Proceedings of the 4th International Conference on Educational Innovation and Philosophical Inquiries
DOI: 10.54254/2753-7064/11/20231370

Student Loan: Topic Modelling with Twitter Data

Pinrun Su[1,a,*], Tianran Wang[2,b], and Yichen Pan[3,c]

[1]Dimensions International School, Singapore, 238318, Singapore
[2]UWC ChangShu, Suzhou, 215500, China
[3]The Winchendon School, Winchendon, 01475, US
a. selinasu0107@gmail.com, b. wdoris498@gmail.com, c. brucepanlyy1111@gmail.com
**corresponding author*

Abstract: The study is about citizens' opinions on student loans by analyzing Twitter reactions to Biden's student loan cancellation project using the machine-driven classification of open-ended response (MDCOR) and found it saved research time, increased efficiency, and ensured authenticity and objectivity of data. After putting data into the application, we found that using five analysis topics is appropriate. The topic's content can be predicted by seeking the relevant word for each case. The analysis of five issues related to student loans shows mixed opinions about the impact of loan forgiveness, with some key terms such as "predatory" and "donation" being significant. At the same time, some topics are not directly related to the issue.

Keywords: student loan, topic modeling, text mining, twitter

1. Introduction

Student loan debt is a significant financial burden for the United States of America. It is used to help students who are undergraduates achieve a higher standard of living in the university and benefit their future. Four different types of direct loans are available for federal students: direct subsidized loans, direct unsubsidized loans, direct PLUS loans, and direct consolidation loans. They are aimed at different types of students who are facing different kinds of situations. Compared to private loans from various companies and banks, a student loan has a stable interest rate and is usually lower than the others. [1]. Based on others' research [2], it is evident that many comments show disparate opinions on social media. For example, the research conducted by Manuel S. Gonzalez Canche considered different sorts of students, consisting of students who didn't rely on loans to finance their education; students who relied on student loans and rapidly repaid this debt; and students who relied on loans and had not fully refunded these amounts at time of measurements, give us potential benefit associated with rapid loan repayment in terms of nonexperimental data. Then, the experiment classified participants into different groups, and their backgrounds and vary conditional on the attainment of a four-year degree; using dichotomous variables for analysis, it became clear that policymakers should consider offering borrowers in this debt category low fixed interest rates. Although the experiment used many methods containing cross-decade comparison and two quasi-experimental techniques, it still has some limitations because the salary of students after leaving the school is not considered, and the samples are limited.

DOI: 10.54254/2753-7064/11/20231370

Our research aims to use the data collected from Twitter and analyze the data with Machine Driven Classification of Open-ended Responses (MDCOR) to understand the responses from distinct respondents using text mining techniques. Unlike other research, we used open-ended questions that usually make it hard to understand the tone and opinion that people hold. The primary purpose of this study is to interpret the public's general ideas.

2. Methodology

Our study aimed to understand the different reactions that Biden's student loan cancellation project received on Twitter and the support and opposition of other groups of people. In this way, we could understand the impact of a policy enactment on the trend of online social media comments and thus understand the different attitudes and practices of social media reactions before and after a significant event. In this research paper, secondary data is used as our analysis data. The leading software used is Twitter API and Machine-driven classification of open-ended responses (MDCOR) [2], an extension of Twitter, a software program that legally collects the comments posted under different posts. Our group obtained many comments on this topic by searching for "#studentdebtforgiveness" on this platform as our analysis data. At the same time, MDCOR allowed the group to transform the textual comments into data. This way, we could continuously reclassify and redefine the resulting data to complete our study. In this research paper, the team used quantitative and qualitative data to complete the primary research process. In contrast, our group used the quantitative method to classify the topics and the qualitative approach to summarize the brief idea for each case.

The sample size, i.e., the data we collected, is 26708 comments, all from the Twitter API. Our group cut off data from August 24, 2022, to October 24, 2022, during which time a total of 26708 comments were posted on Cancel Student loans under the tag [3]. After collecting the data, our group imported it into MDCOR and analyzed it. Machine-driven classification of open-ended responses (MDCOR) is a convenient software that allows people to import data directly, and the system automatically generates the corresponding analysis MDCOR can analyze large amounts of data very efficiently, eliminating the need for coding and saving a lot of time while maintaining the accuracy of the results. At the same time, MDCOR can quantify the textual content of the participant's responses (in this research paper, comments) without changing their voices, making it easier for the computer system to process and operate. In addition, MDCOR can provide the user with quantitative and graphical results, providing the researcher with a clear visualization of the results in an efficient combination of multiple formats. (González Canché) [2,4].

