Project Report

Analysis of current inflation sentiments in the US political society (media)

Foundations of Computational Social Systems

Winter Term 2022 | 23

Group Members:

Jignesh Jatakia   jignesh.jatakia@student.tugraz.at   (BA major)
Jigna Jatakia   jigna.jatakia@student.tugraz.at   (IT major)
Artyom Kukushkin   artem.kukushkin.02@gmail.com   (EM major)
Ihor Savchenko   igor.savchenko@tavr.at   (BA major)

Table of contents

1. MOTIVATION AND RESEARCH QUESTIONS

1.1 MOTIVATION

Inflation was a hot topic in US political society, especially in 2022, which was designated as a year of exceptional political and economic volatility marked by historical inflation peaks. Our group is extremely motivated to analyse the inflation sentiments in the US political society (media) in this crucial and important topic. We are a group of four people from various educational backgrounds who wanted to investigate the effects of drastic changes in a country's economic statistics, such as inflation, on human behaviour. We wanted to show the overall importance of inflation while also learning about its role in current political and societal processes in the United States.

Looking the inflation historical data in the United States, the inflation data are low and controlled for so long entire generations of Americans. Year-over-year inflation averaged about 2.3% per month between 1991 and 2019, and exceeded 5.0% only four times, which quite controllable.

During the years 2020 to 2022, millions of people were out of work as a result of the pandemic, the inflation raised rapidly and touched 9.1% in the month of June 2022. Central bankers, politicians and other agencies were worked to lift the US economy out of a pandemic-induced recession. Americans prioritize inflation as the nation's top problem, and US President has stated that addressing the problem is his top priority (DeSliver, 2022).

When a topic is very important and pertains to the country's economy and politics, such trending topics are generally highlighted in the media (news) and social media platforms such as Twitter. We decided to conduct additional research on Twitter. We concentrated on top political journalists and influencers' Twitter accounts in the United States.

1.2 RESEARCH QUESTIONS

As a result, our group is interested in delving deeper into this topic, and we have several questions that can be concretized in the following way.

We planned our research and concentrated on two major categories: inflation and non-inflation (general). We were motivated to test the following hypotheses after first measuring the general characteristics, then comparing them to the characteristics in tweets describing inflation.

H1. The tweets dedicated to inflation have significantly more retweets and likes than average tweets of the same authors within the same period.

H2. The tweets dedicated to inflation are marked by significantly stronger negative sentiment than average tweets of the same authors within the same period.

H3. The medium frequency of inflation tweets correlates with the US inflation statistics in the same period.

2. DATA RETRIEVAL AND GENERAL OVERVIEW

Before we are going into a deep explanation of our overall research, we would like to give a bird view of data retrieval and the methods we have used in our project. We began by collecting Twitter users' (US political pundits/journalists) data (https://twitter.com/i/lists/15084461/members) and saving it in the pundits.csv file. We used twarc2 to collect timelines and saved them as TIMELINE.xlsx.

In the process of data collection and after discussing the matter with our course professor, we slightly changed our approach in comparison with the one presented in our plan document. First, instead of calculating retweets and likes only, we decided to use a compound impact factor for each tweet, calculating it this way:

(retweets + quotes) / number of followers

"Likes" factor was excluded as too ambiguous for a sensitive theme like inflation (people may refrain from likes if the topic scares or disgusts them), and the "number of followers" factor was added as directly influencing the absolute numbers of retweets and quotes.

Moreover, we changed the criteria of sorting out the tweets, setting the following limits: >10000 followers, > 1000 tweets, last tweet no earlier than Feb 1 2023. We analyzed exclusively the tweets of 2022 as the interesting year of significantly high inflation and political turbulence.

After gathering the data, we assessed sentiments using the VADER (GeeksforGeeks, 2021) method. We quantified both sentiment (qualitative data) and impact (quantitative data). We wanted to test our hypotheses and provide answers to our research questions, so we (additionally to the methods we studied in the course) used the Chi-square test (JMP Statistical Discovery LLC, 2023) to test sentiment qualitative data for independence between the inflation sample and the non-inflation sample. In addition, for impact analysis (qualitative data), we computed means for the inflation sample and the general population and compared them using a Z-test (Zach, 2023) (Frost,2023). Finally, we discovered a month-by-month correlation between the inflation rate and general sentiment in year 2022.

