Spring 4-28-2023

Using Machine Learning to Measure Political Polarization on Social Media

Veronica Cagle

Using Machine Learning to Measure Political Polarization on Social Media

Veronica Cagle
University of Mary Washington

Introduction

There have been a lot of discussions regarding U.S. politics that the country is becoming more politically polarized[8,15,23]. There is a lot of disagreement over the topic of political polarization. Depending on what group is being measured can also impact expert opinions. We can look at the public[14] or we can look at party elites[25]. There is the potential for political polarization in many different groups of people. I focused on the public and if we as a society have become more polarized over the years. Some claim that we have, especially considering political events in the past few years while others disagree[1]. Polarization creates a divide between us and makes it difficult for us as a society to reach a consensus. In a democracy, it is important for us to come together to make a decision that benefits everyone.

With an increase in social media sites for information to spread and the public being able to communicate more freely with each other, there is a way for polarization to be measured[2]. Online discourse is increasingly important and openly available for researchers, or anyone interested, to look into. Social media should reflect some of the public's opinions and allows for the ability to measure the degree to which polarization has changed or stayed the same[27].

Polarization itself is a very nuanced topic. For me to be able to understand political polarization, I had to perform a Literature Search. This entails a search of published work to find similar research to what I am interested in. I needed to see the work that had already been done, so I could separate myself and have a unique finding to present. I spent a good amount of time reading about political polarization and how to measure it[13,18,22]. As I read through many papers, I realized that there is not a simple agreed-upon definition for polarization. I, along with three other researchers on my team, came together to discuss our definition of polarization. I took into consideration the papers that I had read as well as my own idea of what I thought polarization meant. We spent a lot of time discussing our own definitions before deciding on our definition. We landed on the idea that polarization can be broken down into style, purpose, and mindset. The style can be described as combative, while the purpose is to win an argument, and the mindset is stubborn. On the opposite side, not polarized has a calm style with the purpose of learning and forming opinions and keeping an open mind. Because polarization is such a complex topic, this was the best attempt at creating a definition that I could have in my mind when reading through public thought.

As mentioned earlier, social media has increased the data that is available. This data is most often text data that users have typed and posted on a website[7,13,22]. To go through all of this data, I went through the process of text mining[29]. This is a way of examining large amounts of text and finding patterns and useful insights about the data. Related to text mining, I used many techniques

of natural language processing to process text data. Natural language processing is a type of linguistic analysis that helps a model understand text data[6]. It tries to understand language in the same way that humans do and pick up on which words are more important to the context than others. I essentially wanted to train a model that could understand the text as well as a human could.

With access to text data online and my definition of polarization, I needed to create a model that could predict whether the text that I give it is polarized or not polarized. This is classification. I need to get labeled data that states whether the text is polarized or not polarized and feed it to a model to try to pick up on the patterns of what makes something polarized. Then, the model should be able to take unlabeled data and predict polarized or not polarized.

An approach to natural language processing known as sentiment analysis aims to identify the emotional tone of the text[19]. This process takes something like a movie review and tries to identify if the user had a positive, negative, or neutral feeling about the movie they wrote the review on. A relatively simple task for a human being is a hard task for a model to be able to do. Sentiment analysis is a classification task that is used often to be able to understand people's feelings on a subject. The classification task of polarized or not polarized is a little more complicated as there is so much involved in the definition of polarized.

Dataset & Preprocessing

There are many social media sites where people express their opinions on a variety of topics. As I was going through the literature search, I found that a lot of people use Twitter[7]. A downside to Twitter is that there is a character limit and does not allow for someone to potentially say as much because of that limit. A social media site that does not have much research done on it, especially regarding political polarization is Reddit. Reddit is a website made for discussions on any topic you want. It is split into subreddits, which are communities related to a more specific topic. You can find a subreddit on any hobby or activity that you are interested in. Regarding my research, there is a multitude of subreddits on political topics[28]. When you sign up for Reddit, you create a username that allows users to stay anonymous. A benefit to this is that when users are anonymous, they are likely to state their opinions more freely without judgment from those who might know them. Reddit has an open API that I used to download the comments on various subreddits. For all of these reasons, I decided to focus on Reddit as my social media platform to measure political polarization.