During the data preparation phase, the user must select the data he wants to analyze locally from the computer. In this step, a certain number of data may be automatically excluded from the system. The user can choose to download the files of the excluded data and evaluate them further. The next step is optional, i.e., remove the most common words; the significance of this step is that too many common words are meaningless and may cause the system to reduce its classification efficiency. For example, when the group analyzed the move to cancel student loans, terms such as "student" and "loan" may become familiar because they are mentioned many times in the comments. If they are removed, the system can be improved to some extent. If they are eliminated, the system's efficiency can be somewhat improved. The next step is to select machine learning sampling parameters, which can take a long time, depending on the computer system. After this step, the researcher can identify the ideal number of codes by choosing "5. Execute Metrics.", through which the researcher can decide how many topics there are in total. At the end of this step, since the resulting values are entirely dependent on the After this step, since the resulting value is completely computer-dependent, the researcher needs to decide again whether to use this value or to choose a more appropriate value in the ideal case. Ultimately, many data analysis and visualization charts can be obtained by pressing the execute button. (González Canché)

DOI: 10.54254/2753-7064/11/20231370

Overall, the Twitter API is a professional platform that guarantees the authenticity and accuracy of the data obtained without any missing or omitted comments. The platform ensures that every word is considered, and the random selection process ensures that the data is objective and unbiased, without personal bias or preference. At the same time, using MDCOR saves us a lot of research time and increases the group's overall efficiency. Although the machine may not be able to interpret the text with 100% accuracy, and there may be a few minor errors, this software platform is the best choice for us than reading 20,000 comments one by one and analyzing them one by one by hand [5].

3. Result

Based on the picture, Deveaud2014 of 5 topics is close to CaoJuan2009 of 5 cases. Therefore, five issues are more appropriate.

According to the relevant words in topic 1, topic one will likely talk about people's votes to cancel the American student loan.

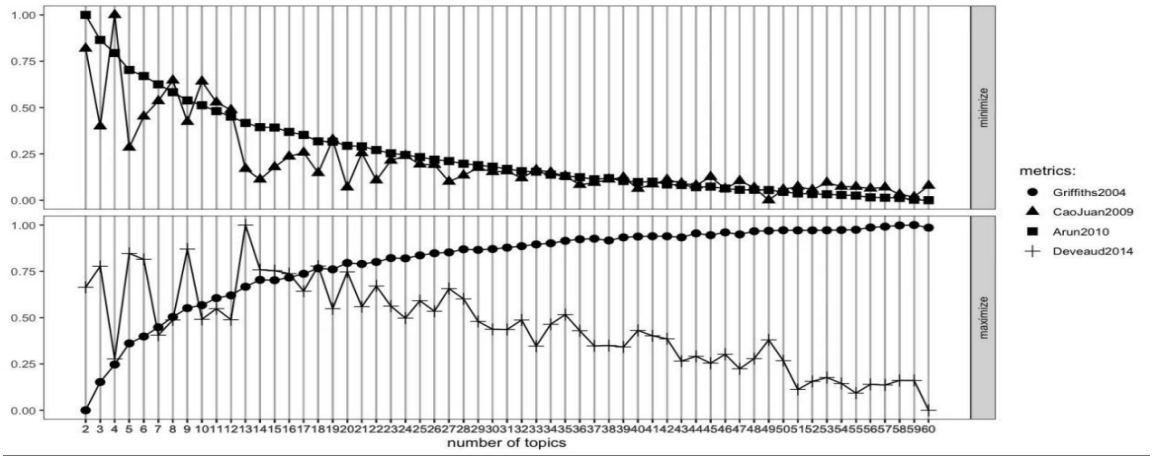

Figure 1: Direct result from MDCOR.

