REDDIT ANALYSIS: WALL STREET BETS

Forum data collection, user network graphs, and entity identification with NLP | 11.16.19

Reddit is a great forum for finding useful information. Many of my Google searches take me to niche community conversations that have already taken place on Reddit, and more often than not I end up learning a lot more about the topic at hand. Recently, I began learning about stock market investing and trading. One reputable Reddit community on stocks, named *Wall Street Bets*, has a strong community of over 730,000 followers, as shown in *Figure 1*. After reading an article in *Business Insider* about a user who made over $100,000 in two trades, I decided to focus my Reddit analysis on this community that labels their followers as "degenerates" and whose description reads, "like 4chan found a Bloomberg terminal."

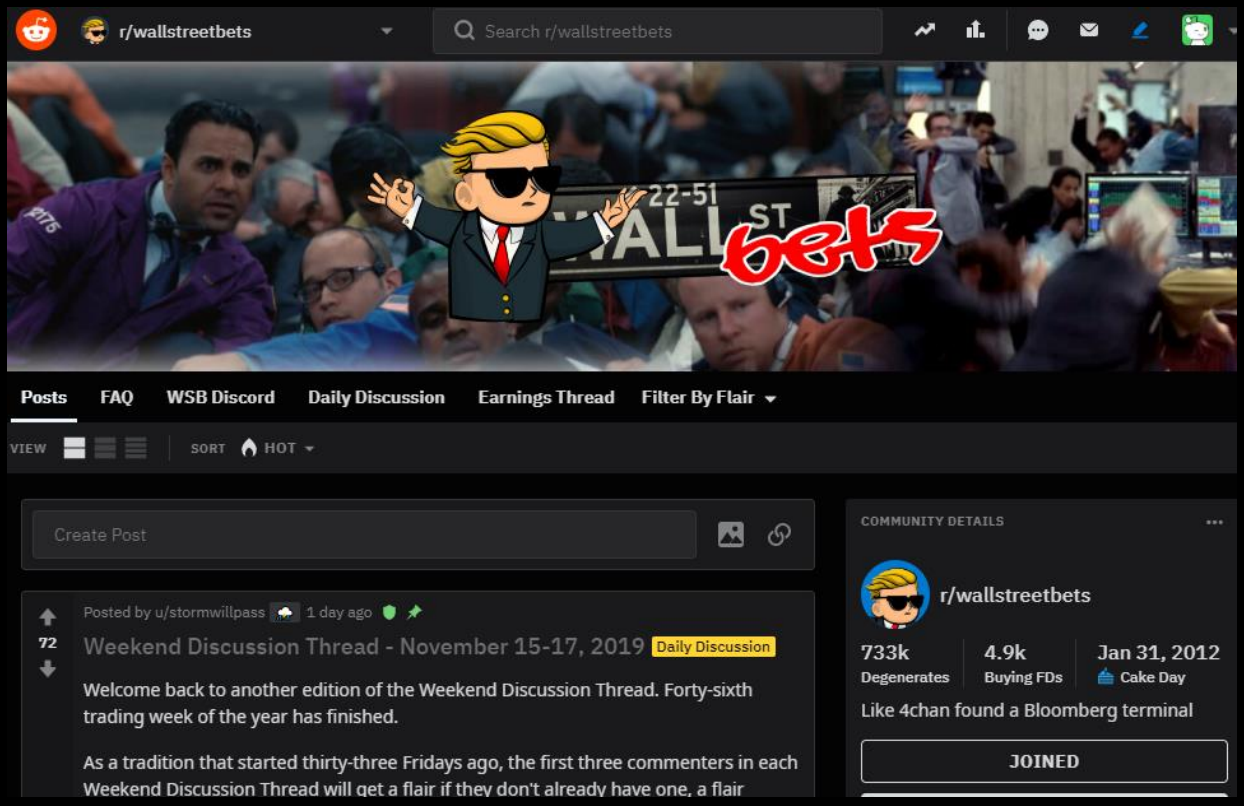

Figure 1. Reddit channel 'Wall Street Bets'

1 | Collecting Forum Posts on Keyword: "Stocks"

Using the R package *RedditExtractoR* is an easy way to collect Reddit posts based on your keyword of choice. For this exercise, I searched the forum for the keyword "stocks". *Figure 2* shows the first results, the date of posting, URL, and the number of comments made on each post. Some posts are about specific companies or industries, and others are about investments in general. The post with the highest comments (682) reads, "can't pump stocks on a Sunday, bro", while the post with the least comments (40) reads, "I just realized Ed Sheeran was telling us to short WeWork. We need to start taking stock tips from him, guys." For my network analysis, I used the latter.