I searched through the various politically themed subreddits and found 25 that seemed to have a decent number of posts and comments. I tried to include a variety of subreddits from Democrat to Republican and those in between. I also have some subreddits that are focused on issues, such as gun control. Some examples of subreddits that I used are r/liberal, r/republican, r/progun, and r/moderatepolitics. I decided to download comment "threads", which are hierarchically nested

structures of user-generated replies to posts. There is an initial post where a user starts a discussion. I took the top-level comments on that post and the replies to that top-level comment. I wanted to have a manageable amount of data, hence only taking the top-level comments and their replies. Reddit was launched in 2005 and the addition of subreddits was added in 2008. Some of the subreddits were created in 2008 and some were created in later years. I collected threads from each subreddits formation to 2022. I was able to collect 6,263,110 threads in total.

In order for a model to be able to perform classification, there should be labeled and unlabeled data. Labeled data in my case, is the threads labeled as polarized or not polarized. I did not have any labeled data as I had just downloaded the text from Reddit. This project started with a team of four. Each of us individually labeled threads as polarized or not polarized. After everyone labeled the same threads, we came together and reviewed the ones that were not unanimous. We discussed why someone thought it was polarized or not polarized and discarded the ones that we disagreed on. This was a long process as polarization is very tricky to distinguish. We were able to create a dataset of 152 labeled threads.

A main concern was the lack of training data. With over 6 million threads downloaded, it would be pretty difficult for a model to learn the nuances of polarization from 152 labeled data points. Due to this, I recruited and led a team of seven undergraduates and one professor. I coordinated and planned meetings with this team to try to explain to them the definition of polarization. I had everyone label the same 30 threads to start and make sure that they understood. With some more explanation and review of the threads that we disagreed on, I felt confident in their ability to label the data. I had them label the thread as polarized, not polarized, or other. I discarded the “other” ratings and any thread that did not have at least four unanimous ratings. I was able to get an additional 370 labeled threads from this group, bringing the total of labeled threads to 522.

To measure the level of inter-rater agreement, I used Fleiss Kappa. Fleiss Kappa is a statistical test used when there is a categorical rating involved[11]. The test measures the degree of agreement between raters over what would be expected by random chance[26].A positive number indicates that there is agreement between raters. As seen in Figure 1, most of our agreement falls between 0.4 and 0.6 which indicates a moderate strength in agreement[26]. With the complex topic of polarization, it is pretty impressive that the overall Fleiss K was 0.47806. I was confident to move forward with the labeled data that I had gained from the team based on these results.

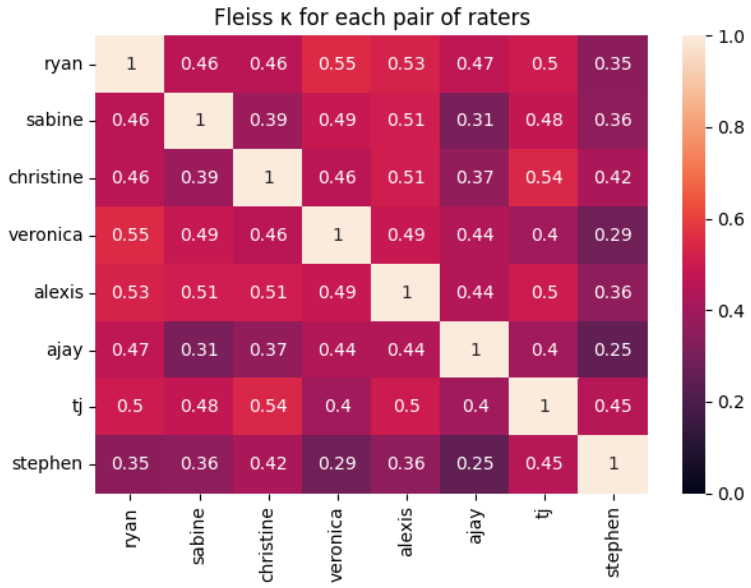

Figure 1: Fleiss Kappa for raters

To run a classification model, it is essential to have training and testing data. Training data is what you feed the model to learn and pick up on patterns in the data. The testing data is what you use to see how well the model is performing. It is necessary to have a different dataset to train and test on otherwise the model would be able to learn all of the existing data and give false hope of great performance. I use the testing data to see how accurately my model is able to predict what category a thread belongs to. Cross-validation is another method to train and test the model. It is a resampling method that takes different portions of the data to test and train a model on different iterations[5]. The training data includes the threads that my team was able to label.

