The Journal of Social Media for Learning 2023

SocMedHE: More than a conference?

Scott Turner[1] and Sarah Honeychurch[2]
[1]Canterbury Christ Church University, UK
[2]University of Glasgow, UK

Abstract

Using SocMedHE as a case study, in this paper we provide some examples of extracting and analysing information from tweets and we introduce some example tools for doing this. We also use these tools in order to explore some different ways in which we can play with this type of data. This paper is an extension of a conference presentation to SocMedHE21 (Turner 2021a).

SocMedHE: More than a conference?

Introduction

Social media provides us with a number of ways to work with our students. We can also use look at social media interactions as data and use tools in order to investigate and explore uses of social media. These tools also offer us creative ways of creating visualisations to represent social media use. Two central aims of this paper are:

- To highlight some of the tools that are available for looking at Twitter data and to show the reader that in most case no special programming or mathematics skills are needed;
- To use these tools to gain some insights about the community around the SocMedHE related hashtags.

Data Sets

Three data sets were used to explore hashtags around SocMedHE. We used TAGS in order to analyse interactions in data sets 1 and 2 and NodeXL in order to analyse data set 3.

(i) Data Set 1 covers the use of #SocMedHE or #SocMedHE19 from 27th March 2019 to 6th May 2021 containing 3041 unique tweets (Turner, 2021b).

(ii) Data Set 2 covers tweets from 6th December 2021 to 24th June 2022 using the hashtag #SocMedHE21 (Turner, 2022).

(iii) Data Set 3 represents a network of 146 Twitter users whose recent tweets contained "#SocMedHE20", or who were replied to or mentioned in those tweets, taken from a data set limited to a maximum of 18,000 tweets. This network was obtained from Twitter on Thursday, 17 December 2020 at 17:41 UTC. The tweets in the network were tweeted over the 8-day, 4-hour, 37-minute period from Wednesday, 09 December 2020 at 12:52 UTC to Thursday, 17 December 2020 at 17:29 UTC.

Tools and Approaches

In this section we briefly describe the tools and approaches that we use in this paper.

TAGS

The main tool used in this analysis is called TAGS, which is a tool developed by Martin Hawksey (Hawksey, 2022). TAGS is a way of capturing tweets for a particular hashtag, initially up to the previous 7 days of setting up TAGS and capturing them as an archive in Google Sheets. One of the great features is once you set it up, you can leave it to collect data every hour and add it automatically to the archive. This approach was used for all the data sets used in this paper.

Scott Turner[1] and Sarah Honeychurch[2]
[1]Canterbury Christ Church University, UK, [2]University of Glasgow, UK

Setting up a TAGS sheet is relatively easy. All you need to do is to go to the website, follow the "Get TAGS" link and choose which version you want to use. It is worth reading and following the guidance on the page about the app not being verified. You will need a Twitter account that you can use in order to link to TAGS. In summary the steps are:

- Go to https://tags.hawksey.info/ Create a new spreadsheet via "Get TAGS". Select "TAGS v6.1"
- Make a copy
- In box 2 enter the search term or terms e.g. #SocMedHE21
- Use the TAGS pull down menu and select "Run Now!"
- Run through a whole load of authorization
- In TAGS menu select "Update Hourly"
- Change the settings on the share button so all can view it.

You can see our TAGS settings to see what is needed. Figure 1 shows the first page of the TAGS sheet after the hashtags you want to consider (box 2) have been running for a while.

