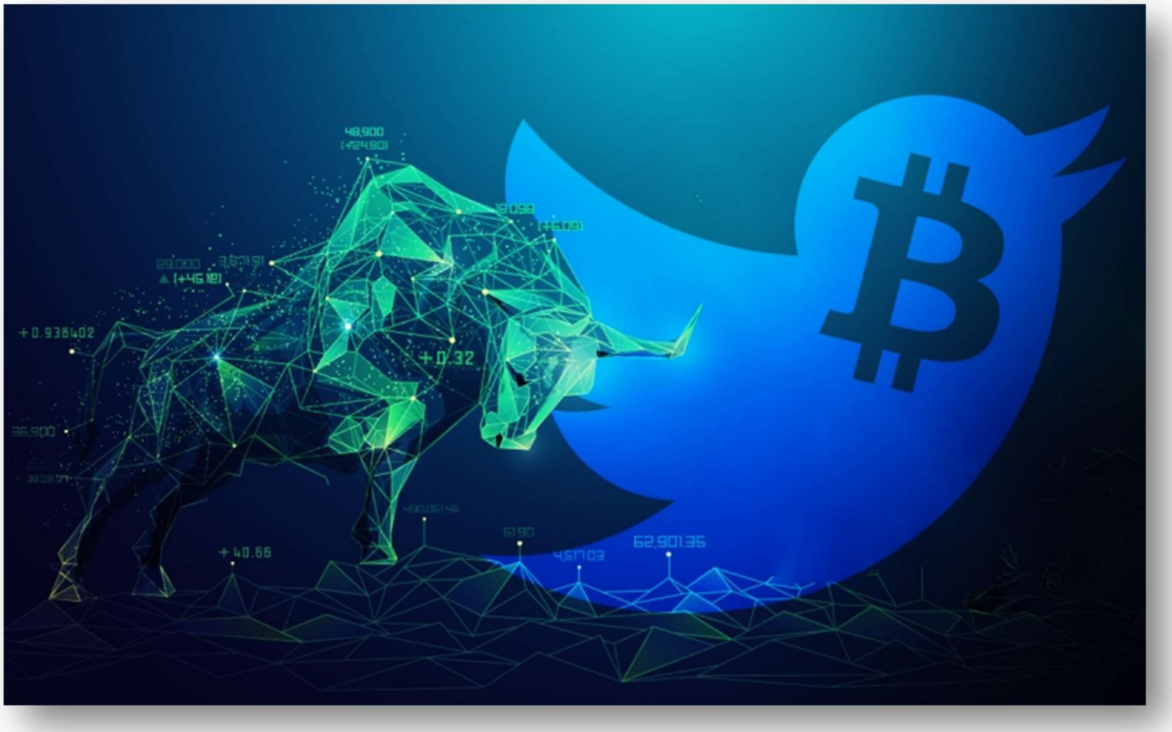
Technocolabs
BUILD WITH AI

TECHNOCOLBS DATA SCIENCE INTERNSHIP

MAJOR PROJECT REPORT



TITLE: Cryptocurrency (Bitcoin) Price Prediction Based on Twitter Sentiments.

You can now predict #BTC price by a single tweet. Don't wait!!

Hit a Tweet....

ABSTRACT:

Many research has shown that real-time Twitter data can be used to predict market movement of securities and other financial instruments. The goal of this project is to prove whether Twitter data relating to cryptocurrencies can be utilized to develop a model to predict the crypto coin price. By way of supervised machine learning techniques, our team will outline several machine learning pipelines with the objective of identifying cryptocurrency (bitcoin) market movement. Our approach to cleaning data and applying supervised learning algorithms such as logistic regression, Decision Tree Classifier and Linear Discriminant Analysis (LDA) to predict bitcoin price based on twitter sentiments with prediction accuracy exceeding 70%.

INTRODUCTION:

Cryptocurrency is an alternative medium of exchange consisting of numerous decentralized crypto coin types. Since its inception in 2009, the Bitcoin has become a digital commodity of interest as some believe the crypto coins' worth is comparable to that of traditional fiat currency. Our method for determining the price prediction whether the price will go up or down bases on sentiments of users involves correlating prices with one of today's most popular social media sources, Twitter. The advantages of using Twitter include having access to some of the earliest and fastest news updates in a concise format as well as being able to extract data from this social media platform with relative ease. Our model strategy applies supervised machine learning algorithms including logistic regression, Decision Tree Classifier and Linear Discriminant Analysis (LDA) to determine whether the price of BTC (digital currency) will increase or decrease within a predetermined time interval and will also shows the sentiment of the tweet entered by user.