Before building a model and giving the training data to it, there were some pre-processing steps. This is unstructured data where there is no real format. It is the sentence(s) that the user typed on Reddit, and I am transforming it into structured data that has a readable and understandable format that can be analyzed to identify patterns.

Data that was gathered directly from social media is a bit messy. I had to clean the data that I had gathered. There were some threads that contained bot messages. I removed those. Converting all the text to lowercase and removing punctuation are common practices used as well. Some other pre-processing steps were looking into stemming as well as the removal of stop words[16]. Stemming reduces a word to its stem so it reduces the total number of words that are input into a model[24]. For example, the word likes and liked could be reduced to the stem like without losing the meaning of the words. Stop words are commonly used words that tend to be less important. Some stop words are “the”, “and”, “ is”.

Methods

Machine learning algorithms cannot take text as input. The text needs to be converted into vectors of numbers. I decided to use a bag of words model. This is a way of extracting features from the text so that textual data can be used with models[12]. It looks for the presence of a word and does not take into account the order of the words[12]. This model finds all the unique words in a corpus and then can use a method to indicate the presence of the word. Different methods include count, freq, tf-idf. Using these methods, the vectors are created to be inputs.

Word embeddings are a representation technique that allows for words with similar meanings to have similar representations[4,20]. For example, the word president would have a very close representation to the word leader or vice president. Embeddings look at large volumes of text and learn which words tend to appear with other words and infer that they are close in meaning. Words closer in the vector space are expected to be closer in meaning.

There are many classification algorithms that I had the option of using. I decided to use a neural network[3] as they have recently set the standard for accuracy in text mining. I built a multilayer perceptron with two 64-unit layers. A neural net takes the vectors representing data as input. It then goes through a series of layers that I define to search for patterns in the text. Weights are randomized before learning begins and as the input is fitted to the model, the weights are learned and changed. These weights are multiplied and added together as the data is moved to the next layer. Eventually, a number from 0 to 1 is outputted. 0 indicates not polarized and 1 indicates polarized.

In an attempt to reduce the number of input variables and improve the performance of a predictive model, I looked at some features. I looked at the average word length, frequency of in-thread quotes, and frequency of links. Adding these in as features aided in the goal of helping the model pick up on polarized texts[9]. I also looked at lexical diversity which is the ratio of unique words to the total number of words. Another feature I looked at was n-grams. N-grams are a sequence of n-words. A unigram has n=1, so it is only one word. A bigram (n=2) is two adjacent words. For example, some common bigrams I saw were Sleepy Joe and fake news. I looked at the most informative 1,000 unigrams and bigrams present in the text.

I ran a program that tested using different features to be able to see what the best accuracy I could get was. Depending on what features I used, the classifier performed with an accuracy between 75% and 80% on a test set. With the disagreement between human raters on polarization, it is a pretty good performance from this classifier.

Discussion

I ran the classifier on the 6 million threads and computed the percentage of threads in each month that are classified as polarized. Figure 2 shows this plot, and you can see that there is not a strong

pattern indicating a rise in polarization overall on Reddit. There are a lot more spikes towards the left side of the plot likely due to fewer threads because Reddit was new at the time. The more data points there are, the more representative it is of the whole population. As the law of large numbers states, as the sample size increases, the mean gets closer to the average of the whole population[10]. With less points, the data is more prone to those spikes. I think it is interesting to look at the spikes in polarization and see if there is a political event that lines up with that spike. Right around 2021, there is a big spike. I speculate that this is right around the time that President Biden was elected and caused there to be an increase in polarization.