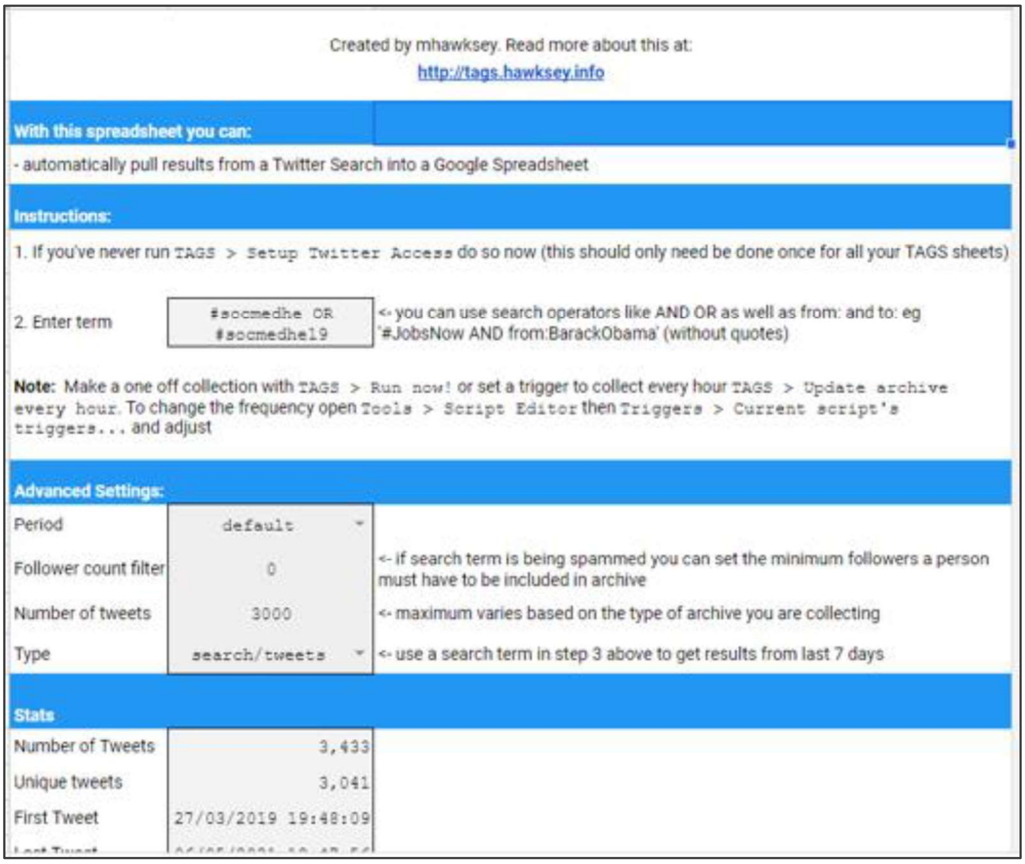


Figure 1 Screen shot of front TAGS page

There are further links (see Figure 2 Overleaf) that allow us to visualise the data a bit in particular TAGSExplorer.

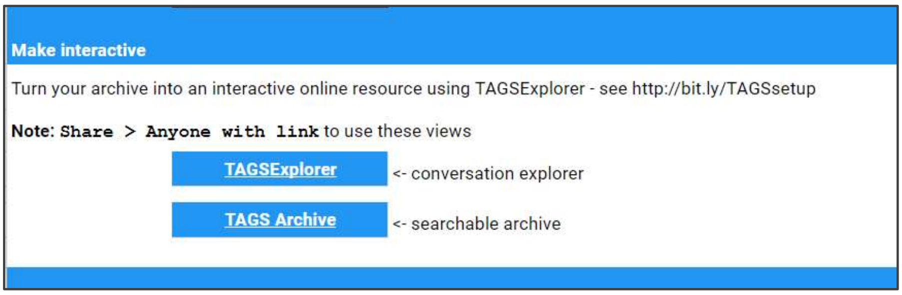


Figure 2 Making it interactive

Figure 3 shows all the Twitter accounts (these are called nodes) using the particular hashtag and connects those together where people have replied to each other.

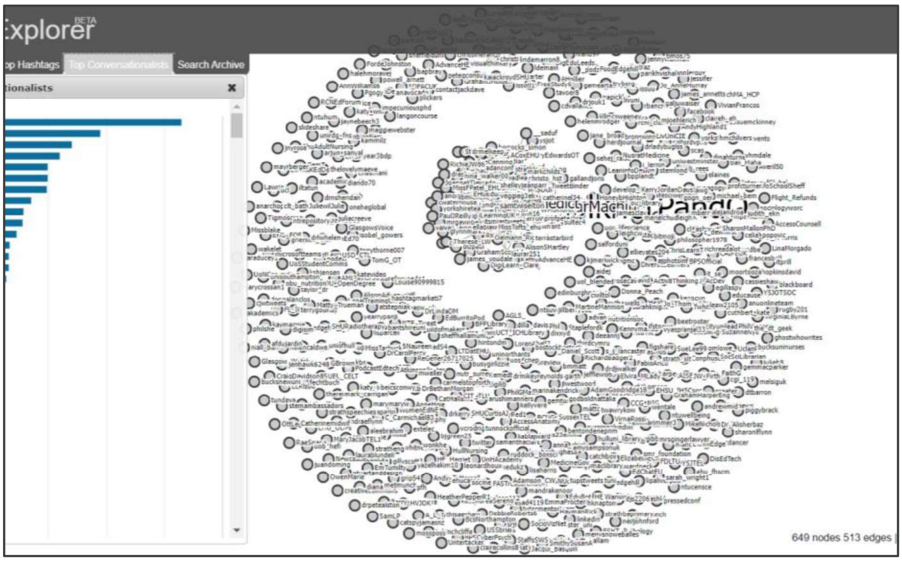


Figure 3 Dataset 1 with replies and nodes

Figure 4 shows the links are made the same data where the is a mention, reply or retweet.