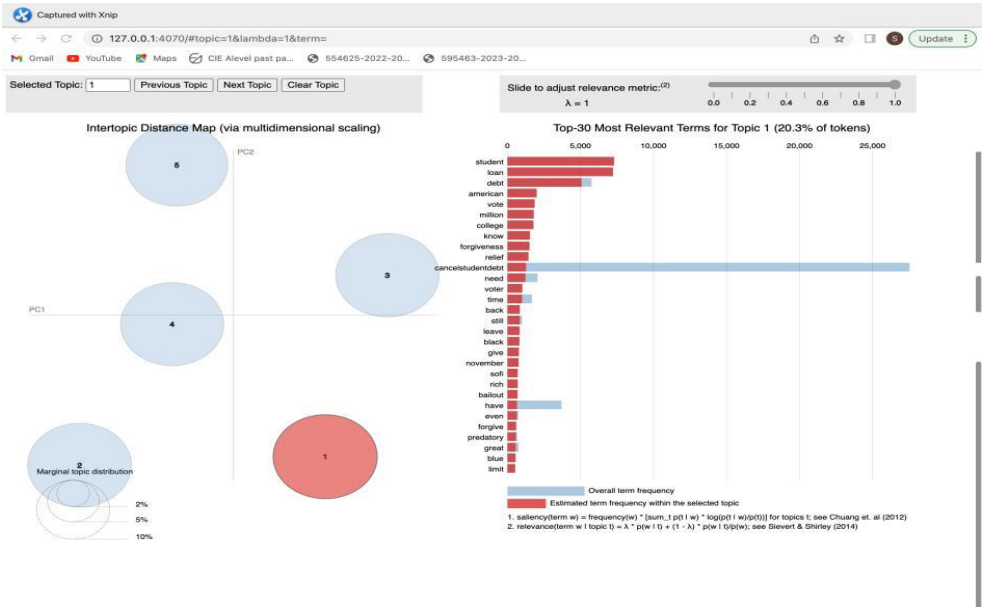

Figure 2: The current president, Joebiden, affects student loan forgiveness.

DOI: 10.54254/2753-7064/11/20231370

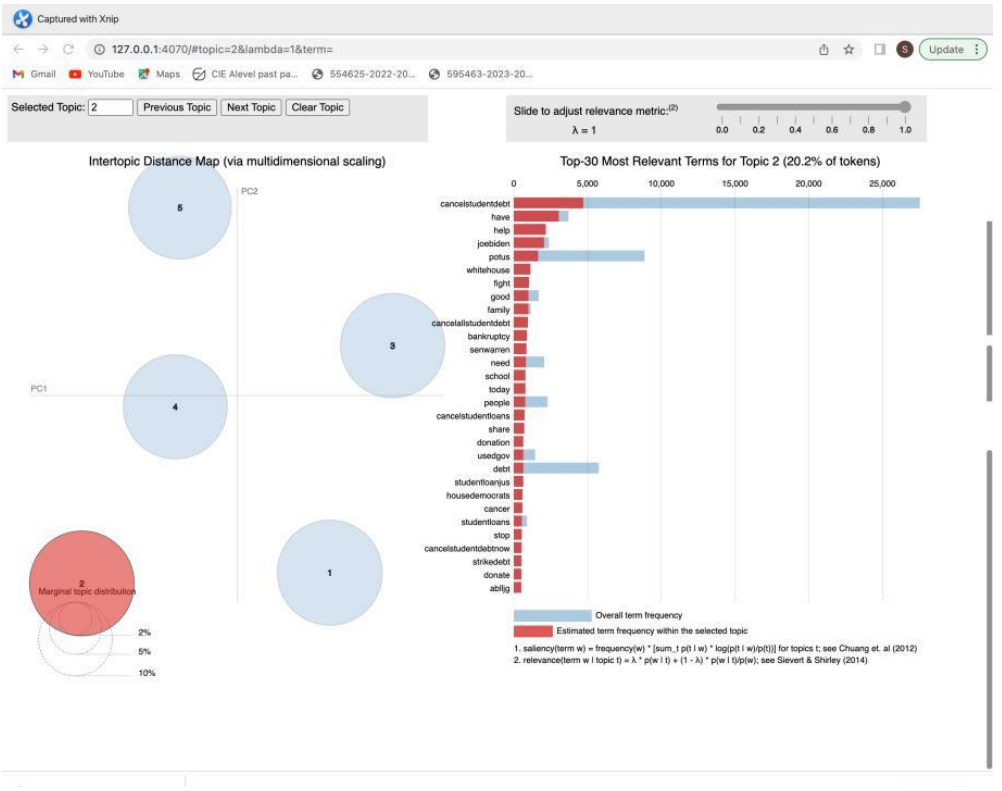

Figure 3: Focuses on the emergency of canceling student debt immediately and discusses the student loan's effect on education.