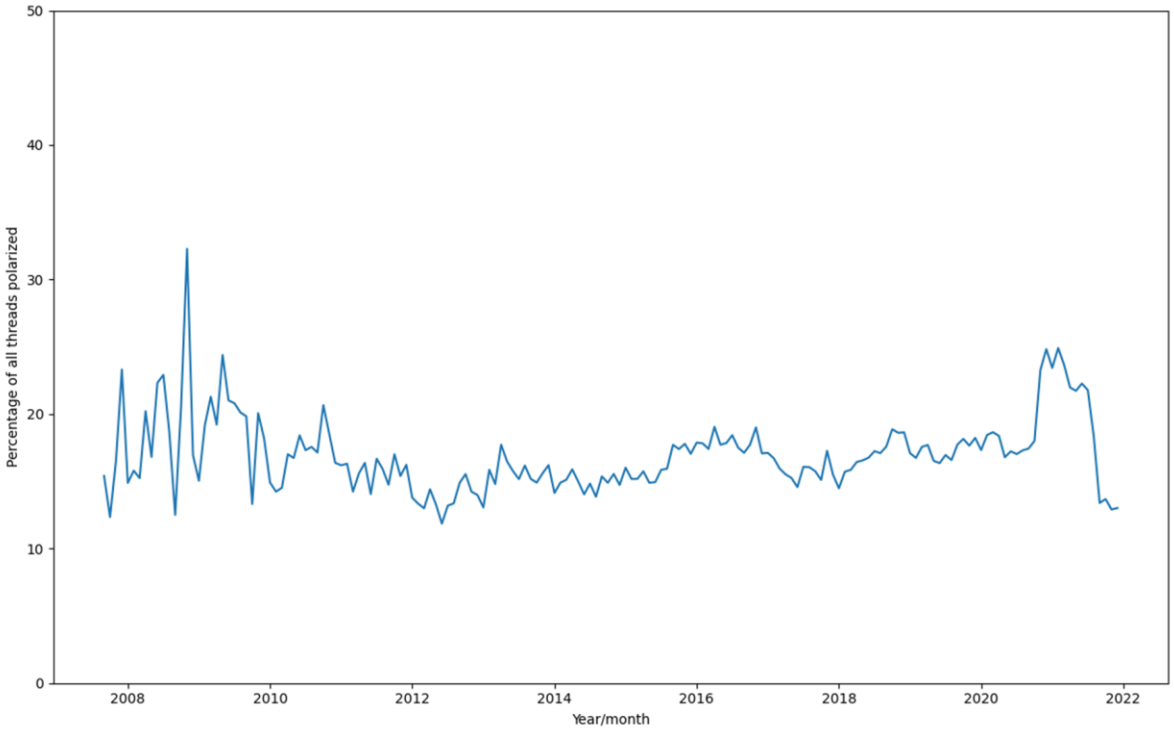

Figure 2: Overall Percentage of Threads Polarized by Month

Figure 3 is similar to Figure 2 but is broken down into subreddits. There are a lot more spikes likely because there are fewer threads. When looking at specific subreddits, like r/liberal, which is the red line in the upper middle pane, there is a consistent rising trend. For many of the subreddits in the upper left pane, it appears they are slowly rising in polarization as well starting in 2014/2015. If I were to make speculation about why, my guess would be because Donald Trump was running for President. Again, in the upper left panel, we see a purple line for r/republican. There is a significant dip in polarization right around 2016. As there was a Republican President elected right around then, it would make sense for that subreddit to decrease in polarization right around that time.

