Data Science II Project A:

Real-time Dashboard of Bitcoin

1. Background and Overview:

Cryptocurrency is the most well-known implementation of blockchain technology and has shown massive growth in the last year attracting the attention of global investors. Bitcoin, the largest cryptocurrency, has a market capitalization of $183.07 billion USD as of March 6, 2018, while the entire cryptocurrency market had a $27 billion USD capitalization in 2017.[1] This immense growth seems to indicate that cryptocurrency is well on its way to becoming a trillion-dollar industry. Given the wide-scale adoption, the need for applications and tools giving an overview cryptocurrency related data has risen. The scope of this project is a real-time updated dashboard of Bitcoin, as it currently makes up 42% of the total cryptocurrency market.[2] The dashboard being presented in this report is useful to gain an overview of the Bitcoin market and surrounding sentiment. As well, a predictive classifier is included which can be used to make buy or sell decisions involving Bitcoin.

2. Main Features of Dashboard:

The dashboard uses both structured and unstructured data, including Twitter data, news articles, and pricing information to give the user a general overview of current and historic Bitcoin activity. The historic price chart and current price, is useful for identifying where Bitcoin is currently sitting relative to the past. The chart showing a count of "#Bitcoin" and "#Ethereum" usage pulled from Twitter gives insight into the social media buzz surrounding this specific cryptocurrency, and some indication on the impact of any concurrent events. The word cloud generated from news sources can show general market sentiment from reputable news organizations which is useful if users of the dashboard have limited time to sift through various news websites. The Naïve Bayesian

[1] Cointelegraph.com. (2018). Available at: https://cointelegraph.com/news/combined-crypto-market-capitalization-races-past-800-bln [Accessed 6 Mar. 2018].

[2] Coinmarketcap.com. (2018). Available at https://coinmarketcap.com/charts/ [Accessed 6 Mar. 2018].

Classifier uses Google News to determine if the returns for the day will be positive or negative. A labelled screenshot of the indicators from the dashboard is provided below:

a. Indicators

i. Current price of bitcoin (BTC) updated every 10 seconds

ii. Daily recommendation from Classifier

iii. Price trend graph of Bitcoin

iv. Frequency graph of #bitcoin and #ethereum usage

v. Word cloud related to bitcoin showing most used terms

b. Source of data

i. The current price of Bitcoin in USD is obtained from the webpage https://min-api.cryptocompare.com/data/price?fsym=BTC&tsyms=USD

ii. Historical prices of Bitcoin in USD are retrieved using the following API and stored in a list until further processing for dashboard. https://min-api.cryptocompare.com/data/histominute?fsym=BTC&tsym=USD&limit=2000

iii. Python library Tweepy was used to access the Twitter API, specifically pulling #bitcoin and #ethereum data and a frequency count.

iv. Instructions on how to count frequencies of the Twitter hashtags were found from https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/.[3]

v. Google News was used for the classifier to find common words http://bit.ly/2oZTKL5

vi. Data from news articles were pulled from CoinsNews to generate the word cloud. The APIs used are

https://newsapi.org/v2/top-headlines?**sources=crypto-coins-news**&apiKey=API_KEY

https://newsapi.org/v2/everything?sources=crypto-coins-news&apiKey=1d656ac0916147bf8d28e1dcda71266a

[3] Marcobonzanini.com. (2018). Available at https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/ [Accessed 19 Feb. 2018].

c. Challenges Faced and Their Resolutions

The team faced many challenges in each stage of building the dashboard.

i. Accessing Twitter data: As the nature of Twitter data is unstructured, this posed problems when trying to extract insights about cryptocurrencies from Twitter. The initial thinking behind using Twitter was to complete a sentiment analysis that would provide a summary of the feelings around certain cryptocurrencies using their names as hashtags over time to see if any trends emerged. However, this became problematic after extracting the words commonly associated with Twitter because many accounts that use these hashtags are advertising some sort of contest and only a small percentage of tweets using these hashtags were actually using them in such a way that would provide insights. Many hours were spent filtering out tweets from spam accounts and limiting the number of nontrivial words being entered into the model. However, it was decided that looking at the use of certain hashtags over time would be more conducive to exploring trends than a sentiment analysis. The sentiment analysis idea was put to rest.

ii. Visualizing and implementing the data in a real-time dashboard: While there are many options of tools to use for data visualization, the requirement that the dashboard must run on Ubuntu limited the options available. The team eliminated Tableau, Qlik sense and SAS as options and had to find out a suitable option that is compatible with Ubuntu. Dash, a python framework, was used instead.[4]

iii. Challenges with Plot.ly & Visualization. The team encountered many barriers when implementing the dashboard and deciding how to display the data. Initially we decided to use Plot.ly to create individual graphs and send them to an online profile, then combine the graphs to create a user-friendly tool. The problem with this was the group's unfamiliarity with the tool causing confusion implementing multiple graphs. Furthermore, Plot.ly becomes less useful as the graphs become more complex. Any insight to investors would be limited to static charts, which can easily be accessed online in more complete detail. To combat this, it was decided that the python framework Dash should be used to build analytical web applications.

