INTERNATIONAL HELLENIC UNIVERSITY

Real Time Crypto Sentiment Analysis

Athanasios Samaras

SID: 3308210038

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

January 2023

THESSALONIKI – GREECE

Athanasios Samaras

SID: 3308210038

Supervisor:	Prof. Christos Berberidis
Supervising Committee Members:	Assoc. Prof. Vassilios Peristeras
	Assist. Prof. Ioannis Magnisalis

I hereby declare that the work submitted is mine and that where I have made use of another's work, I have attributed the sources according to the Regulations set in the Student's Handbook.

January 2023

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in Data Science at the International Hellenic University.

This year, on the 3rd of January Bitcoin, has its thirteenth anniversary. Such an essential mark on the calendar, if anyone could consider that after that day thirteen years ago, more than 12.500 cryptocurrencies were released, according to CoinMarketCap, in 2021.

This trend created new job descriptions, such as crypto traders, crypto analysts, crypto developers, and enthusiasts. As anything significant to the public audience nowadays it's posted on the web, cryptocurrencies do not differ. People are trying to follow the trends regarding cryptocurrencies, learn about the news, and even watch some technical breakdowns to help them make their next move.

In this paper, CryptoFeel.eu is being introduced, connecting the dots related to the Cryptocurrency News, Cryptocurrency Prices and Analytics, Voice of The Crowd, and Sentiment Analysis, all in real-time. CryptoFeel.eu aims to set strong pillars, related to future research in the field of real-time decision-making on cryptocurrency, taking into consideration the news and the sentiment analysis on social media.

Real time decision making based on the Crypto sentiment analysis, by the time that this thesis was written, was totally missing from the market. Since this market operates 24/7 and bad or good news may have an impact on the market, this tool is now here, tracking the prices and the news 24/7 as well. In the first 2 months, this project has collected & analyzed more than 600.000 Tweets and more than 1.000 News & Articles from Yahoo Finance News, performing Sentiment Analysis on the Tweets and Abstractive Text Summarization & Sentiment Analysis on the Yahoo Finance News. These numbers are promising a great in-depth view in the important news that might have an impact on the prices.

Athanasios Samaras

January 2023

Keywords: CryptoFeel.eu, Sentiment Analysis, Cryptocurrency, NLP, Summarization, Roberta, Bert, T5, Pegasus, Abstractive, Extractive, Evolution Index;

Athanasios Samaras

January 2023

Acknowledgements

I would like to thank Dr Christos Berberidis the supervisor of this research, for their guidance, support and feedback. Additionally, I would like to thank my friends and family for the conversations that we had and created an out of the box mentality.

Contents

1 Introduction

Cryptocurrencies are famous, this is not a secret [1]–[3]. One side of this story is that blockchain, a disruptive technology that cryptocurrency is heavily based on, supports the idea of decentralized finance. That means that in this system, there is no mediation need for the traditional financial system [4], [5]. Cryptocurrencies, also known as "crypto assets" or simply "cryptos," are digital assets intended to function as trade mediums, similar to traditional currencies. They use robust cryptography, as their name suggests, to safeguard exchange. This cryptography, in short, encrypts the raw message at the source, sends the encrypted text to the recipient, and decrypts the content at the destination using the same secret key.[6] They lack an intrinsic worth today, like most conventional money also known as “fiat”, and transactions determine their value.[7]

The other side of previous story puts some emphasis on the risks that cryptocurrencies include, such as, the anonymity and privacy that can be connected with illegal acts related to Ponzi schemes, money laundering and so on [8]. Adding an extra layer to this part, this environment exists due to some critical reasons or issues and resolving these could have a great impact on all cryptocurrencies. These reasons are the regulations that are not in place along with the emerging technologies related [5], [8].

Beyond this, though, they are increasingly becoming a part of the digital transactional infrastructure, and their underlying technologies hold promise for providing solutions in various business processes. Processes, like control over lengthy chains of transactions that depend on trust and are vulnerable to fraud, along with managing supply chains and protecting intellectual property rights. [7]

In today’s world, more and more data is posted on social networks, and this could be great content to analyze in order to measure the engagement effectiveness [9]. Having that in mind, enthusiasts and crypto investors do not differ, as any type of news that could be found in micro-blogging websites could help on drawing a different strategy [10].

