Klima DAO

# TWITTER ANALYTICS REPORT

| Submitted By: | Subject: Data Science | Date Submitted: |
| --- | --- | --- |
| Ruiz Rivera | | Monday, April 4th, 2022 |

## INTRODUCTION

Since its launch in August 2021, Klima DAO has been on a mission to democratize climate action by incentive demand for carbon assets.

In the eight months since the project's launch, its users and stakeholders have collaborated to trap 12.9 *million* tonnes of carbon and continue to help bootstrap the Regenerative Finance (ReFi) ecosystem.

As a climate-tech startup in its infancy, it's vastly important for the DAO to spread awareness and educate people about its brand. In essence, there is a need to monitor the brand's health on Twitter, one of the largest platforms for public discourse, as it continues building itself as a mainstay in the public's consciousness.

Concurrently, more participation in the ReFi ecosystem directly equates to increasing Klima's likelihood of succeeding years out into the future.

*And with that in mind, we'll emphasize that our* ***business goal for this project*** *is to leverage Machine Learning to find and recommend actionable insights that might better optimize brand visibility in the Twittersphere.* The **target variable** for this study will be the number of retweets that a tweet receives since **retweets** are excellent for multiplying a brand's visibility amongst broader networks of users. There are two aspects of retweets that this project will hone in on:

1. The factors that contribute most to whether or not a tweet gets retweeted and...
2. The degree to which a tweet goes viral through retweets.

.

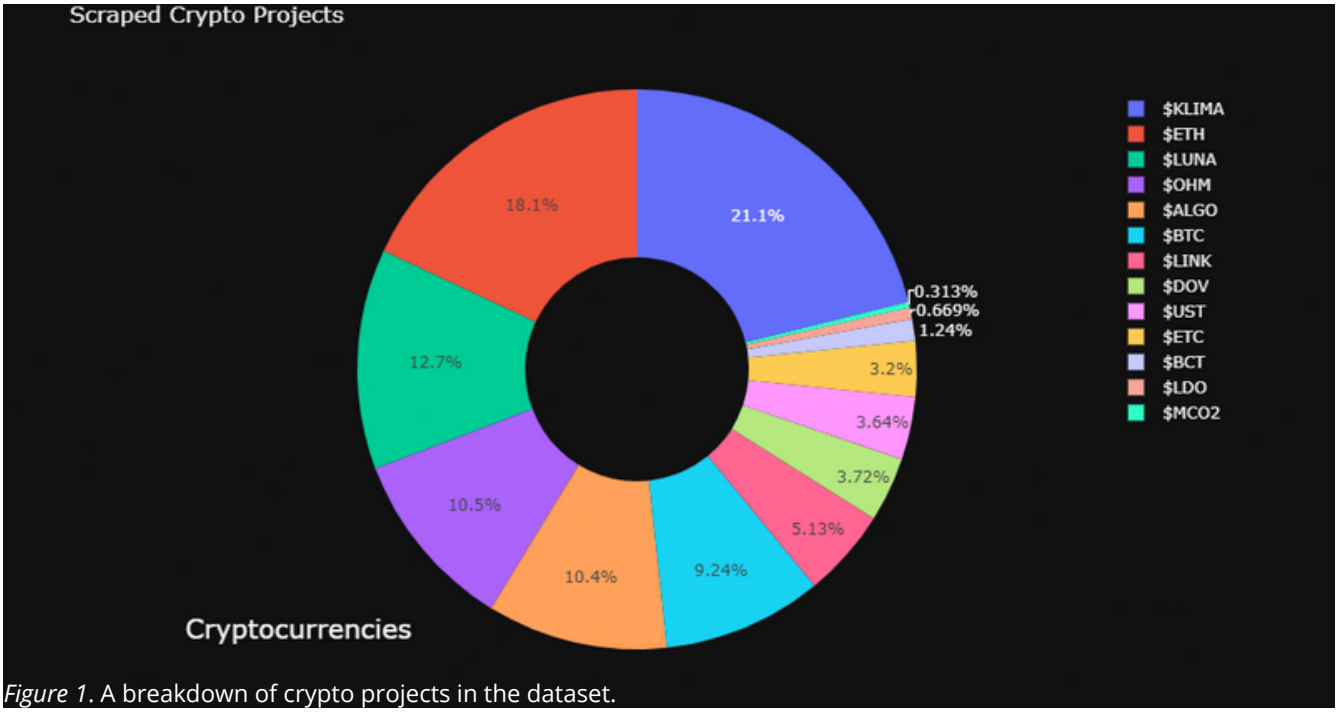



*Figure 1*. A breakdown of crypto projects in the dataset.