The two approaches for training the model involves using direct text, like tweets from Twitter users and using third party open-source sentiment analysis APIs to rate the positivity and negativity of words within each post.

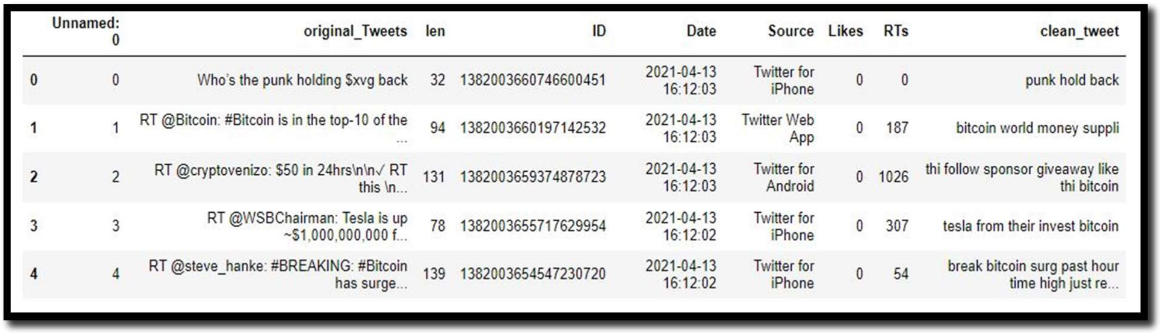
DATA:

In order to create model for the learning algorithms we utilize Tweepy - an open-source Python library for accessing the Twitter API. The keyword, bitcoin, is searched in real time and tweets containing this token is placed into a text file. Additional data being collected for each post containing the keyword includes the user ID, a unique identifier which cannot be changed, and a time stamp. In addition, the prices of the cryptocurrency (bitcoin) is collected for last 22 hours with interval of 5 minutes via the Yahoo Finance API and placed into text files to create a price history.

While tweets are collected in real time, to clean the data, the following procedure is carried out.

- The first step is to remove all non-alphabetic characters.
- The second step is to remove duplicates.
- Stop words are subsequently removed from tweets based on membership in the "stop words" corpus of the Natural Language Toolkit.
- The next step was stemming and lemmatization.

"The average length of the tweets collected was approximately 126. Whereas the maximum number of Re-Tweets encountered were 11002."



	Unnamed: 0	original_Tweets	len	ID	Date	Source	Likes	RTs	clean_tweet
0	0	Who's the punk holding $xvg back	32	1382003660746600451	2021-04-13 16:12:03	Twitter for iPhone	0	0	punk hold back
1	1	RT @Bitcoin: #Bitcoin is in the top-10 of the ...	94	1382003660197142532	2021-04-13 16:12:03	Twitter Web App	0	187	bitcoin world money suppli
2	2	RT @cryptovenizo: $50 in 24hrs\n\n✓ RT this \n...	131	1382003659374878723	2021-04-13 16:12:03	Twitter for Android	0	1026	thi follow sponsor giveaway like thi bitcoin
3	3	RT @WSBChairman: Tesla is up ~$1,000,000,000 f...	78	1382003655717629954	2021-04-13 16:12:02	Twitter for iPhone	0	307	tesla from their invest bitcoin
4	4	RT @steve_hanke: #BREAKING: #Bitcoin has surge...	139	1382003654547230720	2021-04-13 16:12:02	Twitter for iPhone	0	54	break bitcoin surg past hour time high just re...

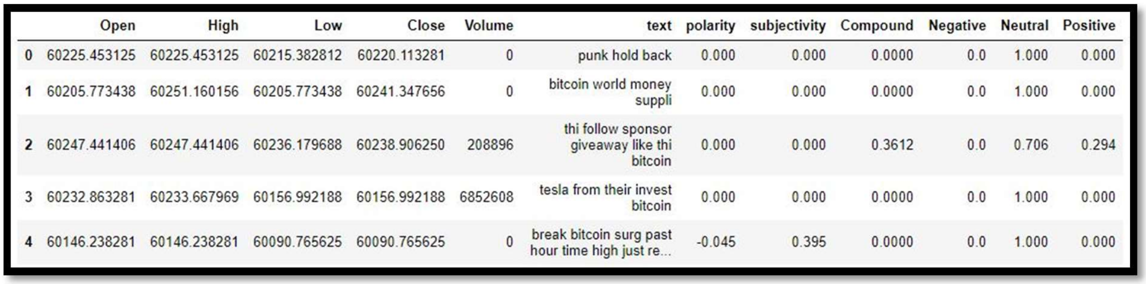
Fig1: *Shows a dataset where clear difference between [original_Tweets] and [clean_tweet] can be seen. With help of [nltk], [stemming] and [lemmatization] techniques able to clean all the collected tweets.*