For this reason, some of the most well-known platforms, Twitter and Reddit, have been already studied, in previous research, in the sense of checking any possible correlation to

the prediction of price variations that are often observed in cryptocurrencies [10]–[16]. Real Time analysis though, is finally here with CryptoFeel.eu

1.1 Research Objectives

This research aims to contribute to the current literature on cryptocurrency prices and sentiment analysis from social media in real-time. Considering that there are days with high growth in this field, we shall not skip to seeking the motivation behind investors' decisions, considering the market's volatility.

The main tools that were used in this study are Beautiful Soup with CoinGecko and Yahoo Finance News and Snscrape Library. CoinGecko is the world's largest independent crypto data aggregator at this time, and it also includes an API that can be used to provide all relevant information regarding cryptocurrencies.

Beautiful Soup is a Python Programming Language library that helps get data out in XML, HTML, and other markup languages. Beautiful Soup Library was a key to getting a snapshot of the data available in CoinGecko and Yahoo Finance News. For this research and for creating this tool, no data from APIs was analyzed, only data that was scrapped through the Web.

The tweets were gathered using the Snscrape Python Library and, all along, were merged to provide valuable insights. Some tweets included noise or low engagement scores. Most of them have been cleaned. However, additional filtering options are provided to end users, cleaning their results even more.

Sentiment Analysis on Twitter was performed using the RoBERTa, a Robustly Optimized BERT Pretraining approach [17] which classifies the tweets as Positive, Neutral or Negative. Yahoo Finance News were scrapped through Beautiful Soup Library as well, summarized via Pegasus, where like an extractive summarization, one output sentence is generated out of the important sentences that are masked from an input source[18], and measured the sentiment using a distilbert-base-uncased-finetuned model.

Moreover, all tools that are mentioned above, were used to create an environment, where researchers and enthusiasts can get a real time snapshot of the price, social media sentiment, Yahoo News impact, and the correlation using the filters and keywords of their choice.

2 Literature Review

This section examines the broad knowledge and recent research on cryptocurrencies, including what they are and how they have changed up to this point. Additionally, the value of Twitter sentiment analysis and News impact is being studied, mentioning techniques used for sentiment analysis and text summarization.

2.1 Cryptocurrencies

Due to the recent adoption of cryptocurrencies by governments and international corporations, cryptos have experienced tremendous growth in popularity. They were growing more popular due to use cases like El Salvador's approval of Bitcoin as legal cash on September 7, 2021, and Twitter's introduction of a new feature on September 23, 2021, which enables users to send tips to other users using Bitcoin Lightning wallets.

The most well-known cryptocurrency is Bitcoin, whose description and application were released in an anonymous paper in 2008 by an individual who goes by the pseudonym, Satoshi Nakamoto. By using a Proof-of-Work user consensus process, Bitcoin was able to offer security and address issues like double spending and the threat of hacker assaults. The goal of this peer-to-peer system was to do away with the necessity for and expense of financial institutions' mediation. Other than Proof-of-Work a Proof-of-Stake could have been introduced. The latter requires from the participants to acquire a crypto stake behind a new block, in contrast to the proof of work which requires from the users to generate a valid new block using their computers and electricity.

"An electronic payment system based on cryptographic proof, instead of trust, allowing any two willing parties to deal directly with each other without the need for a trusted third party," is how Nakamoto characterizes Bitcoin in his whitepaper. According to Lansky's [19]definition, cryptocurrencies are decentralized digital currencies that can satisfy all six of the following criteria.

- There is no need for a central authority because consensus is reached decentralized.
- The ownership of bitcoin units is documented.

- When the production of new cryptocurrency units is permitted, the system will establish the guidelines and requirements for their creation and ownership.
- Only cryptography may be used to establish ownership of cryptocurrency units.
- It is legal to trade cryptocurrency units in order to shift ownership. For a statement of the transaction to be issued, evidence of ownership of the units must be provided.
- When numerous transactions that alter the ownership of the same cryptographic units are scheduled to take place simultaneously, the system only completes the first transaction.

