A useful method for postgraduate of chemical engineering in understanding the public opinions of public events related to their major: A case study of health communication and public opinions of COVID-19 on Twitter in Germany

Ningning Zhao[1], *Xiaojiao* Yu[1], *Zhicheng* Zhang[2,*], *Zhong* Yu[1], *Binghua* Yao[1], *Jinfen* Niu[1], *Nailiang* Liu[1], and *Junpeng* Li[1]

[1]Xi'an University of Technology, School of Science, 710054 Xi'an, Shaanxi, China
[2]Xi'an International Studies University, School of Journalism and Communication, 710028 Xi'an, Shaanxi, China

Abstract. In order to adapt to the latest development of higher education in the era of big data, improve the sensitivity for postgraduate of chemical engineering in the public opinions of public events related to their major, this study took "the pandemic of COVID-19 on Twitter in Germany" as an example and investigated the health communication and public opinions on COVID-19 based on the Python scripts and Twitter streaming application programming interface using statistical analysis and chi square feature selection. The provided method is helpful to improve the insight, prediction, scientific literacy and innovation ability of postgraduate of chemical engineering.

1 Introduction

With the development of information technology and the continuous improvement of information construction, significant changes have taken place in the technical and social environment. Employees in various fields are facing huge challenges, which also puts forward higher requirements for cultivation of high-level talents in colleges and universities as well as the higher education development. The postgraduate stage is one of the main stages to focus on cultivating research capabilities such as problem analysis and problem solving. Improving their scientific literacy and introducing the new research ideas and methods bred in the era of big data into the postgraduate education system, which are conducive to expanding the vision of the postgraduate, and stimulate them to adopt new research ideas and methods to solve scientific and practical problems in their chosen fields, and promote the improvement of the innovation ability of postgraduate students.

* Corresponding author: zhangzhicheng@xisu.edu.cn

At present, the world is in a critical period of the COVID-19. The social public opinions brought by the COVID-19 will have an important impact on the insight, prediction and research innovation ability of the postgraduate students of chemical engineering, especially the postgraduates in biochemical engineering. The postgraduate students of chemical engineering should be able to obtain the source of big data related to their major, have a comprehensive vision of big data, be able to fully understand the public opinions brought about by the COVID-19 through reasonable methods, analyse and use big data to solve key problems and problems in the practice, promote scientific research innovation and development of science and technology.

World Health Organization (WHO) announced a Public Health Emergency of International Concern (PHEIC) on January 30, 2020 [1]. China National Health Commission confirmed that the virus was human-to-human transmissible [2]. WHO announced "COVID-19" as the name of this new disease on 11 February 2020. ICTV announced "severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)" as the name of the new virus on 11 February 2020 [3]. Social distancing was widely used by governments aiming to reduce physical contact between people. In this case, more and more social interactions move online, the conversation around COVID 19 has continued to expand, with increasing users turning to social media. Twitter has become central to allow us to stay connected during the crises.

Twitter data provides us with a very good opportunity to analyse the perception and reaction to a public disease outbreak among the same linguistic and cultural group of populations.

In this study, 1.5 million Twitter posts including tweets, retweets and replies were obtained employing Python scripts and Twitter streaming application programming interface. On which basis, the health communication on COVID-19 has been demonstrated and public opinions to the pandemic of COVID-19 on Twitter in Germany has been investigated using statistical analysis and chi square feature selection.

2 Methodologies and results

We obtained all the Twitter data via Python programming scripts from January 20, 2020. For this study, we ran Python scripts to access all the samples of tweets in German using the topics including coronavirus, COVID-19, SARS-CoV-2, Wuhan coronavirus and nCoV. The time frame was January 20 to August 2(Coordinated Universal Time).

We calculated the basic descriptive statistics including the number of posts, the number of posts per user, retweets, hashtags and embedded URL links. We estimated the Pearson's correlation between the proportion of tweets with hashtags and the proportion of tweets with embedded URL links.

We retrieved COVID-19 related Twitter data in five key topics: coronavirus (N = 948525), COVID-19 (N = 665590), SARS-Cov-2(N = 60974), Wuhan coronavirus (N = 1143) and nCoV(N = 2476). More than 200 thousand unique users joined health communication of Covid-19 during that period of time on Twitter. The mean number of posts per user ranged from 1.81 to 5.75. However, the distribution for the number of posts per user is 1.0 for all 5 corpora. Across the five corpora, as the proportion of tweets with hashtags increased, the proportion of tweets with URLs increased ($r = 0.986$, $p < 0.005$). More than 60 percent of tweets are embedded with URLs in total.

