# What does the Discord Says: Case Study from Ocean Protocol Community

*A Submission to Ocean Data Challenge hosted through Desights AI*

**Dominikus Brian 钟鸿盛 | domi@dreambrook.tech**
**www.dreambrook.tech**

# Introduction

In this comprehensive case study, we performed exploratory analysis on Ocean Protocol Discord Server to uncover insights regarding the community growth and how it affects key ecosystem metrics such as $OCEAN token price and also dynamics of member communication and activities. To begin with, here let us briefly describe the data input for this study. The dataset taken as input for this analysis spans across 1,645 days starting with first entry from Author named *sheridan_oceanprotocol who* first *Joined the Server* on *2019-08-20 19:55:00* to entry from Author named *wanderclyffex* whose latest recorded post in the dataset were *"Thank You* 👍*"* on *2024-02-20 18:04:00.* In total the "Raw Data" used to bootstrap the analysis we perform here contain a total of 84,754 entries, including a substantial 15,696 entries from GitHub bot feeds mainly for new commits from Ocean Protocol's GitHub Page. For each row entry, of the Raw Data there is 7 items included, namely: 'Channel','AuthorID', 'Author', 'Date', 'Content', 'Attachments', and 'Reactions'. To complement upon this dataset, we also downloaded $OCEAN token price daily data starting from 2019-05-04 to 2024-03-07. In the following sections we will highlight our key findings associated with general trend and patterns observed in the dataset both in different timescale and categories. We elaborately explore and analyze the correlation between $OCEAN token price and daily server metrics. Alongside the analysis we present numerous interesting rankings, leaderboard, and categorization that can be useful to understand community dynamics and groups. Thematic discussions were analyzed in order to extract prototypical topic and questions that is recurring throughout the dataset. Most importantly we also utilized various metrics and processed data in this study to then build a scam alert analysis procedure and classification model, followed with a forecasting model that capable of predicting next-day server statistics based on historical data of the past 3 days or 1 week. In the Key Findings section, we present the main findings of this report. Visualization and Discussions narrate through the insights uncovered in this work, followed with Methodology section filling in details of method and techniques used. We then conclude the work along with providing several suggestions on how the tools and pipeline developed here can be further extended to build useful technology such as Discord Bot and AI assistant for community server.

# Key Findings

**General trends and patterns.** Activities within the Ocean Protocol discord server grew rapidly from 2019-2022, after peaking during the summer months of 2022, the community seeing a decline in activities in year of 2023. Data tracking by months revealed that between the months of April to August users are more active compared to early or late during the year. Daily activities count indicated that early and last days of the month often accompanied with moderate surge of activities. Hourly activities are at minimum around 6-12AM UTC. Activities during $5^{th}$ -$8^{th}$ July 2022 exhibit record breaking numbers in both number of new users joining the server and active user posting. This significant up surge seems to traced back to the week in which Ocean V4("Onda") just went live and Ocean Data Bounty program was announced.

**Correlation between $OCEAN token price and daily server metrics.** In order to quantify daily server activities that may correlate to the $OCEAN token price, we introduce 5 quantifiable metrics, namely: n_author, n_channels, n_activities, n_words_ave, Sentiment_value. Statistical analysis and correlation studies imposed on these metrics and price data shown that Sentiment_value and n_words_ave is the two most correlated metrics with Pearson correlation of 0.63, and 0.46, respectively. Sentiment_value in particular demonstrated a superb usefulness when it comes to price forecasting, the trend in which a flux of increasingly positive Sentiment_value commonly occurs several days before $OCEAN token price increase was then observed.

**Rankings, leaderboard, and category clustering.** These exciting visual / characteristic depictions of the server dynamics are capable of providing quick overview of what the whole community landscape looks like, and also to find important key players or opinion leaders in crucial to the server activities. We cluster these findings based on mainly channel or author distinction. Several users that also part of Ocean Protocol Ambassadors such as *blockchainlugano, kreigdk, dotunwilfred.eth, zippy1979,* and *robinlehmann* are extraordinarily active and engaging in all aspects of activities within the server. Other user such as *alexcos20*, *white_rider*, *wanderclyffex, innopreneur,* and *cesar_cripto,* are several that are the most responsive toward answering questions and providing explanation across community. There's also users shown to be very active for reason such as routine "gm/good day" message (e.g., *bhavingala*) and by numerous conversation with Ocean GPT (e.g., *lukas85#6738*)

**Thematic discussions and questions.** Toward distinguishing different topic or genre within the server activities we utilized combination of both text analysis and also the latest of large language model technologies. Tokenization embedding and text processing were performed on whole dataset and also on subset of the raw data such as for FAQ related content and scam related one. Token count shown that words such as "building", "academy", "tutorial", "information", "contribute" are one of the most prominent across the tokenized content data, indicating the notion of BUIDL culture amongst members that joined the server. Visual results are presented via word cloud format. Topic wise, discussions around NFT ownership, Wallet and API tools, Development Tools, also Testnet related Resources are frequently observed.

**Scam analysis and classification model.** Noting that the dataset didn't explicitly contain a labeled "scam" data subset, therefore in this work we developed a first-degree approximation to this by curating a "scam alert" data. The scam alert data comprise of data that are highly associated with the occurrence of scam incident, i.e., conversation that happens before or after the scam. Based on this dataset using a RandomForest classification algorithm we develop a classification model that produce either “Possible Scam Detected” or “Safe and Sound” given a string of text as input. The RandomForest classifier developed achieved a decent f1-score (metrics for accuracy in classification) of 0.95 for detecting possible scam and 0.99 for identifying non-scam. Associated confusion matrix obtained from the training and test were also shown.

**Daily server statistics forecasting model.** To maximize the utility of all the developed datasets and analysis protocol, we also go a step further by building a forecasting model for the price and 5 server daily activity metrics. The goal for this forecasting is targeted toward predicting the next-day daily server statistics based on the past 3 or 7 days historical information. The developed Machine Learning (ML) models performed adequately well for predicting price and Sentiment_value with RMSE and MAE of 0.034 and 0.019 for price prediction using past 7 days data and RMSE and MAE of 0.055 and 0.042 for Sentiment_value prediction.

# Discussions

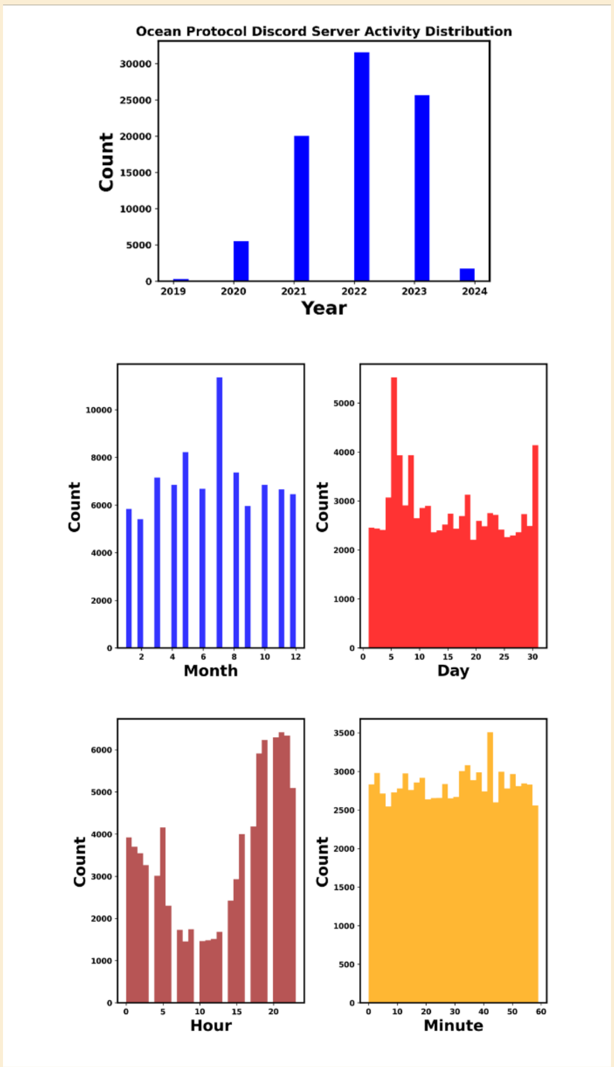



**Figure 1.** Temporal Distribution of the Ocean Protocol Discord Server Activity by Year, Month, Day, Hour, and Minute.

We first began with checking the temporal distribution of the Ocean Protocol Discord Server activities. This is performed by clustering the entries within the raw data into histogram bins based on year, month, day, hour, and minutes. The results of which is shown in Fig.1. We noted that within the data shown there’s a significant portion of it that were of bot generated activities. In our preliminary data exploration, we have also conducted histogram analysis for data without several of the most active bots such as the GitHub bot and the MEE6#4876 Bot along with the OceanGPT, OceanDiffusion#4502 and awesomeQA#0813.

