Topic Modeling and Sentiment Analysis of Web3.0 Tweets (1475 words)

Akshat Arvind

Indiana University

I. Introduction

The internet we are familiar with started in early 2000s. It is the form of internet we use today. Nowadays, people not only consume content but create it as well. This era of internet has been termed as Web2.0 which is an evolved version of the early World Wide Web where people could only consume data from the internet. Web3.0 is the future of web, where creators have full ownership of their data and content. In current scenario, big social media companies can sell user’s data or censor it as they have ownership of user’s data. Web3.0 aims to solve these problems by creating decentralized form of the web. It is based on certain principles like openness, decentralized, censorship-resistant, immutable, trustless and permissionless.

Twitter hosts a strong Web3.0 community and thousands of tweets are made on Web3.0 including topics like NFT, Blockchain, Cryptocurrency, Metaverse etc. Previous research work has been done to analyze the phases of development in Web 3.0 (*Kreps, D. and Kimppa, K. (2015), "Theorising Web 3.0: ICTs in a changing society"*) and gauging the merits of Web3.0 using social, economic and cultural lenses (*Chohan, Usman W. and Chohan, Usman W., Web 3.0: The Future Architecture of the Internet?*). This study aims to extend research in this field and perform topic modeling and sentiment analysis on tweets related to Web3.0.

II. Research Questions

The tweets collected and analyzed were used to answer the below questions: -

i) What are the most discussed topics in Web3.0 and how the topics have changed from 2021 to 2022?
ii) What is the overall sentiment related to Web3.0 and how has it changed from 2021 to 2022?

III. Data

The data for this study has been collected as tweets from Twitter using ‘snscrape’. Since the analysis demanded access to historical tweets, ‘snsscrape’ package proved to be a better choice to avoid twitter’s rate-limiting and time period limitations on their API. Total 2000 tweets were collected split equally across two timeframes – Jan 2021 to Aug 2021(1000 tweets) and Sep 2021 to Mar 2022 (1000 tweets). The collection criteria was set as the hashtag Web3.0, as it is the umbrella term covering tweets on various topics related to Web3.0. All the tweets made related to Web3.0 have the hashtag #Web3.0 as it makes the tweets available to a huge audience.

IV. Methods

i) Topic Modeling

For this study, topic modeling was conducted using the LDA (Latent Dirichlet Allocation) technique with the help of ‘gensim’ package in python and the exploration of generated clusters from LDA were done using ‘pyLDAvis’ package in python. Also, before topic modeling the collected tweets were cleaned, and stop-words removed using the NLTK library available in python. This helped in improving the data being used for topic modeling.

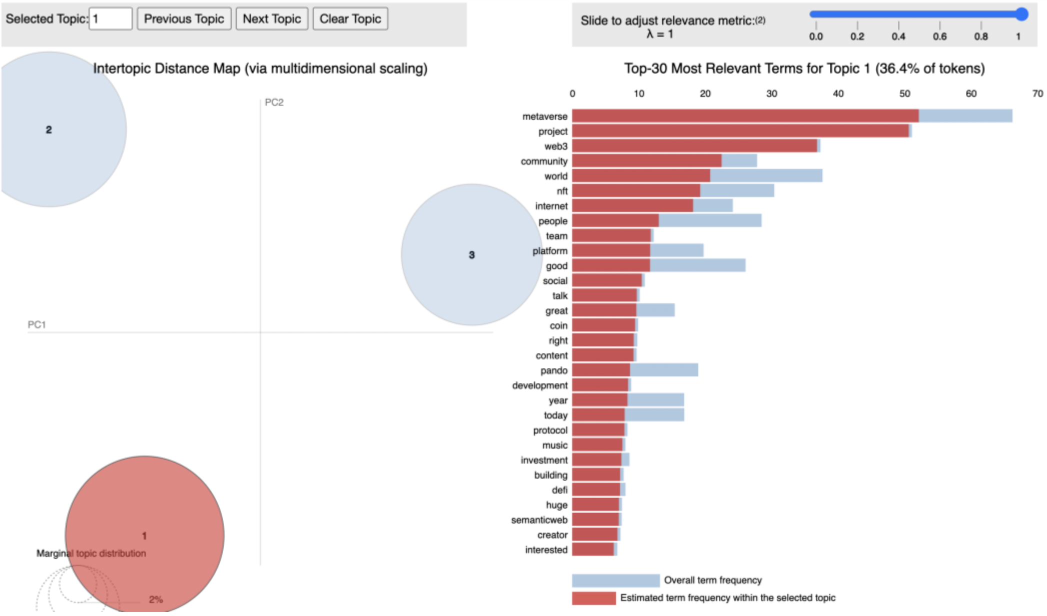
Steps involved in Topic Modeling are as follows: -