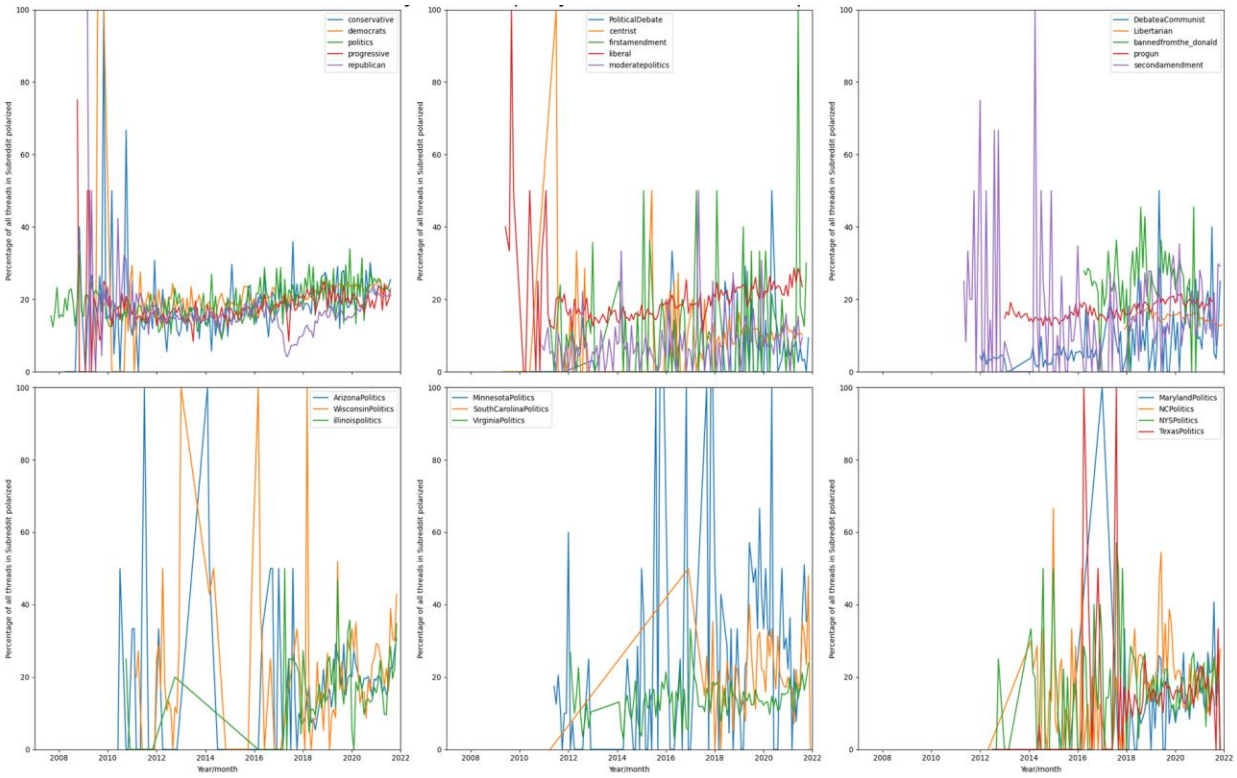

Figure 3: The percentage of threads polarized per month, by subreddit

Conclusion

This paper described the disagreement over whether society has become more polarized. The solution to this problem was building a multilayer perceptron neural network to measure political polarization on social media. The model performed with an accuracy between 75% and 80% on an independent test set. When run on threads from all 25 politically themed subreddits, there does not seem to be an increase in polarization overall on Reddit. However, there is an indication of an increase in polarization when you look at individual subreddits.

Figure 2 shows that while there is not necessarily an overall increase in political polarization, there are spikes that can be matched up with political events. There is a large spike right around the beginning of 2021. While this is just my speculation, the January 6th attack on the Capital occurred in the beginning of 2021 and lines up with the timeline of why there would be an increase in polarization at that time. The beginning of 2021 was also when President Biden was elected which could have impacted this spike in polarization. As seen in Figure 3, r/republican has a decrease in polarization right around 2016. I speculate that this is because of the election of Donald Trump, a Republican candidate. Some of the spikes and dips in polarization can be connected to political events.

It seems that overall, as a society, we are not becoming more politically polarized, but certain groups or areas of our society are. As a country, we are not all becoming more polarized, but we

need to be aware that some groups may be forming that are becoming more polarized and will start similar events to the January 6th attack.

Acknowledgments
I would like to thank Alexis Kochanski, TJ Davies, Stephen Davies, Sabine Frye, Christine Wehner, Ryan Stewart, and Ajay Mathew for their contributions to this project.

References

[1] A. I. Abramowitz and K. L. Saunders, “Is Polarization a Myth?,” The Journal of Politics, vol. 70, no. 2, pp. 542–555, 2008.

[2] Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2019), “Discovering Political Polarization on Social Media: A Case Study,” 2019 15th International Conference on Semantics, Knowledge and Grids (SKG), 182–189. https://doi.org/10.1109/SKG49510.2019.00038.

[3] Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2020), “Learning Political Polarization on Social Media Using Neural Networks,” IEEE Access, 8, 47177–47187. https://doi.org/10.1109/ACCESS.2020.2978950.

[4] Bengio, Y., Ducharme, R., & Vincent, P. (2000), “A neural probabilistic language model,” Advances in neural information processing systems, 13.