Let's discuss data processing, analysis and our findings in detail before we conclude and critique of our research.

3. DATA PROCESSING

The raw data were filtered using standard Python procedures. This is, for example, how the protected and less popular pundits accounts were removed:

```
pundits = pundits[(pundits['protected']==False)&
(pundits['public_metrics.followers_count']>10000)&
(pundits['public_metrics.tweet_count']>1000)]
```

To find last tweets of each pundit, we had to run an additional twarc timeline procedure, leaving only the latest data element for each pundit:

```
for pundit in pundits.id.values:
    data = twarc_client.timeline(pundit, max_results=100)
    pundit_latest.append(next(data)['data'][0])
```

Then an additional table containing author ids and latest tweet dates was merged with the main pundit table and all "outdated" pundits were removed:

```
pundits = pundits[(pundits['created_at_y']>'2023-02-01')]
```

After that, the main twarc timeline procedure was run. We repeated the procedure 3 times, collecting up to 300 tweets from each of remaining 253 pundits, to make sure every pundit covered the whole research interval of the year 2022.

We reshaped the timeline table into a single column and created simple functions to extract only the data we need (tweet texts, dates, authors and the numbers of followers for each author):

```
def extract_text(x):
    return eval(x)['text']
def extract_date(x):
    return eval(x)['created_at']
def extract_author(x):
    return eval(x)['author_id']
def find_followers(x):
```

```
return pundits[pundits.id==x]['public_metrics.followers_count'].iloc[0]
```

We got the data frame "pundits timelines new", or ptnew, and sorted out all tweets not posted in 2022. Thus our data – in total, 30182 tweets - were ready for analysis.

4. Data analysis

We divided our tweets into inflation-themed (211) and non-inflation-themed (29971) dataframes. We used the words "inflation", "prices" and "gas" as the most relevant to the current inflation wave to make a division. An example in Python:

pt_new["contains_inflation"] = pt_new["tweet"].str.contains(r'\b(inflation|prices|gas)\b')

For each dataframe, we ran the VADER analysis, using -0,3 and 0,3 as the negativeness and positiveness limits, respectively. We narrowed the "neutrality" to get better sensitivity: political pundits as public people cannot allow to be overemotional in their tweets (i.e. using expletives or strong slang).

In the end, all our tweets were divided by two independent criteria: sentiments (negative, positive, neutral) and theme (non-inflation, inflation). These are qualitative characteristics, so we performed a chi-square test to find out: if the change of theme influences the sentiment distribution in our tweets? We formulated 2 hypotheses: H0 – there is no influence and H1 – there is influence

This was our contingency table, already with relative values:

	Non-inflation	Inflation
Positive	0.345155	0.297170
Negative	0.183338	0.245283
Neutral	0.471507	0.457547

The chi-square test for these data and (3-1)*(2-1) = 2 degrees of freedom has produced the test-value of 0,013 with a critical value of 5,99. This means H0 holds with a very high probability (about 99%). We cannot say that there is any influence of a tweet theme on its sentiment.

For the quantitative analysis, we used our impact formula (see Chapter 2 above). Here is the example of filling the dataframe with the calculated impact values:

pt_inflation["impact"] = 0
for i in range(len(pt_inflation['raw'])):
 pt_inflation["impact"][i]=
(eval(pt_inflation["raw"][i])["public_metrics.retweet_count"]+
eval(pt_inflation["raw"][i])["public_metrics.quote_count"])/pt_inflation["followers"][i]

This time, we compared the mean impact values of all 2022 tweets (taken as a "general population") and specifically of inflation-themed tweets ("a sample"). We had a H0 (inflation mean does not differ significantly from the population mean). And a H1 (there is significant difference). For this purpose, we used a classical one-sample Z-test. Our mean impact value for inflation-themed tweets (0,0227) turned out to be roughly two times higher than the general population mean (0,0118). Still, large standard deviations do not allow to prove a statistically significant difference. Our Z-test has shown the p-value of 0,513, which means that the impact change can be explained by random factors with about the same probability as by the tweet content.

As the final part of our analysis, we got some interesting results by building a scatterplot with two variables: mean tweet sentiment for every month of 2022 and an official inflation level in the US for the corresponding month. The scatterplot has shown a moderate negative correlation between two valuables with a coefficient of -0,53. A clearly visible outlier in April was seen in the chart.