1. The collected tweets stored in a CSV were read using the ‘pandas’ library and the column containing the tweets was used for the analysis, also rows which had null values in tweet column were dropped.
2. Tweets were cleaned using a Python function which eliminates punctuations and numbers from the tweets.
3. Using the ‘NLTK’ library, stopwords were removed from the tweets to improve the quality of data in the tweets.
4. The tweets were then passed through the process of lemmatization, which involves removing inflectional endings from a word and return the base or dictionary form of the word. Lemmatization was done using the ‘spacy’ library available in python and the ‘en_core_web_sm’ package which is a small English pipeline trained on written text.
5. The tokenized_tweets were then used to create a dictionary and a document term matrix.
6. Using the LDA model, inside the ‘gensim’ library the LDA model was trained using the dictionary and document term matrix generated above. The number of topics to be generated was set to 3. One such example of the generated topics and the relevant terms can be seen in Table 4.1.

Topic	Relevant terms in the Topic
1	Metaverse, Project, Web3, Community, world, NFT, Internet
2	Future, Space, Privacy, Money, People, Good
3	Blockchain, Crypto, Technology, Network, Metaverse

Table 4.1 Generated Topics and the Relevant Terms

7. To better visualize the results from Topic Modeling, python library 'pyLDAvis' was used. One such example of the visualization can be seen in Figure 4.1.



Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 1

Figure 4.1 LDA visualization using pyLDAvis

ii) Sentiment Analysis

Sentiment Analysis was conducted using VADER tool available as a module in python library. VADER analyzes the sentiment of text and assigns a compound score which ranges from +1 to -1. The scores generated were used to classify tweets into categories like very positive, positive, neutral, negative, and very negative.

Steps involved in Sentiment Analysis are as follows: -

1. The collected tweets stored in a CSV was read using 'pandas' library and converted to dataframe. Rows with null values were dropped from the dataframe and the column

containing the tweets was cleaned through a Python function which eliminated punctuations and numbers from the tweets.

2. Each tweet in the dataset was analyzed by the SentimentIntensityAnalyzer method under vader.Sentiment Module. The generated scores were used to create 5 categories shown in Table 4.2. The generated scores were added as a new column in the dataframe to represent the sentiment associated with the tweet.

Compound Score	**Sentiment Category**
Greater than 0.5	Very positive
Greater than 0 and less than or equal to 0.5	Positive
Equal to Zero	Neutral
Less than zero greater than or equal to -0.5	Negative
Less than -0.5	Very Negative

Table 4.2 Sentiment Categories based on Compound Score

3. To present the results, the ‘countplot’ method inside the ‘seaborn’ library in python was used to generate the overall sentiment distribution for the tweets from both timeframe. One such example is shown in Figure 4.2.