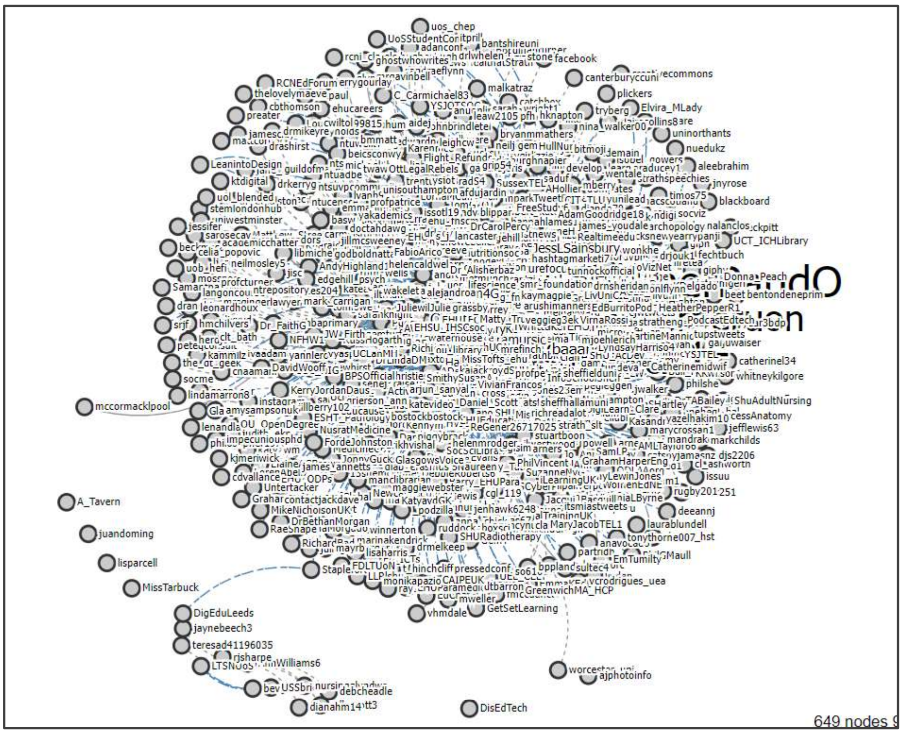


Figure 4 Dataset 1 with replies, tweets, mentions and retweets

Gephi

Gephi is a tool for analysing and visualising networks, rather than a tool solely for social media. It is available for free at https://gephi.org/ In this paper it is used to visualise and further process data collected from other tools.

Sentiment Analysis

Sentiment Analysis is a widely used approach of looking at data, usually textual, and categorizing words into positive, negative, and neutral words to gain a sense of how for a particular group a phrase or - as in this case a hashtag - is viewed. In this paper a piece of Python code, which is a modified version of one from Wintjen M (2020), is applied to a Comma Separated Variable (CSV) file of tweets in order to perform a basic sentiment analysis. The code is included as Appendix A.

NodeXL

Nodexl has free and paid version (Social Media Research Foundation, 2022a). The Pro/education (paid) version was used in this work. NodeXL is essentially an Excel add on that can capture data from a variety of different Social Media platforms. It also offers a repository to store the dataset and images the NodeXL Graph Gallery (Social Media Research Foundation, 2022a). Figure 5 shows some example graphs that can be found online when you search for #SocMedHE. Dataset 3 covers the actual conference specifically using NodeXL.

