Radicalization of Social Media Users

Jun Young Jeong[1], Jeesoo Kwon[2], Betzabeth Narvaez[3]

Yonsei University
[1]bryanjung1@gmail.com,
[2]kwonjeesoo20@yonsei.ac.kr
[3]betzynarv@yonsei.ac.kr

Abstract

As social media has become an ever present presence in our daily lives, there is a need to acknowledge the dangers and effects of the constant social media content consumption. One of the most dangerous effects is the radicalization of a user's beliefs through social media. The constant repetition of certain content on social media can lead to a user's beliefs to be reinforced and radicalized. This is extremely dangerous as it can lead to real-world repercussions as users may go on to act on their radicalized beliefs. By finding a way to quantify the radicalization of a user, we could use the information to deter the further radicalization of that user.

Keywords: Social Media, Radicalization

I. Introduction

1.1 Purpose of Study

The purpose of our project is to solve the radicalization of a social media user's belief that occurs due to social media. By radicalization of a user's beliefs, we define it as when a user's hate towards a subject increases over time. Radicalization of a user's belief through social media can occur in many ways, such as through a social media filter bubble. A filter bubble occurs when a user is constantly recommended very similar content on their social media feed due to the user's past behavior. The content recommendation algorithm in social media uses the user's past behavior to recommend new content, but it can quickly lead into repeating the same content recommendations over and over again and thus creating a bubble of content, views, and beliefs that the user will get stuck in. Being stuck in a filter bubble can be dangerous as it can lead to reinforcements of extreme views and beliefs which can lead to the radicalization of a user. To prevent this and to promote consumption of media from several perspectives, we decided to try to find a way to quantify the radicalization of a user.

Currently, the topic of radicalization through social media is only discussed in niche groups and there is not a lot of public attention on this topic. By conducting this project, we can bring more attention to the effects of it and the dangers of social media. Through this project, we could find more concrete information and confirmation of the radicalization of a user which could then be used to help prevent and deter the further radicalization of a user's belief.

1.2 Rationale

Regarding the data, there were 6 data sources in total with a single group of five variables and a control variable. The five variable data were selected with a hashtag #Trump on Twitter. The reason behind this choice was because we needed data with the account which is already showing the case of radicalized tweets or the tweets with strong emotions. With that data, the group thought that such accounts can later be analyzed to check if the users were always like that or if they became radicalized as the time passed. Also, #Trump was chosen as a keyword for the account research procedure because we thought searching with a political topic would have much more users with radicalized or extreme beliefs than any other topics. Moreover, for a control variable, we were looking for an account which does not reveal extreme emotion in its tweets in order to compare from other users' emotion analysis. This is why we chose a bot account, which simply posts the link and article titles from a certain website, without showing any strong emotions throughout its tweet.

1.3 Research Question

How can we quantify a degree of radicalization in social media users?

1.4 Methodology

Our analysis for our project included topic modeling and sentiment analysis. The reason for using the method of topic modeling is because we want to specifically focus on the topic the user is displaying signs of radicalization towards. So to find the overall topic of the user's tweets, we needed to use topic modeling, as it would be too difficult to go through thousands of tweets and determine the topic ourselves. For the topic modeling, we specifically used the Latent Dirichlet Allocation technique. This technique, also called LDA in short, is a representative algorithm for topic modeling. LDA assumes that documents are made up of a combination of topics, and that they generate words based on the probability distributions. With the data, the LDA will trace the process back by which the document was created. To use this technique for our project, we tokenized the words from tweets first with LDA and ran the actual LDA process to gain the topic of each Twitter account's tweets.

To quantify and check the radicalization of a user, we used sentiment analysis of the user's tweets. We picked the method of sentiment analysis because radicalization is usually displayed through strong emotions towards a topic, strong emotions that were previously not there. So since we are analyzing emotion, we went with sentiment analysis. Furthermore, to increase the quality of our sentiment analysis, we used the sentiment predictor from AllenNLP. AllenNLP is an open-source deep learning library for Natural Language Processing (NLP). We decided to use AllenNLP because it would not only be convenient to use with the result of high quality data but also provide better results compared to rudimentary sentiment analysis conducted through other python libraries.