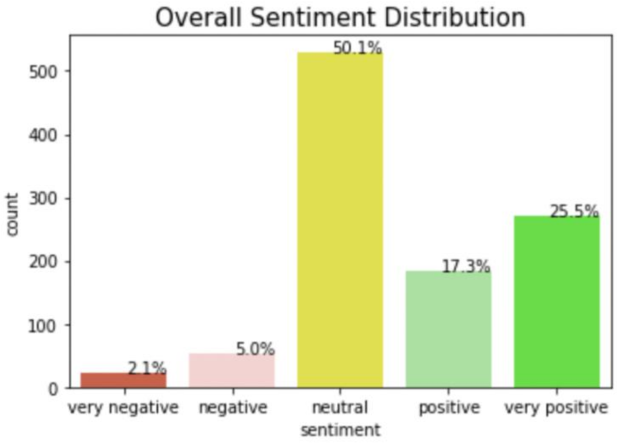


Figure 4.2 Overall Sentiment Distribution

V. Results

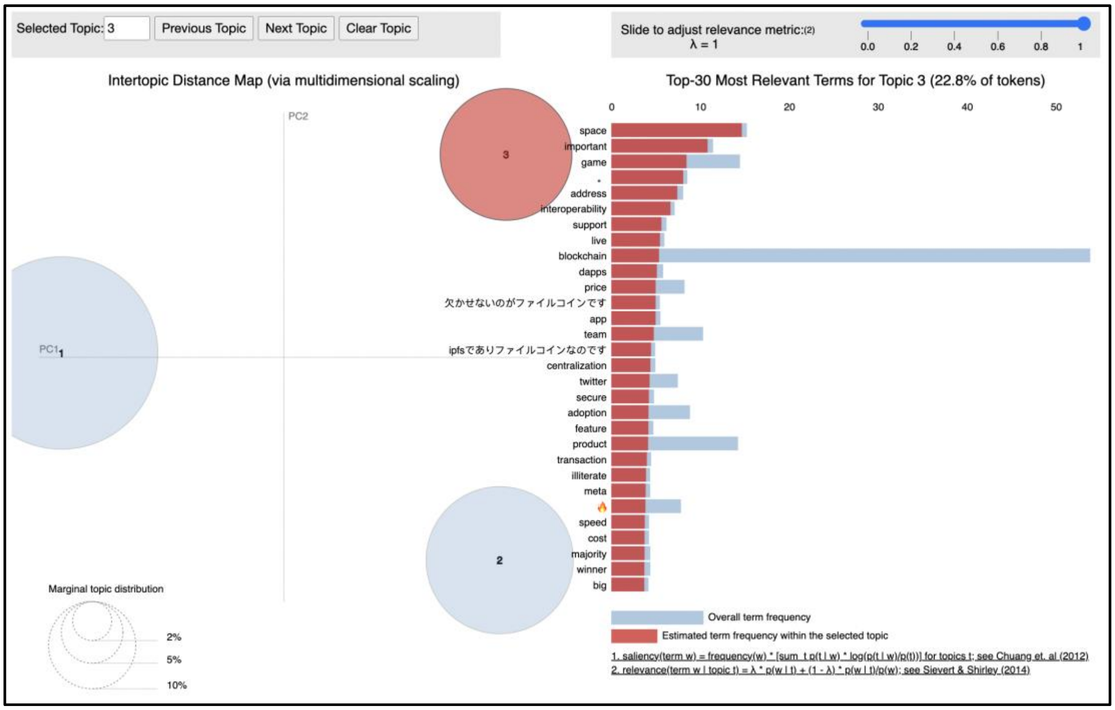
i) Topic Modeling on Tweets from January 2021 to August 2021

The generated topics and the relevant terms from the tweets from January 2021 to August 2021 are shown in Table 5.1.

Topic	Relevant terms in the Topic
1	Project, Internet, Future, Blockchain, DeFi, Bitcoin
2	Network, Wallet, Blockchain, Money, Application
3	Space, dApps, Address, Game, Interoperability

Table 5.1 Generated Topics and Relevant Terms

Figure 5.1 below shows the topic visualization for tweets from January 2021 to August 2021 generated using pyLDAvis.



Selected Topic: 3

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5.1 Interactive Topic Visualization

ii) Topic Modeling on Tweets from September 2021 to March 2022

The generated topics and the relevant terms from the tweets from September 2021 to March 2022 are shown in Table 5.2.

Topic	Relevant terms in the Topic
1	Metaverse, Project, Web3, Community, world, NFT, Internet
2	Future, Space, Privacy, Money, People, Good
3	Blockchain, Crypto, Technology, Network, Metaverse

Table 5.2 Generated Topics and Relevant Terms

Figure 5.2 below shows the topic visualization for tweets from September 2021 to March 2022 generated using pyLDAvis.

