



# Tickr

Stock Price Predictive Analysis Project

# Presentation Overview

Here's what we have for you today:

1. Project goals.
2. Real world applications.
3. A.I.-based algorithms.
4. Program demonstration.
5. Technical challenges.
6. Future work.

# Introduction

Tickr combines natural language processing with machine learning to predict trends and changes in stock prices.

Data sources:

- Past stock price data
- News articles
- Tweets



So  
Inspiring

“It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers... They would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to expect the machines to take control.”

– ALAN TURING



01

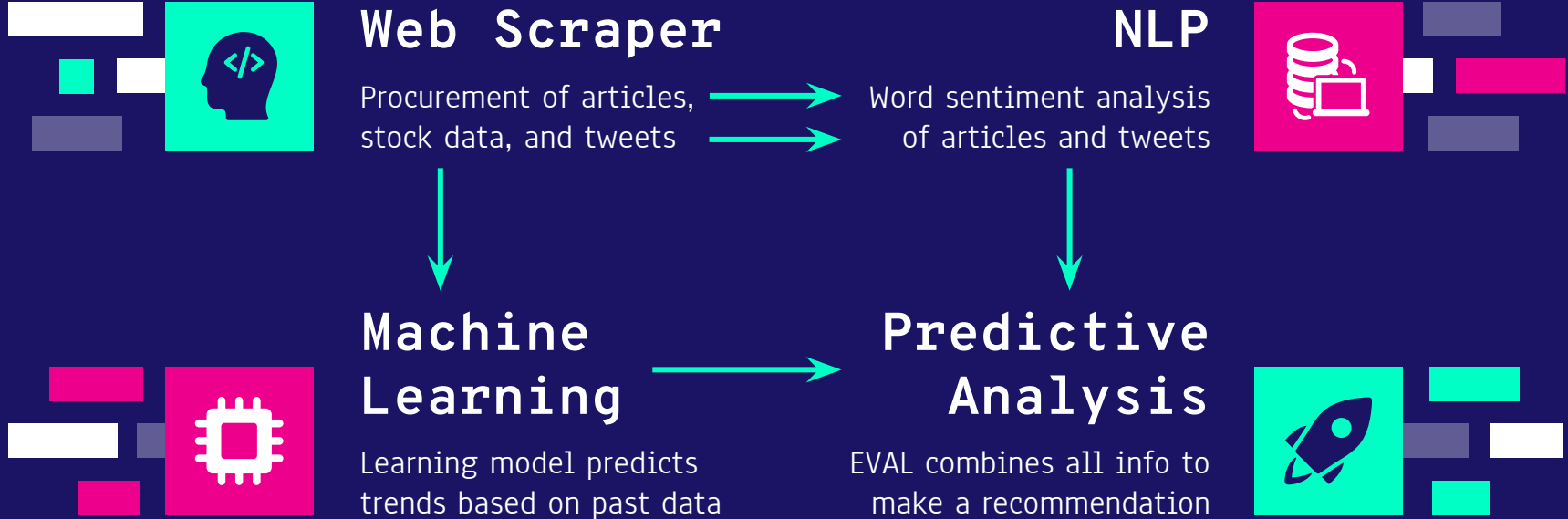
# PROJECT GOALS



## Main Objective:

Using word sentiment analysis, clustering, and predictive machine learning, we aim to provide the user with a reliable recommendation to buy, hold, or sell stocks and cryptocurrencies.

# How the Pieces Fit



# Web Scraper

News was sourced from CNBC.com.

A database of article links was created by searching for a given company on CNBC.com and collecting all links in the search results.

Our scraper followed the links in the database and the contents, dates, and titles of these articles were collected using the BeautifulSoup library.

The contents of the articles were saved to speed up run-time in the future and are used to cluster news articles.



# Yahoo Finance

Yahoo Finance has historical data containing past values of stocks.

We used the Yahoo Finance API to access this data and collect it for our project.

These values were used in the machine learning and the article clustering.

# Natural Language Processing

Natural Language Toolkit (NLTK) sentiment analyzer:

Classifies news articles and tweets as positive, negative, or neutral.