The comparison shows that the general trend is preserved as shown in Fig.1. Here we also highlight the key factor that to a certain degree skew the peak signals in activities during year of 2022, Month of July, and also day 6. These extraordinary counts are due to sudden surge in activity throughout the time window between 5-8$^{th}$ July 2022. Next, we investigate the activity distribution not merely in a single time space but also in combinations with each of the 21 channels available in the discord server at the time of dataset extraction.
The results of these analysis is aggregated in Fig.2, and also given in full throughout Appendix A1.

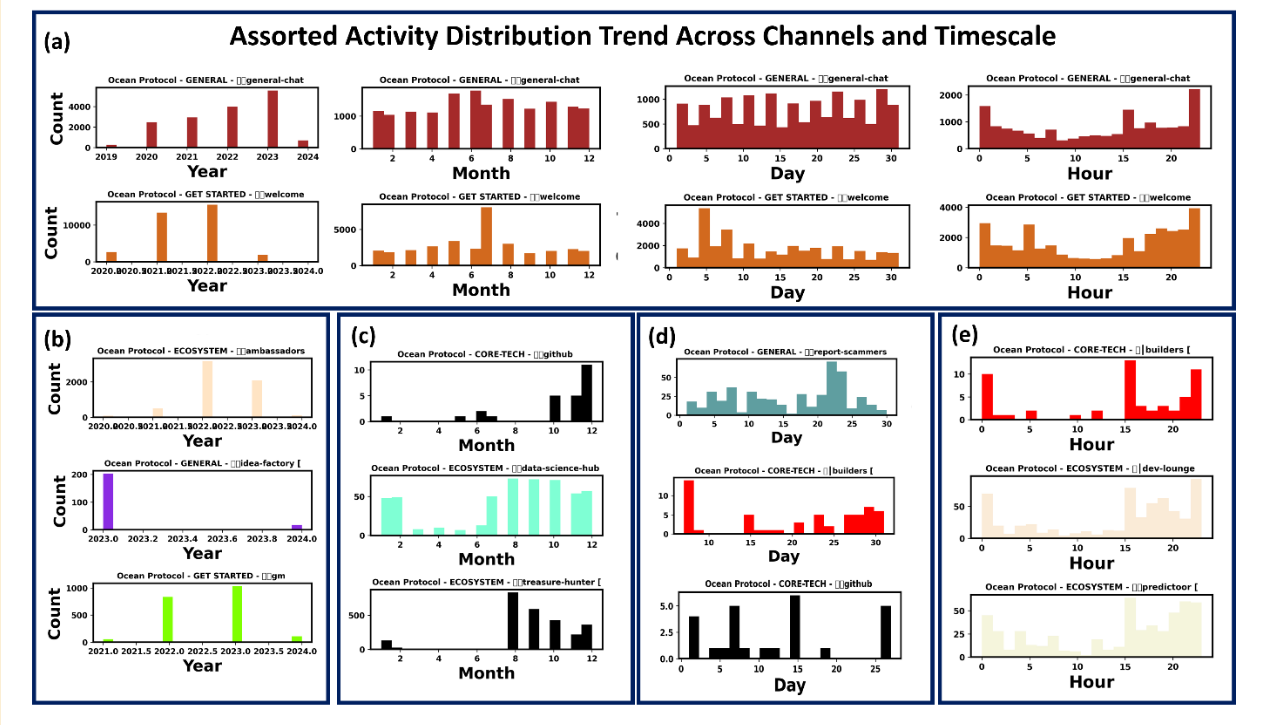



**Figure 2.** Activity Distribution Trend across both channels and timescale.(a) For all timescale focusing on the general-chat and welcome server, and snapshot from (b) yearly, (c) monthly, (d) daily, (e) hourly activities across different channels.

Careful Inspection of Fig.2 brought to light several interesting patterns, in which some of the activity are similar across channels (e.g., on hourly basis), and more dispersed on yearly, monthly, and daily basis for different channels. Channels such as Core tech builders are typically very active early of the month and not so much the rest of the month. Github data is more pattened among weeks with peaks of activities every other week or so, followed with huge push at the end of the year. Data for general chat and welcome channel indicated highly periodical and regular pattern. This is also supported, as shown in Fig.3 by the fact that these two channels are the most active by activities count , with 33530 and 15955 activity counts, respectively. Fig.3 also shows the ranking leaderboard of activities by author (including both bot and non-bot) followed with remarks regarding the author identifiable characteristics.

## Ranking Leaderboard

*(a) By Channel*

| Rank | Channel Name | Activities Count |
|---|---|---|
| 1 | Ocean Protocol - GET STARTED - 👋 \| welcome [7278... | 33530 |
| 2 | Ocean Protocol - GENERAL - 💭 \| general-chat [612... | 15955 |
| 3 | Ocean Protocol - CORE-TECH - 🚄 \| github [7739269... | 15722 |
| 4 | Ocean Protocol - ECOSYSTEM - 🧬 \| ambassadors [76... | 5956 |
| 5 | Ocean Protocol - ECOSYSTEM - 🧬 \| treasure-hunter... | 2594 |
| 6 | Ocean Protocol - GET STARTED - 🌞 \| gm [911643560... | 2038 |
| 7 | Ocean Protocol - GET STARTED - 🖐 \| introduce-you... | 1808 |
| 8 | Ocean Protocol - ECOSYSTEM - 🎨 \| ai-fun [1088460... | 1724 |
| 9 | Ocean Protocol - GENERAL - 🎉 \| announcements [72... | 1373 |
| 10 | Ocean Protocol - GET STARTED - 😵 \| ask-the-ai [1... | 707 |
| 11 | Ocean Protocol - ECOSYSTEM - 💻 \|dev-lounge [971... | 613 |
| 12 | Ocean Protocol - ECOSYSTEM - 🤖 \| predictoor [115... | 532 |
| 13 | Ocean Protocol - ECOSYSTEM - 📊 \| data-science-hu... | 512 |
| 14 | Ocean Protocol - ECOSYSTEM - 🚜 \| data-farming [9... | 511 |
| 15 | Ocean Protocol - GENERAL - 🚨 \| report-scammers [... | 458 |
| 16 | Ocean Protocol - GENERAL - 🐦 \| tweets [116232715... | 254 |
| 17 | Ocean Protocol - GENERAL - 💡 \| idea-factory [108... | 220 |
| 18 | Ocean Protocol - GENERAL - 🗣 \|ocean-lounge [108... | 93 |
| 19 | Ocean Protocol - ECOSYSTEM - 📈 \| traders-lounge ... | 83 |
| 20 | Ocean Protocol - CORE-TECH - 🌊 \|builders [10887... | 56 |
| 21 | Ocean Protocol - CORE-TECH - 🚀 \|tech-announceme... | 15 |