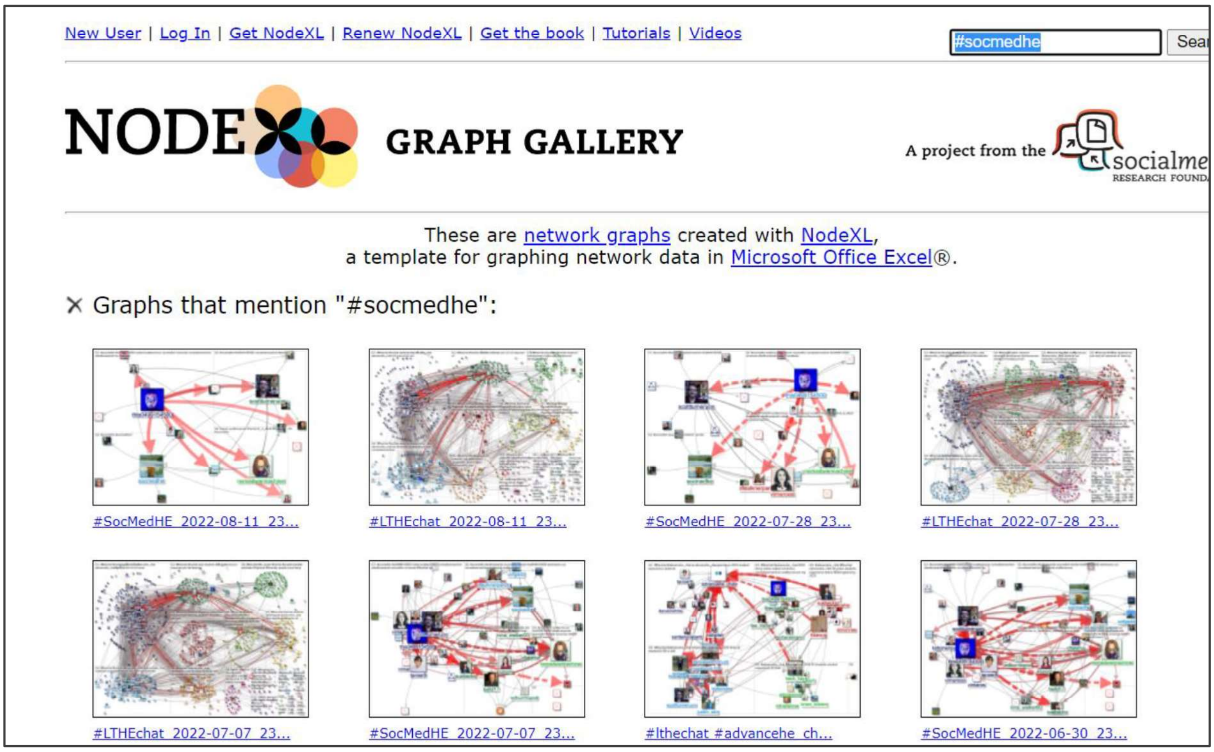


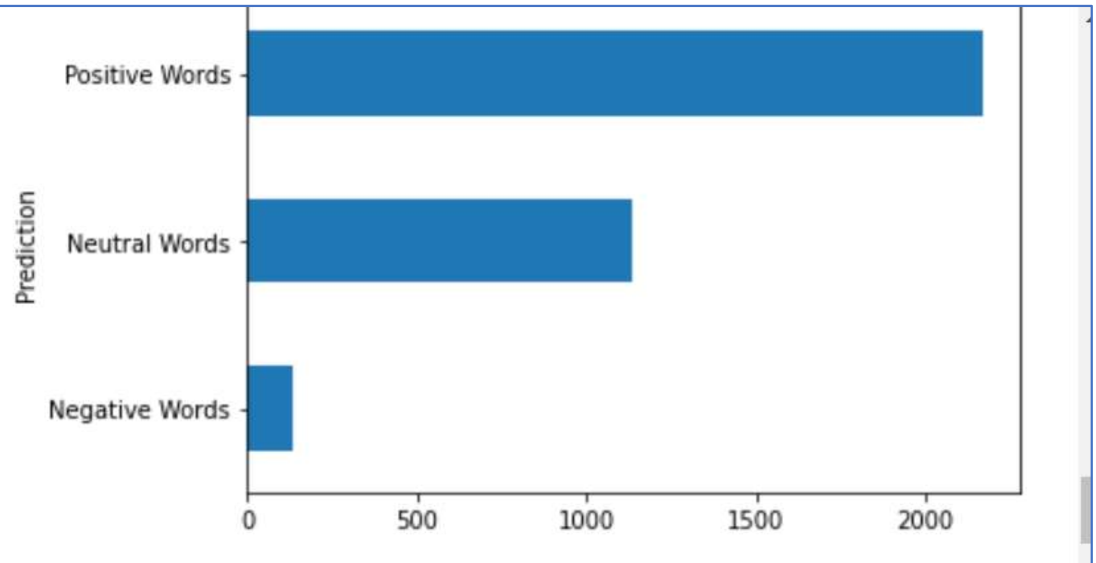
Figure 5 Examples of NodeXL graphs that use the hashtag #SocMedHE

Data Set 1: TAGS and basic sentiment analysis

This data set was used to explore TAGS as a tool and to apply a relatively simple form of sentiment analysis. Figure 1 and Figure 2 above show the settings for the TAGS search for covering the use of #SocMedHE or #SocMedHE19 from 27th March 2019 till 6th May 2021. This duration was selected to provide a relatively large collection of tweets and to cover one SocMedHE conference.

