ANALYSIS OF REDDIT POSTS ABOUT CRYPTO MARKET AROUND FTX SCANDAL
(1500 words)

Akhila Sakiramolla
Fall 2022

1. Introduction

FTX is one of the largest cryptocurrency exchanges in the world, valued at $32B at the start of 2022. It's founder, Sam Bankman-Fried (SBF), is one of the most influential persons in the crypto industry. The company has various major financial companies as its investors and was largely successful but recent turn of events led to the firm filing for bankruptcy on November 11 and SBF facing investigations by the Justice Department and Federal Regulators. Consequently, other crypto companies are being destabilized and widespread distrust is sweeping the industry. This has again raised huge concerns about the industry which is largely unregulated. All of this happened in a span of two weeks, starting from November 2 when CoinDesk published a report pointing out inconsistencies in about SBF's company balance sheets (Mark, 2022). With this study, I aim to understand how this event alone affected the sentiments around crypto industry and identify the major topics of discussion in the industry to understand the repercussions.

Reddit is a social news website and forum where content is socially curated and promoted by site members through voting. Reddit contains hundreds of subcommunities, called subreddits. Each subreddit focuses on a specific topic, such as technology, politics, or music. In recent years, Reddit has become a go-to source of investing advice - including discussions about cryptocurrency. The most popular subreddit for investors looking to invest in cryptocurrency is r/Cryptocurrency. There are over 5.8 million members in this Reddit group, and it has been active since 2013 (Brook, 2022)

2. Research Question

The research has been conducted to answer the following questions by analyzing the reddit posts and comments regarding the crypto market amidst the FTX scandal:

1) How are users on most popular crypto subreddits reacting to the recent fall of FTX and its corresponding effect on the market?
2) What are some of the most prevalent themes/topics in these discussions?

3. Data

3.1 Data collection

For this study, I collected data from Reddit using PRAW, which is python module that can be used to access Reddit's API. I used the authorized instance to get more flexibility for accessing the posts and comments. I gathered all the data from the subreddit **'Cryptocurrency'**. I used search queries like **'FTX'** and **'SBF'** to filter for the posts and collected 20 posts for each subquery. I then collected around 6500 comments under these posts. I collected information like post user id, post

created time, post body, comment id, comment body, etc., I wrote a python function that searches using the query, stores all the fields in dictionaries and converts them into a data frame.

Search queries – FTX, SBF

Time period – November 1,2022 to Dec 8, 2022

3.2 Data preprocessing

For data preprocessing, below are the steps I have followed to get the comment content:

1. Converted text to lowercase
2. Removed emojis and special characters
3. Removed links/URLs
4. Removed punctuation marks
5. Removed numbers

	id	created	body
0	yxq8s2	2022-11-17 13:27:43	here is the twitter threaddirect link to the d...
1	yuorb4	2022-11-14 04:25:57	ftx is the biggest collapse of a corporation s...
3	yxkat4	2022-11-17 07:58:17	a tornado cash developer is still in jail with...
4	yrvdes	2022-11-10 23:40:00	as most of you are already aware the current ...
5	z0w281	2022-11-21 10:55:43	ftx official twitter released an update yester...
...	...	...	...
20269	z92lir	2022-11-30 20:42:13	dudes gonna go down as the greatest scammer to...
20270	z92lir	2022-11-30 20:42:13	all those politicians who got donations are st...
20271	z92lir	2022-11-30 20:42:13	yet the mainstream media is surprisingly suppr...
20272	z92lir	2022-11-30 20:42:13	will crypto die
20273	z92lir	2022-11-30 20:42:13	democrats knew too all that us taxpayer printe...

6717 rows × 3 columns

Fig 1 – Sample comments data

4. Methodology

4.1 Lemmatization

To get all the unique tokens in the comments, I performed tokenization, which breaks down the given text into the smallest component of a sentence called a token. Afterward, I performed lemmatization, which reduces words to their base words, followed by the removal of stop words. These words do not have any significance in search queries. For this purpose, I used a python library called **'Spacy'**, which is an open source library for Natural Language Processing.

4.2 Removal of less frequent words

After looking at the data, I realized that there are many words that are very frequent but did not contain any useful information. I wanted to remove these words in order to get meaningful topics from the data. For this purpose, I looked at the bigrams and trigrams in the data and implemented a TF-IDF model to return the words that are highly frequent but not very useful. I removed these words from the corpus before the analysis. I used a python library called **'Gensim'** to analyze the bigrams and trigrams and also to implement TF-IDF model.