II. Literature Review

What is radicalization?

According to the Cambridge dictionary, 'radicalization' is defined as 'the action or process of making someone become more radical (or extreme) in their political or religious beliefs.' Such radicalized behavior could be dangerous for an individual as these extreme views may result in a further contorted belief and eventually could lead to a serious accident regarding the belief. The literature below suggests the extremists in the United States and their relationship between social media.

The Use of Social Media by United States Extremists

From 'The Use of Social Media by United States Extremists,' Michael Jensen, Patrick James, et al., have collected the data from 479 people who exposed radicalized behavior between 2005 and 2016 and their social media activities and cast the relationship between the two factors. The data reveals some main points regarding this issue. Firstly, many U.S. extremists are being impacted greatly by the online social media, such as Youtube, Facebook, and Twitter, by accelerating their radicalization process. The literature states that "in 2016 alone, social media played a role in the radicalization processes of nearly 90% of the extremists in the PIRUS data." To be specific on this point, the data shows that in 2005, the average radicalization duration of an U.S. extremist was about 18 months, whereas in 2016, the average was at 13 months. Second, an individual who participates in the extremist activities by himself/herself (known as the 'lone actors') shows more activity on the social media than the group participants, with the social media impacting the lone actor 88.23% of their radicalization in 2016. Like such points, social media does play a significant role on the users and their radicalization process. As the paper also mentions, there needs to be a way to quickly identify if a social media user has begun the radicalization process and provide any support that could alleviate such symptoms if the user is being radicalized.

What is the role of social media?

Nowadays with the prevalence of social media in our daily lives, social media plays an important role in the radicalization of a user's belief. It is now an essential tool used by extremists and radicals to spread their beliefs and recruit people for their movement. Users themselves may not be completely aware of their exposure to this extreme content and may become influenced by it and in turn normalize and agree with these beliefs and behaviors thus the user becoming radicalized. The following piece shows how social media plays a part in the radicalization of individuals:

Isolation and Social Media Combine to Radicalize Violent Offenders

Daniela Hernandes and Parmy Olson come together to write a piece on how isolation and social media combine together to radicalize violent offenders. They introduce the concept that radicalization is 'driven by a need to matter and be respected' and that usually violence is a means to that. They point out that between 2005 to 2016, social media was an important factor in the radicalization of 50% of individuals in extremist groups. They also mention that the data shows that Islamist and far-right extremists, which also include white nationalists, are the most likely to be radicalized through social media. They also made the important statement that social media is accelerating the pace at which people can be radicalized. Furthermore, their piece includes research results for the National Consortium for the Study of Terrision and Response to Terrorism, conducted by the University of Maryland with the U.S. Department of Homeland Security. That research concludes that social media is just part of the radicalization process now, with Facebook being used by 64.5% of U.S. extremists between 2005 and 2016, Youtube being used by 30.6% of U.S. extremists, and Twitter being used by 23.4% of U.S. extremists between the same years.

III. Study Design and Methodology

3.1 Purpose of Research

The purpose of this research and project is to try to find a way to quantify the radicalization of a user through social media.

3.2 Ethical Consideration

In our project, there is one important ethical consideration that should be taken in account, which is a privacy concern. Due to the nature of our research and project, a user may feel that their privacy is invaded. In our research process, we will have to select a user and then obtain all their posted tweets and then analyze them for hateful speech. A user may not want all their tweets to be analyzed since some users tweet as an outlet with no real motivation so those tweets may be not representative of that user. Another possible ethical consideration that should

be noted is our inaction towards tweets or posts that may contain harmful words and that should have been reported.