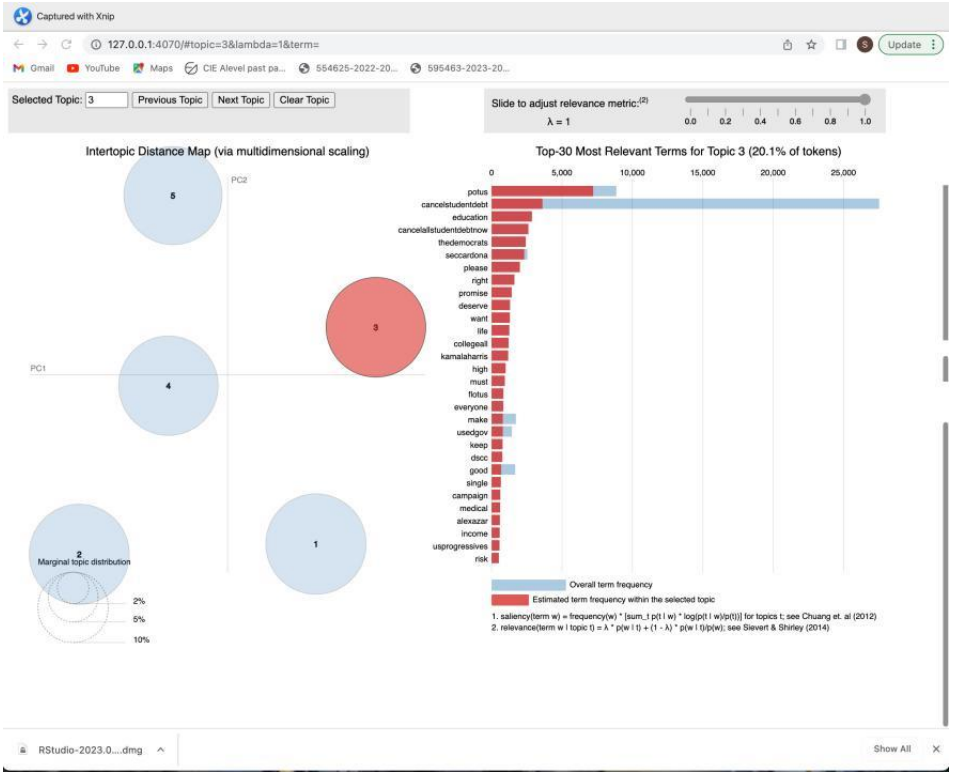

Figure 4: Underlines the significance of canceling student debt by showing the incredible effect it may bring if the policy is implemented.