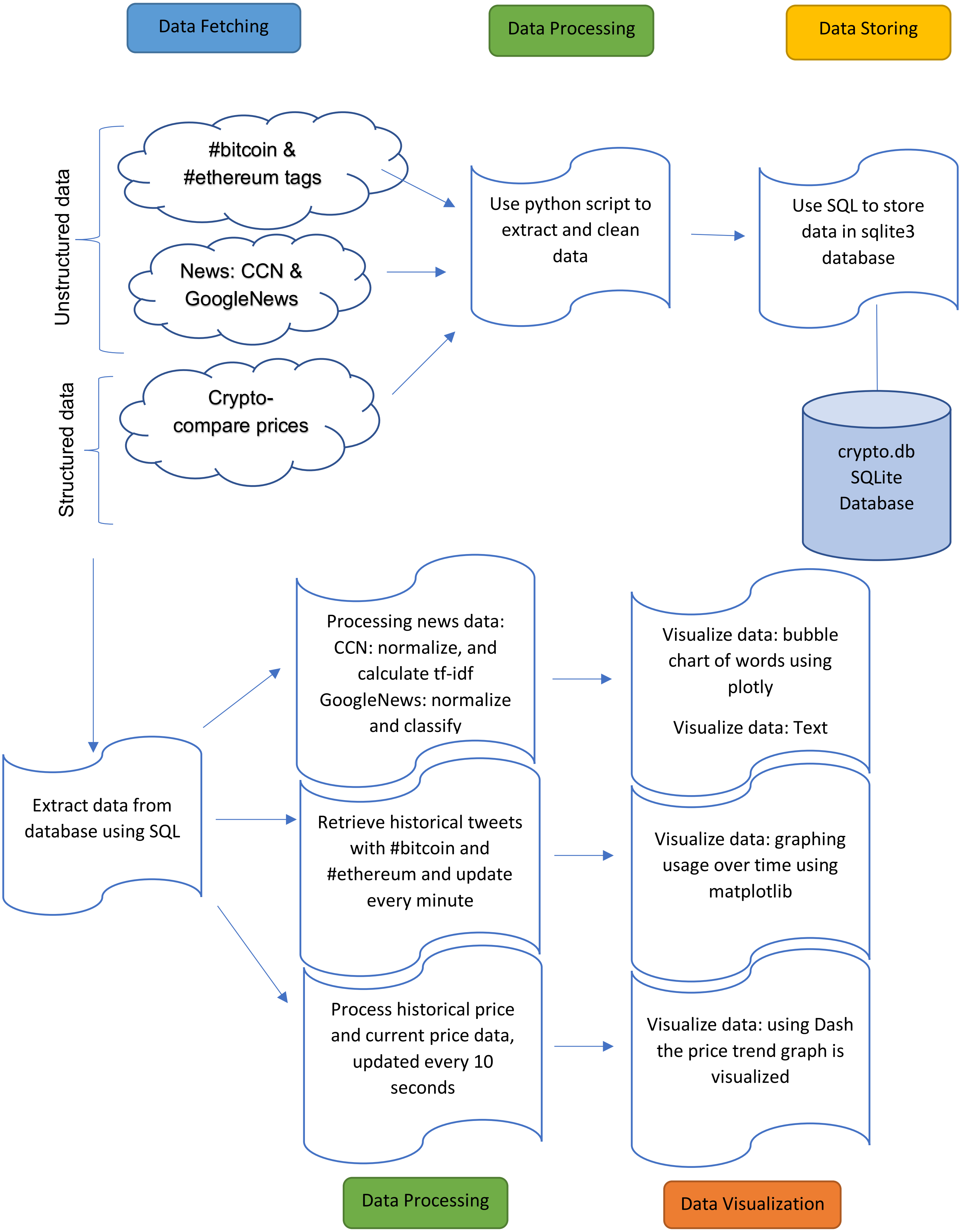
[4] Dash.Plot.ly. (2018). Available at https://dash.plot.ly/live-updates [Accessed 25 Feb. 2018].