Splitting our research question into our two subcomponents is an excellent opportunity to apply binary and multi-class NLP/Machine Learning classification techniques to the problem set. In addition, our findings may also help extend our understanding of the social and algorithmic constructs governing social media platforms like Twitter or Reddit.

Since early February 2022, we've been scraping tweets weekly through Twitter's API to build a new and relevant dataset. In total, we've managed to collect roughly 55,000 tweets relating to Klima DAO and other prominent crypto projects, such as Ethereum, Bitcoin, and Terra. Please see Figure 1 for a breakdown of different cryptocurrencies captured in the dataset. And the bulk of the data cleaning process was spent on simple tasks such as deleting unnecessary columns, duplicate rows, processing text data, and correcting for multicollinearity. During this stage, the main pain points were extracting the rich information within the nested dictionaries of specific columns and feature engineering them so they could be key features in our predictive models.

Our feature engineering also involved employing a VADER sentiment analyzer to integrate a tweet's sentiment as one of our predictive features. We also developed new features by aggregating information from the existing columns. For example, "potential_reach" was a metric we developed in lieu of page views to capture the total number of people who could have seen a tweet based on the combined follower count of the original author and the retweeter. We also created a feature to capture the total engagement that a tweet receives, based on likes and retweets, and another that mentions the number of user mentions in a tweet, among many other features. Lastly, we transformed our two target variables in this analysis stage by transforming retweet counts into a boolean value, denoted as "RT_binary," and into tiers to categorize how viral a tweet became, as represented by our "retweet_tiers" feature.

In our exploratory data analysis phase, we investigated high-level trends in our dataset to better understand our predictive features.

.