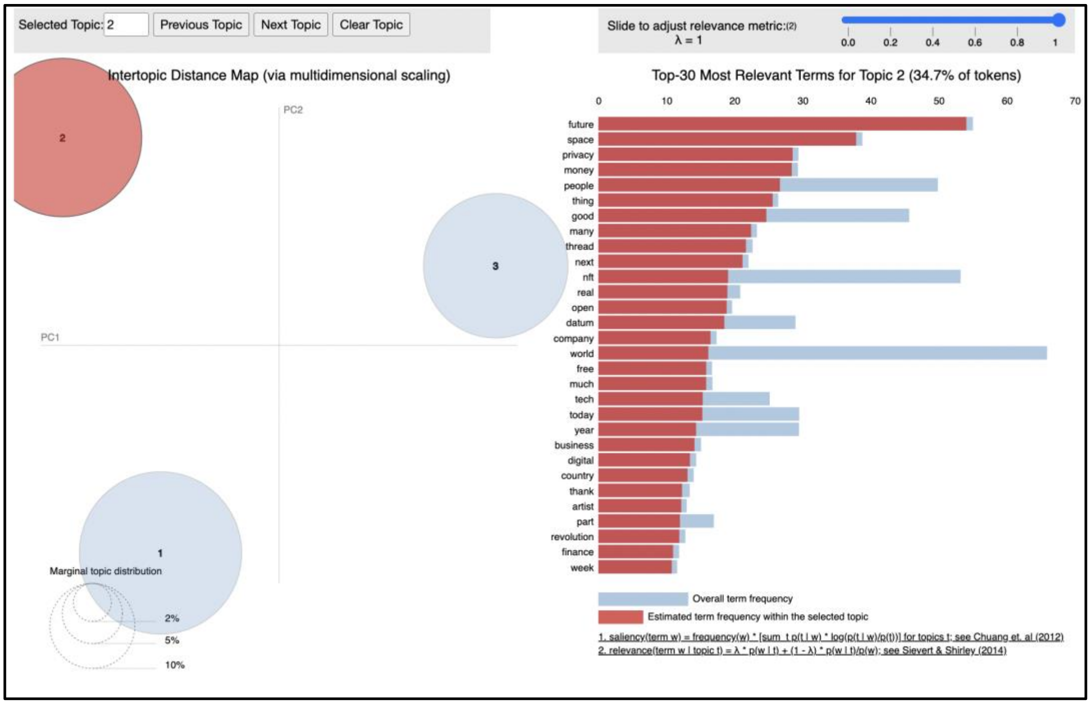

Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
3
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (34.7% of tokens)
0
10
20
30
40
50
60
70
future
space
privacy
money
people
thing
good
many
thread
next
nft
real
open
datum
company
world
free
much
tech
today
year
business
digital
country
thank
artist
part
revolution
finance
week
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5.2 Interactive Topic Visualization

Discussion

After analyzing the topics, the results indicate that in early 2021, people discussed and shared information on Blockchain, DeFi, Bitcoin and how Web3.0 is the future. The results from Sep 2021 to Mar 2022 indicate newer terms like Metaverse, NFT Projects, Crypto and discussion on Web3.0 learning communities which are very active on twitter. One interesting result is the occurrence of term ‘Future’ in results from both timeframes which indicates there is much more to Web3.0 that is yet to come.

i) Sentiment Analysis on Tweets from January 2021 to August 2021

The chart Figure 5.3 below shows the overall sentiment distribution of Tweets on Web3.0 from January 2021 to August 2021.