[5] Berrar, Daniel, "Cross-Validation," (2019): 542-545.

[6] Chowdhary, KR1442, and K. R. Chowdhary, "Natural language processing," Fundamentals of artificial intelligence (2020): 603-649.

[7] Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2011), “Political Polarization on Twitter,” Proceedings of the International AAAI Conference on Web and Social Media”, 5(1), Article 1. https://ojs.aaai.org/index.php/ICWSM/article/view/14126.

[8] D. Baldassarri and P. Bearman, “Dynamics of political polarization,” American sociological review, vol. 72, no. 5, pp. 784–811, 2007.

[9] Deng, X., Li, Y., Weng, J., & Zhang, J. (2019), “Feature selection for text classification: A review,” Multimedia Tools and Applications, 78(3), 3797–3816. https://doi.org/10.1007/s11042-018-6083-5.

[10] Erd, Paul, "On a new law of large numbers," J. Anal. Muth 22 (1970): 103-l.

[11] Fleiss, J. L. (1971), "Measuring nominal scale agreement among many raters," Psychological Bulletin, Vol. 76, No. 5 pp. 378–382

[12] Harris, Z. S. (1954), “Distributional structure,” Word, 10(2-3), 146-162.

[13] Hemphill, L., Culotta, A., & Heston, M. (2016), “#Polar Scores: Measuring partisanship using social media content,” Journal of Information Technology & Politics, 13(4), 365–377. https://doi.org/10.1080/19331681.2016.1214093.

[14] J. H. Evans, “Have Americans’ Attitudes Become More Polarized?—An Update,” Social science quarterly, vol. 84, no. 1, pp. 71–90, 2003.

[15] L. Boxell, M. Gentzkow, and J. M. Shapiro, “Cross-Country Trends in Affective Polarization,” National Bureau of Economic Research, Working Paper 26669, Jan. 2020.

[16] Luhn, Hans Peter, "Key word-in-context index for technical literature (kwic index)," American documentation 11.4 (1960): 288-295.

[17] Makrehchi, M. (2016), “Predicting political conflicts from polarized social media,” Web Intelligence, 14(2), 85–97. https://doi.org/10.3233/WEB-160333.

[18] Marozzo, F., & Bessi, A. (2018), “Analyzing polarization of social media users and news sites during political campaigns,” Social Network Analysis and Mining, 8(1), 1. https://doi.org/10.1007/s13278-017-0479-5.

[19] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams engineering journal 5.4 (2014): 1093-1113.
[20] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013), "Distributed representations of words and phrases and their compositionality," Advances in neural information processing systems, 26.
[21] Morini, V., Pollacci, L., & Rossetti, G. (n.d.), "Capturing Political Polarization of Reddit Submissions in the Trump Era".
[22] Nithyanand, R., Schaffner, B., & Gill, P, (2017), "Online Political Discourse in the Trump Era, " ArXiv:1711.05303 [Cs]. http://arxiv.org/abs/1711.05303.
[23] Pew Research Center, "Political Polarization in the American Public," Pew Research Center - U.S. Politics & Policy, 2014. https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/.
[24] Porter, Martin F, "An algorithm for suffix stripping," Program 14.3 (1980): 130-137.
[25] R. L. Claassen and B. Highton, "Policy Polarization among Party Elites and the Significance of Political Awareness in the Mass Public," Political Research Quarterly, vol. 62, no. 3, pp. 538–551, 2009.
[26] Scott, W. (1955),"Reliability of content analysis: The case of nominal scale coding," Public Opinion Quarterly, Vol. 19, No. 3, pp. 321–325.
[27] Serrano-Contreras, I.-J., García Marín, J., & Luengo, O. (2020), "Measuring Online Political Dialogue: Does Polarization Trigger More Deliberation? Media and Communication," 8, 63–72. https://doi.org/10.17645/mac.v8i4.3149.
[28] Soliman, A., Hafer, J., & Lemmerich, F. (2019), " A Characterization of Political Communities on Reddit," 259–263. https://doi.org/10.1145/3342220.3343662.
[29] Tan, Ah-Hwee, "Text mining: The state of the art and the challenges," Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases. Vol. 8. 1999.