The process then followed by **"Calculating Sentiment Polarity and Subjectivity".**

- The subjectivity shows how subjective or objective a statement is.
- The polarity shows how positive/negative the statement is, a value equal to 1 means the statement is positive, a value equal to 0 means the statement is neutral and a value of -1 means the statement is negative.

from textblob import TextBlob : performing NLP function to detect Polarity and Subjectivity.

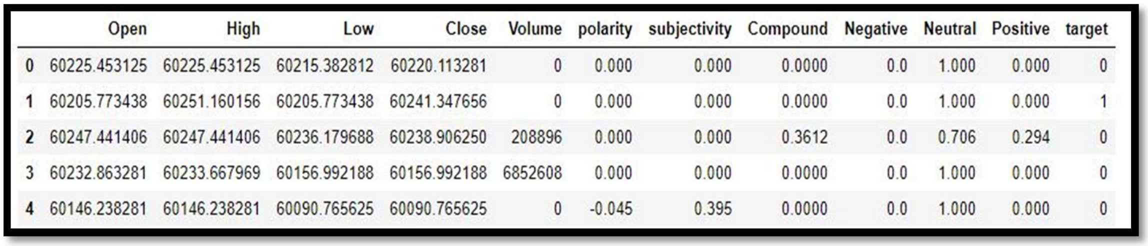
Later using **"Sentiment Intensity Analyser"** a function created to get sentiment scores i.e negative, positive, neutral and compound. Where the compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1(most extreme positive).



	Open	High	Low	Close	Volume	text	polarity	subjectivity	Compound	Negative	Neutral	Positive
0	60225.453125	60225.453125	60215.382812	60220.113281	0	punk hold back	0.000	0.000	0.0000	0.0	1.000	0.000
1	60205.773438	60251.160156	60205.773438	60241.347656	0	bitcoin world money suppli	0.000	0.000	0.0000	0.0	1.000	0.000
2	60247.441406	60247.441406	60236.179688	60238.906250	208896	thi follow sponsor giveaway like thi bitcoin	0.000	0.000	0.3612	0.0	0.706	0.294
3	60232.863281	60233.667969	60156.992188	60156.992188	6852608	tesla from their invest bitcoin	0.000	0.000	0.0000	0.0	1.000	0.000
4	60146.238281	60146.238281	60090.765625	60090.765625	0	break bitcoin surg past hour time high just re...	-0.045	0.395	0.0000	0.0	1.000	0.000

Fig2: *Shows a dataset where can be seen the sentiments (polarity & subjectivity) of each tweet has been identified.*

Both the datasets Tweet Dataset and Price Dataset merged together and a "target" column has been made by first identifying the price index i.e. (Latest Closing price – Last Closing price). If the difference of price index is negative the target column will get value of [0] which means Price Down and if the difference of price index result as positive then the target column will get value as [1] which means Price Up.



	Open	High	Low	Close	Volume	polarity	subjectivity	Compound	Negative	Neutral	Positive	target
0	60225.453125	60225.453125	60215.382812	60220.113281	0	0.000	0.000	0.0000	0.0	1.000	0.000	0
1	60205.773438	60251.160156	60205.773438	60241.347656	0	0.000	0.000	0.0000	0.0	1.000	0.000	1
2	60247.441406	60247.441406	60236.179688	60238.906250	208896	0.000	0.000	0.3612	0.0	0.706	0.294	0
3	60232.863281	60233.667969	60156.992188	60156.992188	6852608	0.000	0.000	0.0000	0.0	1.000	0.000	0
4	60146.238281	60146.238281	60090.765625	60090.765625	0	-0.045	0.395	0.0000	0.0	1.000	0.000	0

Fig3: *Shows a dataset where can be seen the sentiments (polarity & subjectivity) of each tweet has been identified as well a target column has been introduced with two values [0] and [1].*