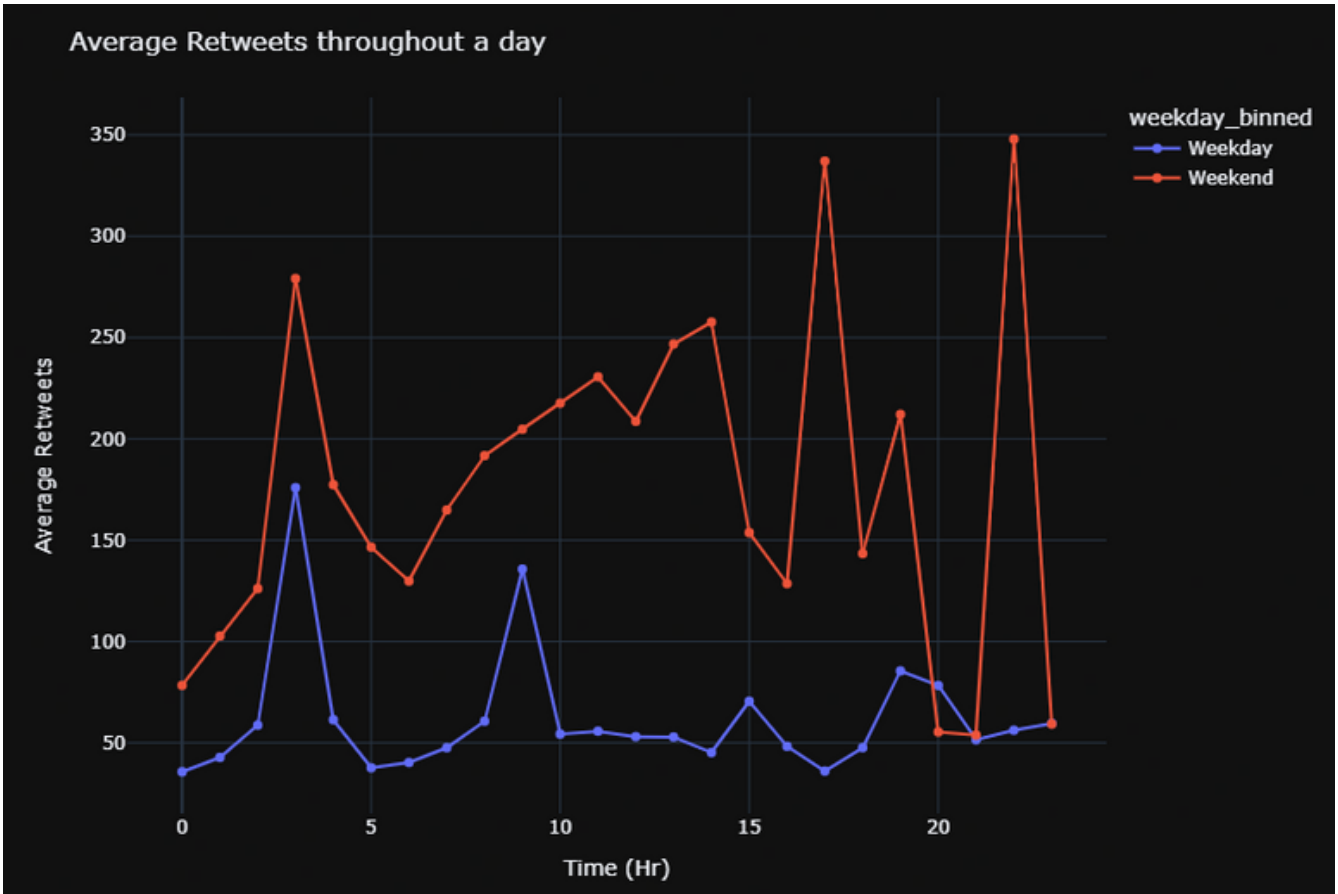



*Figure 2.* Crypto-Twitter hotspots throughout the week

For example, in Figure 2, we found that crypto Twitter was generally much busier during the weekend than the weekdays and that there are also several hot time slots where we see spikes in activity throughout the week. Since its launch in August 2021, Klima DAO (decentralized autonomous organization) has been on a mission to democratize climate action by incentive demand for carbon assets. In the eight months since the project's launch, its users and stakeholders have collaborated to trap 12.9 million tonnes of carbon and continue to help bootstrap the Regenerative Finance (ReFi) ecosystem.

Also, our engineered “potential reach” metric was handy in helping us gauge the brand’s health in terms of the number of people who may have come across Klima tweets over time. For example, in Figure 3, we can track the impact (and the lulls) of specific marketing campaigns amongst the Twitter community by cross-referencing them with the date on the chart.