TAGS has many useful features including visualisations of the connections between people's tweets in terms of tweets and replies (figure 3) and with replies, tweets, mentions and retweets (Figure 4). In Figure 3 we can see that subgroup is formed of those replying to each other (linked with a solid line) and this does seem to move those to the centre of the graph. This can be seen as indicative of greater engagement, whether it can also be viewed as indicative of these nodes as having greater influence is less clear.

Another feature of TAGS is that it includes a second sheet which is an archive of all the collected tweets. If we just treat the content of the tweet as just text, we can collect them as a text file and perform a textual analysis on them in order to find out if the words used are considered to be positive, negative or neutral according to a standard repository classifying them. In this case we used the lexicon VADER (Valence Aware Dictionary and Sentiment Reasoner) which is believed by some authors (e.g. Lamberti, 2022) to be especially attuned to social media. The final output compares the number of positive, negative, and neutral words (see Figure 6 Overleaf).




```
[8]: print(the_data.groupby('Prediction').size())

Prediction
Negative Words     133
Neutral Words     1132
Positive Words    2168
```

Figure 6 Results of sentiment analysis

So, for these tweets (and by implication the conference as #SocMedHE and #SocMedHE19 were the hashtags analysed), we can see that 2168 positive words were used compared with only 133 negative words. This would suggest that the overall tone of the tweets was positive.

Data Set 2

To produce the data, TAGS was used collect all uses of the #SocMedHE between 6th December 2021 and 24th June 2022, saved to the archive (the second sheet in TAGS) as a CSV file. To focus on just mentions in the tweets the CSV file was edited to include just the person mentioning and who they mentioned. The data was then imported into Gephi and analysed. Details of this procedure are available in the video at Turner (2020). The key point really is we have just sorted the data into just those connecting via mentions. Looking at mentions for SocMedHE (Figure 7 Overleaf) the number of links between individuals Is indicated by the thickness of the line (so the higher the number of tweets, the thicker the line).

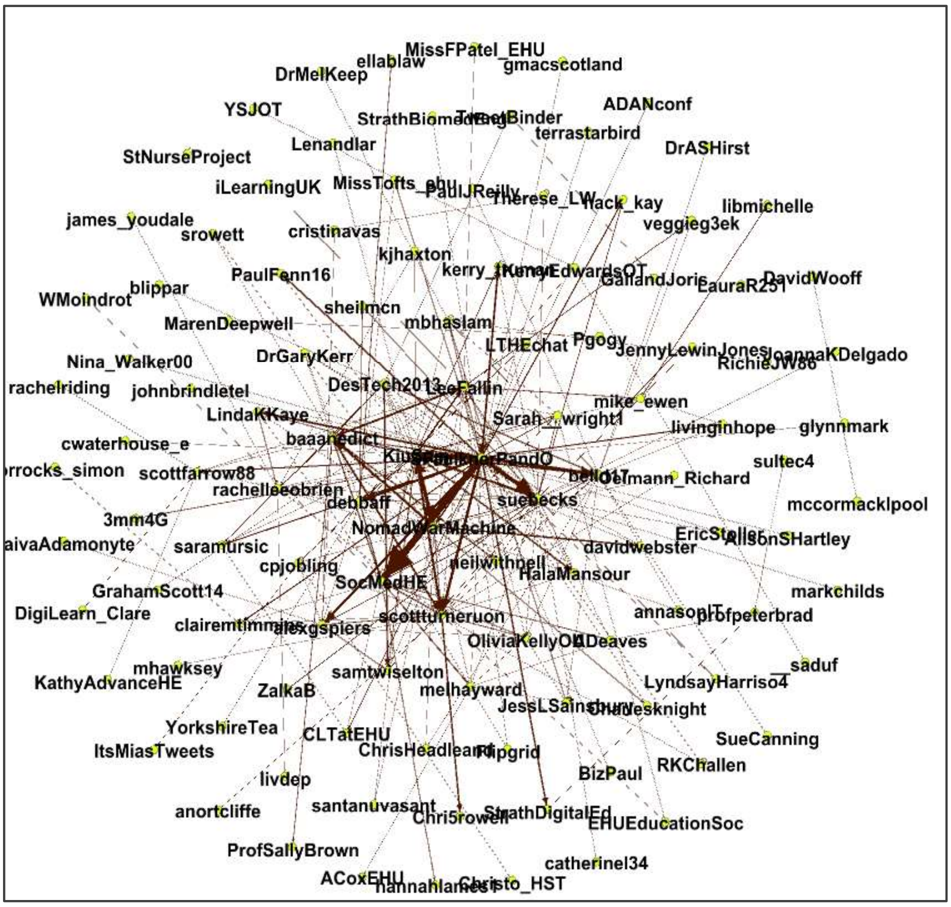


Figure 7 SocMedHE mentions

Figure 7 shows the mentions between 6th December 2021 to 24th June 2022 analysed in Gephi and displayed in a Fruchterman Reingold layout (Hansen, D. L. et al., 2019). One way of understanding these connections is think of them as being like lots of springs which are being pushed towards those they have more connections with and away from those with fewer connection with. So, the centre nodes are more likely to connected with each other, though not necessarily all of them to each other.