iv. Challenges for the classifier: In order to predict whether bitcoin prices would increase or decrease based on frequent words used in the news, historical news and prices needed to be accessed. As most APIs only pull news for the current day, neither an API nor the familiar JSON format could be used. Instead the web was scraped using the BeautifulSoup package. It was a steep learning curve. Scraping the news articles and the dates from the HTLML document proved very difficult, in part since none of the group members had worked with HTML before. After a lot of hard work and frustration, ten bitcoin news article descriptions were pulled from Google News every day in 2017. Running this particular script took about 20 minutes each time. There were significant challenges cleaning the HTML documents and associating them to Bitcoin price data. The association was eventually achieved by converting dates into formats Python could work with. In implementing the classifier, understanding how the training function wanted data was a bit of a challenge. The data was eventually converted into the correct list of dictionaries format, and the classifier was trained. Originally the NLTK Naive Bayesian Classier was trained against three target values (>5%, -5% to 5%, and <-5%). These were coded as 1, 0 and 2 respectively. Unfortunately, the classifier performed terribly. The exercise was redone with two buckets (positive and negative changes in price) and got better results.

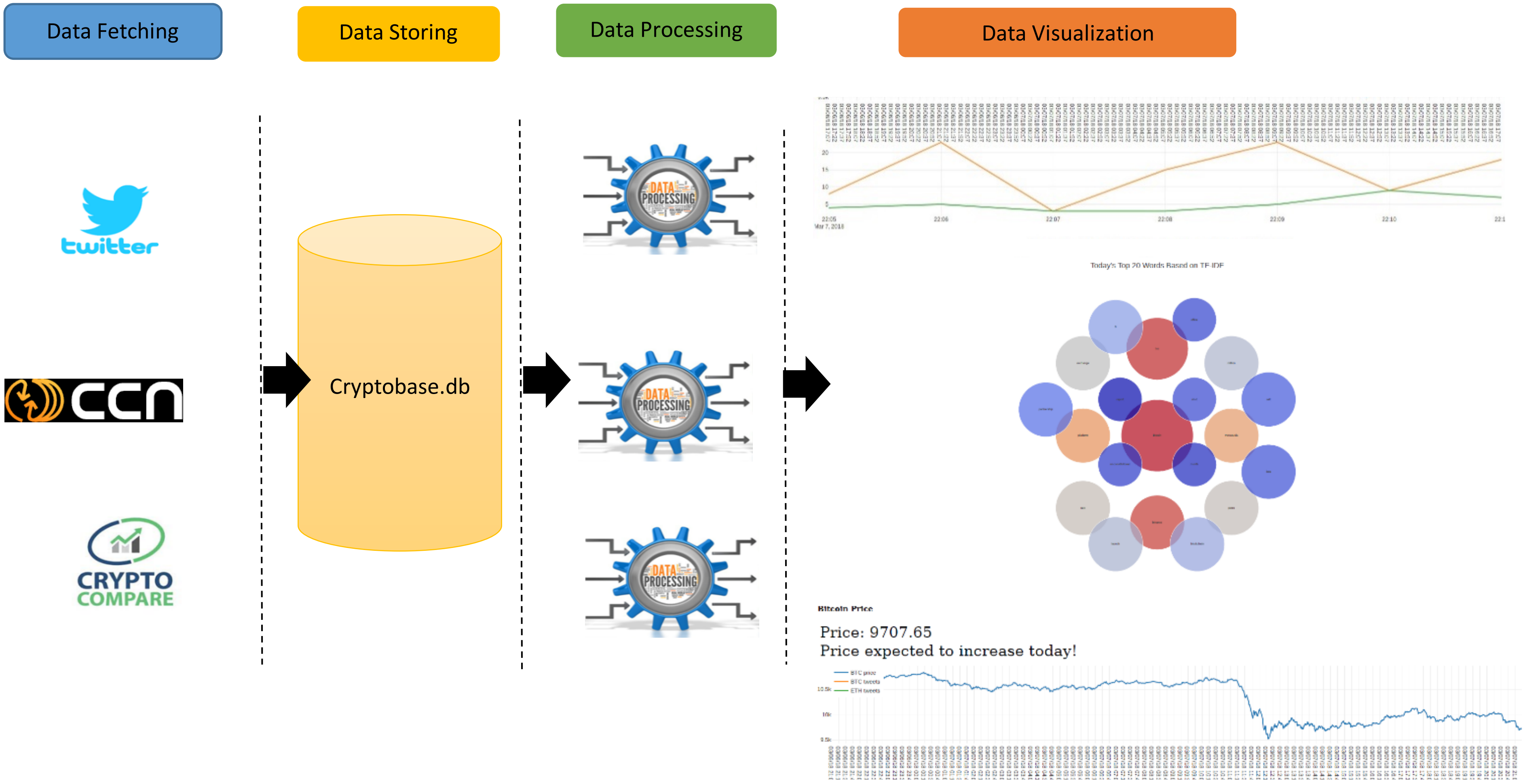
d. Data Processing:

i. The Bitcoin prices, Twitter data, and news articles are read in with APIs to Python where they are then stored in a sqlite3 database (crypto.db) using SQLite. Next, the data is extracted from the database using SQLite and python script is again used to process the data. Unstructured text data related to news is normalized i.e tokenizing, adding stop words, stemming and lemmatizing are done. All normalized words are counted and then the term frequency–inverse document frequency (tf-idf) is calculated, which is a numerical statistic reflecting how important a word is to a document in a collection or corpus. Unstructured tweets containing #bitcoin and #ethereum are counted on a per-minute basis and this data can be live-streamed onto the dashboard. Structured data of historical bitcoin prices are visualized real-time using Dash, updating every minute and the current price every 10 seconds.

The next page contains the data processing pipeline.


Data Fetching
Data Processing
Data Storing
Unstructured data
#bitcoin & #ethereum tags
News: CCN & GoogleNews
Structured data
Crypto-compare prices
Use python script to extract and clean data
Use SQL to store data in sqlite3 database
crypto.db SQLite Database
Extract data from database using SQL
Processing news data: CCN: normalize, and calculate tf-idf GoogleNews: normalize and classify
Retrieve historical tweets with #bitcoin and #ethereum and update every minute
Process historical price and current price data, updated every 10 seconds
Visualize data: bubble chart of words using plotly
Visualize data: Text
Visualize data: graphing usage over time using matplotlib
Visualize data: using Dash the price trend graph is visualized
Data Processing
Data Visualization

A high-level architecture of this entire process is given below-



Price: 9707.65

Price expected to increase today!