3.3 Participants

Although the people we collected Twitter data from did not choose to participate in our analysis, we will list those users here. There were a total of 6 users we collected tweets from, those included:

- Ben Shapiro
- Charlie Kirk
- Ryan Fournier
- Stinchfield 1776
- VicToensing
- The Tech Platform

The first 5 users were found by exploring the #Trump. Additionally, the reason for specifically picking those 5 users is because we needed to pick users that were already displaying signs of radicalization, which in this case meant they were showing strong emotion, ie. hate, towards a topic. The reason for this is because after seeing how negative and hateful the users tweets are right now, then we would analyze all their tweets and determine if the users were always like this, as in having that level of hate towards a subject, or if they slowly became like that due to radicalization. For the last user, we decided to go with a user, or bot, that did not display any sort of strong emotion in their tweets to check how different the result of the analysis would be for that user. We picked The Tech Platform specifically because they, which is a bot, automatically posted the link and article title for any article published in the The Tech Platform website. So this user did not display any sort of emotion since it only tweets tech article titles.

3.4 Study Design

Our project followed the study design of Casual Design. We wanted to test our hypothesis which was that we could quantify ‘hate’ in a user’s tweets and use them to determine if radicalization has occurred through social media. We wanted to check that when we plotted the sentiment of a user’s tweets relating to a topic, if there was a trend of the sentiment line going very negative then it could indicate the radicalization of a user. The independent variables being the user, and the sentiments of the filtered tweets, while the dependent variables are the topic word found for each user based on their past tweets and their plot of the filtered tweet’s sentiments.

3.5 Study Limitation

The major limitation of this study is that a user may get radicalized due to outside sources, not due to social media. A user can start consuming media and content from sources such as the television, newspaper, and books, and start getting their extremist views reinforced from there. Users can also get radicalized by meeting groups and or clubs outside social media. There are many outside variables that could affect the user's belief and lead them to start posting what they think on social media. In our project, we are assuming that social media is the only source of media and content that the user is consuming, so any radicalization that occurs of a user's belief is a direct result of the social media filter bubble that keeps reinforcing those views. However, we know that that is not always the case.

IV. Data Analysis

4.1 Data Pre-Processing

After collecting all the posted tweets from each Twitter user mentioned in the 'Participants' section, we cleaned the tweets to be used for further analysis. The first step in the data preprocessing process to remove everything except text for each tweet. This meant having to remove all links, emojis, numbers, and symbols from the twitter text. After having done this, we saved the newly cleaned tweets in new csv files to be easily used for the next steps in the analysis.

4.2 Description of Data

Prior to the data preprocessing, the collected tweets from the users were downloaded in a csv file with a total of around 3,300 tweets for each user. Each file had 6 columns containing an 'id', 'created_at', 'text', 'likes', 'in reply to', and 'retweeted' column. The 'id' column identified the tweets in chronological order. The 'created_at' column had the data for when the tweet was created, with the format of 'year-month-day time'. The 'text' column contained the text for the tweet. The 'likes' column contained the number of likes, and the 'in reply to' contained a username if the tweet was in reply to someone else's tweets. Additionally the 'retweeted' column contained a boolean value specifying if the tweet had been retweeted.

After cleaning the data, for the next steps in the analysis, we only focused on the 'id' and 'text' columns. The 'text' column would be used in the topic modeling and sentiment analysis, while the 'id' column would be used in the plotting of the tweet's sentiment for each user.

4.3 Topic Modeling

The first step in our analysis process was to get the topic of each collected tweet. The reason for this is because in the later step of sentiment analysis, we needed to filter out tweets to

only conduct sentiment analysis on the tweets related to the found topic. To find the topic for each tweet we used the python libraries NLTK and Gensim and the topic modeling technique called Latent Dirichlet Allocation.

When starting the topic modeling process, we first had to further prepare each tweet text to be used in the topic modeling process. This was done by tokenizing, lemmatizing, and using the NLTK English stopwords to filter out any unnecessary words in the text. To make this easier when performing this for multiple users, we created a 'prepare_text_for_lda' function which would take in the text and perform the mentioned steps. After preparing the text, we got the topic for each tweet using the Gensim LDA model, then the function returned the overall topic for all the tweets for the user. Then we used the function on each user and saved each found topic for use in the sentiment analysis.