Applying a different layout (see Figure 8) we find there are a few groups that are unconnected but these are very small, so we can focus on the main group. Visually there seems to be a central 'hub' with most of the connections, followed by a few smaller hubs connected to the main hub. It implies that to get to anyone in the group it takes 6 or fewer hops and on average 3 (see Table 1). This means that this is a well-connected group.

Diameter: 6   Average Path length: 2.936280272748648

Table 1 Average path between nodes

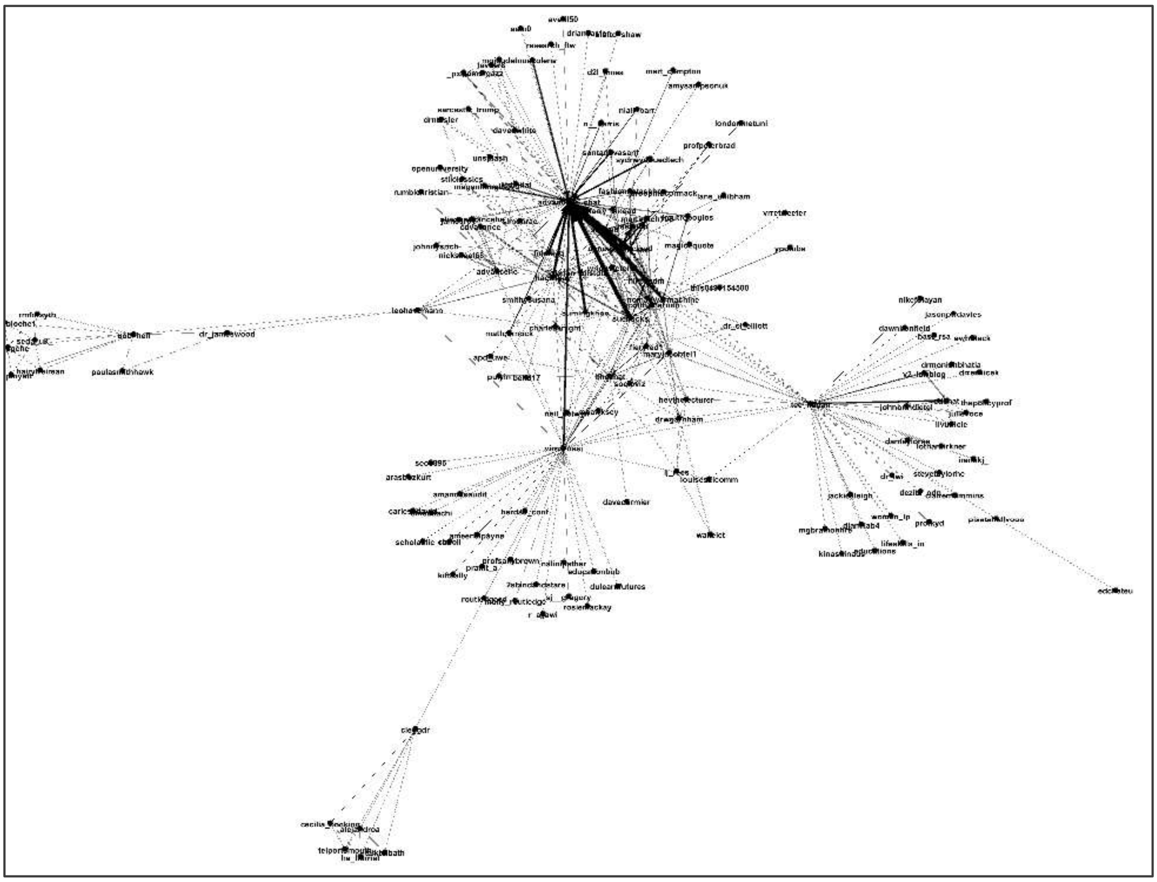


Figure 8 Fruchterman Reingold layout

Data Set 3

The last tool discussed in this paper is NodeXL. As mentioned earlier, there is both a free version and paid education version, and it is the the paid education version that is used here to produce both the visualization (for example Figure 9) and data about the network (e.g. Table 2).

