

## Progress Report Crypto Sentiment Analysis

For our team course project, we are currently making strides for the model architecture, data preprocessing, and planning to start working on the data visualization as soon as this step has been completed. We have lost a group member (Rachel Lee), but are still working and on track to get everything we're intending to do done. We have made the change to not make our crypto market cap sentiment truly live as this would probably require using a web-server, rather we're going to look at a delta T of time to compare our sentiment based on the changes of that time-span.

In regards to the model architecture and the work being done for it: we've begun designing the basic architecture to make use of Naive Bayes and Linear SVM from scikit learn. If time permits, we're hoping to experiment with neural models and incorporate time-series (in hopes to truly make our project live). Currently our implementation entails the model being tuned using a validation set, and may possibly apply a K-fold cross validation. Once our model is fully trained and the hyperparameters are tuned, we will pull tweets for a specified cryptocurrency over T time and sanitize the text for input into our model. We're currently comparing these deltas in general sentiment with the market cap of that coin to test correlation.

We've also started to make process on our pre-processing step where we've acquired several labeled datasets we will use to train our sentiment model. All our datasets are composed of tweets and have a sentiment label. Some are focused on general sentiment for a variety of popular cryptocurrencies. We're still working on sanitizing the datasets using NLTK libraries to remove punctuation and stopwords, as well as lemmatize different sense of words (specific to crypto polarity). From this we can experiment with different vocabulary sizes and test different representations of the text. We'll be working on fine tuning this so as to not omit words that may be of importance for our crypto vocabulary.

As of now we have not begun working with our visualizations, but we're intending to use some form of matplotlib still to visualize the model we'll train to give accurate depictions of polarity based on the vocabulary we've created of our datasets. We're hoping to also include some interesting metrics as well as model accuracy (if we get the time) in comparison to actual results pulled. As of now we want to include pie charts and graphs to describe polarity in comparison to multiple currencies, and how they relate from changes in the market with sentiment found from Twitter. Our goal is to give visualizations that will help the average person understand what this sentiment is breaking down, and why this information can be useful.