Display Twitter mood about companies for each prediction.

Cluster news articles by word frequency tables and predict change based on their correlations.

# Article Clustering

For each company we have a database of articles and the dates they were published.

We find the change in stock price in the 24 hours after each article was published.

We cluster the articles based on word frequencies.

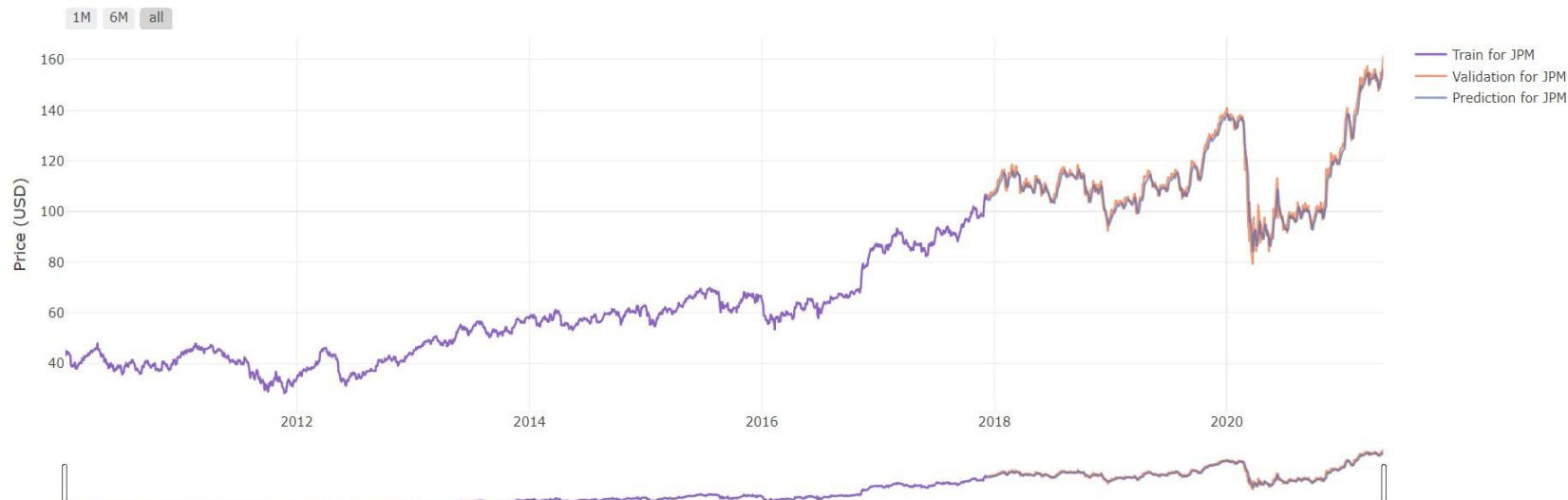
The mean change of the cluster the latest article is sorted into is the prediction for how the stock price will change in 24 hours.

# Machine Learning

## LSTM Model: How It Performed

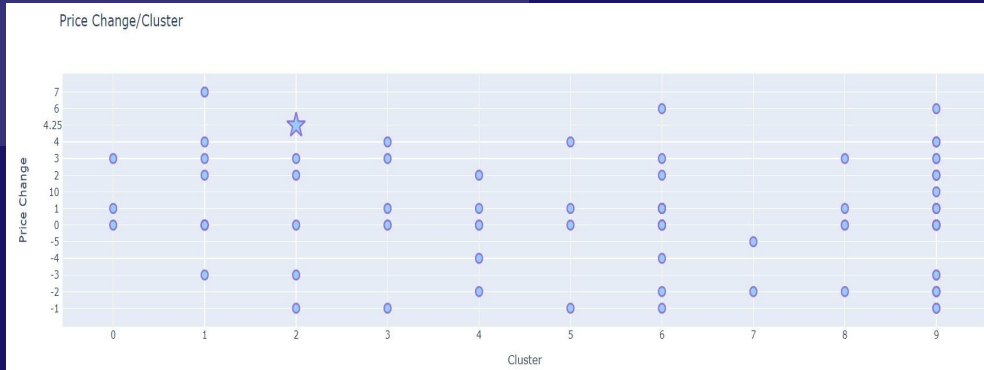
JPMorgan

Training and Testing Data



# Predictive Analysis

- Show user how the sentiments around the stock/crypto are being perceived.
- User may make conclusions based on sentiments.