	date	num_comments	title
1	16-11-17	213	Today I broke through the $10,000 milestone after my first 2 months of trading! I d like to thank the entire V
2	03-09-17	159	6 Month Anniversary. You can see exactly where i stopped listening to r/Robinhood and started YOLOing all
3	11-09-19	642	Study proves day trading stocks is not profitable, 97% lost money and only 0.4% earned more than a bank t
4	18-08-19	682	Can t pump stocks on a Sunday bro
5	16-11-19	106	'An overextended market losing its engines': A notorious bear says tons of red flags are piling up around stc
6	13-11-19	125	WTF is the news on $DIS? stock is mooning
7	12-11-19	100	Renewable Energy Stocks
8	11-11-19	40	I just realized Ed Sheeran was telling us to short WeWork, we need to start taking stock tips from him guys
9	30-10-19	113	Twitter ($TWTR) announces it will ban political ads. Stock down 3% AH
10	02-11-19	164	UBER stock is gonna tank on Monday AH into Tuesday. Get the puts ready.
11	12-11-19	55	Base your Stock Portfolio on your Astrological Sign, what could go wrong?
12	01-10-19	395	TD Ameritrade cuts fees. $0 stock, ETF fees. $0.65/contract for options.
13	04-11-19	77	Can do we some type of match making where one person buys a call and his opponent buys a put and they
14	15-11-19	73	Barclays makes the case for an Uber stock 'double'

Figure 2. Reddit posts on the keyword "stocks"

2 | User Network Plot

To create a relatively small network of post interactions, I use *RedditExtractoR* to filter the results and return the URL of the post with the minimum number of comments. This post with the least comments about Ed Sheeran and WeWork is shown in *Figure 3*. After extracting the content related to this post, I use the R package *dplyr* to plot the network of commentary as shown in *Figure 4*. According to the plot, there are at least five main contributors to the post, each with their own separate, but related, conversation. A quick glance at the nodes of this interactive graph show that all conversations about this post are humorous with no real stock information.

Figure 3. Reddit "stock" post

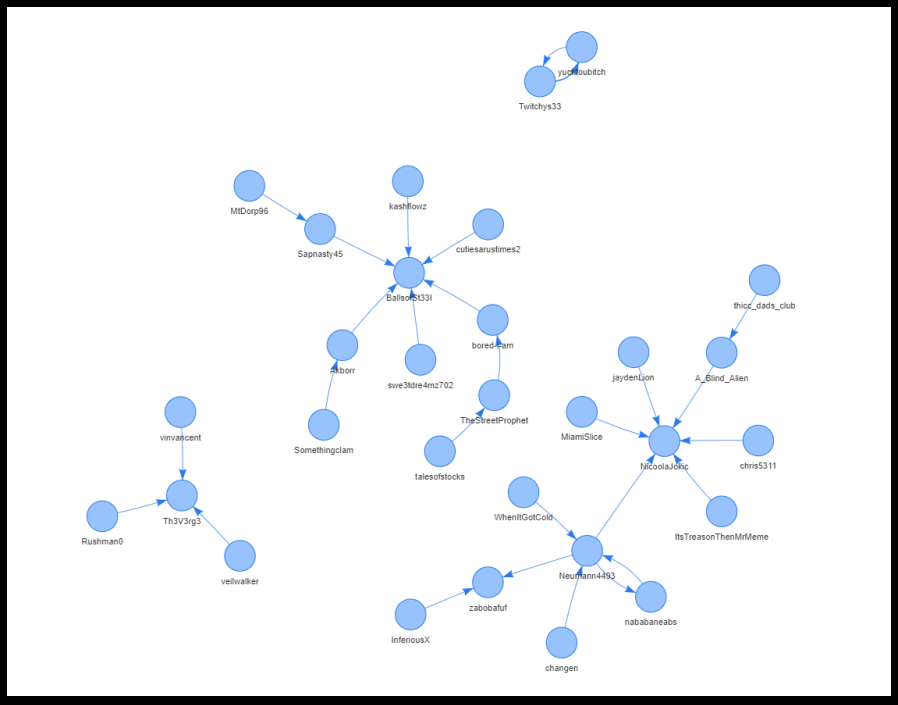

Figure 4. User network of commentary on the "Ed Sheeran/WeWork" post

3 | Extracting Named Entities

To try and probe these conversations for specific content, I use the R package *openNLP*. After creating annotators that run a Maxent algorithm, I can loop through the comments to list any suspected entities. *Figure 5* shows the extracted entities, their type, post number, entire

comment and position in the comment. For this post, the entity extraction of the comments is 50% accurate; four entities were correctly identified and the other four were either slightly or completely mistaken. Of the 4 correctly identified entities, 3 entities were people and 1 was a monetary reference. This indicates that the NLP package is better at recognizing people among named entities. Of the 4 incorrectly identified entities, 2 were of locations, indicating that locations are harder to identify for *openNLP*.

In conclusion, most replies to this comical subreddit post about a celebrity are, in fact, more comical posts about celebrities. Go Reddit. These results are probably typical of the conversations that lurk around the subreddit community known as *Wall Street Bets*.

	Post	Type	Entity	Position	Comment
1	11	Location	Buy	1	Buy windowless vans.
2	12	Person	Adam Neumann	1	Adam Neumann should be the face of this sub. He is full blo...
3	25	Money	$1M margin	46	But you can bet your ass he s old enough for $1M margin
4	28	Person	Adam Neumann	203	All I know is that Ashton Kutcher is a shill who thinks just be...
5	28	Organization	Ashton Kutcher	20	All I know is that Ashton Kutcher is a shill who thinks just be...
6	34	Person	Jerry	1	Jerry's neighbor. The fat mail dude.
7	37	Date	://youtu.be/EwzD8U4u76k?t=61	21	My DD on this: https://youtu.be/EwzD8U4u76k?t=61
8	4	Location	Chumbawamba	36	He was listening to the chick from Chumbawamba instead ...

Figure 5. Extracted entities from the Reddit post comments.