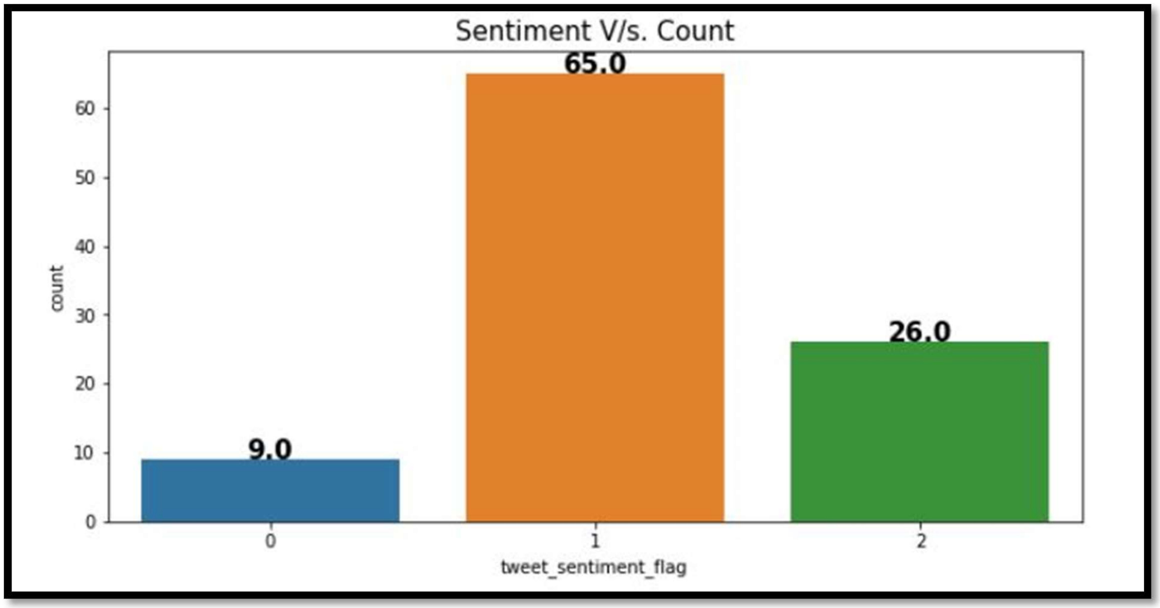


Fig4: *The Bar Chart shows, Negative Sentiment Count: 9.0, Neutral Sentiment Count: 65.0, Positive Sentiment Count: 26.0.*