- Make final recommendation to the user based on Reinforcement Learning Model predictions (from last slide) with the Clustering predictions.



02

REAL WORLD APPLICATIONS

# Stock Market Forecasts

- The applications for an AI that can successfully predict changes in the stock market are obvious.
- Many academics suggest that it is impossible to predict exact prices, as it is quite unlikely that models can predict Black Swan Events, or take into account many different and apparently “uncorrelated” variables.
- Our goal is to predict the short term value of a stock/cryptocurrency, acknowledging the limitations of our model and relying on research that suggests that news articles and press releases are correlated to stock prices.



03

# A.I.-BASED ALGORITHMS



# K-Means Clustering

- Used to cluster news articles based on the relative frequency of words in each article.
- Database of news articles were clustered based on word frequency. The cluster the latest article belonged to was used to predict the change in price over the following 24 hours.

# Natural Language Processing

- Words carry context and meaning that machines do not understand inherently.
- We can define the meaning of words for machines using word sentiment analysis.
  - Some sentiment analysis processes check for words and their frequency of use in a body of text against “positive” and “negative” sentiment lists.
  - NLTK uses this technique.
- Sentiment analysis can determine the “mood” of a text or collection of texts.

# Machine Learning Algorithm

- A Reinforcement Learning model was used for predicting the future values, based on the Long Short-Term Memory (LSTM) architecture. The LSTM is a recurrent neural network architecture.
- We chose the LSTM networks because they are well-suited to classifying, processing, and making predictions based on time series data.
- The data was preprocessed, then split into training, testing, and validation sets. Once that was done, a Sequential model was created with two LSTM layers, one Dropout layer, and two Dense layers.
- Root Mean Squared Error and R-Squared values were calculated in order to determine if the model is fitting the data correctly.