4.3 Sentiment Analysis

Sentiment analysis can help us gather insights regarding the context and analyze the mood and emotions of the general public. In order to better understand various events and their impact, these sentiments can be used. **TextBlob** is a Natural Language Processing library in Python. It uses **Natural Language ToolKit (NLTK)** to perform various tasks. It provides easy access to a wide variety of lexical resources and allows users to work with categorization and classification.

Words are measured by their intensities and semantic orientations to define sentiments. A text message is usually represented as a bag of words. An average of all the sentiments is calculated after assigning individual scores to all the words. Using TextBlob, we can determine the polarity and subjectivity of a sentence. The polarity of a sentiment lies between [-1,1], -1 defining a negative sentiment and 1 defining a positive sentiment. A text's subjectivity measures how much information is based on fact and opinion. Subjectivity indicates a higher level of opinion rather than factual information in a text (Shah, 2020)

I installed **TextBlob** module from **textblob** library available in python. Each comment is analyzed by TextBlob(text).sentiment.subjectivity and TextBlob(text).sentiment.polarity methods to get subjectivity and polarity respectively. Based on polarity scores, the comments can be classified as follows:

1. Positive sentiment: polarity > 0
2. Neutral sentiment: polarity = 0 and
3. Negative sentiment: polarity < 0

We observe that on the whole majority of the users have a positive sentiment towards posts that discuss about the FTX scandal and SBF (Fig 2). I also looked at the distribution of comments with respect to polarity and subjectivity, and observed that majority of the comments have polarity between -0.5 and 0.5 and have subjectivity between 0.4 and 0.8 (fig 3)

I further wanted to understand the trends of these comments over time. I created buckets for each week from 1st November 2022 to 9th December 2022 and plotted number of posts with different sentiments over the weeks and their average polarity (Fig 4 and 5). We observe that there are maximum number of comments in the week of 11/08 – 11/13, indicating maximum engagement,

as that was the week where the scandal started unravelling and FTX filed for bankruptcy. The comments have reduced over time with lowest number in the week of 12/05 – 12/07, as the discussions regarding this topic started dying down. But the sentiment seems to be fairly consistent throughout the entire time.