4.4 Sentiment Analysis

The next step in our analysis was to conduct a sentiment analysis on the tweets relating to the found topic. To do that, we created a function that would filter tweets based on the found topic and then it would save the filtered tweets in a list. After this, to conduct the sentiment analysis, we used the Allen NLP library's 'sent_predictor' function on each tweet in the list, which was then saved in another list. This sentiment list of tweets would then later be used for creating a plot of the sentiment of each filtered tweet for each user.

4.5 Analysis of Sentiment Plot of Users

In order to see the trend of sentiment for the found topic for each user, we decided to plot the sentiment of the filtered tweets. We did that by using the sentiment list of the filtered tweets for each user, then plotting each sentiment by chronological order. The resulting plot had a y axis of sentiment with an x axis of the tweet chronological order and showed the trend of sentiment through time for the tweets relating to the topic found for a specific user.

V. Validity and Reliability of Research

5.1 Validity of Research

Due to the nature of our analysis, in that due to that fact we are analysing the radicalization, we had to select participants that were already displaying signs of radicalization, ie. hate speech, it is inevitable that there will be at least some bias. Another aspect that impacted the analysis is that we also chose users who had a long history of tweets, meaning they were not new users rather they were already established users. If this analysis were to be replicated in other circumstances, such as using new users that displayed no signs of radicalization, the results could come out completely different and incorrect. Otherwise, with the resources and analysis

methods we used, we could determine that there is at least the potential to see the trend of radicalization of a user.

5.2 Reliability of Research

The analysis techniques used in our project are already established techniques that have been used many times in other projects before, therefore we know the analysis techniques used are reliable. Although we know there are several other techniques we could have used for our analysis, we concluded that techniques used were best for us as they were reliable, had easily available documentation and tutorials, and were within our skill level.

VI. Conclusion, Suggestion and Further Developments

6.1 Conclusion

After our conducted analysis, one result we got from Ben Shapiro's tweets is that the chosen topic using LDA with the get_topic function is "people", as shown in the Appendix Figure 3. This means that for most of the tweets on his timeline, he had 'people' as a subject.

Also, when we ran the get_sentiment_timeline function using the AllenNLP (elaborated in Appendix Figure 2), we could get a plot like shown in Appendix Figure 4. This plot represents the sentiment of a user's- which in this case is Ben Shapiro's- tweets along the timeline. The x-axis represents the timeline, and the y-axis represents a sentiment analysis using the AllenNLP; positive numbers reveal the positive sentiment, and negative ratings reveal the negative sentiment. To analyze Ben Shapiro's sentiment, it is showing a positive sentiment on average, as the drawn lines are above the zero. However, there are some periods of time between points 2 to 4 on the x-axis in the graph when the user had a decline in his sentiment, but began to have a positive sentiment again after period 4.

This analysis was showing different results than what we were expecting; like explained in 3.3 Participants in this paper, because the five users were picked as we thought they were already radicalized, we were expecting the plot graph to be at a negative level. However, Ben's graph was showing a positive sentiment analysis from tweets. We then asked ourselves the reason behind this; the sentiment was only analyzed on topics filtered by the topic generator function. Thus, we can understand that Ben Shapiro was posing a positive stance regarding the word 'people.'

From the results, we found that there is potential to see a trend of 'hate', or at least a trend of negative and positive feelings towards the found topic of the user. This could be used to signal when a user started becoming radicalized by checking their plot of the sentiments of their

filtered tweets. By seeing the time when their sentiment starts to go negative, we can infer that that is when they started showing signs of radicalization. If this project is further developed and automated, it could work as a tool to alert users when they start to show signs of radicalization, which then can be used for helping deter the further radicalization of the user.