DOI: 10.54254/2753-7064/11/20231370

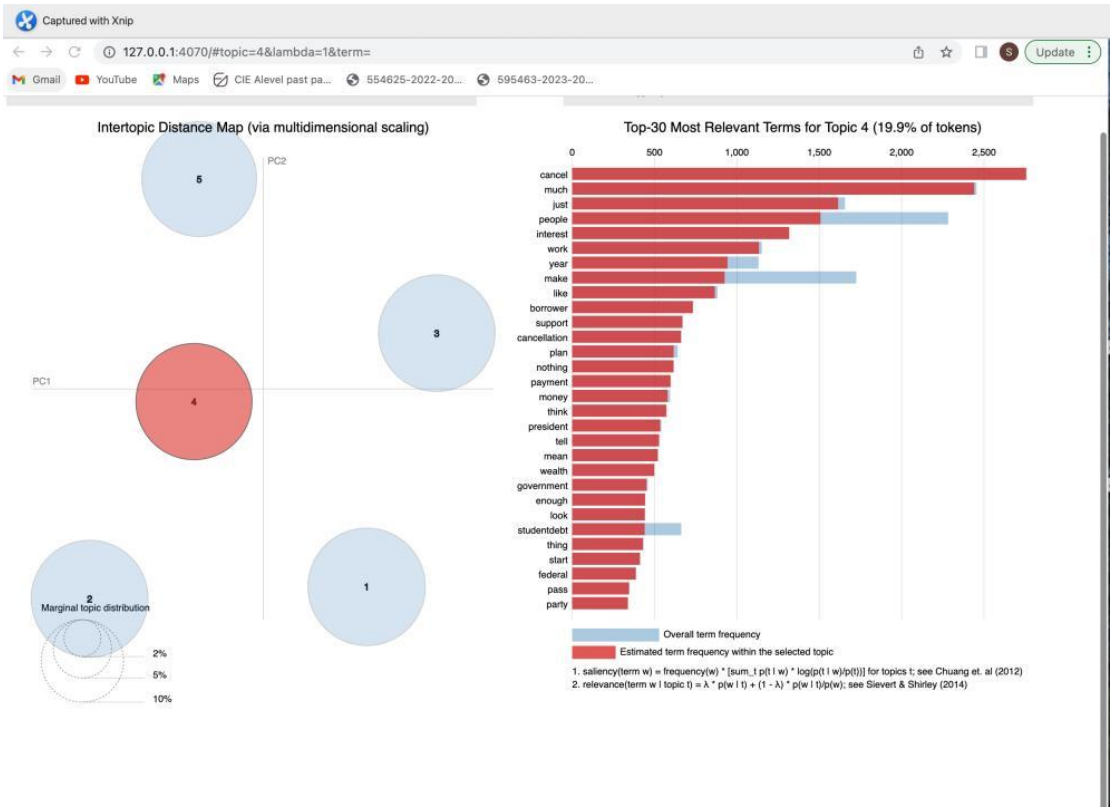

Figure 5: Topic 5 is about that besides America, international influence might appear after student loan cancellation.

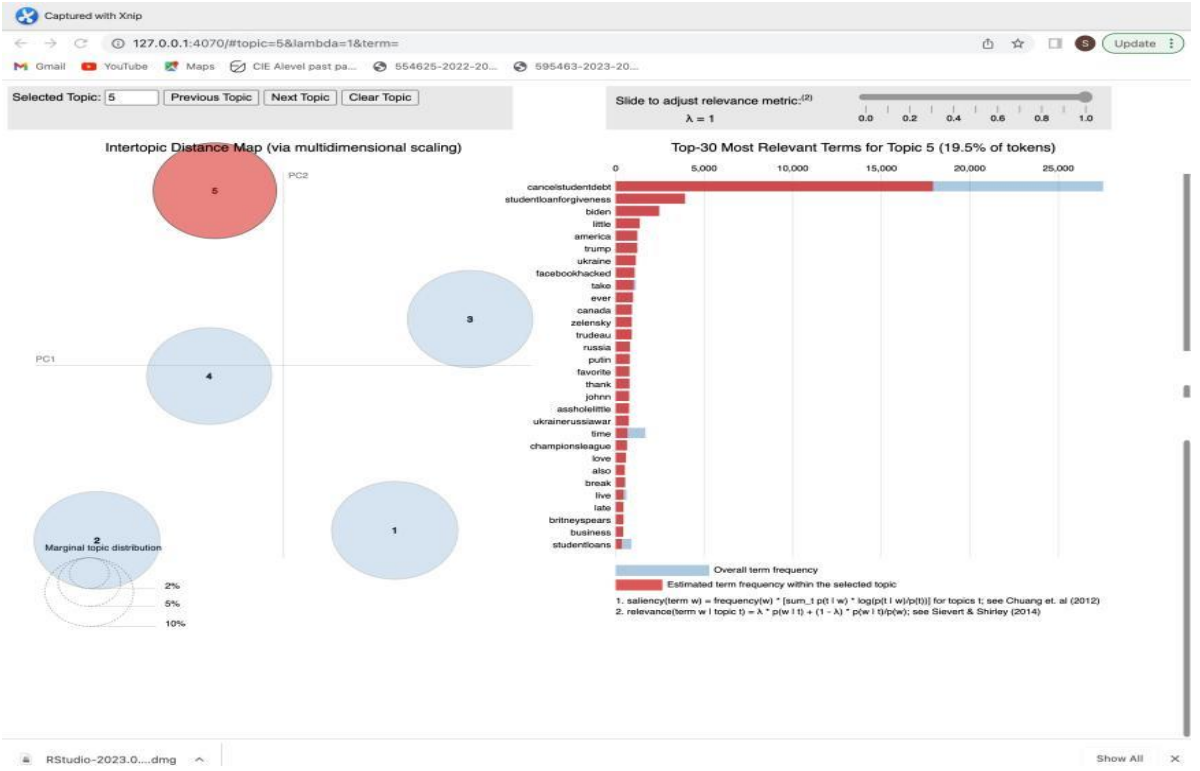

Figure 6: 19.5% of tokens from the intertropical distance map.