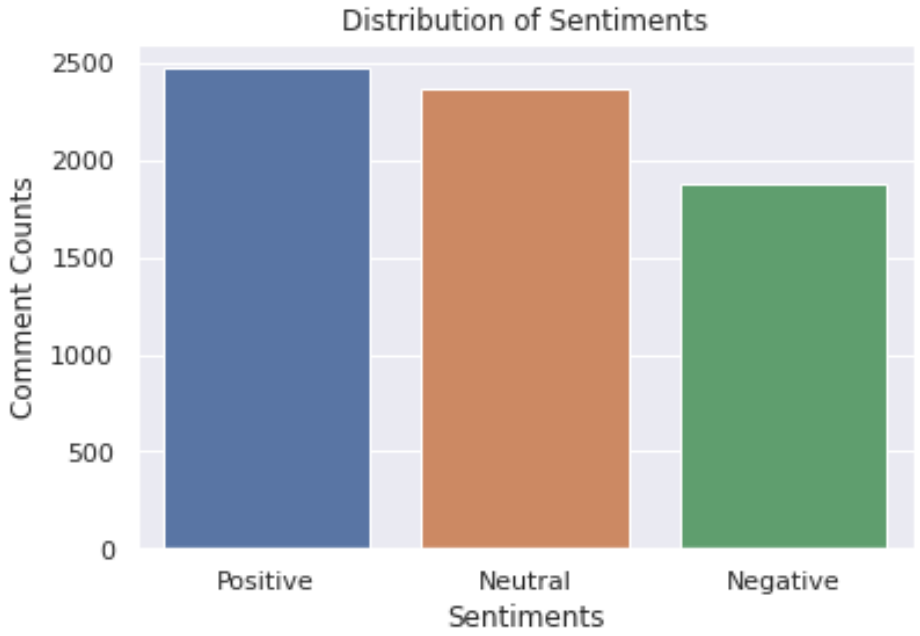

Fig 2 – Overall distribution of sentiments

Fig 3 – Distribution of comments with respect to subjectivity and polarity

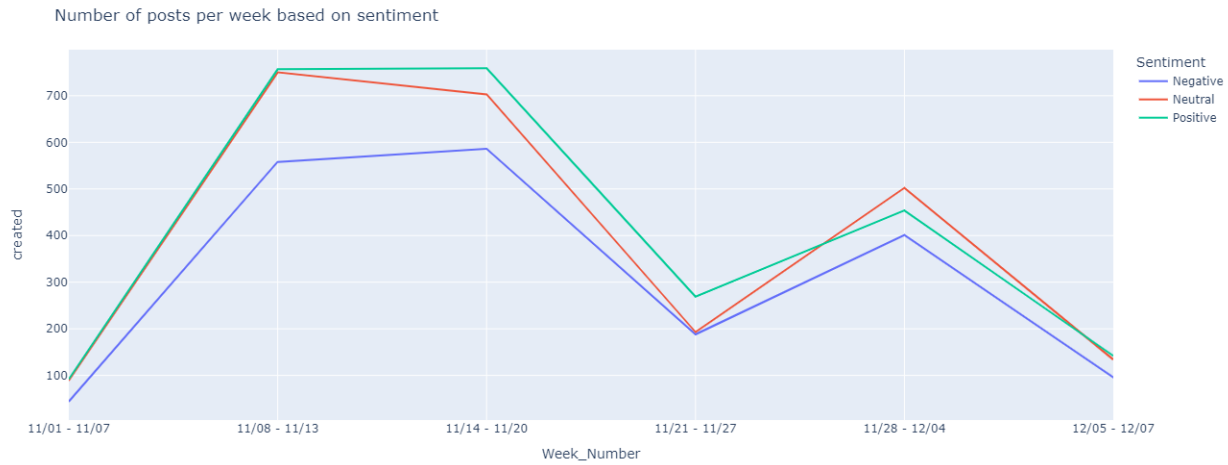

Fig 4 – Number of posts per week

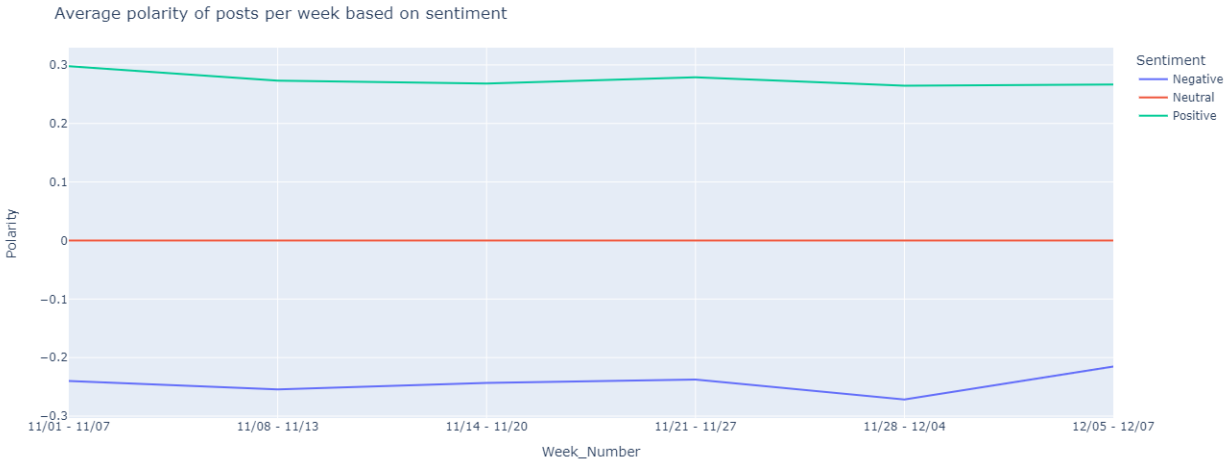

Fig 5 – Average polarity per week

4.4 Topic Modelling

Topic models can be described as a type of statistical language models that can help uncover hidden structures in a text corpus. Using these models, we can tag abstract topics in documents that best represent the information contained in them. They are very useful to perform various functions like clustering documents, organizing large data blocks, retrieving information from unstructured text, and selecting features (Kapadia, 2019).

One of the most popular topic modeling methods is latent Dirichlet allocation. Documents contain various words, and topics contain certain words. On the basis of words contained in the document, the LDA aims to find topics to which the document belongs. A similar set of words will be used in documents with similar topics. This enables the documents to get the probability distribution over the latent topics. I used the python library **'Gensim'** to implement the LDA model. LDA model requires a stream of documents, a dictionary containing word IDs and topics, and a

dictionary containing word mappings. Additionally, we can specify the number of iterations and passes. In order to determine whether the topics are interpretable, we will have to visualize the topics after they have been trained by the LDA model.