Then we analysed retweets to test participation in each corpus. We ranked the number of retweets for each tweet and extracted all user names. The number of retweets indicates the influence of its tweets. Thus, we identified the most influential twitter participants in the COVID-19 discussion. We chose the top 30 users ranking the top of the retweet list from each corpus, and categorized them into the following categories of users:

Media: Twitter profiles of media, reporter, anchor and journalists.

Political: Twitter profiles of government agencies, political groups or parties and politicians.

Medical: Twitter profiles of hospitals, health care organizations and the medical practitioners.

Others: Twitter profiles that do not belong to the three categories referred to above.

We found that the categories of the top 30 users differed between the corpora. In coronavirus and nCoV and SARS-CoV-2 corpora, the number of media users are 11(36.67%), 12(40%) and 8(26.67%) separately. While in COVID-19, nCoV and SARS-CoV-2 corpora, seven (23.33%) 11 (36.67%) and ten (33.33%) were medical users respectively. Political users take up 16.67 percent (5 political users) and 20 percent (6 political users) in Coronavirus and COVID-19 corpora.

Keywords selected will be able to present the public opinion and trend of discussion. We ran Python scripts to calculate the weeks when the most tweets were published. The week number in this study will follow ISO-8601 standard. After preliminary analysis, we found that the largest number of tweets came in the 12th week (N = 207858). Then we ran chi-square feature selection algorithm to select the keywords in the 12th week when the public were more willing to share their ideas online [4].

We performed a chi-square feature selection analysis to pick the keywords that are most unique to a specific corpus and on a specific date. Chi-square feature selection is very helpful in machine learning by testing the relationship between the features [5]. Chi-square test measures how expected count E and observed count Q deviates each other, testing the independence of the two variables, following the equation as below:

$$x_c^2 = \sum \frac{(Q_i - E_i)^2}{E_i} \tag{1}$$

In the equation, c is the degree of freedom, Q is the observed value, E is the expected value.

The algorithm vectorizes the corpus into a sparse matrix, calculating the scores of each word in the bag-of-words model [6]. The null hypothesis of chi-square test is that each word in the bag-of-word model is independent and there is no relationship. A high score of chi-square value tells that the hypothesis of independence is incorrect, indicating the particularity of the word to a specific corpus. We picked the keywords with the highest scores of chi-square values, ranking on top of the list calculated. German words were translated into English through Google translation service. Table 1 showed the top 20 most unique words in each corpus. The higher it ranks, the more unique it is.

Table 1. Top 20 keywords for each corpus, selected by chi-square feature selection.

	Coronavirus	**COVID-19**	**nCoV**	**SARS-CoV-2**	**Wuhancoronavirus**
1	Trumpfail	Bum	SOS	June	Condemn
2	Market	Cacao	Cannabidiol[1]	Stock[2]	Friend
3	Heidekreis[3]	Sicker	IFSG[4]	Stupid	Trump
4	Pollination	Kiss	Sick	EAU5	Mask
5	Radical	Sending prayer	Made[6]	BAföG[7]	Blackout
6	VDGN[8]	Cold	Dead	Wonnemar9	CCP[10]
7	Member	Eat	Zombie	Good luck	Pandemic
8	Foreign	Explanation	South America[11]	DGB[12]	People
9	Senator	Disruption	Sars	Borkheide[13]	WHO[14]

10	CVP[15]	Quarantine area	CoV	MAYDAY[16]	China
11	ASB[17]	Pollination	Virus	AStA	November
12	Velo[18]	Senator	Coronavirus	Diakonia[19]	Tenant
13	Najaf[20]	SOS	Gene sequencing	LOINC[21]	Lie down
14	Mühlenkreiskliniken[22]	Foreign	Inspect	Weight loss	Local
15	China	AStA[23]	Relaxed	Billion[24]	Pneumonia
16	Elbow	Goodbye	Genom	Extra time	Make
17	Vitamin	Pray	Genomic	Muslim	Mainstream
18	Disinfectant bottle	Shetland[25]	Wordag[26]	Save[27]	Market
19	Amberg[28]	Manufacture	Concern	Daily	Over
20	Delta[29]	Tripsdrill[30]	Lean back	Bruck[31]	Media