After Bitcoin was introduced, the cryptocurrency industry exploded with investments in the years that followed, as seen by the sharp rise in its market cap and transaction volume. The fact that the supply of bitcoins is limited, with a maximum of 21 million projected to finally enter circulation through mining until the year 2.140, is an additional feature that should be taken into account. But it can be divided into "Satoshi's, SATS", where 1 Satoshi = 0.00000001 Bitcoin. (CoinMarketCap, 2021)

CoinMarketCap (2021) estimates that as of January 7, 2018, the total market value of cryptocurrencies increased from $5.5 billion at the beginning of 2015 to $820,2 billion. On September 9, 2021, the estimated market capitalization of all cryptoassets was over $2.3 trillion, a rise of more than 418%.

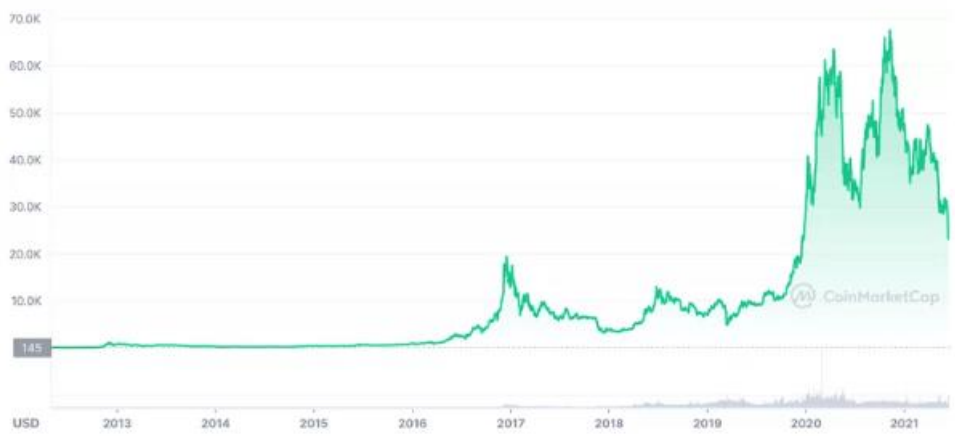


Figure 1: BTC Price over the years: CoinMarketCap

It would be helpful to compare the size of the market to the overall worth of gold, which the World Gold Council estimates to be roughly US$11.3 trillion.

2.1.1 The crypto-world

It is impressive to see how quickly Bitcoin alternatives like Cardano (2017), Litecoin (2011), and Ethereum (2015) have grown in popularity. More than 12.500 cryptocurrencies exist as of now (CoinMarketCap, 2021). This attempt to address various issues or give more effective solutions than Bitcoin's Proof-of-Work consensus process, which requires much energy to maintain network security.

According to Vitalik Buterin, the creator of Ethereum, three fundamental problems must be solved when creating blockchains. These three factors—decentralization, security, and scalability—combine to generate the idea of the "Blockchain Trilemma." The problem of how to accomplish all of these goals without making any sacrifices is still open. (CoinMarketCap, 2021)

Although cryptocurrencies have undoubtedly gained much attention, they have also come under fire because some of their distinctive characteristics, which offer anonymity to all parties, can be exploited for illegal activities, including money laundering, tax fraud, and Ponzi schemes.[8]

Major cryptocurrency exchanges like Coincheck, hacked in 2018 for $530 million, and Mt. Gox, felled in 2014 for $400 million, show the need for increased security.

According to some studies like this [5] ,which look at how the announcement of plans to build KODAKCoin increased the company's stock price, in the unstable state of the cryptocurrency market, the volatility of the cryptocurrency values is fed by media releases.

This was also made clear by Walmart's phony collaboration announcement with Litecoin on September 13, 2021, which was exposed immediately. The news was released by a company that provides communication services, and Litecoin's official account later tweeted it.