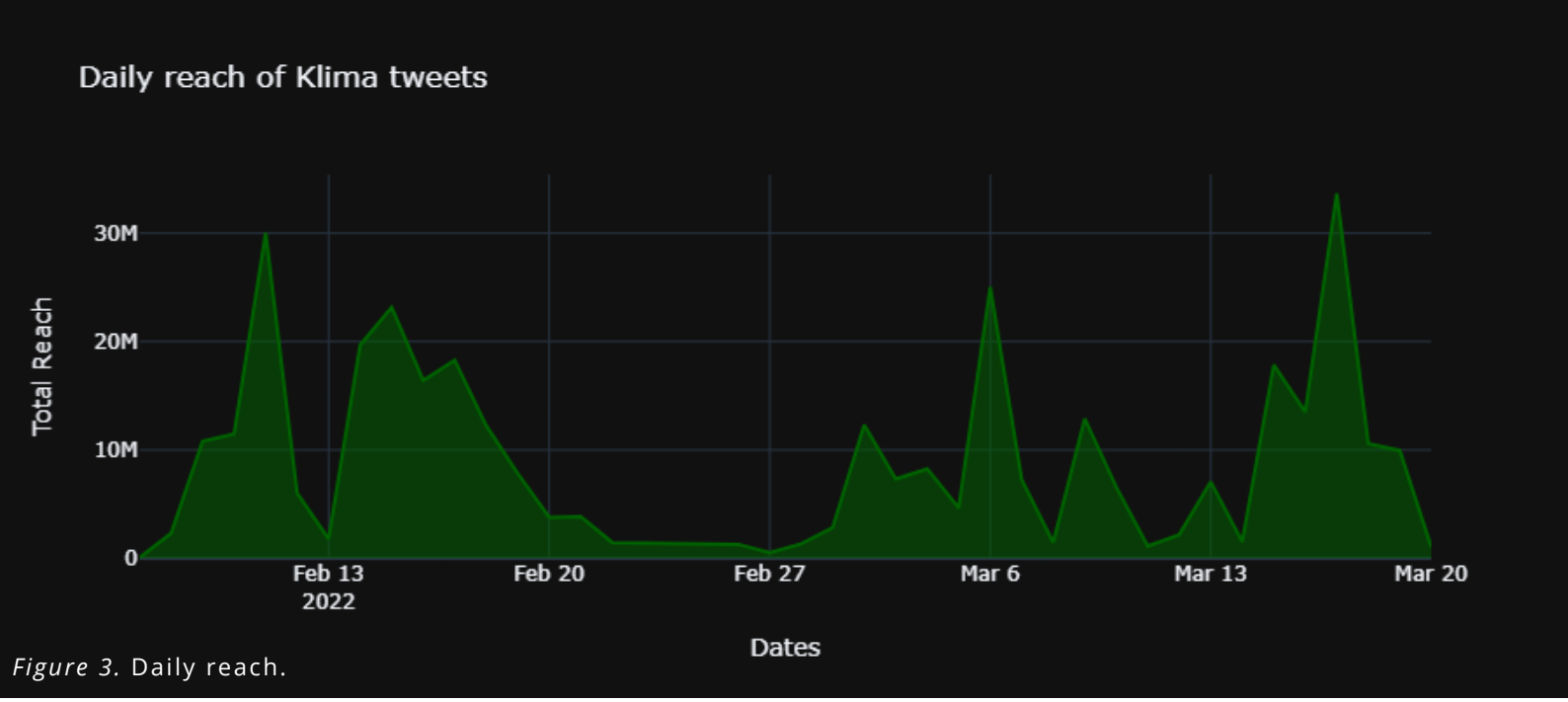
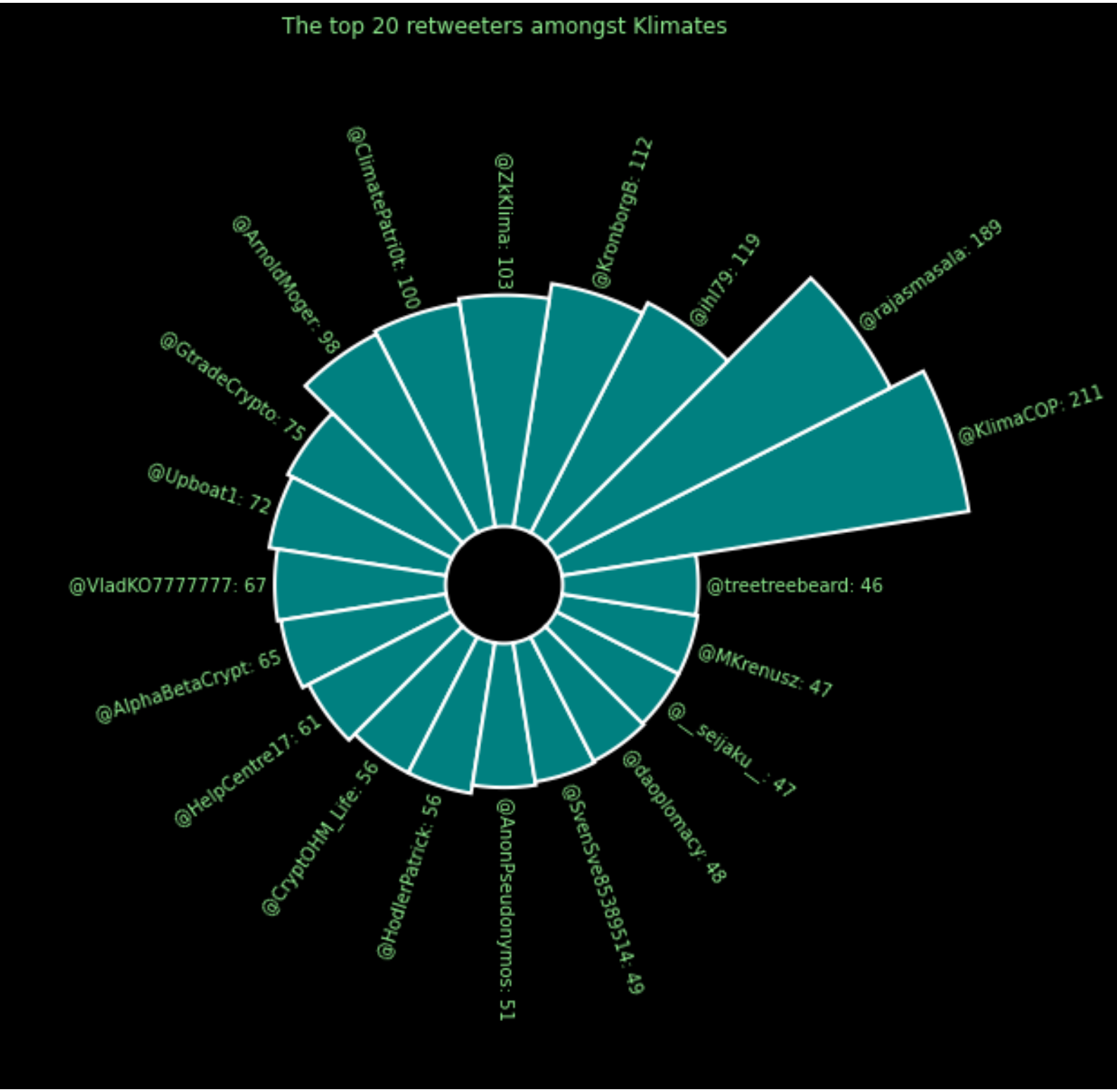