3. **Diagram of the dashboard:**

<u>“Wireframe” Drawing vs. Implemented Dashboard</u>

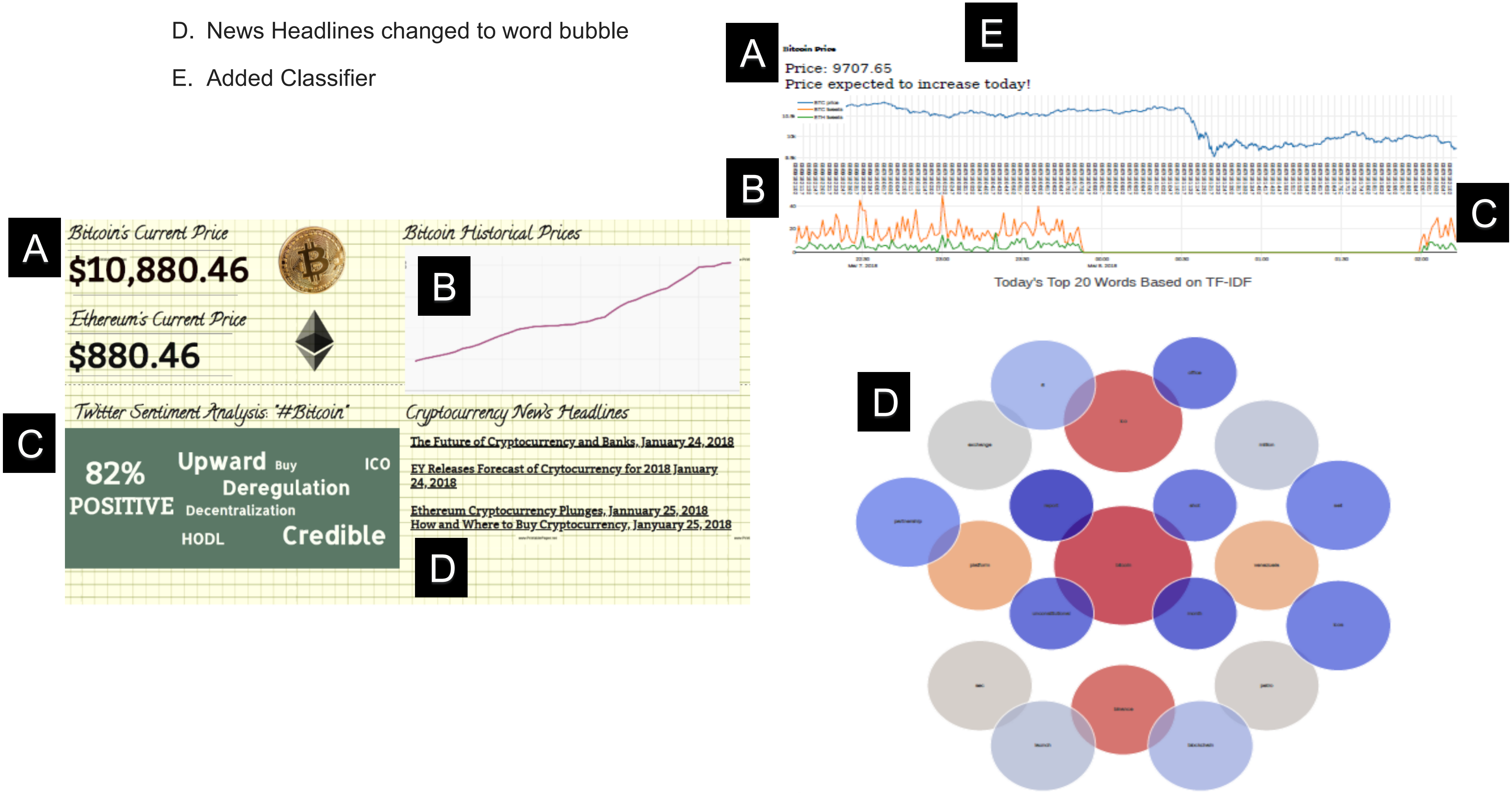
A. Bitcoin Price Data stayed the same

B. Historic Price Data stayed the same

C. Twitter Sentiment changed to frequency count of #Bitcoin and #Ethereum

D. News Headlines changed to word bubble

E. Added Classifier


A
Bitcoin's Current Price
$10,880.46
Ethereum's Current Price
$880.46
B
Bitcoin Historical Prices
C
Twitter Sentiment Analysis: "#Bitcoin"
82% POSITIVE
Upward
Buy
ICO
Deregulation
Decentralization
HODL
Credible
Cryptocurrency News Headlines
The Future of Cryptocurrency and Banks, January 24, 2018
EY Releases Forecast of Crytocurrency for 2018 January 24, 2018
Ethereum Cryptocurrency Plunges, Jannuary 25, 2018
How and Where to Buy Cryptocurrency, Janyuary 25, 2018
D
A
Bitcoin Price
Price: 9707.65
Price expected to increase today!
E
B
C
Today's Top 20 Words Based on TF-IDF
D

4. Training A Predictive Classifier

To train a predictive classifier, the group scraped Google News articles related to Bitcoin for every day in 2017. Relevant aspects of the HTML file were identified (date and article description). Each date scraped provided 10 article descriptions. The group obtained historical Bitcoin prices and matched it to the Google news articles using the dates, calculating price changes day over day. The words in the news articles were normalized and assessed based on TFIDF. All the words were combined into one category and the top 500 (based on TFIDF) were chosen as potential features for the Naïve Bayesian Classifier. For each normalized Google News article, we tested whether or not that word appeared in the list of features. We then matched each article to a target value. Originally we started with three target values (as detailed in the challenges section), but after that method failed, all articles were coded as 0 or 1 (decrease in price or increase in price from previous day, respectively). We balanced the 0s and 1s and separated the balanced set into training and testing. We trained the model using the NLTK Naïve Bayesian Classifier model and calculated accuracy. The 500 feature words gave us an unsatisfactory accuracy but by decreasing the number of features to 200 we were able to increase accuracy significantly by limiting overfitting. We use the classifier to predict whether or not the price of Bitcoin will increase from yesterday based on the words used in the current day's Google News results and display that prediction on the dashboard.

5. Resources Used

Cointelegraph.com. (2018). Available at: https://cointelegraph.com/news/combined-crypto-market-capitalization-races-past-800-bln [Accessed 6 Mar. 2018].

Marcobonzanini.com. (2018). Available at https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/ [Accessed 19 Feb. 2018].

Coinmarketcap.com. (2018). Available at https://coinmarketcap.com/charts/ [Accessed 6 Mar. 2018].

Dash.Plot.ly. (2018). Available at https://dash.plot.ly/live-updates [Accessed 25 Feb. 2018].