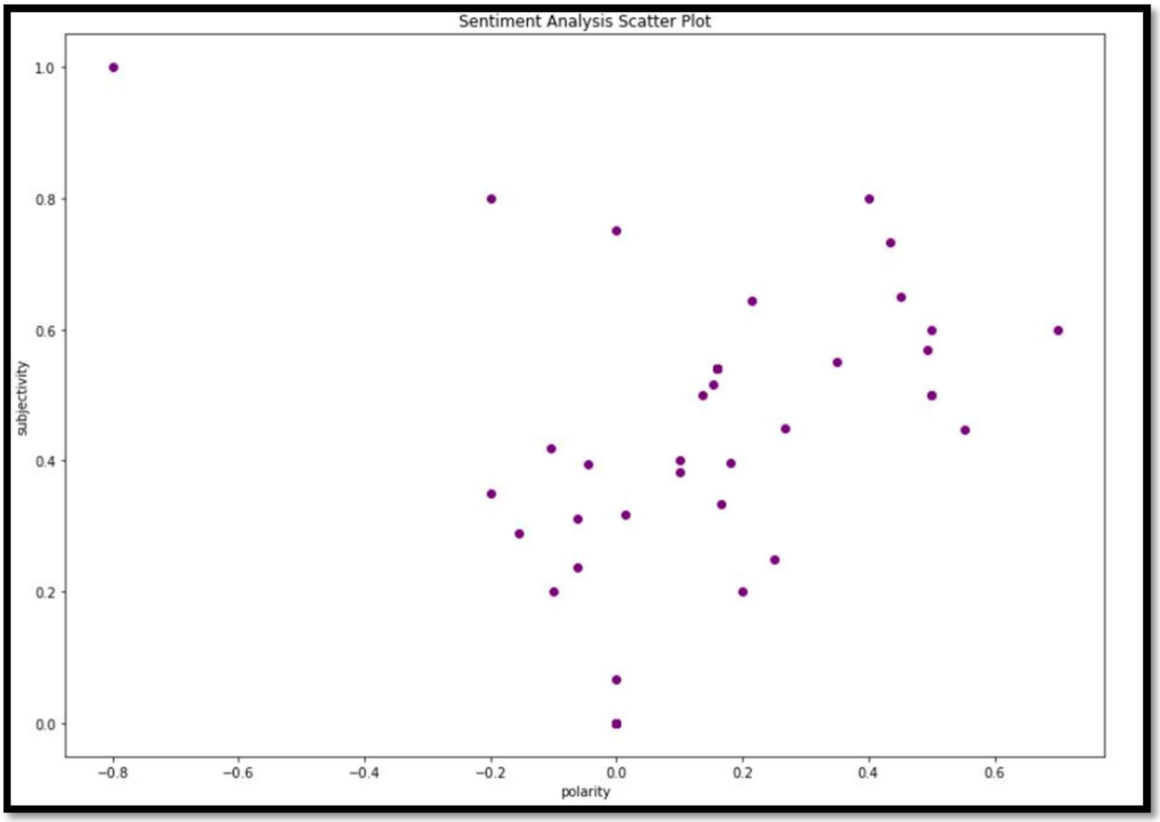


Fig5: *The Scatter Plot shows that more of the sentiments captured from the tweets are titled towards positive with more opinion related tweets rather than factual.*

MODEL BUILDING:

from tpot import TPOTClassifier : TPOT is a python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming. It will automate the most tedious part of machine learning by intelligently exploring thousands of possible pipelines to find the best one for the data.

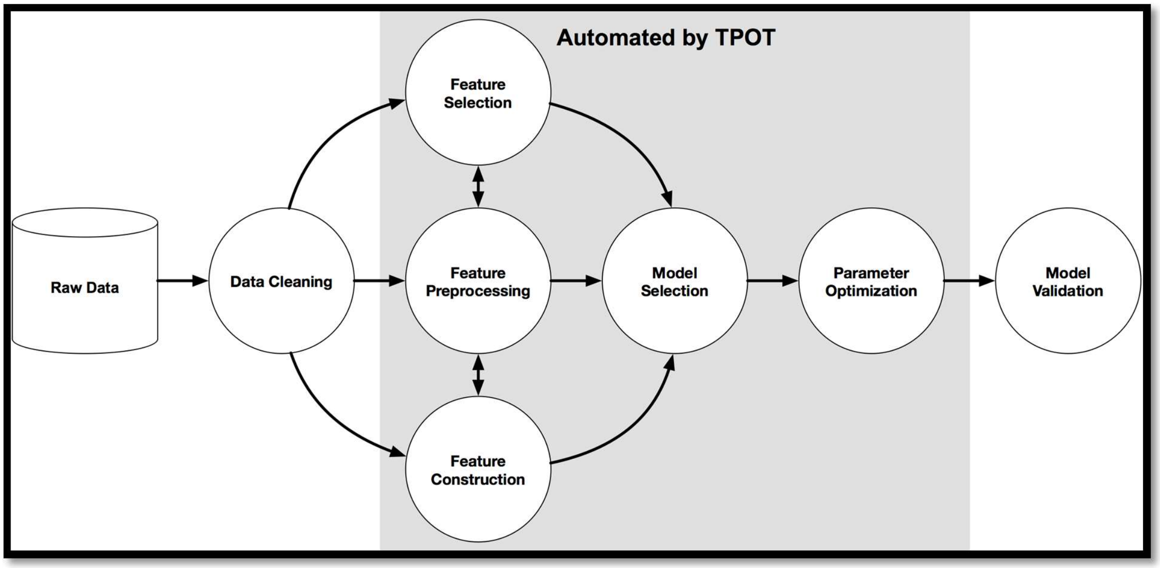
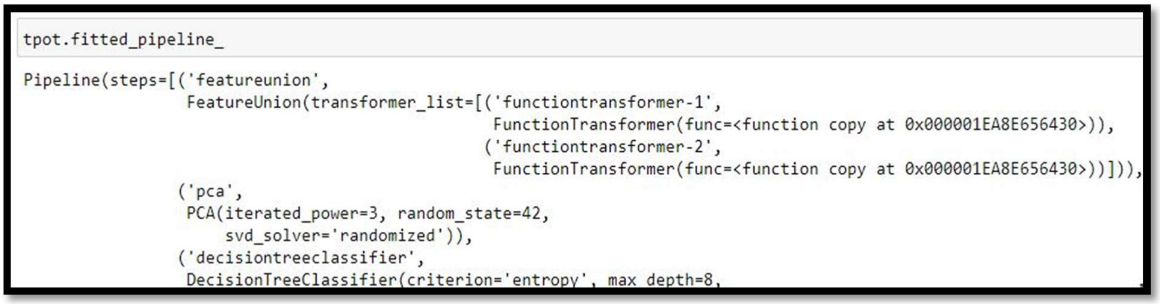


