XRP Price Prediction using NLP and LSTM Network Model

[Document subtitle]

Jayden Dzierbicki

2023

Contents

[Git-Repository 2](#_Toc132395002)

[Overview 2](#_Toc132395003)

[WebCrawler 2](#_Toc132395004)

[Data Wrangling 2](#_Toc132395005)

[Machine Learning 2](#_Toc132395006)

[Conclusion & Lessons Learned 2](#_Toc132395007)

# Git-Repository

Documentation and code can be located in a Git-repository:

<https://github.com/jaydendzierbicki/webcrawl-long-short-term-memory-network>

# Introduction and Proposed Model

The rise of cryptocurrency and blockchain technology has piqued the interest of investors worldwide. Cryptocurrency is a digital currency that uses strong cryptography to verify and protect financial transactions. Ripple Labs is a technology company that developed the Ripple payment protocol and exchange network with a focus on crypto solutions for business and central banks, utilizing their XRP coin. However, XRP, like any other cryptocurrency, experiences significant price fluctuations, posing risks and uncertainties for investors of all levels. To aid investors in making informed decisions, market movement prediction systems have emerged. Traditional supervised learning algorithms have been used in the past to predict changes in cryptocurrency based on historical price data, but predicting price changes can be challenging due to the efficient market hypothesis, which states that the market always follows a haphazard pattern. As a result, investors often actively monitor discussion boards to try to elicit insights that could point to an incoming market movement. However, this task can be overwhelming, and data science can be utilized to develop an automated tool to assist investors. To tackle this challenge, we propose a model which employs natural language processing (NLP) sentiment analysis to forecast price movements within a specific timeframe using the long short-term memory (LSTM) model which has been used successfully in other applications such as predicting bitcoin and stock price (Vo, 2019; Lahmiri & Bekiros, 2019; Kang et al., 2022; Shahzad et al., 2021). The focus is on XRP, owing to its vibrant community, significant market presence, and ongoing developments in the SEC Vs Ripple case in the United States, which has gained widespread media attention. The model aims to encompass novice, intermediate, and experienced investors by collecting data from an array of sources, such as Yahoo Finance and Investing.com, using web crawling and scraping techniques. The data is preprocessed and converted to prepare it for the LSTM network, utilizing Word2Vec to allow the algorithm to comprehend the context of words within the text corpus. However, ethical and legal considerations must be taken into account when gathering and scraping data. By undertaking data gathering, preprocessing, and model development, the proposed model aspires to predict XRP price movements using sentiment analysis and LSTM networks, delivering valuable insights into cryptocurrency price fluctuations.

## Model Workflow

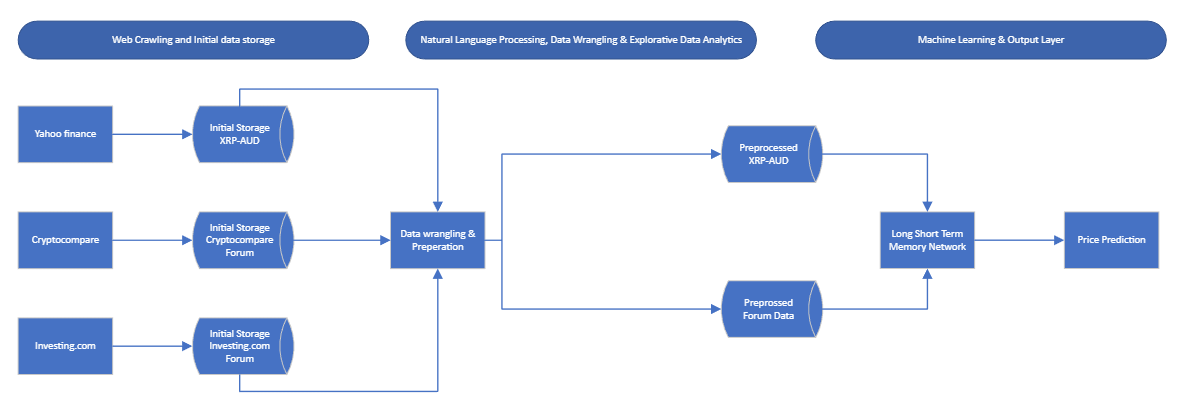
In Figure 1, we present an overview of the data pipelines we propose to achieve our objective. These pipelines comprise three key phases: web crawling and initial data storage, NLP combined with data wrangling and explorative data analytics (EDA), and machine learning along with the output layer. Throughout the report, we will discuss each of these phases in detail. Through utilizing a LSTM network model we hypothesis that we will be able to predict XRP price movement with some level of accuracy similar to what we have seen with bitcoin and Ethereum (Vo, 2019), allowing us to contribute further to this field of research.