*(b) By Author*

| Author | Activities Count | Is Bot? | Remarks |
|---|---|---|---|
| **MEE6#4876** | 20986 | Y | Utility Bot |
| **GitHub** | 15696 | Y | GitHub Bot |
| **Deleted User** | 5853 | N | Users that Left/Deleted from the server |
| **blockchainlugano** | 1207 | N | Ocean Protocol Ambassador Lead |
| **kreigdk** | 1126 | N | Ocean Manager \| Elite Ambassador |
| **dotunwilfred.eth** | 1113 | N | Makes a lot of official Announcement for challenges and winners, answer community questions. |
| **bhavingala** | 1016 | N | Routine GM/ Good Day |
| **zippy1979** | 904 | N | Ocean Protocol Ambassador Lead |
| **robinlehmann** | 856 | N | Co-founder at DataUnion \| Ocean Protocol Ambassador |
| **lukas85#6738** | 802 | N | Chat with OceanGPT |
| **doteth** | 677 | N | Ocean Protocol Team |
| **OceanDiffusion#4502** | 637 | Y | Image generation Bot |
| **bigimeyagazzz** | 542 | N | Contributor \| Blockchain Dev |
| **birususama** | 536 | N | Community Manager |
| **white_rider_** | 516 | N | Contributor |
| **awesomeQA#0813** | 416 | Y | QA Bot |
| **qazim** | 387 | N | Consistent GM |
| **alexcos20** | 371 | N | Ocean Core Team , VP of Engineering, From Romania, French Speaker. |
| **OceanGPT#0740** | 331 | Y | GPT Bot for Ocean |
| **mickssp** | 329 | N | Answer community questions, Routine GM, Share official post |
| **denkobetona** | 317 | N | Active in Discussion, Bulgarian Speaker |
| **wanderclyffex** | 301 | N | Ocean Protocol Team |
| **hanadi.fh** | 299 | N | Community Manager, Post interesting events. |
| **jesusdatalatte** | 296 | N | Ambassador, Spanish Speaker |
| **innopreneur** | 290 | N | Founder of Desights AI, Detail answers to community questions. |
| **zephilno_1** | 277 | N | Community Discussions |
| **cesar_cripto** | 267 | N | Ask and Discuss questions in forums, participate in Data Challenge, Spanish Speaker |
| **lrt8723** | 267 | N | Ambassador, Spanish Speaker. |
| **Alexpao#9344** | 258 | N | Community Manager, Post events, Female |
| **Ocean Protocol • TweetShift** | 250 | Y | Tweet Related Bot |

*(C) By Aggregated Channel Type*

| Channel Type | Activities Count | Total |
|---|---|---|
| CORE-TECH[0-2] | 56+15+15722 | 15,793 |
| ECOSYSTEM [3-10] | 1724+613+83+<br>512+511+532+<br>5956+2594 | 12,525 |
| GENERAL [11-16] | 1373+254+220+15955+93+ 458 | 18.353 |
| GET-STARTED [17-20] | 2038+33530+1808+707 | 38,083 |

**Figure 3.** Activity Ranking Leaderboard (a) by channel, and (b) by author, including both bot and non-bot authors, (c) by Aggregated Channel Type.

In Fig.4.we plotted the temporal evolution of the Ocean Protocol Discord Server Daily Stats. In this temporal evolution data, the anomaly we mentioned above regarding data around early July 2022 are evidently present such that appearing in Fig. 4(b) and (c). 3 Days that account for most of this anomaly are $5^{th}$ (349 users 3207 activities), $6^{th}$ (516 users, 888 activities) and $8^{th}$ July (277 with 1675 activities). Too look into this, we peer through the ocean report for that particular week i.e. https://blog.oceanprotocol.com/introducing-the-ocean-data-bounty-7eb1e3bf3c0a. We also noticed that particular week, Ocean V4("Onda") just went live and Ocean Data Bounty program is announced. Interestingly Many of these sudden user spikes were then deleted after sometime.

Notably from comparing Fig.4 (a) and (d) we also observed interesting pattern of flux in sentiment values originating from excitements or verbosity in communication in the discord server. Following this flux, almost immediately a price hike can be observed following the same pattern. This is an interesting indication of cause-and-effect relation between server activity and the daily market price of $OCEAN Token. In Fig. 4(f) the number of daily active channels are increasing steadily, this is mainly due to newly created channels along the years, and also due to newer member then have more channels to visit and post into.

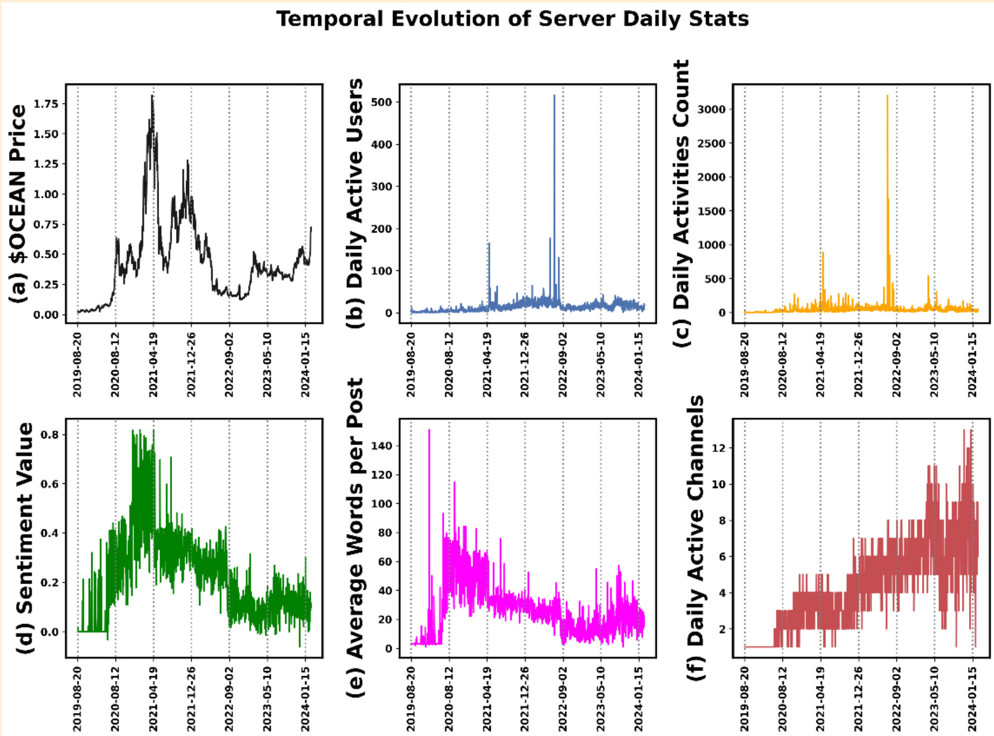



**Figure 4.** Temporal Evolution of Ocean Protocol Discord Server Daily Stats.

To further quantify and analyze this correlation, we performed statistical analysis by means of Pearson correlation studies and histogram counts, along with price-colored grid plot scatters, as shown in Fig.5. The correlation between Price and Sentiment_value is the most apparent with positive correlation of 0.63, this can be vividly understood by looking at the bottom left most panel showing the scatter plot between the two variables. In which, we can observe that increasing hype or positive sentiment are also almost always associated together with price increase. Beyond correlation with prices, we also identified logical and strong positive correlation of 0.77 between n_author (i.e., Number of Daily Active users) and count of activities. Other correlation worthy of mention is the n_words_ave and Sentiment_value with positive correlation of 0.78. This indicate that when the community discussion tone become increasingly positive, on average, each post contains more words.