Fig6: *The TPOT provides us with the Python code for the best pipeline it found.*



```
tpot.fitted_pipeline_

Pipeline(steps=[('featureunion',
                 FeatureUnion(transformer_list=[('functiontransformer-1',
                                                 FunctionTransformer(func=<function copy at 0x000001EA8E656430>)),
                                                ('functiontransformer-2',
                                                 FunctionTransformer(func=<function copy at 0x000001EA8E656430>))])),
                ('pca',
                 PCA(iterated_power=3, random_state=42,
                     svd_solver='randomized')),
                ('decisiontreeclassifier',
                 DecisionTreeClassifier(criterion='entropy', max_depth=8,
```

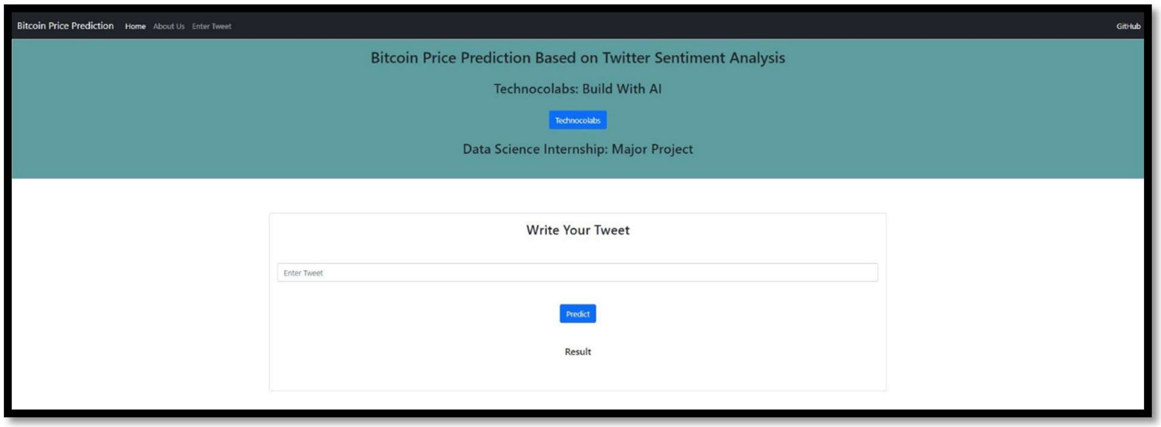
Fig7: *The TPOT provides 3 best fitted pipeline for our model to predict the bitcoin price based on twitter sentiments.*

The Decision Tree Classifier Model gave the accuracy of 55%. The next best fit model given was PCA (Principal Component Analysis) but as our dataset is of supervised machine learning and PCA is used for unsupervised machine learning algorithms, i.e. PCA ignores class labels. Thereby the best alternative for PCA is LDA (Linear Discriminant Analysis) most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications best for supervised machine learning algorithms and make assumptions about normally distributed classes.

LDA gave the accuracy of 70%.

DEMO:

API Link: https://bitcoinpredictionapi.herokuapp.com/



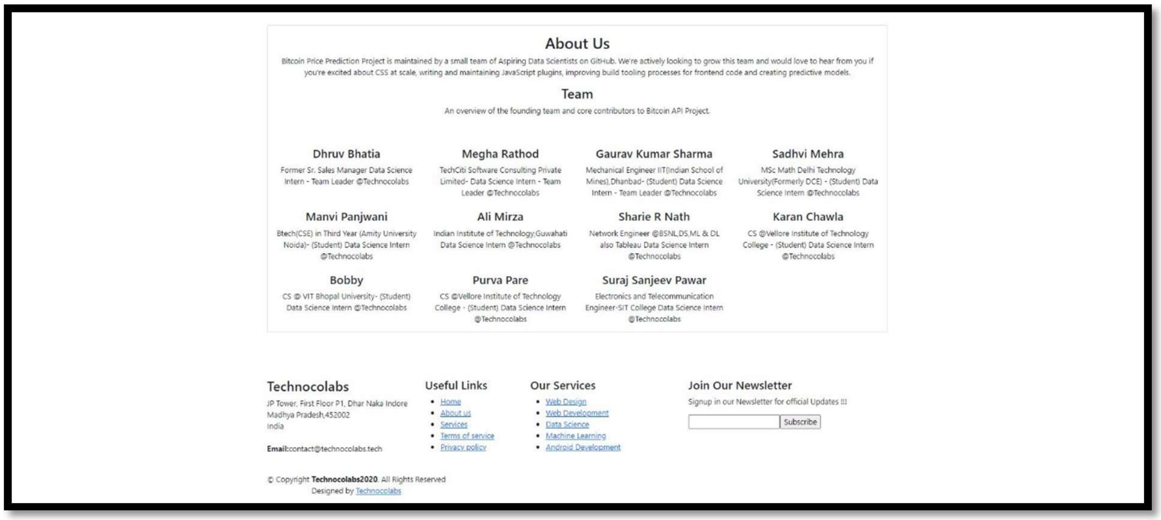
Bitcoin Price Prediction Based on Twitter Sentiment Analysis

Technocolabs: Build With AI

Data Science Internship: Major Project

Write Your Tweet

Result



About Us

Bitcoin Price Prediction Project is maintained by a small team of Aspiring Data Scientists on GitHub. We're actively looking to grow this team and would love to hear from you if you're excited about CSS at scale, writing and maintaining JavaScript plugins, improving build tooling processes for frontend code and creating predictive models.

Team

An overview of the founding team and core contributors to Bitcoin API Project.

Dhruv Bhatia
Former Sr. Sales Manager Data Science Intern - Team Leader @Technocolabs

Megha Rathod
TechCiti Software Consulting Private Limited- Data Science Intern - Team Leader @Technocolabs

Gaurav Kumar Sharma
Mechanical Engineer IIT(Indian School of Mines),Dhanbad- (Student) Data Science Intern - Team Leader @Technocolabs

Sadhvi Mehra
MSc Math Delhi Technology University(Formerly DCE) - (Student) Data Science Intern @Technocolabs

Manvi Panjwani
Btech(CSE) in Third Year (Amity University Noida)- (Student) Data Science Intern @Technocolabs

Ali Mirza
Indian Institute of Technology,Guwahati Data Science Intern @Technocolabs

Sharie R Nath
Network Engineer @BSNL,DS,ML & DL also Tableau Data Science Intern @Technocolabs

Karan Chawla
CS @Vellore Institute of Technology College - (Student) Data Science Intern @Technocolabs

Bobby
CS @ VIT Bhopal University- (Student) Data Science Intern @Technocolabs

Purva Pare
CS @Vellore Institute of Technology College - (Student) Data Science Intern @Technocolabs

Suraj Sanjeev Pawar
Electronics and Telecommunication Engineer-SIT College Data Science Intern @Technocolabs

Technocolabs
JP Tower, First Floor P1, Dhar Naka Indore
Madhya Pradesh,452002
India

