**Problem statement:**

Identify the intent of tweet using one of the ML model.

**Solution:**

The process identifying the intent of tweet includes the following steps:

1. Importing the dataset

2. Explore the data and visualize the data

3. Preprocessing the text

4. Representation of the text

5. Building machine learning models

6. Evaluating the models

**1. Importing the Dataset:**

The dataset speaks about NFT (non-fungible tokens) .NFTs can represent real-world items like artwork and real estate.They can be sold and bought on NFT marketplaces. Promoting NFTs on social media platforms is a great way to get an audience and buyers for your digital art. The commonly used social media platform is twitter.

The given dataset has 120000 such tweets. In which 96364 tweets are labeled i.e., target column identifies the intent of the tweet – Community, giveaway, appreciation, whitelist, presale, donk, pinksale, interested and launchingsoon. 31089 tweets are unlabeled.

We use these 90000+ tweets as train dataset for training the model and remaining rows as test dataset where the task is to assign the intent of the tweet.

Pandas is used to import thedataset.

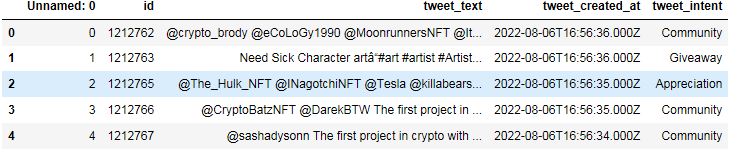


Fig: Train Dataset

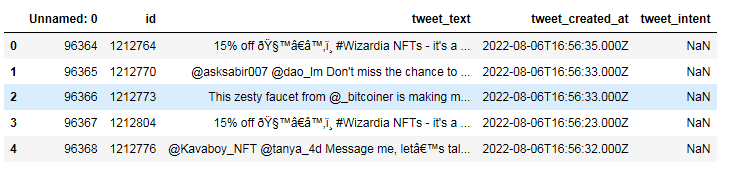


Fig: Test Dataset

**2.Exploratory Data Analysis (EDA):**

In our case the dataset have imbalanced data, the majority classes might be of our great interest. It is desirable to have a classifier that gives high prediction accuracy over the majority class, while maintaining reasonable accuracy for the minority classes. Therefore, leave it as it is.

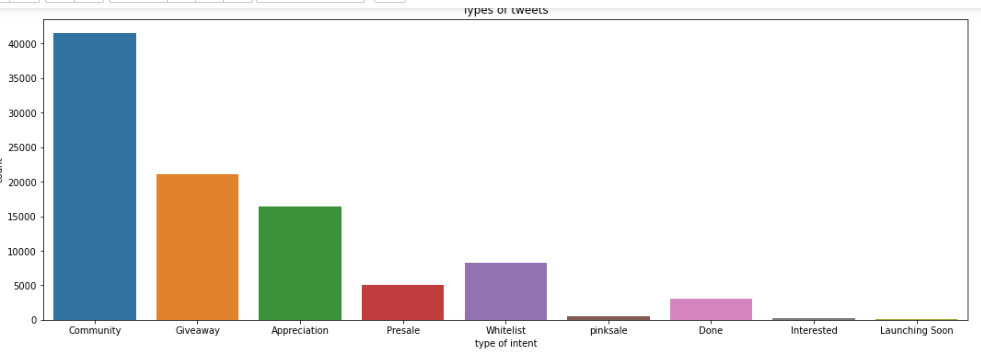
****

Fig: Target distribution

Seaborn library was used to visualize the above graph.

**Word Clouds:** The word clouds indicate the frequency of words in our dataset. They provide us with an overview of the data in our dataset. Below two word cloud are examples.



Fig: Visualization of all keywords associated with community



Fig: Visualization of all keywords associated with Appreciation

**3.Preprocessing the data:**

Then pre-processed the tweets involves the following:

**A.**

1**.** Make all the text into lowercase

2. Removal of URLs,

3. HTML Tags are being cleaned

4. Taking out punctuations

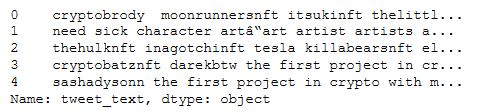


Fig: Cleaned text

 "re" module was used to do above task.

**B.** Tokenization: It basically refers to splitting up a larger body of text into smaller lines.

Used nltk.tokenize.RegerpTokenizer(r’\w+) – outputs all the words as tokens.

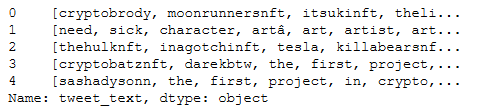


Fig: Tokenized text

**C.** Stopwords Elimination: To remove all common pronouns

Library - from nltk.corpus import stopwords

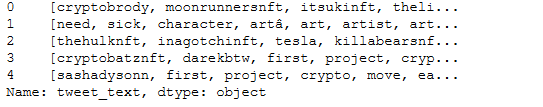


Fig: Text after removal of stopwords

After completing above steps, the text is combined, the preprocessed text looks like the below.

And the text is now ready to be transformed into vectors.

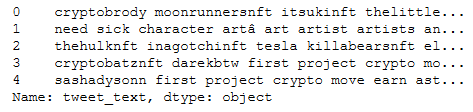


Fig: Preprocessed text

**4. Text Representation:**

**TF-IDF:** Tfidf Vectorizer is the base building block of many NLP pipelines. It is a simple technique to vectorizer text documents — i.e., transform sentences into arrays of numbers — and use them in subsequent tasks.

I have used sklearn.feature\_extraction.text.TfidfVectorizer to calculate a tf-idf vector for each of tweet text.

**5.Model Building**:

The following classification algorithms are used:

1. Logistic Regression,
2. Multinomial Naive Bayes,
3. Support Vector Machine and
4. Random Forest

To implement these classifiers, the Scikitlearn machine learning library was used.The libraries are:

from sklearn.linear\_model import LogisticRegression

from sklearn.naive\_bayes import MultinomialNB

from sklearn.svm import SVC

from sklearn.ensemble import RandomForestClassifier

**6.Evaluation:**

After preprocessing and transforming the text into a suitable format for the prediction model. We must assess the model's performance, or how well the model predicts the outcome. I have used accuracy performance metric to predict the performance.

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 0.95 |
| Multinomial Naive Bayes | 0.89 |
| Support Vector Machine | 0.95 |
| Random Forest | 0.95 |

Table: Comparing of results of different classification algorithm

**Conclusion:**

The aim of the project is multi-class text classification of tweet. Based on given text as an input, I have identify the intent of the tweet. I begin task with data analysis and data pre-processing from our dataset. And used Tfidf Vectorizer to transform sentences into arrays of numbers. Used different machine learning algorithms to get more accurate predictions and choose the most accurate one for our issue. The following classification algorithms have been used: Logistic Regression, Multinomial Naive Bayes, Support Vector Machine (SVM), Random Forest. Finally Random Forest was chosen to identify the tweet intent.