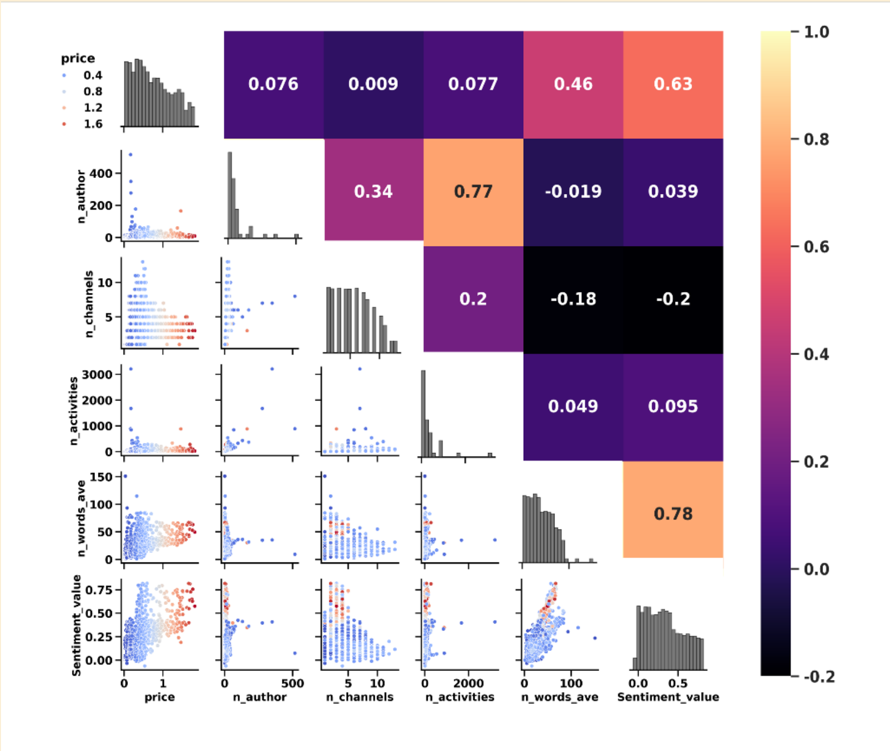



**Figure 5.** Correlation and Distribution of Daily Price (snapped at UTC 00.00) and Server Daily Stats (added up to the end of the day(in UTC Timezone))

Next, we progress toward categorization of members/user within the discord server. In Fig 6. We show one possible useful categorization based on activity counts, how long a user been joining the Ocean Protocol discord server, how many channels that the users participated in, as well as the overall mean sentiment value of each user. By activity counts, user is categorized as either Join Only, Participatory, Active, Very Active, or Super Active, based on their Activity Counts. The number of days in server is calculated based on how many days since the user have made their first appearance (e.g., Joined the Server) and the last time they post in the Discord server. Based on this, user is distinguished into Newcomers, Older Guys, Early Members, and Founding Members. Notably, the distribution of these two categorizations is skewed toward the lower end, owing to the fact that most users are either Join Only and Participatory which also generally belong to the Newcomer category. By Channel participations, we introduce 3 types of categories, so-called Visitors, Travelers, and Explorers. Typically, most users are merely involved in 1-3 servers, and mostly is within the general-chat and welcome channel. Only a small fraction of users ascends to the category of Travelers with 4-6 Channels been participated in and even less with mere 24 users with Channel participation more than 9. Based on Sentiment value we can simply divide the dataset into a rather negative toned entries, neutral, and also positive toned one. The sentiment value quantification were performed using pretrained text blob python package.

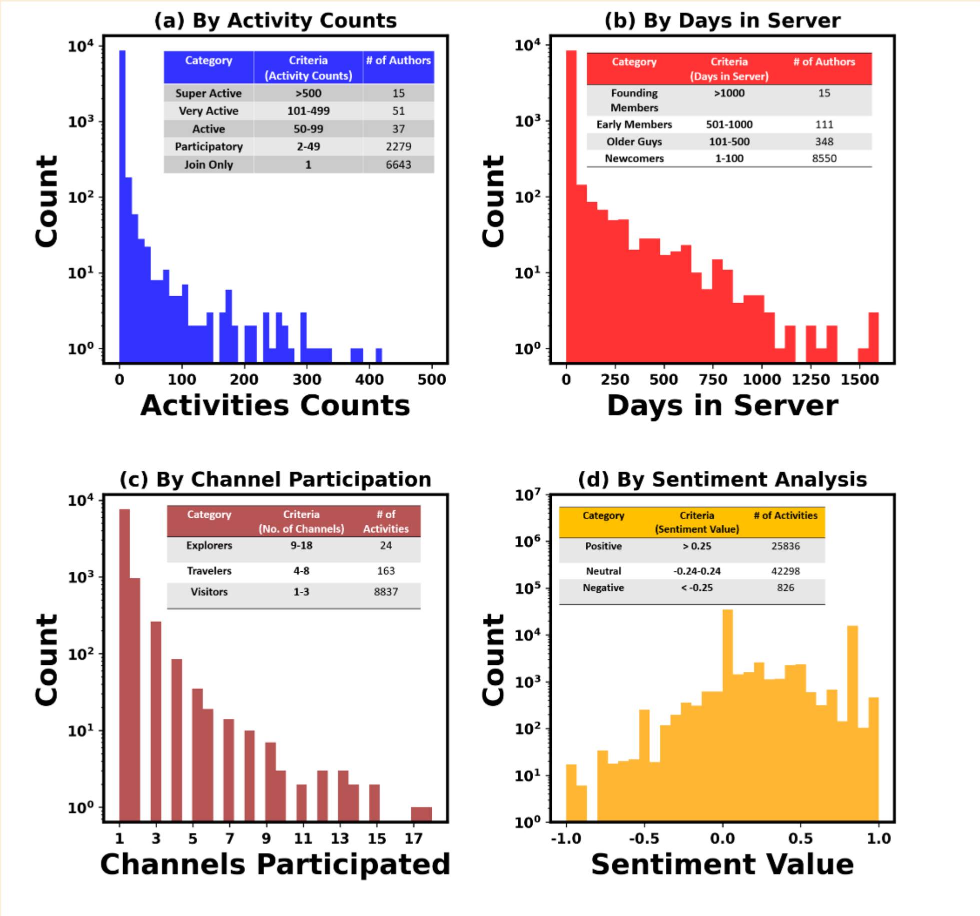



**Figure 6.** Classification of Members by Activity and Other metrics

To understand context of the community activity better, not only quantitatively but also qualitatively, we now dig deeper into the text content in the dataset. To start with, in Fig.7. we first present the word clouds based on raw data text content, and also a processed (cleaned from stopwords, non-alphanumeric symbols, etc.)

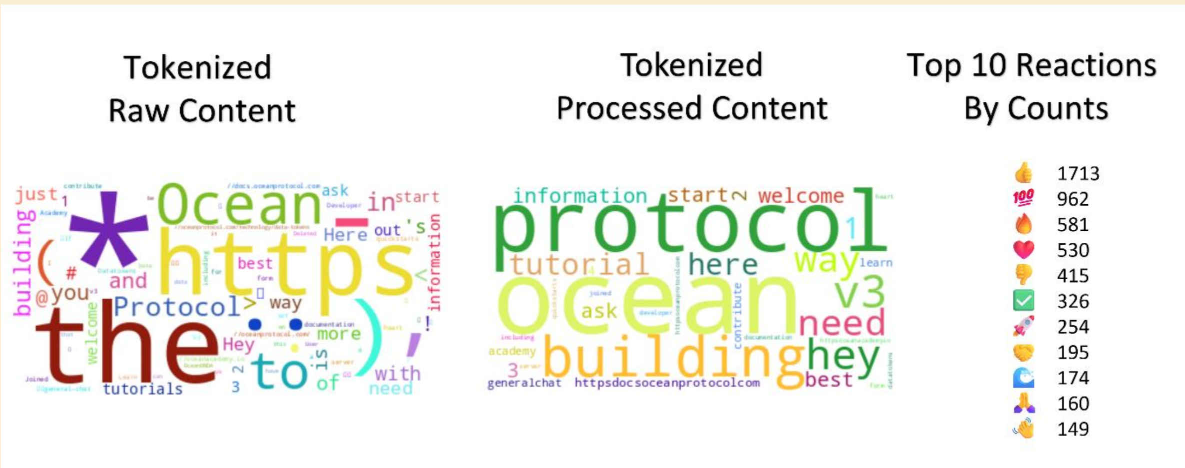



**Figure 7.** Word Cloud of Tokenized Raw Content, Tokenized Processed Content, and Emoticon Reactions used frequently in the Ocean Protocol Discord server.