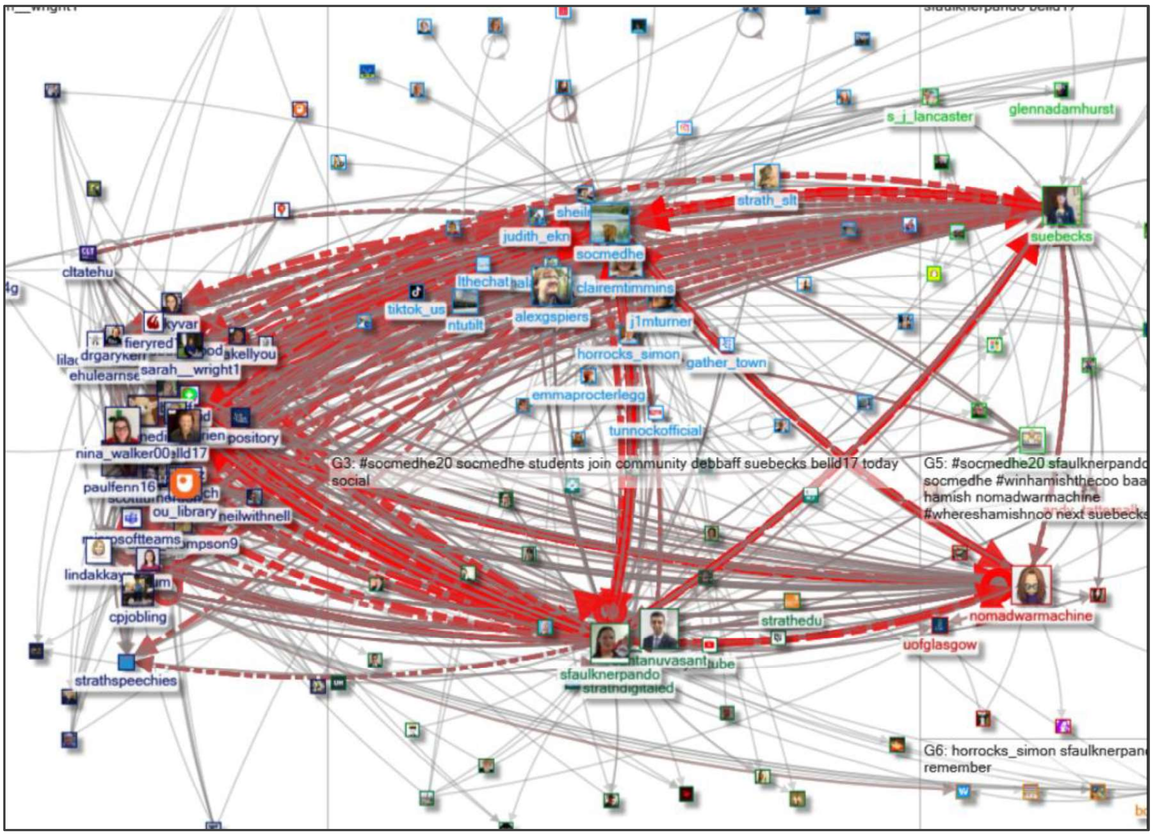


Figure 9 NodeXL network graph

NodeXL was applied to the hashtag "#SocMedHE20" over the 8-day, 4-hour, 37-minute period from Wednesday, 09 December 2020 at 12:52 UTC to Thursday, 17 December 2020 at 17:29 UTC, that is around the day of the conference itself.

Vertices : 146
Unique Edges : 468
Edges With Duplicates : 3411
Total Edges : 3879
Number of Edge Types : 5
Mentions : 1484
MentionsInRetweet : 1187
Replies to : 474
Retweet : 592
Tweet : 142
Self-Loops : 145

Maximum Vertices in a Connected Component : 146
Maximum Edges in a Connected Component : 3879
Maximum Geodesic Distance (Diameter) : 5

Table 2 Statistics NodeXL also produces alongside the network graph.

Visually it shows a lot of strong connections between people on Twitter (in Table 2 Vertices) and covers mentions, retweets and tweets without mentions. As in Table 1, in Table 2 we can see in most cases fewer than 5 hops are needed to go from one person to another. During this time period 146 people were using the hashtag or mentioned in the context of the hashtag producing 3879 tweets.