An interactive inter topic distance plot can be created using **PyLDAvis**, a popular visualization library, in which every topic is represented by a circle and its size represents its prominence in the corpus. We can use this plot to identify the most frequently used words in each topic and also understand how one topic is related to another. Moreover, the plot offers the option to adjust the relevance metric value to highlight the most frequent words in every topic for better interpretation. I implemented this separately on comments with positive and negative sentiments in order to understand what kinds of themes exist in different sentiments (Fig 6 and 7).

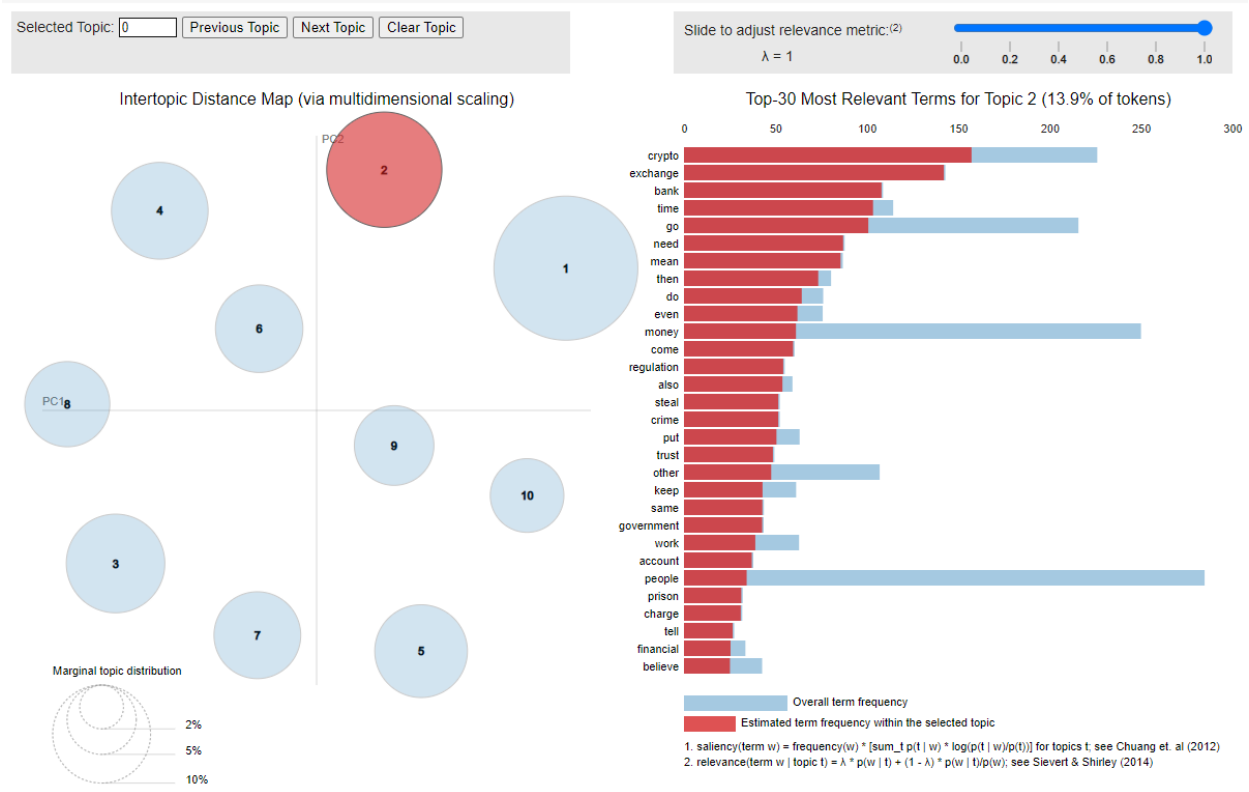

Fig 6 – Topic modelling on negative sentiment comments

5. Results

Upon analyzing the topics, I made the following observations:

1. Negative sentiment – In topic 2, we can see words like regulation, government, prison, etc., indicating that people having these discussions maybe pushing for government intervention in the crypto industry and introduce regulation.
2. Positive sentiment – In topic 2, we can see words like invest, trust, start, etc., indicating that people maybe encouraging to invest in crypto despite the scandal as this is the only company that's impacted.