The tokenized content data provide us with a sneak peek toward the most frequent keywords, or text/symbol appearing throughout the server. Before moving forward, let us also revisit the qualitative aspects of the sentiment value we discussed above. To this end, let us first peer into the negative example. Authors that have activity post associated with this negative sentiment is shown in Fig.8. below.

| Author | Sentiment Score | Total Activities |
|---|---|---|
| maesa3577 | -0.637500 | 4 |
| .phucson | -0.500000 | 1 |
| quandinhhh | -0.500000 | 1 |
| HypnoJimmy#0379 | -0.500000 | 1 |
| .0xace | -0.500000 | 1 |
| waterandpani_14049 | -0.468750 | 2 |
| muhammadqasim92 | -0.400000 | 1 |
| ish343434 | -0.350000 | 1 |
| hrdl1 | -0.333333 | 2 |
| flyycypher | -0.312500 | 1 |
| gotothemoon0504 | -0.312500 | 1 |
| smartvic771 | -0.266667 | 3 |

**Figure 8.** Screenshot of few top most Author with highly negative Sentiment score

Text content that contained keywords like the F* word, Sorry, Error, fails, invalid, lose, and so on are highly likely to be designated with negative Sentiment score. Positive sentiment is on the line of keywords such as good, great, happy, love, best, welcome, holiday, and such. Meanwhile interactions such as, hello, hi, welcome, and so on are in the realm of neutral sentiment.

Noting of impossibility to manually read over or analyze line-by-line the text content within the dataset, we turned our attention to one of the latest hot technologies in AI, i.e., Large Language Model. By first producing the meshed content text strings, data subsets, such as those containing specific keywords, or from specific channels. We then feed these smaller subsets of text into GPT-4 model to provide insights on the general topics and conversational theme contained within the data subset. Here we focused on describing part of the result obtained for the question related data subset. It was revealed that in general most questions in the community revolve around several core topics such as Asset Security related to NFT or Stake. A number of questions were also specifically addressed toward various wallet related concerns or API issue. Development tools, such as that related to data marketplace and also the compute-to-data concept are also appearing here and there within the data subset. Last but not least, testnet related resources also turned out to be one of the most questioned topics.

Spam Classification is one of the very useful applications we can build based on the dataset available. Nevertheless, since the raw dataset is not properly labeled when it comes to scam related data. We work toward this problem via first degree approximation through identifying and curating a data subset that correspond to occurrence of scam instance. One source of this curated spam data is the report-scammer channel, another is by screening across all content for post that contains the following keywords: ['spam','scam','fraud','deleted','banned']. The curated scam dataset could then serve as a first order approach to scam alert classification.

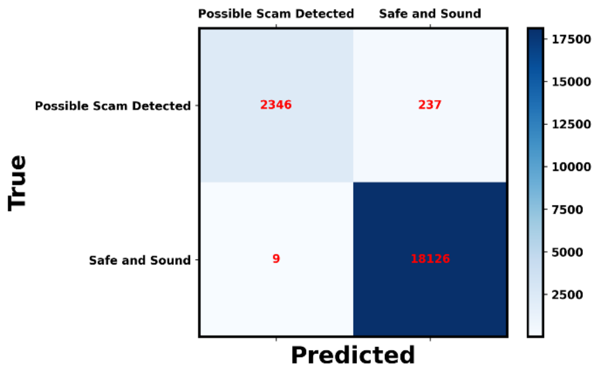
The curated scam dataset was first preprocessed by doing: lowercasing, removing non-alphanumeric characters, doing tokenization, stopword removal and also lemmatization. Then we split the data into training and test set with ratio of 70:30. The model used for this work is Random Forest Classification as implemented in Scikit-Learn package. Other details are provided in Methodology section. Evaluation of the model performance is based on the f1-score metrics. The classification model was designed to output 'Possible Scam Detected' for text that have high likelihood to be associated with Scam activity (that is text that likely to appears before or after scam post), and to output 'Safe and Sound' otherwise. The developed model based on RandomForest classifier achieved a modest outcome with significantly high accuracy, and a Confusion Matrix that indicated minimum False Positive or False Negative as shown in Fig. 9. Noting of how diverse text representation can be, the Classifier model may or may not easily overfitted. Thus, we run a simple deployment test to see its usability, as shown in Fig. 9 (c), out of 5 random deployments, 1 instance turned out to be doubtful, since a knowledgeable human can easily identify the context, but sometimes for the classifier model the strings of text may seems to be innocuous and deceptive.

**Random Forest Scam Alert Classifier**

*(a) Classification Report*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Possible Scam Detected | 1.00 | 0.91 | 0.95 | 2583 |
| Safe and Sound | 0.99 | 1.00 | 0.99 | 18135 |
| accuracy | | | 0.99 | 20718 |
| macro avg | 0.99 | 0.95 | 0.97 | 20718 |
| weighted avg | 0.99 | 0.99 | 0.99 | 20718 |

*(b) Confusion Matrix*




*(c) Test Deployment using Ocean Protocol Discord Server Conversations Not in the original dataset*

| Test Text | Classifier Output |
|---|---|
| "Hey everyone, I made a big mistake and got scammed by https://ocean-events.net/ Does anyone have any advice on what I can do now? All Oceans were transferred out of my wallet." | 'Possible Scam Detected' |
| "It looks really cool! All in one place and it's really easy to keep track of what's going on)" | 'Safe and Sound' |
| "@Mike-Alphaday Did you do it yourself?" | 'Safe and Sound' |
| "How did you get scammed ?@T-Dog" | 'Possible Scam Detected' |
| "@turnerraymond. he invited me to another Discord group. They asked for my wallet there... " | 'Safe and Sound' |

**Figure 9.** Result of the classification report and confusion matrix for Scam Alert classification based on RandomForest algorithm.

Last but not least, we proceed with developing and training forecasting models for predicting daily server activities and also $OCEAN token price. Since the data that are readily available for this forecasting task are in the form of time series, we here decide to employ a LSTM approach (Long-Short Term Memory) for the model building.

The forecast model will generate the "Forecasted Snapshot of Tomorrow Daily Server Stat", based on historical data of previous days. After running several trials we stick on with choosing 3 days and 7 days as the appropriate time window parameters used to train the LSTM model. The model were trained individually for each properties: n_author, n_channels, n_activities, n_words_ave, Sentiment_value, and $OCEAN price. The best models obtained after scanning through a sizable hyperparameters are then selected and combined together into Model_Set_3 (for 3 days time window input) and Model_Set_7 (for 7 days time window input). The LSTM model performance were measured based on root mean square (RMSE) and mean absolute error (MAE) metrics. Result for both the metrics comparison and test deployment is shown in Fig. 10. Here, we noted that the developed ML models work adequately well for two target properties, price and sentiment_value, while are off by quite a bit with the rather stochastic or unpredictable data of n_authors, n_channels, n_activites, or n_words_ave. Improvement of the models would be an interesting task to be performed in the future.