Figure 1: A schematic representation of the pipeline, from left to right: Data is scraped from various sources and saved into landing tables in MySQL. The data undergoes wrangling, preparation and EDA before being stored in preprocessed tables in MySQL. The two input layers, consisting of preprocessed XRP-AUD data and preprocessed forum posts, are fed into a Long Short-Term Memory (LSTM) network for price prediction.

# Web crawling

Web crawling and scraping are terms that are often used interchangeably, but they represent distinct concepts and serve unique purposes. When combined, they can be powerful tools for data extraction and analysis. Web scraping involves extracting data from a specific website, where the target site is known, and the data is obtained by inspecting the HTML elements of the webpage (Perez, 2023). On the other hand, web crawling refers to the automated process of navigating the internet to download or index content from multiple websites or URLs. Many sites implement technologies, such as Cloudflare to prevent unauthorized crawling (Gillis, 2022). It is essential to review a website's terms of service before engaging in web crawling or scraping activities, as some sites may have specific rules or restrictions in place which could hinder the automation.

## Methodology

In each domain, we employed a similar extract-transform-load (ETL) process. Data was extracted from the target website and underwent minor transformations, such as adding an ID and splitting datetime into date and time components, before being loaded into a designated MySQL table. This method offers efficient data storage and management, particularly for large datasets. MySQL Workbench provides scalability, accommodating growing data volumes and seamless Python integration for further analysis. By leveraging these storage benefits and scalability, our data infrastructure is robust and ready for future expansion or the inclusion of new domains, acting as a central repository.

## Domains and workflow

The goal of this section is to collect data from three sources for our LSTM network model’s two primary input layers: one containing combined online forum posts, and the other featuring XRP-AUD financial time series data. We sought posts from two forums to minimize biases and include a diverse user base of novice, intermediate, and experienced investors. If time had permitted, we would have attempted to include social media data, similar to past studies (Vo, 2019). Table 1 summarizes the explored domains, covering terms of service considerations, available data, and website limitations. For investing.com, we emailed the website owners, informing them of our intent to scrape data and ensuring compliance with their terms of service. All data was scraped according to the terms and was publicly available at the time of writing for anyone to access.

*Table 1: This table provides an overview of the two websites/domains targeted by the investingCrawler and cryptocompareCrawler web crawlers. It outlines the specific terms of service for each website, the available data that can be scraped, and the limitations or challenges faced during the data extraction process.*

|  |  |  |  |
| --- | --- | --- | --- |
| Website/Domain | Terms of Service | Available Data | Limitations |
| <https://www.cryptocompare.com/coins/xrp/forum> | Not prohibited in terms and conditions | Date of post  Post by user | Dynamic page which requires ability to scroll down, unsure how fat data goes back |
| <https://www.investing.com/crypto/xrp/chat> | Prohibits scrapping usernames, and requires prior approval | Date of post  Post by user | Need to navigate multiple pages automatically and adhere to terms of services |

### Forum Data Scrapping

The investingCrawler and cryptocompareCrawler are custom web crawlers designed to scrape comments and their respective dates from specific websites. Although they share several similarities in terms of the technology components and libraries they utilize, they target different websites with distinct structures, and their methodologies for accessing the data differ. The following comparison table highlights the similarities and differences between the two crawlers (table 2):