*Figure 3.* Daily reach.

Lastly, in Figure 4, we looked into the community members who were individually the biggest drivers of retweet rates and the Klima DAO brand. The insights gathered in the exploratory phase helped us build intuition for our predictive models' features.

Now that we've prepared our dataset and filtered out the features that are best suited for our predictive models, we can move on to the modelling phase. This involves splitting our dataset into three parts to avoid data leakage - Training, validation, and testing. Because we're interested in predicting retweet rates amongst Kima tweets, we decided to use it as our test set since it also makes up a portion of the total dataset that we would normally see anyways in a train-test split (21.1%). We also ran a 75/25 train-validate split on our non-Klima tweets for our predictive models to train on. As we did the splits and looked into the distributions of the target variable, we found some cause for concern in the spread of our retweet tiers. For example, if we examine Figure 5, we'll find that the distribution of posts retweeted 10-49 times vastly favours Klima tweets as it represents 30.6% of its total spread, while only 19.1% of non-Klima tweets are in this tier. In contrast, only 0.3% of Klima tweets ever reach 500+ retweets, whereas 9.2% of non-Klima tweets have achieved this mark so far. These significant differences between the training and the test may be a problem for us when running our multi-class predictive models.




For our binary target, the results are similar between both datasets. Our baseline for our binary modelling will be 62% which is the score one can get if they predict that every tweet gets retweeted. And for our multi-class models, the baseline target will be 30%.

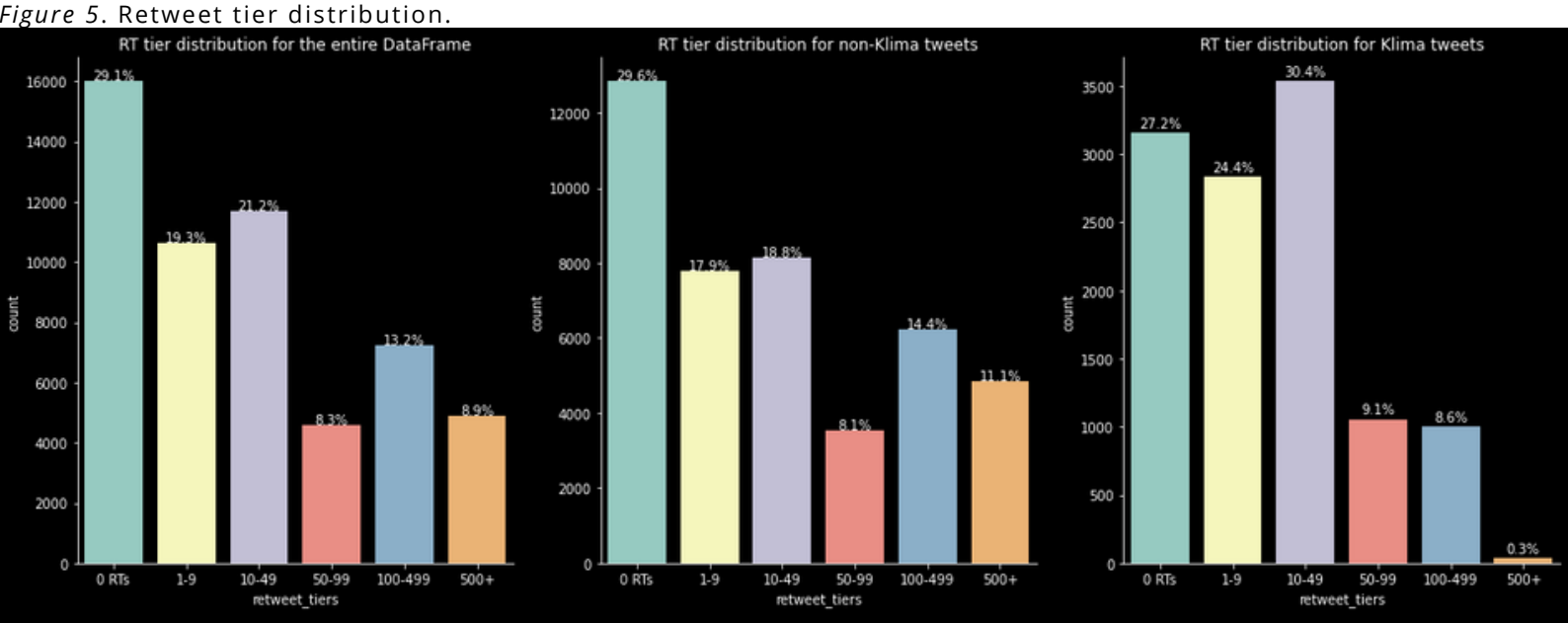
| Predictors | Model | Top Predictors | Accuracy (%) |
|---|---|---|---|
| num | Logistic Regression | [neutral, weekday, hour] | 65.14 |
| num | Random Forest | [likes, mentions, likes given, following] | 83.86 |
| num | Linear SVM | | 64.58 |
| num | Gradient Boosting | [likes, mentions, likes given, neutral] | 83.67 |
| num | XGBoost | [likes, mentions, following, likes given] | 83.56 |
| text | Naive Bayes | | 56.32 |
| num+text | Logistic Regression | | 56.26 |
| num+text | Random Forest | [likes, mentions, following, likes given] | 80.67 |
| num+text | Gradient Boosting | [likes, mentions, likes given, following] | 83.3 |