(a) LSTM Models Metrics Comparison

| Target Properties | Model Set | RMSE | MAE |
|---|---|---|---|
| Price | Model_Set_3 | 0.0327 | 0.0205 |
| | Model_Set_7 | 0.0339 | 0.0195 |
| N_channels | Model_Set_3 | 1.952 | 1.559 |
| | Model_Set_7 | 1.750 | 1.347 |
| N_author | Model_Set_3 | 3.791 | 2.989 |
| | Model_Set_7 | 3.395 | 2.549 |
| N_activities | Model_Set_3 | 12.97 | 11.15 |
| | Model_Set_7 | 14.93 | 12.96 |
| N_words_ave | Model_Set_3 | 5.915 | 4.907 |
| | Model_Set_7 | 5.842 | 4.976 |
| Sentiment_value | Model_Set_3 | 0.05586 | 0.042 |
| | Model_Set_7 | 0.05478 | 0.042 |

(b) Live Deployment (Model_Set_7 ; 5 Random Prediction)

*Predicted*

| price | n_author | n_channels | n_activities | n_words_ave | Sentiment_value |
|---|---|---|---|---|---|
| 0.020495 | 2 | 1 | 2 | 2.076720 | 0.002244 |
| 0.034423 | 0 | 1 | 0 | 2.083507 | 0.002148 |
| 0.071286 | 3 | 1 | 3 | 3.170236 | 0.197921 |
| 0.457622 | 3 | 3 | 3 | 3.168078 | 0.305909 |
| 0.178940 | 2 | 2 | 3 | 2.351841 | 0.074015 |

*Actual*

| price | n_author | n_channels | n_activities | n_words_ave | Sentiment_value |
|---|---|---|---|---|---|
| 0.017328 | 2 | 1 | 2 | 3.000000 | 0.000000 |
| 0.034805 | 2 | 1 | 2 | 3.000000 | 0.000000 |
| 0.078122 | 8 | 2 | 44 | 37.772727 | 0.147167 |
| 0.446491 | 22 | 6 | 49 | 27.428571 | 0.260756 |
| 0.180890 | 8 | 6 | 16 | 13.937500 | 0.099740 |

**Figure 10.** Forecasting model for daily server statistics and $OCEAN token price.

Along with this, we want to also provide some suggestion regarding possible future work to expand the tools utilized and developed for this case study to be deployed in building a specialized discord bot for scam alert and server activity prediction. The bot would then also have capability to perform all the above analysis on-the-fly, as the server and community grows.

# Methodology

### Data Cleaning, Analysis, and ML Model Construction

Here we provide some further details regarding the tools used for Data Cleaning, Analysis, and ML Model Construction. Overall, we use our in-house developed python tools called DreamBrookPy, but also incorporate several commonly used specific task module such as text blob and nltk for text analysis, also scikit-learn and keras for machine learning purpose. For correlation analysis we utilized the Pearson correlation coefficient that measures the linear relationship between two datasets. Ranging from -1 to 1, a value of 1 implies a perfect positive linear relationship, -1 implies a perfect negative linear relationship, and 0 implies no linear relationship. The Pearson Correlation Coefficient between two variables $X$ and $Y$ is defined as:

$$r_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}$$

In the above $\bar{X}$ and $\bar{Y}$ are the means of the $X$ and $Y$ variables, respectively.

Random Forest Classifiers are an ensemble learning method commonly used for classification (and regression) tasks that operate by constructing a multitude of decision trees at training time. When we train Random Forest Classifier model, it creates a diverse set of decision trees, each based on a random subset of the training data and features. This randomness ensures that the trees are uncorrelated, which helps in reducing the model's variance. When making predictions, the classifier takes the majority vote from these trees to decide the final class. In developing our Random Forest classifier, we have specified the following GridSearch parameters given in Fig. 11. (left panel).

```
p_grid =  {
'bootstrap': [True],
'max_depth': [10, 20, 50, 100],
'max_features': [2, 3, 5, 10],
'min_samples_leaf': [3, 4, 5],
'min_samples_split': [8, 10, 12],
'n_estimators': [100, 200, 300, 1000]
}
```

```
param_list_1 = [3,7,14] #window size
param_list_2 = [32] # batch size
param_list_3 = [2] # lstm depth
param_list_4 = [0.2,0.3] # dropout rate
param_list_5 = [50] # N_unit
```

**Figure 11.** Hyperparameter for the (left) Random Forest Classifiers and (right) LSTM models used in this work.

The LSTM model was developed by searching for optimum hyperparameters based on a comparably smaller space as shown in Fig.11 (right panel). Primarily, the window size in term of number of days in the history retain is optimized.

# Conclusion

In conclusion, throughout our analysis of Ocean Protocol Discord server data from August 2019 to February 2024, we reveal several distinct trends and patterns in user engagement and activity that ought to be useful for understanding the community dynamics and need. The key findings of this works span across 6 aspects:

(1) General trends and patterns where we discover $5^{th}$ -$8^{th}$ July 2022 exhibit record breaking activities most likely owing to launch of Ocean V4("Onda") and Ocean Data Bounty program.

(2) Strong correlation between $OCEAN price and Sentiment_value fluctuation serves as a good signal to justify trends of token price.

(3) Clustering of Users by various categories help distinguish different user niche and type.

(4) Questions across the community often revolve around the development tools and also relevant integration with the Ocean Protocol system.

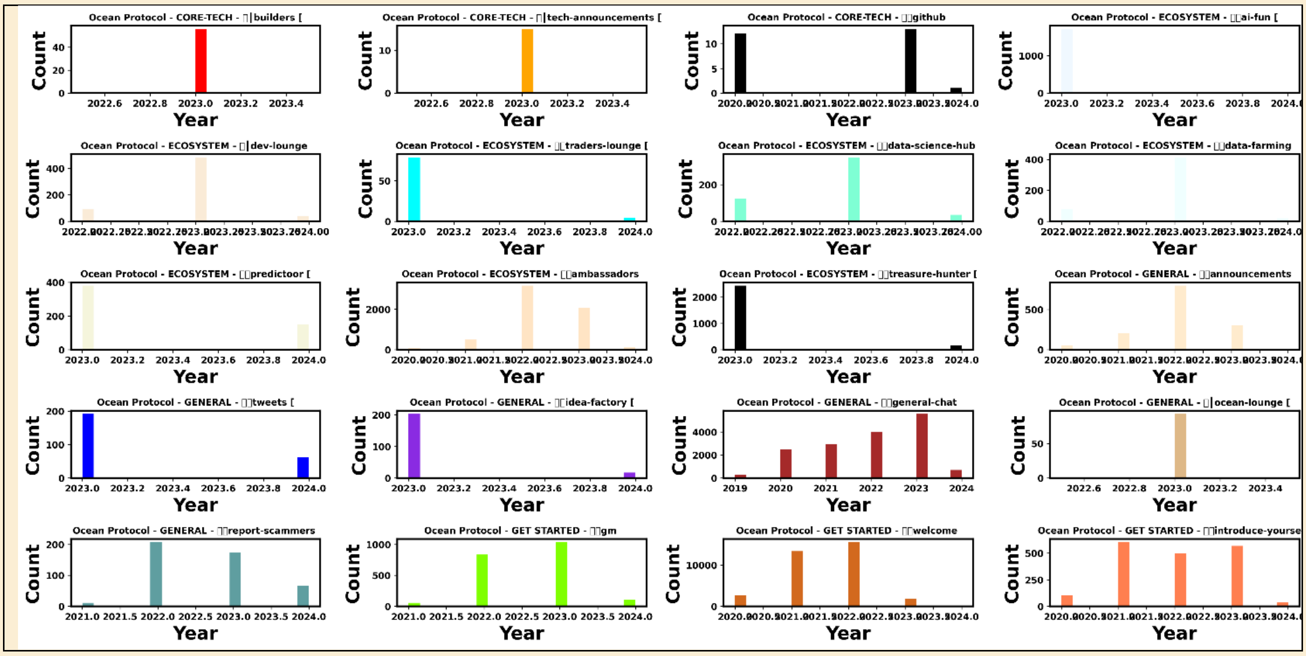
(5) Scam Alert system that we developed based on Random Forest classifier in this work perform modestly in providing a scam detection model.

