know what people in different cities of the Netherlands think about #MeToo	1	2	3	4	5
I would like to know what my friends think about #MeToo	1	2	3	4	5
I would like to follow how the news around certain celebrities evolves (e.g. Harvey Weinstein)	1	2	3	4	5
I think men are more positive regarding #MeToo	1	2	3	4	5
I think there are more retweets than original tweets about #MeToo	1	2	3	4	5
I think there are more replies than original tweets about #MeToo	1	2	3	4	5
I would like to know (more) about male versions of #MeToo such as #MenToo	1	2	3	4	5
I think #MeToo is more popular in America than in Europe	1	2	3	4	5

Do you have any further remarks?

Thanks for filling in this survey!

B – Survey results

Table 2 – Results of survey questions

	I think #MeToo is **not** happening a lot in my industry (e.g. study/job)	#MeToo is less popular in western countries	I think there are more original tweets than retweets about #MeToo	I think there are more replies than retweets about #MeToo	I think women are more negative regarding #MeToo	I would like to know what my favourite famous persons (e.g. actors, singers etc.) think about #MeToo	I think there are more original tweets than replies about #MeToo
AVG	3,043478	1,782609	2,347826	3,173913	2,434783	2,608696	2,26087
STDEV	1,021508	0,850482	0,71406	0,834058	0,945135	1,117592	0,751809
	3	1	2	4	1	2	1
	4	2	3	2	2	1	4
	4	1	3	3	1	2	3
	4	3	4	3	4	3	3
	2	3	2	2	3	3	2
	2	1	2	3	4	3	2
	4	2	1	2	3	4	1
	3	1	2	2	2	2	4
	4	1	2	3	3	1	2
	4	1	3	4	2	1	2
	2	2	3	4	2	4	2
	1	2	4	4	2	3	3
	2	1	2	3	2	2	2
	3	2	2	4	2	4	2
	5	4	2	4	4	3	2
	2	3	2	4	2	1	2
	2	2	2	4	4	4	2
	4	1	2	3	2	2	2
	3	2	2	4	3	4	2
	3	1	2	2	2	3	2
	3	2	2	2	3	1	2
	2	2	3	3	2	3	3
	4	1	2	4	1	4	2

I think #MeToo is just a hype	I do not want to know more about #MeToo	I think #MeToo is more popular in Europe than in America	I think women are more positive regarding #MeToo	I would like to know what people in the Netherlands think about #MeToo	#MeToo is more popular in western countries	I would like to know what people in different countries around the world think about #MeToo	I think #MeToo is happening a lot in my industry (e.g. study/job)	I would like to know more about the spreading of the hashtag	I think there are more retweets than replies about #MeToo
2,304348	2,409091	2,217391	3,565217	3,478261	3,869565	3,521739	2,869565	2,73913	3,217391
1,145536	0,73414	0,671262	0,895752	0,790257	0,868873	0,994053	1,01374	1,009833	0,902347
1	3	2	5	4	5	3	4	2	4
4	4	2	4	2	4	4	2	2	3
1	1	3	5	4	4	4	1	2	3
4	3	2	2	3	2	3	2	4	3
2	3	3	3	3	4	3	4	3	4
2	2	4	2	4	4	4	3	5	3
3	3	2	4	2	4	2	2	2	5
3	2	2	4	4	5	4	3	3	4
2	2	2	3	2	4	2	2	2	4
4	3	2	4	3	4	2	2	2	2
2	1	2	3	4	4	5	3	4	3
1	2	2	4	4	4	3	5	3	2
3	3	2	4	3	4	2	4	3	4
4	2	2	4	4	4	4	3	3	2
2	2	2	2	5	2	5	2	4	3
1	2	3	4	4	3	5	4	4	4
1	2	2	2	4	2	4	4	1	2
3		1	4	3	5	4	2	1	4
1	2	3	4	4	4	4	3	3	2
2	3	2	4	3	4	4	2	2	4
1	3	3	3	3	4	4	3	3	4
2	3	2	4	4	4	2	4	2	3
4	2	1	4	4	5	4	2	3	2

I think there is a difference in feelings about #MeToo between men and women	I think there are certain countries where people react more negative on #MeToo than in the Netherlands.	I would like to know what people in different cities of the Netherlands think about #MeToo	I would like to know what my friends think about #MeToo	I would like to follow how the news around certain celebrities evolves (e.g. Harvey Weinstein)	I think men are more positive regarding #MeToo	I think there are more retweets than original tweets about #MeToo	I think there are more replies than original tweets about #MeToo	I would like to know (more) about male versions of #MeToo such as #MenToo	I think #MeToo is more popular in America than in Europe
4,217391	4,130435	3,304348	3,391304	2,73913	2,478261	3,913043	3,652174	3,434783	3,869565
0,951388	0,457697	0,97397	0,838783	1,251086	0,845822	0,733178	0,775107	1,121123	0,919701
5	4	2	4	2	2	5	5	4	5
5	5	4	4	1	2	4	3	4	5
5	3	4	3	4	1	3	3	3	3
4	4	3	4	2	4	3	3	4	4
4	4	3	3	3	4	4	4	3	3
5	4	4	4	2	4	4	3	5	4
3	5	1	1	1	2	5	4	1	5
5	4	3	3	2	2	4	2	4	4
1	4	2	2	2	2	4	4	2	4
3	4	2	3	4	3	4	4	4	3
5	4	4	4	5	3	3	3	4	4
4	4	4	3	4	4	2	4	4	4
5	5	2	4	1	2	4	4	3	5
4	4	4	4	4	2	4	4	4	4
4	4	5	4	4	3	4	4	5	4
5	4	4	2	2	2	4	4	5	3
5	4	4	4	1	2	5	4	1	2
5	4	3	4	2	2	4	4	3	5
4	4	4	4	4	3	4	2	4	2
4	4	3	3	4	2	5	5	2	4
4	4	3	3	3	2	4	4	3	3
4	4	4	4	4	2	3	3	4	4
4	5	4	4	2	2	4	4	3	5

Table 3 - Age of subjects

18-25	25-35
18	5

Table 4 – Gender of subjects

Female	Male
9	14

Table 5 – Subjects had heard from #MeToo via the following sources

Facebook	News	Instagram	Twitter	Tumblr	Conversations	Everywhere
18	19	6	3	2	2	1

Table 6 – Subjects use the following social media

Facebook	Twitter	Instagram	Snapchat
22	2	14	12

Table 7 – Subjects use hashtags as often as follows

Never	Daily	Weekly	Monthly	Yearly
9	0	3	8	3

Table 8 – Subjects had heard of the following celebrities or happenings

Kevin Spacey	Times Up	Alyssa Milano	MoiAussi	WomenWhoRoar	Harvey Weinstein	McKayla Maroney	The Silence Breakers	YoTambién	Tarana Burke	None
20	8	6	6	4	15	4	3	2	1	0