5. Conclusion

In conclusion, our group conducted research on the inflation sentiments in US political society by analyzing tweets from top political journalists and influencers' Twitter accounts in the United States. We had three research questions and formulated hypotheses based on them. We retrieved data from Twitter users and used a compound impact factor for each tweet to calculate the quantitative data. We also used the VADER method to assess the qualitative data. After analyzing the data, we found that tweets dedicated to inflation had visibly more social impact than average tweets of the same authors within the same period. Additionally, we found that the tweets dedicated to inflation were marked by slightly stronger negative sentiment than average tweets of the same authors within the same period. Lastly, we discovered a moderate month-by-month correlation between the inflation rate and general sentiment in the year 2022. Overall, our findings suggest that the issue of inflation had some impact on the sentiment expressed on Twitter during this period, indicating the importance of this topic in US political society.

Although visual data show definite changes between the sample and the rest/general population, these changes are not statistically significant (we cannot reject the null hypothesis).

Correlational analysis shows moderate negative correlation between the inflation level and general sentiments in the Twitter expert group. This is logical, but the question still needs more data for precise analysis.

When analysing the scatterplot, we paid a particular attention to April 2022. In that month, there was a strongly positive mean sentiment value, which made April a clear outlier in our dataset. After searching the Internet, they would have found information that in April, Elon Musk bought the Twitter company. That effect could cause such a mood swing in the market.

6. Critique

We tried to provide clear explanation of their hypotheses and methods used for data collection, sentiment analysis, and hypothesis testing. The use of VADER sentiment analysis and statistical tests such as the Chi-square test and Z-test to analyze the data was appropriate and demonstrated an analytical approach.

However, there are some areas where the text could be improved. In the analysis, we took a period of one year. We got the result, but it's not enough. In order to draw a more accurate conclusion, we should take a longer period and analyze the results over the past few years.

One should take into account the imperfection of data gathering and analysis methods (e. g. a tweet praising an external article about inflation is deemed „positive", while the article subject itself is clearly negative). A method to study inflation based on three most popular words is also not ideal. There might be error cases of two types: falsely assigning a non-inflation tweet into an inflation group (a tweet concerning gas producing or gas leaks has nothing to do with gas prices) as well as vice versa (a tweet describing hard economic situation, but not directly mentioning inflation or prices). A deeper textual analysis is needed for better results.

The study only focuses on Twitter data and may not reflect the sentiment of the broader U.S. public. And since Twitter is not aimed at all groups of society, the data is not complete. To improve, we can take data from a broader Twitter group, or from other social networks like Facebook or Reddit.

It is worth noting that we performed a cross-sectional and not a longitudinal study (Fall, 2018). Our research was carried out in the middle of February, 2023, concerning the events of 2022. We used a specific group of Twitter users and studied their characteristics at the time of our research. But the composition of the expert group tends to change over time: some pundits lose popularity, retire or even die, while some new pundits appear and gain popularity and social impact. If we decide to repeat the study with the same methods, the further we are from 2022, the more inadequate the results would be (i. e. we cannot make sound assumptions about a distant past based on the current group of pundits, who might not have been active in that past). In the next step, it would be interesting to carry out a longitudinal cohort research, concentrating on sentiment changes in tweets of specific persons over a long period of time.

References

DeSilver, D. (2022, June). Retrieved from https://www.pewresearch.org/fact-tank/2022/06/15/in-the-u-s-and-around-the-world-inflation-is-high-and-getting-higher/

Fall. (2018). Retrieved from https://academic.oup.com/poq/article/82/3/470/5098458?login=true

Frost, J. (2023). Retrieved from https://statisticsbyjim.com/hypothesis-testing/z-test/

GeeksforGeeks. (2021). Retrieved from https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/

JMP Statistical Discovery LLC. (2023). Retrieved from https://www.jmp.com/en_au/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html

Zach. (2023). Retrieved from https://www.statology.org/z-test-python/

Appendix

The whole code with supplementary materials can be found on Github:

https://github.com/IgorFeanfox/Foundations-of-CSS

Bar.bmp – bar diagram showing the distribution of sentiments in inflation and non-inflation tweets

pdf_cdf.bmp – probability density and cumulative distribution functions of a chi-square distribution, with an actual value and a critical value marked

scatter.bmp – scatterplot showing the correlation between mean monthly tweet sentiments and inflation