*Table 2: This table presents a comparison of the investingCrawler and cryptocompareCrawler web crawlers, highlighting the key differences and similarities in their technology components, target website structures, methodologies, and data storage.*

|  |  |  |
| --- | --- | --- |
| Feature | Investing.comg Crawler | Cryptocompare Crawler |
| Libraries | Selenium, Chrome WebDriver, BeautifulSoup, pandas | Selenium, Chrome WebDriver, BeautifulSoup, pandas |
| Target Website Structure | "commentInnerWrapper" class | "post-content" class |
| Data Location | "span" tags with "js-date" and "js-text" classes | "div" tags with "content-body" and "item-ago ng-binding" classes |
| Methedology | Iterating through pages | Scrolling through pages for a set duration |
| Data Storage | Pandas DataFrame, CSV file and stored in mySQL | Pandas DataFrame, CSV file and stored in mySQL |

Both crawlers leverage popular Python libraries such as Selenium, Chrome WebDriver, BeautifulSoup, and pandas. Selenium and Chrome WebDriver are used to load and interact with dynamic web pages, making them ideal choices for handling JavaScript rendering. BeautifulSoup is employed to parse and extract data from the HTML content of a webpage, while pandas is utilized to store the scraped data in a DataFrame, allowing for easy manipulation.

The investingCrawler targets a website with a specific structure, where comments and dates are stored within elements of the "commentInnerWrapper" class. The data is located within specific elements ("span" tags) with corresponding class names "js-date" and "js-text." The crawler follows a sequence of methods, including initializing the crawler, iterating through the pages, extracting data from the page, saving the DataFrame to a CSV file, and closing the WebDriver session. It is designed to crawl all 1222 pages of the targeted website with a wait time of 15 seconds set using the implicitly\_wait() method to ensure the content is loaded before extracting the data.

On the other hand, the cryptocompareCrawler targets a website with a different structure, where comments and dates are stored within elements of the "post-content" class. The data is located within specific elements ("div" tags) with corresponding class names "content-body" and "item-ago ng-binding." The crawler's methodology includes initializing the crawler, scrolling through the pages for a set duration, extracting data from the page, saving the DataFrame to a CSV file, and closing the WebDriver session. It is designed to scroll through the targeted website for a specified duration, loading and extracting data from the dynamically loaded content, with a scroll interval of 6 seconds set to ensure the content is loaded before extracting the data. In summary, the investingCrawler and cryptocompareCrawler web crawlers, while sharing similarities in technology components, differ in their target website structures and methodologies.

### Historic Price Data

To acquire the historical price data of XRP-AUD, which is essential for our model, we utilized the **yahoo\_fin** package in Python. By executing the command **get\_data("XRP-AUD", start\_date="01/01/2017", end\_date="15/04/2023", index\_as\_date = False, interval="1d")** in Python, we successfully retrieved daily XRP-AUD data ranging from November 10, 2017 to April 14, 2023. This data contained information on date, opening price, highest price, lowest price, closing price, adjusted closing price, and volume. We then processed the data and imported it into MySQL. This serves as a vital component, as it represents the actual price for a specific day in our machine learning model. It is important to note that the specific time zone employed is not explicitly stated; however, we hypothesize that the data is reported in UTC, as suggested by the XRP price page (XRP AUD (XRP-AUD) price history; historical data 2023).

At this point, as shown in Figure 1, we have successfully acquired and stored data from three sources in preparation for data wrangling, pre-processing, and EDA as we move towards building our LSTM network model.

# Data Wrangling

# Machine Learning

# Conclusion & Lessons Learned

# Reference

G

Gillis, A.S. (2022) What is a web crawler? everything you need to know from techtarget.com, WhatIs.com. TechTarget. Available at: https://www.techtarget.com/whatis/definition/crawler (Accessed: April 18, 2023).

P

Perez, M. (2023) What is web scraping and what is it used for?, ParseHub. Web Scraping Blog (Tips, Guides + Tutorials) | ParseHub. Available at: https://www.parsehub.com/blog/what-is-web-scraping/ (Accessed: April 18, 2023).

X

XRP AUD (XRP-AUD) price history &amp; historical data (2023) Yahoo! Finance. Yahoo! Available at: https://finance.yahoo.com/quote/XRP-AUD/history?p=XRP-AUD (Accessed: April 18, 2023).