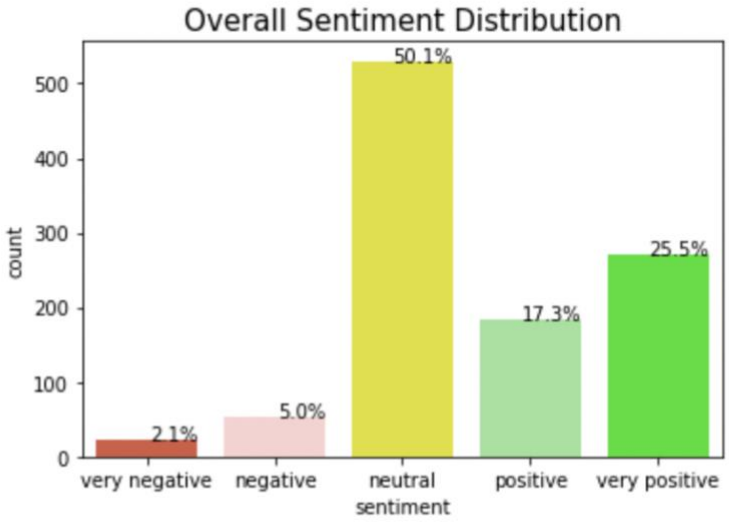


Figure 5.3 Overall Sentiment Distribution

ii) Sentiment Analysis on Tweets from September 2021 to March 2022

The chart Figure 5.4 below shows the overall sentiment distribution of Tweets on Web3.0 from September 2021 to March 2022.

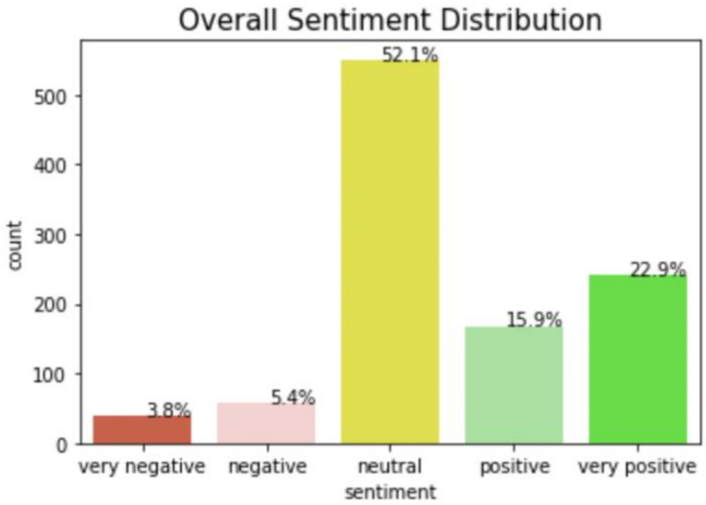


Figure 5.4 Overall Sentiment Distribution

Discussion

After analyzing the sentiment distribution, the results indicate that majority of the tweets are of neutral sentiment in both the timeframes because of the information related tweets on Web3.0 which are tagged as neutral sentiment. Around 40 % of the tweets contain positive sentiments which indicates that people are excited to learn and explore more about Web3.0. The tweets with negative sentiments are less then 10% in both the timeframes.

VI. Limitations

Though careful measures were taken to improve the quality of data being used for the analysis, there could be possibility of noise in the data from tweets containing spams like NFT Project promotions, Ethereum Giveaways etc. A more robust filtering method to target such spam tweets could significantly improve the quality of the data and produce better results.

VII. Conclusion

In this research, topic modeling and sentiment analysis was conducted on 2000 tweets collected from twitter from two time frames i.e., January 2021 to August 2021 (1000 tweets) and September 2021 to March 2022 (1000 tweets).

The results from topic modeling indicate how new technologies are being evolved in Web3.0 space. In 2021, the major topics being discussed included Blockchain, DeFi, Bitcoin etc. and by the end of 2021 and beginning of 2022 there were new technologies like Metaverse, NFT, Cryptocurrency were being discussed. This shows how quickly the Web3.0 space is evolving.

The results from sentiment analysis from both the timeframes indicate that around 40 % of the tweets contain positive sentiment on Web3.0 and less than 10% tweets have negative sentiment. The remaining 50% tweets are neutral as majority of the tweets are informational in nature and do not have any sentiment related to it.

After this study it can be stated that

- Newer technologies keep on emerging in Web3.0 space, with Metaverse and NFTs trending in current times.
- The sentiment associated with Web3.0 in people is majorly positive.

VIII. References

Kreps, D. and Kimppa, K. (2015), "Theorising Web 3.0: ICTs in a changing society"

Chohan, Usman W. and Chohan, Usman W., Web 3.0: The Future Architecture of the Internet? (February 7, 2022).

Owa, D.L.M. (2021) Identification of Topics from Scientific Papers through Topic Modeling. Open Journal of Applied Sciences, 11, 541-548. https://doi.org/10.4236/ojapps.2021.114038

Vasiliev, Y. (2020). Natural Language Processing with Python and SpaCy: A Practical Introduction. No Starch Press.