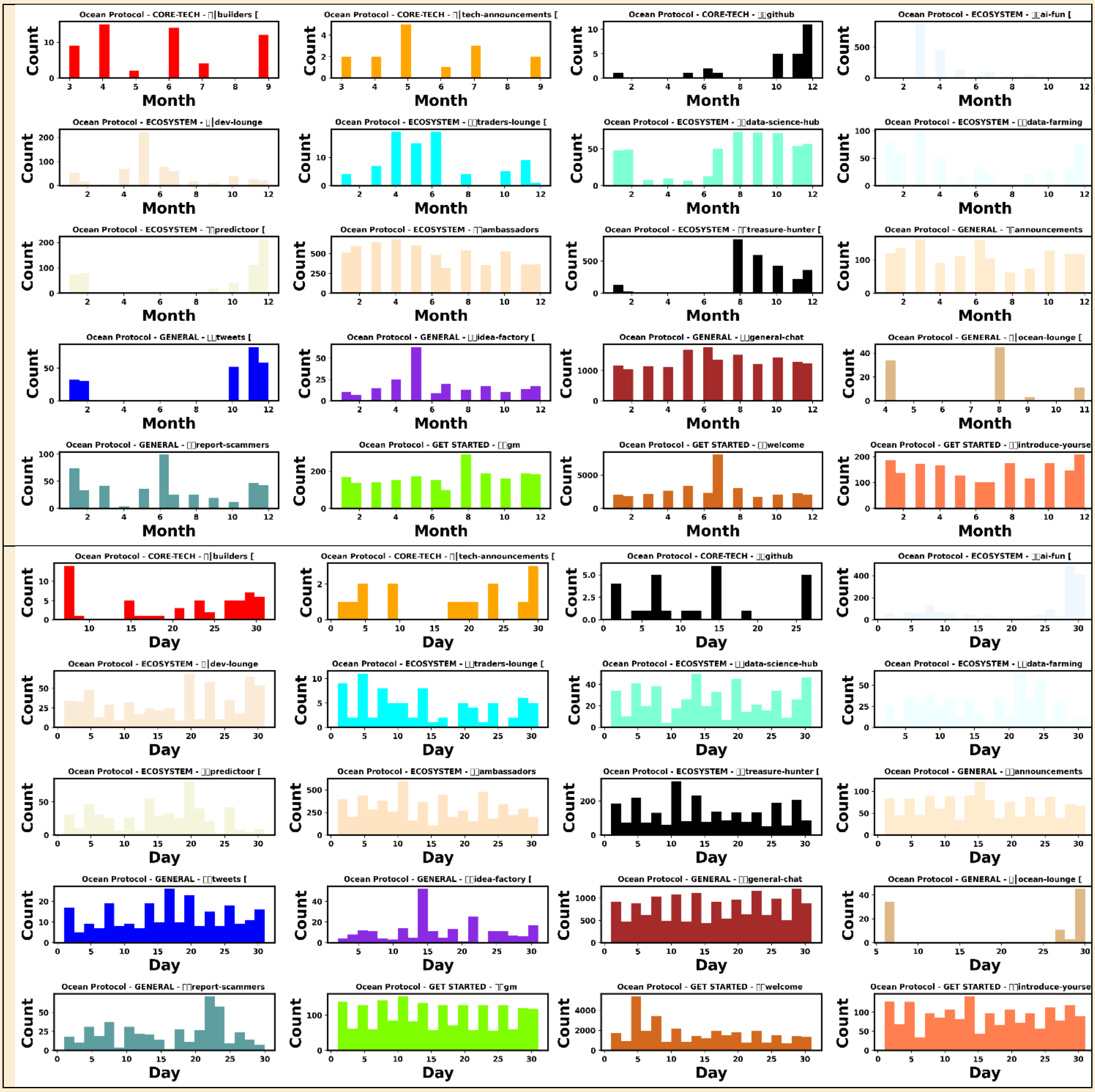
(6) ML model based on LSTM developed here performed well for price and Sentiment_value, and could serve as foundational basis in developing more complex model that predict the wider spectrum of server activities.

# Appendices

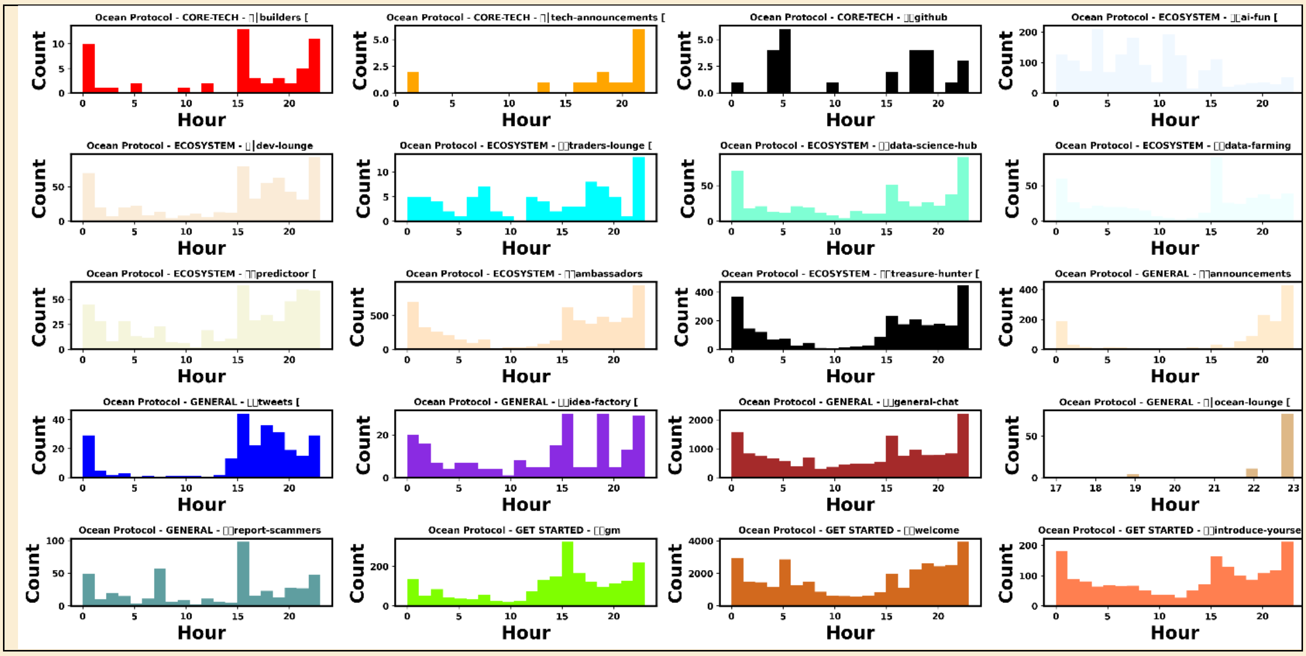
## Appendix A1

Full Data Distribution by Timescale and across all Channels.