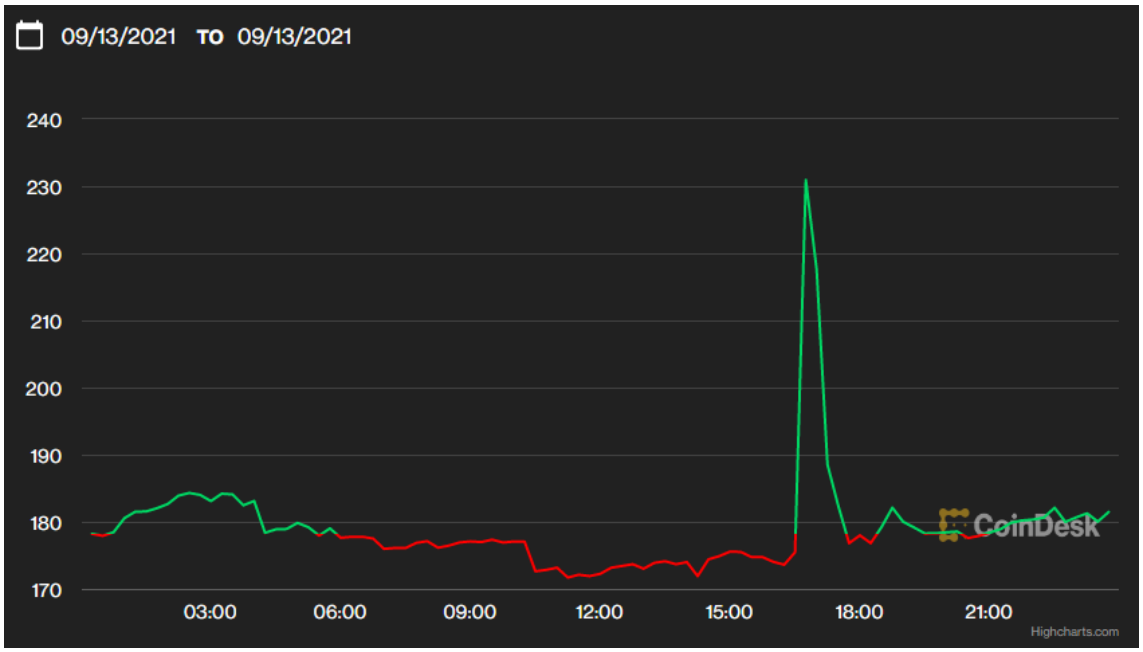


Figure 2: 13 September 2021 Litecoin Prices. Coindesk

This resulted in the price of Litecoin immediately, which rose by more than 30% after the announcement, only to drop back down after the phony news was exposed. The situation lasted about 20 minutes, and Walmart responded by rejecting the information in a statement.

2.1.2 The Twitter-world

People often talk about cryptos on Twitter, and when talking about cryptos, it's ideal to know with whom or what cryptos could be connected.

Tweets are studied [20] to impact the number and average value of Dogecoin transactions, but Litecoin transactions affect tweets. The same study proves that Dogecoin whale behavior impacts tweet volume, but there is no correlation between Litecoin whales and tweet volume. The proposed [20] ARIMA (0,0,0) models' predicting error was 0.08% (with Litecoin) and 0.22%. (with Dogecoin). So, those are just the beginning of scientific discoveries that can inspire the development of a trading robot based on these insights.