Notes: 1. Cannabidiol oil might benefit patients infected with COVID-19; 2. Stock toilet paper; 3. The name of a place in Germany; 4. German Infection Protection Act; 5. European Association of Urology; 6. The virus appeared originally; 7. Abbreviation of Bundesausbildungsförderungsgesetz, Germany's Federal Training Assistance Act; 8. German land usage organization; 9. A recreation center in Germany closed temporarily during pandemic; 10. China communist party; 11. Governments tightened the measure to prevent pandemic in south America; 12. German trade union confederation; 13. The name of place in Germany; 14. World health organization; 15. The Christian Democratic People's Party of Switzerland; 16. A pop music band; 17. Charity emergency rescue organization; 18. Pro-cyclist organization; 19. A public service organization of the protestants; 20. An Iraqi city which is locked down because of COVID-19; 21. Logical Observation Identifiers Names and Codes,a medical research organization; 22. Medical organization; 23. University student union; 24. Billion dollars; 25. A place of UK where suffered serious covid-19 pandemic; 26. A German doctor; 27. To save people's life; 28. A German city; 29. An airliner; 30. The oldest park in Germany; 31. A German city.

The words 'SOS', 'China' and 'Market' appeared in different corpora. Most of the words are specific to this outbreak (i.e., 'Coronavirus', 'Virus','Gene Sequencing'), we found that this pandemic had potential influences on various aspects of society. AStA referred to a student union. Tripsdrill is a famous and popular park in Germany and DGB referred to German Trade Union Confederation. A pop star group with the name of MAYDAY posted tweets cancelled their concerts because of the pandemic. Some medical organizations were involved to confront with COVID-19, including 'EAU' (European association of urology), 'WHO' (world health organization), 'LOINC' (logical observation identifiers names and codes, a medical research organization) and 'Mühlenkreiskliniken' (a medical organization group). We also find keywords 'stock' which indicates people are stocking toilet paper. 'Elbow', 'Disinfectant bottle', 'Cold' and 'Sick' keywords appeared in some educational tweets that helps the public about COVID-19.

3 Discussion and conclusion

First, our findings suggest that the online reactions to the COVID-19 outbreak in Germany differed between various topics. More than 200 thousand German users joined Twitter public communication on COVID-19 in total. Coronavirus and COVID-19 topics received more participants than others. Twitter community is a specific platform for social topics. An understanding of social media culture of a specific topic will contribute to the success of social media communication. Second, media and medical Twitter users play significant roles among Twitter communication about COVID-19 in Germany, who are more willing to join social media communication and receive broad public trust. Beyond that, Twitter communication brings a large number of users in various areas or background. Twitter users are more willing to spread educational information on social media in Germany.

Third, keywords selection found that COVID-19 has numerous influences in various aspects of society and life. Numerous negative keywords were selected. Medical, political, educational and international centers and organizations are involved. Twitter plays an important role in educating the public with millions of tweets and information. Fourth, sources of retweets data identify that professional Twitter users (medical, political and media users) received more trust and their tweets received more retweets by Twitter users.

To conclude, we observed dramatic differences in COVID-19 related Twitter communication between various topics. COVID-19 pandemic requires profound public health policies among countries and states. The results will help nation public health agencies in Germany and across the globe understand their own communities' Twitter communication cultures.

The research method of this paper can be extended to the big data analysis and research in the chemical industry, which helps chemical graduate students to have basic big data literacy and good scientific literacy, and solve key problems and problems in the industry by analysing and using big data, so as to promote scientific research innovation and promote scientific research development.

This work was supported by Research Project on Postgraduate Education and Teaching Reform of Xi'an University of Technology (No. 252042029), the Education and Teaching Reform Research Project of Xi'an University of Technology (Nos. xqj2119, xjy2103, xjy2102, and 251032205), Scientific Research Project of Education Department of Shaanxi Province (Project title: Big data and social media communication; No. 19JK0711), and Scientific Research Project of Xi'an International Studies University (Project title: Big data and social media communication; No. 19XWB09).

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. WHO. Statement. https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov), (2020)
2. WHO. WHO Timeline - COVID-19. https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19, (2020)
3. WHO. Naming the coronavirus disease (COVID-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it, (2020)
4. Y. Yang, J.O. Pedersen. A comparative study on feature selection in text categorization. In: Proceedings of the international conference on machine learning, **412** (1997)
5. Scikit-learn. SelectKBest. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html, (2020)
6. Scikit-learn. CountVectorizer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html, (2020)