DOI: 10.54254/2753-7064/11/20231370

4. Discussion

Starting from the big picture, from the data collected and analyzed in topic 1, under the topic of people voting to cancel the student loan, besides some useless and standard terms, predatory might be an essential term for further investigation and discussion. It should be inferred that the public thinks this phrase confirms the belief that student loans are a take.

Topic 2 appears that donation or donate is a "clickbait" term which means it seems with very high frequency occurred under the topic of the effect of the current president, Joe Biden, on student loan forgiveness. The fact that multiple people are petitioning for student loan forgiveness is a testament to the social impact of student loan forgiveness. Many negative terms appear under this topic. For instance, cancer, bankruptcy, and fighting. This is a testament to the negative social impact that student loans can have.

Topic 3 is about student loans' effect on education. Surprisingly, the POTUS (president of the United States) has the highest number of hits, and the most relevant terms are politics and politicians. The positioning of the final data set does not seem very accurate because of the shift in focus and not much information related to education. But this is the main issue that people are discussing, not the issue of the data that is being compiled.

Topic 4 represents the tremendous effect that student loan forgiveness might bring out if the policy brings to development—omitting useless information that is politically relevant. Most of the relevant information on this topic supports student loan forgiveness. But one of those words, "nothing," may represent some people's view that nothing may change with student loan forgiveness.

Topic 5 is about the possible international impact of eliminating student loans, except in the United States. I think this data set is also off the main topic, for it can be seen that most of the opinions discussed revolve around the Russian-Ukrainian war and the leaders of the two countries, not the international impact that student loans can bring. The rest of the data also does not contribute to the analysis of the topic.

5. Conclusion

In summary, this passage discusses the significant financial burden of student loan debt in the United States, the different types of federal student loans available, and their stable interest rates compared to private loans. It also highlights the varied opinions surrounding student loans on social media. It mentions a research study that explores the potential benefits of rapid loan repayment and suggests policymakers consider offering low fixed interest rates. The passage then introduces a new research approach that uses the machine-driven classification of open-ended responses to analyze Twitter data and understand general public opinions on the issue of student loans. The researchers used secondary data from Twitter API and MDCOR software to collect and classify 26,708 comments. MDCOR was used to transform textual comments into data, and the researchers used both quantitative and qualitative methods to analyze the data. The software allowed for efficient data analysis and provided quantitative and graphical results. The researchers excluded common words to improve the system's efficiency and used machine learning sampling parameters to identify the number of topics. The Twitter API ensured the authenticity and accuracy of the data, and MDCOR saved research time and increased efficiency. While the software may not be 100% accurate, it is a better choice than analyzing comments manually. After getting the result, we can tell that the issue of student loans has been a highly debated topic recently, with many people advocating for the cancellation of student loan debt. To better understand public opinion on this issue, an analysis was carried out on five different topics related to student loans. The findings revealed mixed ideas and perspectives, with some key terms standing out in each topic.

DOI: 10.54254/2753-7064/11/20231370

Acknowledgement

They contributed equally to this work and should be considered co-first authors.

References

[1] *Office of US Department of Education. (n.d.). Federal student loans for college or career school are an investment in your future. Federal Student Aid. Retrieved April 15, 2023, from https://studentaid.gov/understand-aid/types/loans*

[2] *Canché, M. S. G. (2023). Machine-driven classification of open-ended responses (MDCOR): An analytic framework and no-code, free software application to classify longitudinal and cross-sectional text responses in survey and social media research. Expert Systems with Applications, 215, 119265.*

[3] *IQVIA company. (n.d.). What is text mining, text analytics, and Natural Language Processing? What is Text Mining, Text Analytics and Natural Language Processing? Linguamatics. Retrieved April 15, 2023, from https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing*

[4] *Robinson, J. S. and D. (n.d.). 6 topic modeling: Text mining with R. 6 Topic modeling | Text Mining with R. Retrieved April 15, 2023, from https://www.tidytextmining.com/topicmodeling.html*

[5] *Phat Jotikabukkana. (n.d.). Social media text classification by enhancing well-formed text trained ... Retrieved April 14, 2023, from https://www.researchgate.net/publication/316030904_Social_Media_Text_Classification_by_Enhancing_Well-Formed_Text_Trained_Model*