Forecasting Bitcoin Price Using Twitter Sentiment Analysis

Author: Roozbeh Dadashzadeh

This project focuses on the relation between twitter sentiment and bitcoin price fluctuations. There has to be an unidentified correlation between the two of them as our studies suggest. Our approach toward this analysis is based on an article claiming that by gathering 2.25 million tweets they have achieved to 79% accuracy in predicting the direction of bitcoin changes(either increase or decrease). This study consists of 3 different steps:

1.Data collection:

Twitter Data Collection :

Tweets are extracted with a file called twitter_search.py which enables us to access twitter data for any given phrase or hashtag. The process of twitter data extraction has begun from 18th October and continuously is running on our server collecting day to day tweets published under #bitcoin.

For now, this script can handle twitter API limitation(50 reqeuest/15 minutes) and 9-day limitation but soon, we will need premium features of twitter API to collect historical data.

https://git.karoproject.com/roozbehdz/twitter-bitcoin-forecasting.git

Bitcoin Data Collection :

Bitcoin prices are gathered in a .csv file in periods of one minute, five minutes, one hour and one day using Bitmex API ranging from 26th of September 2015 up until now.

2.Preprocessing

This process just begins with converting .json file which we have access to from the server to the pandas data frame. Then we try to clean the data and eliminate bots and identifying influencers through the power of preprocessor and nltk libraries and this cleaned data is sent to be processed based on sentiments using VADER(MIT licensed). This library perfectly outperforms other libraries like bing and NRC.

3.Modeling

Time Series

Using sentiments calculated in the last step, we apply the time series algorithm. In this step, we use ARIMA modeling to predict bitcoin price firstly just considering the prices then considering sentiments and prices altogether.

Tweet volumes

It is mentioned in a paper worked by Abraham et al. that the volume of tweets related to cryptocurrencies also can give us insight about probable bitcoin price fluctuations.in this study, we try to measure this claim.

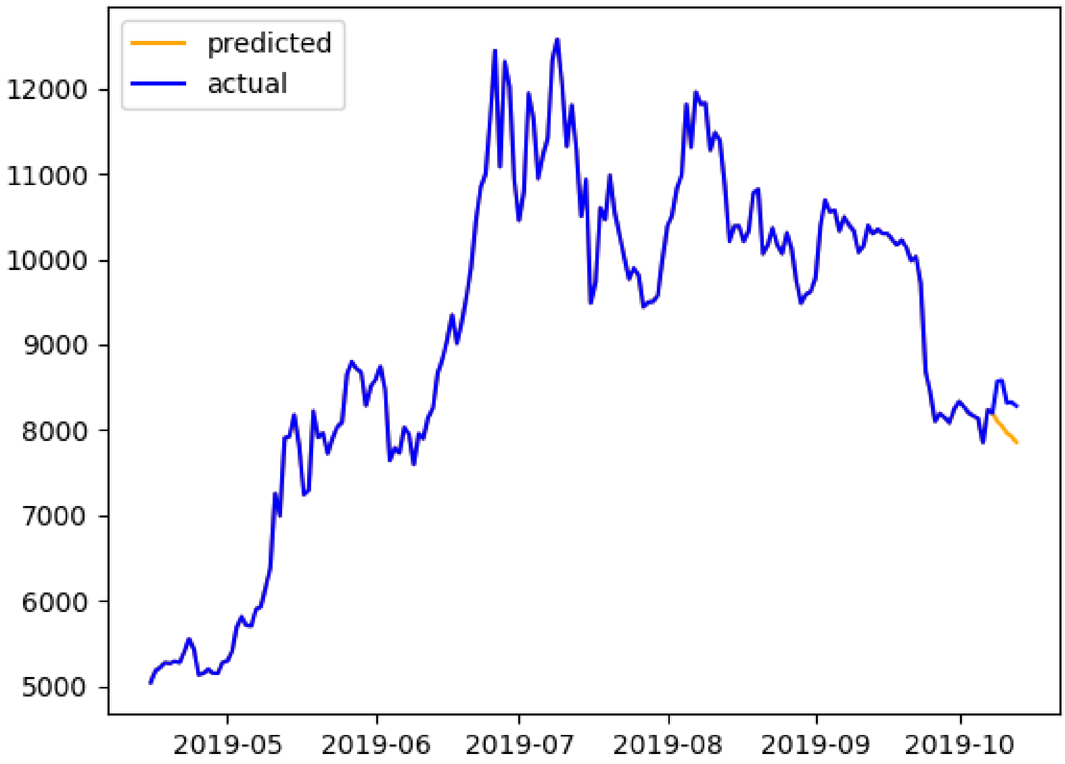
- Problems :

1. Bitcoin prices data tends to be non-stationary therefore it will be difficult to apply time series algorithms. However, adding the sentiments to the recipe can make it much more feasible.
2. Twitter has a specific limitation on the volume of tweets and requests we send to its servers. So we have to wait a long time to have access to the fresh data and this makes it difficult to use this platform for real-time predictions.
3. Another limitation of twitter would be this fact that it doesn't let us have access to tweets for more than 9 days so we have to wait a proper time to gather a rich set of data.

- Outputs :

1. Once, I tried to use sentiments and the prices to predict the upcoming days' prices using a test dataset which has a relatively low number of observations. I think due to the lack of big data, the model couldn't predict the exact numbers but could manage to find the right trend.

2. This result made me wonder what if I can predict the exam numbers with the help of relatively big data? I started collecting data per minute, per hour and per day. Ultimately, I could rich to 44-49 percent accuracy which I find useless because of the accuracy alongside the precision matters.



- To what extent am I familiar with machine learning algorithms and clustering?

Actually, I started learning data science with the acquisition of the fundamentals of supervised learning. I have learned linear regression models so far. Then I realized that to have a better performance in my bachelor's project, I have to learn data mining especially text mining. In the process of getting to know the natural language processing I became familiar with the concept of K-means clustering and I used this method to categorizing Amazon book reviews in a range of zero to 5 stars. Nowadays I'm watching an online Udacity course and attending Dr. Teresa Data science classes held by Tosee Institue.