Conclusion

From the sentiment analysis that we conducted it can be seen that, at the very least, SocMedHE is a very positive conference, this is anecdotally backed up by the comments made verbally during and after the conferences days. The social network analysis that we have outlined, as represented by graphs both in TAGS and NodeXL, seems to show a tightly connected group of tweeters who use the various SocMedHE hashtags to share with each other both during the conference and outside it. In all the data sets there was generally large number of tweets of over 3000. Although most of those are on the days of the conference, there are tweets happening at other times, suggesting that SocMedHE is an active community.

Also, the mentions include a lot of people outside the conference group or being mentioned by people outside of the main group. This is a further potential indication of the community's active nature both sharing the tweets but also encouraging others to take part who not participants on the day.

Twitter data is not difficult to collect through tools such as TAGS and provides a good source of data for educational research and use. As another example, one of the authors has used data set 2 with their Data Intelligence students as a rich data set to investigate further. Because of the availability of tweets it can be argued that Twitter provides a good source for learning opportunities in Higher Education that are not just focused on Computing or Social Media specifically, but discussions around ethics, marketing, social sciences related areas and many more.

There is a lot more than can be investigated further such as how do the types of tweets vary with time, both in type and number, or a deeper textual analysis of tweets to identify themes and further insights.

References

Hansen, D. L., Shneiderman, B., Smith, M. and Himelboim, I. (2019) 'Installation, orientation, and layout', *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*, pp. 55–66. doi: 10.1016/B978-0-12-817756-3.00004-2.

Hawksey M (2022) TAGS – Twitter Archiving Google Sheet [online] https://tags.hawksey.info/ accessed on 15/8/2022

Lamberti, A. (2022) Sentiment Analysis with VADER and Python https://medium.com/artificialis/sentiment-analysis-with-vader-and-python-5b7ac4f3b13b accessed on 18/5/2022

Social Media Research Foundation (2022) NodeXL https://nodexl.com/ accessed on 15/8/2022

Turner, S. (2020) *Going from TAGS to Gephi - YouTube*. Available at: https://www.youtube.com/watch?v=nTZZEHDkqNg (Accessed: 23 August 2022).

Turner, S. (2021a) *#SocMedHE more than a conference : CCCU Research Space Repository*. Available at: https://repository.canterbury.ac.uk/item/8zv0w/-SocMedHE-more-than-a-conference (Accessed: 23 August 2022).

Turner, S. (2021b) use of # SocMedHE or #SocMedHE19 from 27th March 2019 till 6th May 2021 https://bit.ly/SocMedHE19_20

Turner, S. (2022) Use of #SocMedHE21 https://bit.ly/SocMedHE_21 accessed 16/8/2022

Wintjen, M. (2020) *Practical Analysis Using Jupyter Notebook* .Packt Publishing UK ISBN 978-1838826031

Disclosure statement

No potential conflict of interest was reported by the authors.

Appendix A

The file SocMedHE20.csv contains the tweets and the code is a modified example taken from Wintjen M (2020) "Practical Analysis Using Jupyter Notebook" pp 264

```
! pip install nltk

import nltk

import pandas as pd

import numpy as np

%matplotlib inlinefrom nltk.sentiment.vader import SentimentIntensityAnalyzer

anlysr=SentimentIntensityAnalyzer()

nltk.download('vader_lexicon')

from nltk.sentiment.vader import SentimentIntensityAnalyzer

anlysr=SentimentIntensityAnalyzer()

the_data=pd.read_csv('SocMedHE20.csv')

the_data.head()

score_compound=[]

score_positive=[]

score_negative=[]

score_neutral=[]

i=0

while (i<len(the_data)):

   my_anlysr=anlysr.polarity_scores(the_data.iloc[i]['text'])

   score_compound.append(my_anlysr['compound'])

   score_positive.append(my_anlysr['pos'])

   score_negative.append(my_anlysr['neg'])

   score_neutral.append(my_anlysr['neu'])

   i=i+1

the_data['Compound score']=score_compound

the_data['Positive score']=score_positive

the_data['Negative score']=score_negative

the_data['Neutral score']=score_neutral

loop=0

pred_sentiment=[]
```

```
while (loop<len(the_data)):
    if ((the_data.iloc[loop]['Compound score'])>0.3):
        pred_sentiment.append('Positive Words')
    elif ((the_data.iloc[loop]['Compound score']>=0) & (the_data.iloc[loop]['Compound score']<0.3)):
        pred_sentiment.append('Neutral Words')
    else:
        pred_sentiment.append('Negative Words')
    loop=loop+1
the_data['Prediction']=pred_sentiment
the_data.groupby('Prediction').size().plot(kind='barh')
```