Dominikus Brian 钟鸿盛 | domi@dreambrook.tech


Ocean Protocol - CORE-TECH - builders
Ocean Protocol - CORE-TECH - tech-announcements
Ocean Protocol - CORE-TECH - github
Ocean Protocol - ECOSYSTEM - ai-fun
Ocean Protocol - ECOSYSTEM - dev-lounge
Ocean Protocol - ECOSYSTEM - traders-lounge
Ocean Protocol - ECOSYSTEM - data-science-hub
Ocean Protocol - ECOSYSTEM - data-farming
Ocean Protocol - ECOSYSTEM - predictoor
Ocean Protocol - ECOSYSTEM - ambassadors
Ocean Protocol - ECOSYSTEM - treasure-hunter
Ocean Protocol - GENERAL - announcements
Ocean Protocol - GENERAL - tweets
Ocean Protocol - GENERAL - idea-factory
Ocean Protocol - GENERAL - general-chat
Ocean Protocol - GENERAL - ocean-lounge
Ocean Protocol - GENERAL - report-scammers
Ocean Protocol - GET STARTED - gm
Ocean Protocol - GET STARTED - welcome
Ocean Protocol - GET STARTED - introduce-yourse
Count
Hour

# Appendix A2

| Author | Sentiment Score | Total Activities | First Post | Last Post | Days in Server | Channels Participation | Average Words per Post |
|---|---|---|---|---|---|---|---|
| innopreneur | 0.147961 | 290 | 2019-09-15 18:50:00 | 2024-01-31 01:08:00 | 1598 | 7 | 24.779310 |
| Deleted User | 0.006000 | 5853 | 2019-08-28 22:52:00 | 2024-01-04 18:11:00 | 1589 | 12 | 4.010080 |
| alexcos20 | 0.083561 | 371 | 2019-11-13 18:55:00 | 2024-02-06 14:23:00 | 1545 | 9 | 12.471698 |
| sheridan_oceanprotocol | 0.402396 | 8 | 2019-08-20 19:55:00 | 2023-10-26 18:00:00 | 1527 | 5 | 20.750000 |
| robinlehmann | 0.104900 | 856 | 2020-05-26 00:01:00 | 2024-02-08 18:55:00 | 1353 | 14 | 26.308411 |
| MEE6#4876 | 0.654987 | 20986 | 2020-06-27 21:00:00 | 2024-02-20 17:13:00 | 1332 | 8 | 58.283189 |
| birususama | 0.168385 | 536 | 2020-07-20 17:30:00 | 2024-02-19 01:50:00 | 1308 | 9 | 28.177239 |
| okpo_e | 0.589375 | 4 | 2020-08-23 02:53:00 | 2024-01-18 15:09:00 | 1243 | 2 | 17.500000 |
| verahert | 0.057209 | 61 | 2019-12-27 01:03:00 | 2023-05-21 05:38:00 | 1241 | 3 | 10.131148 |
| blockchainlugano | 0.127692 | 1207 | 2020-12-21 03:58:00 | 2024-02-20 05:49:00 | 1156 | 11 | 30.737365 |
| trentmc0 | 0.081330 | 45 | 2020-06-03 19:47:00 | 2023-06-29 11:21:00 | 1120 | 4 | 32.111111 |
| jon77777 | 0.136965 | 37 | 2021-01-28 09:27:00 | 2024-02-20 11:44:00 | 1118 | 3 | 12.918919 |
| sarahsweetchilli | 0.144917 | 185 | 2020-10-10 09:24:00 | 2023-08-29 17:54:00 | 1053 | 5 | 22.378378 |

| Author | Sentiment Score | Total Activities | First Post | Last Post | Days in Server | Channels Participation | Average Words per Post |
|---|---|---|---|---|---|---|---|
| sheridan_oceanprotocol | 0.402396 | 8 | 2019-08-20 19:55:00 | 2023-10-26 18:00:00 | 1527 | 5 | 20.750000 |
| kremalicious | 0.137938 | 98 | 2019-08-20 22:22:00 | 2022-06-30 16:47:00 | 1044 | 1 | 53.193878 |
| brucepon. | 0.000000 | 2 | 2019-08-20 23:42:00 | 2021-11-05 20:53:00 | 807 | 2 | 3.000000 |

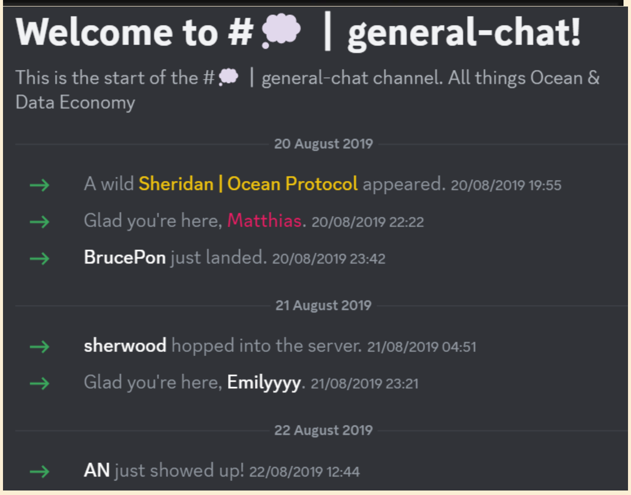



**Figure A2.1:** (left) Screenshot of Author Stats Data (right) 3 first member of Ocean Protocol Discord Channel. Nostalgia Trivia: First message in the server is by User Fermy#6766 saying "Seeping up the first message!" Noteworthy, Bruce Pon is Founder of Ocean Protocol!.

## Top Author Activities by Channels

Channel Name : Ocean Protocol - ECOSYSTEM - | ambassadors

| Author | Activities Count |
|---|---|
| blockchainlugano | 806 |
| zippy1979 | 775 |
| doteth | 343 |
| lrt8723 | 244 |
| MEE6#4876 | 213 |

Channel Name : Ocean Protocol - ECOSYSTEM - | treasure-hunter

| Author | Activities Count |
|---|---|
| kreigdk | 311 |
| blockchainlugano | 263 |
| feno1486 | 139 |
| abdulmazid_ | 121 |
| georgegem | 120 |

Channel Name : Ocean Protocol - GENERAL - 🎉 | announcements

| Author | Activities Count |
|---|---|
| dotunwilfred.eth | 646 |
| Ocean Protocol • TweetShift | 250 |
| birususama | 244 |
| bhavingala | 56 |
| kuerax | 50 |

Channel Name : Ocean Protocol - GENERAL - | tweets

| Author | Activities Count |
|---|---|
| MEE6#4876 | 254 |

Channel Name : Ocean Protocol - GENERAL - 💡 | idea-factory

| Author | Activities Count |
|---|---|
| lukas85#6738 | 48 |
| white_rider_ | 17 |
| kreigdk | 10 |
| nickscavuzzo33 | 10 |
| doteth | 10 |

Channel Name : Ocean Protocol - GENERAL - 💭 | general-chat

| Author | Activities Count |
|---|---|
| bhavingala | 763 |
| kreigdk | 528 |
| robinlehmann | 481 |
| bigimeyagazzz | 475 |
| dotunwilfred.eth | 458 |

Channel Name : Ocean Protocol - GENERAL - |ocean-lounge

| Author | Activities Count |
|---|---|
| cesar_cripto | 16 |
| blockchainlugano | 12 |
| himas. | 10 |
| kreigdk | 10 |
| richard_bisong | 8 |

Channel Name : Ocean Protocol - GENERAL - 🚨 | report-scammers

| Author | Activities Count |
|---|---|
| robinlehmann | 43 |
| ttk100#9960 | 37 |
| eth_pooly | 34 |
| white_rider_ | 24 |
| zippy1979 | 20 |

Channel Name : Ocean Protocol - GET STARTED - | gm

| Author | Activities Count |
|---|---|
| qazim | 193 |
| bhavingala | 172 |
| mantisclone | 124 |
| chefcryptochris | 94 |
| kitblake | 83 |

Channel Name : Ocean Protocol - GET STARTED - 👋 | welcome

| Author | Activities Count |
|---|---|
| MEE6#4876 | 20486 |
| Deleted User | 5330 |
| web3ocean | 154 |
| gareth0411 | 56 |
| doteth | 20 |

Channel Name : Ocean Protocol - GET STARTED - | introduce-yourself

| Author | Activities Count |
|---|---|
| denkobetona | 114 |
| igork_ldn | 64 |
| robinlehmann | 54 |
| white_rider_ | 45 |
| QueenPat#9259 | 44 |

Channel Name : Ocean Protocol - GET STARTED - | ask-the-ai

| Author | Activities Count |
|---|---|
| OceanGPT#0740 | 254 |
| lukas85#6738 | 133 |
| swish24_08 | 89 |
| white_rider_ | 34 |
| awesomeQA#0813 | 29 |

**Figure A2.2** Top Author Activities by Channels.