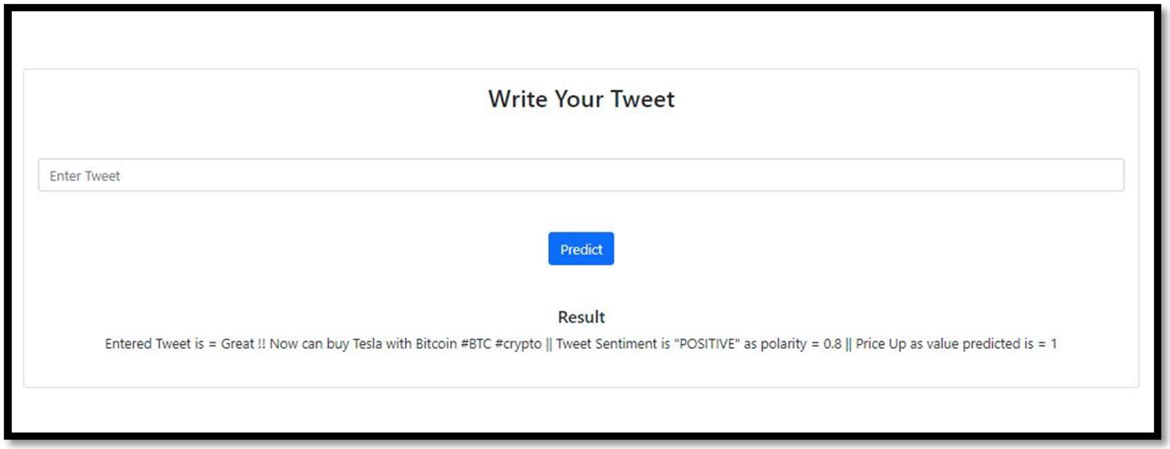
Email:contact@technocolabs.tech

Useful Links

Our Services

Join Our Newsletter
Signup in our Newsletter for official Updates !!!

© Copyright **Technocolabs2020**. All Rights Reserved
Designed by Technocolabs



Write Your Tweet

Result

Entered Tweet is = Great !! Now can buy Tesla with Bitcoin #BTC #crypto || Tweet Sentiment is "POSITIVE" as polarity = 0.8 || Price Up as value predicted is = 1

Fig8: *The API Overall Outlook. Frontend developed with the usage of HTML/CSS/JS/Bootstrap. The Result will not only show the predicted price in form of [0]: Price Down or [1]: Price Up but also shows the sentiment of tweet entered by user by explaining the tweet polarity in terms of [-1], [0] or [+1].*

OVERVIEW:

This is a Flask web app which predicts the bitcoin price based on the twitter sentiments.

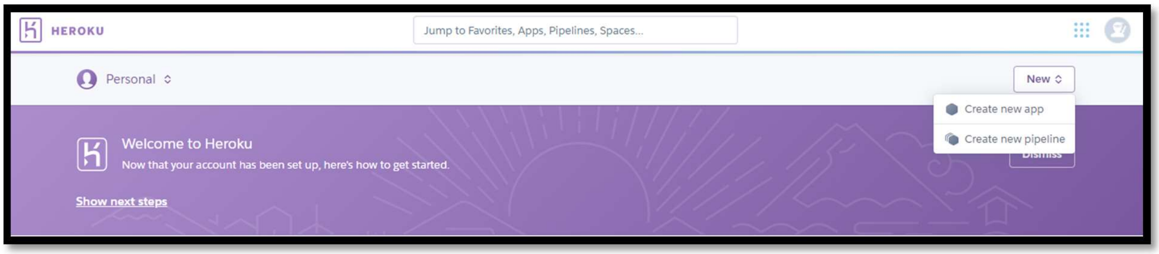
INSTALLATION:

The Code is written in Python 3.6.10. To install the required packages and libraries, run this command in the project directory after cloning the repository:

```
pip install -r requirements.txt
```

DEPLOYMENT ON HEROKU:

Login or signup in order to create virtual app. This can be done either connect GitHub profile or download ctl to manually deploy this project.



Welcome to Heroku

Now that your account has been set up, here's how to get started.

Fig9: *Heroku Login Page Overview: Next step would be to follow the instructions given on Heroku Documentation (https://devcenter.heroku.com/articles/getting-started-with-python) to deploy a web app.*

DIRECTORY TREE:

```
├── flask code
│    ├── static
│    │     ├── styles.css
│    │     ├── main.js
│    ├── templates
│    │     ├── about.html
│    │     ├── homepage.html
│    │     ├── layouts.html
│    │     ├── link.html
│    │     ├── tweet.html
│    ├── __init__.py
│    ├── routes.py
├── Procfile
├── README.md
├── run.py
├── bitcoin_price.ipynb
├── bitcoin.pkl
├── requirements.txt
```

TECHNOLOGIES USED:

MADE WITH PYTHON

Flask web development, one drop at a time gunicorn scikit learn

FUTURE WORK:

- In order to further improve the accuracy of the learning algorithms, additional research can be performed in the area of model accuracy.
- Creating a training set that is completely unskewed could result in lower classification error. In addition, we can formulate a set of words where each element has a high correlation with cryptocurrency (bitcoin) market movement and use this as a basis for training the learning algorithms.

TEAM MEMBERS:

- DHRUV BHATIA
- GAURAV KUMAR SHARMA
- PURVA PARE
- MEGHA RATHOD
- SURAJ SANJEEV PAWAR
- SADHVI MEHRA
- MANVI PANJWANI
- ALI MIRZA
- SHARIE R NATH
- KARAN CHAWLA
- BOBBY

****