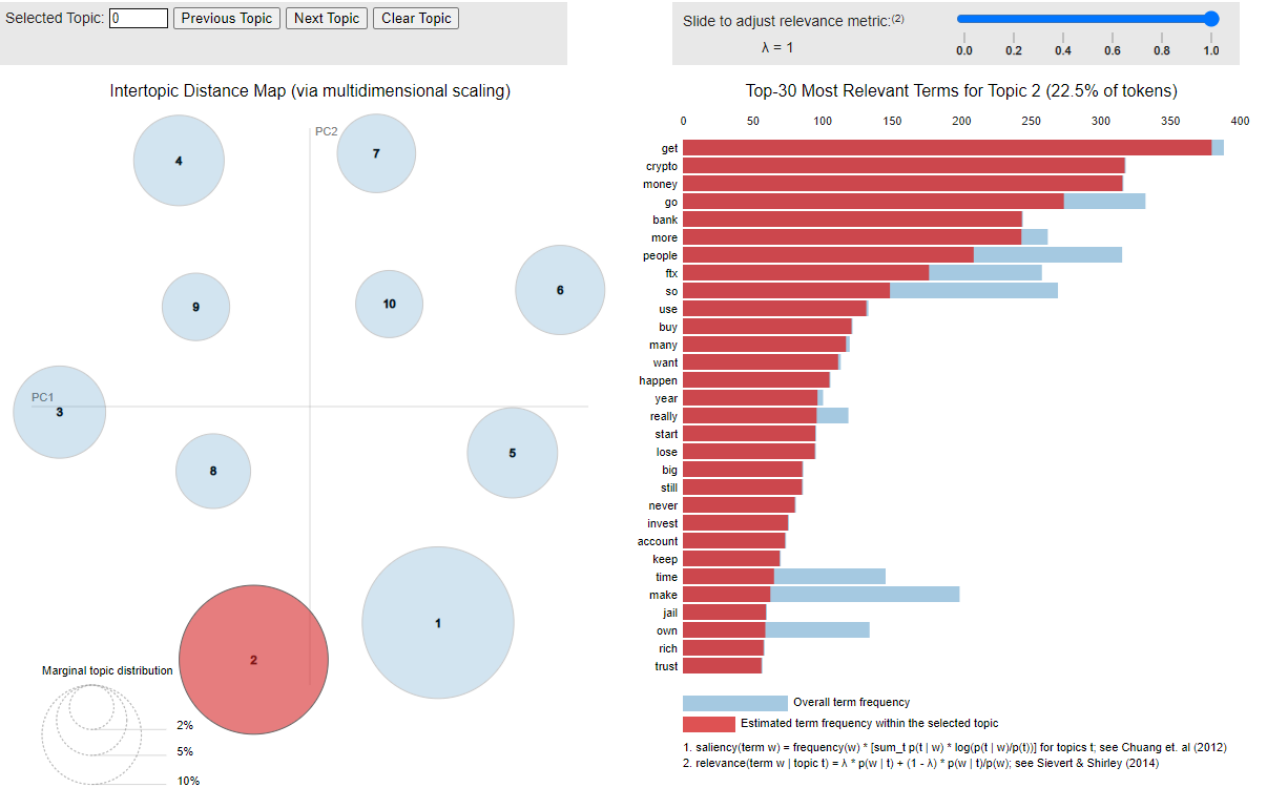

Fig 7 – Topic modelling on positive sentiment comments

6. Conclusions and Limitations

Through Sentiment Analysis and Topic Modeling, I was able to find answers to the study's main themes. We gained an insight into people's reactions to this scandal by observing the variations in comments as the scandal progressed. LDA on the positive and negative comments helped us identify the most prominent and interesting topics in the discussions on reddit.

Although, I was able to address the study's goal using the analysis, there are some limitations. I observed that the topics identified through topic modelling using LDA are able to capture the main themes but have some overlapping topics, as I saw some words repeated in different topics. I will have to do a much detailed analysis to understand the context of these words within specific topics. Topic modelling may identify topics that require domain knowledge in order to affirm the accuracy and interpretation of the topics.

7. References

1. Julian Mark, 2022, Why the FTX collapse has plunged the crypto world into upheaval https://www.washingtonpost.com/business/2022/11/10/ftx-faq-crypto-turmoil/

2. David Yaffe Bellany, 2022, *How Sam Bankman-Fried's Crypto Empire Collapsed* How Sam Bankman-Fried's FTX Crypto Empire Collapsed - The New York Times (nytimes.com)

3. Connor Brook, 2022, Best Crypto to Buy Right Now on Reddit in 2022 https://www.business2community.com/cryptocurrency/best-crypto-to-buy-reddit

4. Parthvi Shah, 2020, Sentiment Analysis using TextBlob https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

5. Shashank Kapadia, 2019, Topic Modeling in Python: Latent Dirichlet Allocation (LDA) https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0