04

PROGRAM DEMONSTRATION





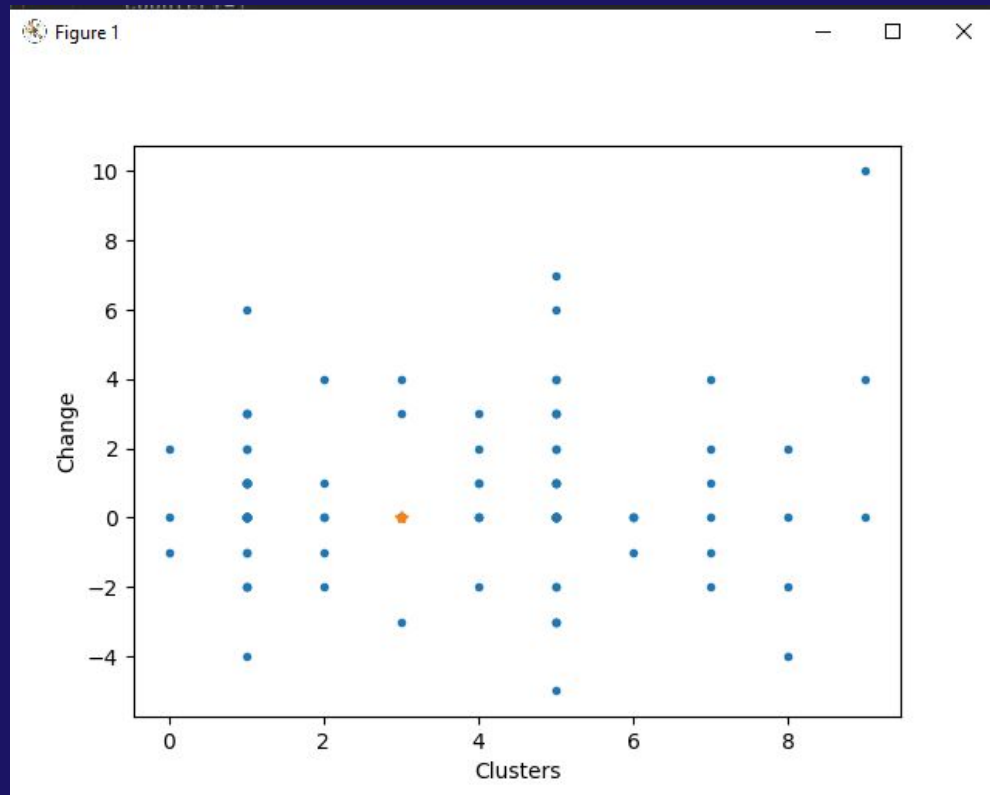
05

# TECHNICAL CHALLENGES

# Article Clustering

It's unclear if article clustering is a good predictor of stock value.

In this graph, we see that within clusters the change associated with each article seems fairly randomly distributed.





# Information Retrieval

- Difficulty finding the information we needed from open sources.
  - Many news sources require subscriptions to access all content.
- Found news sources that were free, legal, and easy to scrape.
  - CNBC, Twitter



# Data Handling

- Unsure how to extract sentiments from news articles.
  - Many libraries claim to have the best technique.
- Chose Natural Language Toolkit (NLTK) for its simplicity and efficiency.
  - Built-in analyzer that classifies text bodies based on the words in the text.

## Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](#) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](#).



- Difficulty accessing tweets from Twitter's API.
  - Requires a developer account. Contacted Twitter, explained our reasons for scraping tweets, and acquired access.
- Understanding the API and rate limits.
  - Studied the API documentation and found tutorials to get the information we needed.

# Scraping HTML

- We scraped HTML code for our articles.
  - How do we separate the article text from the code?
- **SOLUTION:** BeautifulSoup and Selenium libraries.
  - BeautifulSoup has built-in tools to search for HTML tags, like article titles and text bodies.
  - And Selenium has tools for interacting with the browser, such as sending requests, scrolling up/down a page, clicking buttons, submitting forms etc.



## Beautiful Soup Documentation ¶

**Beautiful Soup** is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

# Predictive Modeling

- Difficulty determining how far into the future we should try to predict.
  - The research we came across suggests that it is unfeasible to look farther ahead than a day or two.
- Our trials showed that predictions of trends and price changes were most accurate when we predicted only one day into the future. And even so, we realized that the most important thing about the predictions was getting the trends right, not the actual values.





06

FUTURE WORK

# Future Work

More User  
Controls

More Data

Improve  
Clustering

Weighted  
Sentiments

Long-Term  
Predictions

More  
Companies

# THANKS!

Do you have any questions?

Ryan Farrell

Bernardo Santos

Allen Westgate



CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.

# Resources

## **On NLP:**

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

<https://www.digitalocean.com/community/tutorials/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk>

[NLP in the Stock Market. Leveraging sentiment analysis on 10-k... | by Roshan Adusumilli | Towards Data Science](#)

## **On Bitcoin:**

<https://arxiv.org/ftp/arxiv/papers/2006/2006.14473.pdf>

## **On Bitcoin Price (.csv file from download):**

<https://www.coindesk.com/price/bitcoin>

## **On Dash:**

<https://data-flair.training/blogs/stock-price-prediction-machine-learning-project-in-python/>

## **On LSTM:**

<https://www.youtube.com/watch?v=QIUxPv5PJOY>

## **On Sentiment Analysis:**

Samuels, A., & Mcgonical, J. (n.d.). News Sentiment Analysis. Retrieved April 24, 2021, from <https://arxiv.org/pdf/2007.02238.pdf>

## **On Scraping HTML:**

Richardson, L. (2007). Beautiful soup documentation. *April*.

## **On Machine Learning For Stock Forecasting:**

[How Does Machine Learning Perform in the Stock Market? | by Marco Santos | The Startup | Medium](#)  
[Historical Stock Price Data in Python | by Ishan Shah | Towards Data Science](#)