6.2 Suggestions

One of the suggestions we have after having completed our analysis is that our analysis would benefit from creating a more sophisticated plot when plotting the sentiment of the filtered tweets for each user. Not only would this make it more aesthetically pleasing, it would make the plots easier to read and lead to trends being determined more easily.

6.3 Further Development

There are further steps and developments that one could make in this analysis and overall project. One of those is that we could calculate an actual degree of radicalization for each user. This could possibly be done by using the plot and sentiment list of filtered tweets to get a standard or the average of the sentiments for each user, then checking their present tweets' sentiment and calculating the degree of deviation from that average. This could be further developed so that when a user displays a significant degree of deviation, they could be alerted of their radicalization and be provided with some resources to combat it so that hopefully they can revert their ongoing radicalization through social media.

VII. References

Hernandez, Daniela, and Parmy Olson. "Isolation and Social Media Combine to Radicalize Violent Offenders." The Wall Street Journal, Dow Jones & Company, 5 Aug. 2019, www.wsj.com/articles/isolation-and-social-media-combine-to-radicalize-violent-offenders-11565041473.

Lydia Emmanouilidou, "Arab Uprisings: What Role Did Social Media Really Play?" The World from PRX, www.pri.org/stories/2020-12-17/arab-uprisings-what-role-did-social-media-really-play.

"Radicalization." Cambridge Dictionary, dictionary.cambridge.org/dictionary/english/radicalization.

"The Use of Social Media by United States Extremists." The Use of Social Media by United States Extremists | START.umd.edu, July 2018, https://www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf.

VIII. Appendix

Figure 1▼

```
[17] # return topic for tweets
     def get_topic(tweets_file):
       """ used with the path of a file with only tweets
       and returns the topic for that file"""

       # get tokens for each tweet
       text_data = []
       with open(tweets_file) as f:
         for line in f:
             tokens = prepare_text_for_lda(line)
             if random.random() > .99:
                 text_data.append(tokens)

       # create dict from data, convert to bag-of-words corpus and save both for future use
       dictionary = corpora.Dictionary(text_data)
       corpus = [dictionary.doc2bow(text) for text in text_data]
       pickle.dump(corpus, open('corpus.pkl', 'wb'))
       dictionary.save('dictionary.gensim')

       # find 1 topic
       ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = 1, id2word=dictionary, passes=15)
       ldamodel.save('model3.gensim')
       topics = ldamodel.print_topics(num_words=1)

       # get string topic
       new_topic = topics[0][1].split('"',1)[-1]
       new_topic = new_topic.rstrip(new_topic[-1])
       return new_topic
```

Figure 2▼

```
[19] # make function for allen nlp use
     def get_sentiment_timeline(tweets, topic):
       """ uses tweets of that user combined with the topic and return timeline
       of sentiments for topic-related tweets"""

       # filter tweets based on topic

       filtered_tweets = list(tweets[tweets['text'].str.contains(topic)])

       # get sentiment list
       sentiment_list = []
       for tweet in filtered_tweets:
         sentiment_list.append(sent_predictor.predict(tweet)['logits'])

       # get parsed lsentiment list to use for plot
       parsed_sent = []
       for items in sentiment_list:
         parsed_sent.append(items[0])

       # plot tweets and thier sentiment in chronological order
       len_x = range(len(parsed_sent))
       plt.figure(figsize=(16,16))
       plt.plot(range(len(parsed_sent)), parsed_sent)
       plt.axhline(y=0,color = 'r')
       plt.yticks([-4,-3,-2,-1,0,1,2,3,4])
       plt.show()
```

Figure 3▼

```
]  1 benShapiro_topic = get_topic('/content/sample_data/benShapiro_tweets.csv')
   2 print("Topic of Ben Shapiro's Tweets: ", benShapiro_topic)

Topic of Ben Shapiro's Tweets:  people
```

Figure 4▼

```
1 # get string topic
2 get_sentiment_timeline(new_cleaned_benShapiro_t, benShapiro_topic)
```