Amongst the Logistic Regression, Linear SVM, Naive Bayes, Random Forest, Gradient Boosting, and XGBoost models we fitted to predict the binary outcome of whether a Klima tweet will get retweeted, we found the most success with the latter three Ensemble models. Each of these models achieved a ~83% accuracy on the test set and were consistent about their top predictors being some combination of likes, and user mentions as a separate tier, followed by the number of profiles following and total likes given. What was noteworthy was that the text data hindered our model's preformance capabilities. We assume this is because the models had difficulty generalizing the climate change-related language associated with Klima tweets when only trained on crypto-centric text.

On the other hand, our Logistic Regression, kNN, Random Forest, and Artificial Neural Networks all struggled to predict the target variable amongst Klima tweets, partly because of the distribution differences we pointed out earlier in Figure 5. However, our Boosting Algorithms remained resilient against the adversity and predicted the retweet tiers with roughly 83% accuracy as well, with the same four predictors mentioned earlier being the biggest drivers of the model's decision-making.

.

*Figure 5*. Retweet tier distribution.



And so, with this wealth of insights - what were the key takeaways?

- Generate content that will drive engagement since the number of likes tends to drive up retweets
- Tag other users in your post that you feel are likely to retweet your content. These may be users you already have a connection with or community members who actively engage with others (See Figure 4).
- Engage with the community. Like, retweet, reply, and follow others to build a connection and be centred in the conversations. The more connections, the more likely your message gets amplified.
- Optimize your posting schedule. Post content during busy times in the network, such as during the weekend.

Although these principles can be generalized to any Twitter user, these results were derived by specifically testing against Klima-related tweets and may differ by community. With our research, our goal was to discover strategies that may help Klima DAO and its brand advocates grow its brand visibility and educate the public on the solutions that the DAO can offer.

*Thank you~*

## References


Grisel, O., Lars, B., & Chyi-Kei, Yau. (n.d.).Topic extraction with Non-negative Matrix Factorization and Latent Dirichlet Allocation. Sci-Kit Learn. https://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Kirenz, J. (2021, December 11). Text mining and sentiment analysis with NLTK and pandas in Python. Jan Kirenz. https://www.kirenz.com/post/2021-12-11-text-mining-and-sentiment-analysis-with-nltk-and-pandas-in-python/text-mining-and-sentiment-analysis-with-nltk-and-pandas-in-python/

Nesi, P., Pantaleo, G., Paoli, I., & Zaza, I. (2018). Assessing the reTweet proneness of tweets: Predictive models for retweeting. Multimed Tools Appl, 77. 26371-26396. https://link.springer.com/content/pdf/10.1007%2Fs11042-018-5865-0.pdf

Twitter. (n.d.). Data dictionary: Standard v1.1. Developer platform. https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet