import praw import pandas as pd import numpy as np import re from collections import Counter import matplotlib.pyplot as plt import datetime as dt from pprint import pprint import seaborn as sns #Accessing reddit api reddit = praw.Reddit(client id = "Stvke0o5GYUn2A", client secret = "FsJBa-6RTMbiwat07Q9WgZ8BpWFV6A", user agent= "reddit_scrape", username = "EarthMetals", password = "FeR87fgt**") # hot posts, rising posts headlines = set() for submission in reddit.subreddit("wallstreetbets").hot(limit=1): headlines.add(submission.title) print(len(headlines)) 1 from praw.models import MoreComments times = []comments = [] for top level comment in submission.comments: if isinstance(top_level_comment, MoreComments): continue comments.append(top_level_comment.body) times.append(top_level_comment.created_utc) print(len(comments)) print(len(times)) 436 436 df = pd.DataFrame(comments, times) df.head() df.to csv(r'C:\Syracuse DataScience\IST 736 - Text Mining\Homework\Homework One\wsbdt comments', index=False) import nltk nltk.download('vader_lexicon') [nltk_data] Downloading package vader_lexicon to [nltk_data] C:\Users\deane\AppData\Roaming\nltk data... [nltk data] Package vader lexicon is already up-to-date! Out[87]: True from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA sia = SIA()results = [] for line in comments: pol score = sia.polarity scores(line) pol score['headline'] = line results.append(pol score) pprint(results[:3], width=100) [{'compound': -0.6771, 'headline': '#Ban Bet Lost\n' '\n' $^{\prime}$ /u/Rodrigo87 (0/1) made a bet that BABA would go to 175.0 when it was 165.02 and it $^{\prime}$ 'did not, so they were banned for a week.', 'neg': 0.182, 'neu': 0.818, 'pos': 0.0}, {'compound': -0.7996, 'headline': 'You gotta feel for CRSR gang in 2021.\n' "During one of the biggest bull run ever, they've been absolutely fucked non stop.", 'neg': 0.264, 'neu': 0.736, 'pos': 0.0}, {'compound': 0.2263, 'headline': 'You guys ever do something like say...buy SPY calls for \$0.40, sell them later for ' '\$0.85, think you're a genius selling the top, and then watch them shoot up to about ' '\$3.15?', 'neg': 0.069, 'neu': 0.807, 'pos': 0.124}] df = pd.DataFrame.from records(results, times) df.reset index(inplace=True) df = df.rename(columns = {'index':'timestamp'}) df = df.rename(columns = {'headline':'comments'}) df.head() timestamp pos compound comments neg neu **0** 1.634322e+09 0.182 0.818 0.000 -0.6771 #Ban Bet Lost\n\n/u/Rodrigo87 (0/1) made a bet... **1** 1.634294e+09 0.264 0.736 0.000 -0.7996 You gotta feel for CRSR gang in 2021.\n\nDurin... **2** 1.634298e+09 0.069 0.807 0.124 0.2263 You guys ever do something like say...buy SPY ca... **3** 1.634300e+09 0.333 0.535 0.132 My spy puts will most likely expire worthless ... -0.6855 **4** 1.634305e+09 0.189 0.811 0.000 -0.1027 Is it gay to pay federal taxes? #creating thresholds df['label'] = 0df.loc[df['compound'] > 0.2, 'label'] = 1 df.loc[df['compound'] < -0.2, 'label'] = -1</pre> df.head() timestamp neg neu pos compound comments label **0** 1.634322e+09 0.182 0.818 0.000 -0.6771 #Ban Bet Lost\n\n/u/Rodrigo87 (0/1) made a bet... **1** 1.634294e+09 0.264 0.736 0.000 -0.7996 You gotta feel for CRSR gang in 2021.\n\nDurin... -1 **2** 1.634298e+09 0.069 0.807 0.124 0.2263 You guys ever do something like say...buy SPY ca... 1 **3** 1.634300e+09 0.333 0.535 0.132 -0.6855 My spy puts will most likely expire worthless ... -1 **4** 1.634305e+09 0.189 0.811 0.000 -0.1027 Is it gay to pay federal taxes? df1 = df[['comments', 'label']] df1.label.value counts() #normalized to percent df1.label.value_counts(normalize=True) * 100 43.577982 0 -1 28.440367 1 27.981651 Name: label, dtype: float64 print("Positive Comments: \n") pprint(list(df[df['label'] ==1].comments)[:5], width=200) print("Negative Comments: \n") pprint(list(df[df['label'] == -1].comments)[:5], width=200) Positive Comments: ['You guys ever do something like say...buy SPY calls for \$0.40, sell them later for \$0.85, think you're a genius selling the top, and then watch them shoot up to about \$3.15?', 'Every time I worry about my portfolio I scroll through the daily thread and get a good laugh at all the flami ng dogshit people are bag holding.', 'Burry Twitter deleted. Bull market back in action confirmed. I expected him to need another week of confirm ation to be honest.', 'Watching UPST is like watching a yacht go by and all my friends are on it, waving and saying we told you to g et on.', "I could have been a very rich man swimming in pussy juice if I did the opposite of everything I've done since last year."] Negative Comments: ['#Ban Bet Lost\n\n/u/Rodrigo87 (0/1) made a bet that BABA would go to 175.0 when it was 165.02 and it did not, so they were banned for a week.', "You gotta feel for CRSR gang in 2021.\n\nDuring one of the biggest bull run ever, they've been absolutely fuc ked non stop.", "My spy puts will most likely expire worthless today but at least I'm not bagholding SDC like a true retard.", 'My buddy was loaded to the tits on SPCE with a low cost basis. He held even during the run up to the 60s-70s saying \$100 was his exit target.\n' '\n' 'Today he will wake up and his position will be red for the for first time (not 100% of his cost basis but it should be in high teens). Hundreds of thousands in gains gone. \n' '\n' 'Should I:\n' '\n' '1. Call him to invite him to lunch, check up on him.\n' '2. Text him a pic of Bezos smiling with his stupid ass cowboy hat on.', 'TSLA really is unstoppable isn't it lol \n\nIt's completely broken the shell of a man once known as Burry and just. keeps. going.'] fig, ax = plt.subplots(figsize=(10,10))counts = df1.label.value counts(normalize=True) * 100 sns.barplot(x=counts.index, y=counts, ax=ax).set title('WSB Daily Thread Sentiment-VADER') ax.set xticklabels(['Negative', 'Neutral', 'Positive']) ax.set ylabel('Percentage') plt.show() WSB Daily Thread Sentiment-VADER 40 30 Percentage 20 10 Negative Neutral Positive In [94]: fig, ax = plt.subplots(figsize=(10,10)) sns.lineplot(data=df, x='timestamp', y='compound').set_title('Sentiment Scores Overtime') Out[94]: Text(0.5, 1.0, 'Sentiment Scores Overtime') Sentiment Scores Overtime 0.75 0.50 0.25 0.00 punodwoo -0.25-0.50-0.75-1.00295000 300000 305000 310000 315000 320000 325000 +1.6340000000e9 timestamp #TextBlob from textblob import TextBlob def sentiment(text): try: return TextBLob (text) . sentiment except: return None df1['sentiment'] = df1['comments'].apply(lambda post: TextBlob(post).sentiment) <ipython-input-97-1a138c8115ef>:1: SettingWithCopyWarning: A value is trying to be set on a copy of a slice from a DataFrame. Try using .loc[row indexer,col indexer] = value instead See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret urning-a-view-versus-a-copy df1['sentiment'] = df1['comments'].apply(lambda post: TextBlob(post).sentiment) sentiment_series = df1['sentiment'].tolist() columns = ['polarity', 'subjectivity'] tb = pd.DataFrame(sentiment series, columns=columns, index=df1.index) tb.head() polarity subjectivity 0.000000 0.000000 0.700000 -0.600000 0.250000 0.250000 -0.191667 0.741667 0.416667 0.583333 fig, ax = plt.subplots(figsize=(10,10))sns.scatterplot(data=tb, x='subjectivity', y='polarity').set_title('Polarity by Subjectivity') Out[100... Text(0.5, 1.0, 'Polarity by Subjectivity') Polarity by Subjectivity 1.00 0.75 0.50 0.25 polarity 0.00 -0.25-0.50-0.75-1.000.2 0.8 0.0 1.0 subjectivity comments = df['comments'] fin = tb.join(comments) fin.head() polarity subjectivity comments 0 0.000000 0.000000 #Ban Bet Lost\n\n/u/Rodrigo87 (0/1) made a bet... 1 -0.600000 0.700000 You gotta feel for CRSR gang in 2021.\n\nDurin... 0.250000 0.250000 You guys ever do something like say...buy SPY ca... -0.191667 0.741667 My spy puts will most likely expire worthless ... 0.416667 0.583333 Is it gay to pay federal taxes? #creating thresholds fin['label'] = 0fin.loc[fin['polarity'] > 0.2, 'label'] = 1 fin.loc[fin['polarity'] < -0.2, 'label'] = -1</pre> fin.head() polarity subjectivity comments label 0.000000 0.000000 #Ban Bet Lost\n\n/u/Rodrigo87 (0/1) made a bet... **1** -0.600000 0.700000 You gotta feel for CRSR gang in 2021.\n\nDurin... 0.250000 0.250000 You guys ever do something like say...buy SPY ca... **3** -0.191667 0.741667 My spy puts will most likely expire worthless ... **4** 0.416667 0.583333 Is it gay to pay federal taxes? In [104... df2 = fin[['comments', 'label']] df2.label.value_counts() #normalized to percent df2.label.value_counts(normalize=True) * 100 0 66.055046 Out[104... -1 17.660550 1 16.284404 Name: label, dtype: float64 fig, ax = plt.subplots(figsize=(10,10)) counts = df2.label.value_counts(normalize=True) * 100 sns.barplot(x=counts.index, y=counts, ax=ax).set_title('WSB Daily Thread Sentiment-TextBlob') ax.set_xticklabels(['Negative', 'Neutral', 'Positive']) ax.set_ylabel('Percentage') plt.show() WSB Daily Thread Sentiment-TextBlob 60 50 40 Percentage 30 20 10 Negative Neutral Positive print("Positive Comments:\n") pprint(list(df2[df2['label'] ==1].comments)[:5], width=200) print("Negative Comments:\n") pprint(list(df2[df2['label'] == -1].comments)[:5], width=200) Positive Comments: ['You guys ever do something like say...buy SPY calls for \$0.40, sell them later for \$0.85, think you're a genius selling the top, and then watch them shoot up to about \$3.15?', 'Is it gay to pay federal taxes?', 'Every time I worry about my portfolio I scroll through the daily thread and get a good laugh at all the flami ng dogshit people are bag holding.', 'Jim Cramer the type of dude to order unsalted fries and add salt to them', 'Bought \$30 worth of spy calls for fun, sold for \$40. Paid a \$10 commission to vanguard. No wonder banks are m aking so much'] Negative Comments: ["You gotta feel for CRSR gang in 2021.\n\nDuring one of the biggest bull run ever, they've been absolutely fuc ked non stop.", '#Ban Bet Won\n---\n\n/u/AlmightySnoo made a bet that TSLA would go to 825.0 when it was 811.1 and it did, con grats retard.', 'The bears existence relies entirely on the bulls decision to take profits\n\nBottom feeders and retards', $^{\prime}$ I'm so fucked. I accidentally hit confirm on 1dte SPY PUTS with my last \$1000. There I go $^{\prime}$, "I remember the good old days you buy some OTM options they expire worthless a week later and you shutup and m ove on to the next losing play. Now we have retards buying shares and they won't stop " 'taking about their bags for months on end. \n' '\n' "Why can't can't you retards just buy worthless options like the rest of us?"]