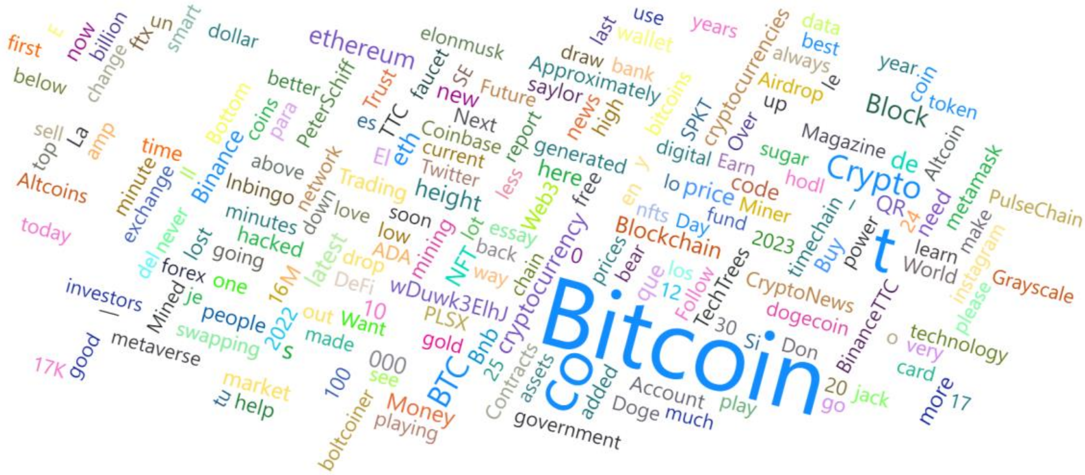


Figure 3: Word Cloud Twitter example

2.2 Sentiment Analysis

Opinion analysis, also known as sentiment analysis or the voice of the crowd/customer or emotion AI, is the field where feelings and opinions are classified, usually with the labels Positive, Negative, and Neutral. The aspects that are analyzed are described in the areas below[21].

2.2.1 Fine-grained sentiment

The customer's feedback sentiment, which makes up most of this activity, is the most straightforward to comprehend in terms of sentiment analysis. It is used chiefly for rating and review analysis. This kind of feedback typically falls into numerous categories, such as the star rating system (1-5, where 1 denotes a highly positive state, 2 a positive state, 3 a neutral state, 4 a negative state, and 5 a very negative one.[21]

2.2.2 Emotion Detection

This category's purpose is implied by its name, which also aids in identifying the emotion that the texts are trying to convey. The common ones include rage, sadness, joy, gratification, fear, panic, worry, or anxiety.[21]

2.2.3 Aspect based

This sentiment analysis method concentrates more on the features of a particular good or service. Let's use an LED television as an example to make it simpler to understand. To better understand the problems with the product and make it better and more practical, the manufacturing business can request input on light, sound, picture quality, or durability.[21]

2.2.4 Intend Analysis

We can learn more about a customer's intent by using this technique. We can comprehend whether a customer only seeks product details or is interested in making a buy. We can take note of, follow, or create a pattern using the intent analysis. Target marketing can take advantage of this information.

2.2.5 Sentiment Analysis - Existing Models

More and more models are populated nowadays, however in [21] 21 state-of the-art-models were analyzed.

Afinn. It was developed by Finn Årup Nielsen, uses a lexicon-based methodology, and includes 3,382 positive and negative words. Each word is assigned a positive or negative score. Afinn's range ranges from -5 to 5.[22]

TextBlob. A lexicon and rule-based sentiment returns the text's subjectivity and polarity and has more than 2,500 words. The polarity range is from -1 (very negative) to 1. (Extremely positive).

GoogleNLP. GoogleNLP, as its name suggests, is a Google-owned service with an easy-to-use API. For one month, Google offers an account without charge. The model is completely opaque to the user, and the sentiment score ranges from -1 to 1.

Vader. VADER is short for (Valence Aware Dictionary and sEntiment Reasoner).

It was developed at the Georgia Institute of Technology by C.J. Hutto and E.E. Gilbert. It is a sentiment model with a lexicon and rules designed specifically for social media texts. Each of its more than 9,000 words was given a number between -4 (very unfavorable) and 4 (extremely positive), and the final score is the average of all 10 scores.[23]

AWS. By offering a variety of services, Amazon is steadily expanding its influence in the machine learning and deep learning industries. One of the services designed exclusively for Natural Language Processing is AWS understand (NLP). Additionally, it is a complete black box for the customer, and Amazon offers a free trial account for one month.

Aigents. Aigents is a "interpretable" model based on "n-grams" that is written in Java and is included in the "aigents-java" distribution at https://github.com/aigents/aigents-java. It includes "out-of-the-box" vocabularies for n-grams linked to positive and negative sentiment. It returns the overall sentiment/polarity of the text based on the frequencies of occurrences of the reference n-grams in the text as well as independent positive and negative sentiment metrics. It has over 8,200 negative and over 3,800 positive n-grams. One of the model's characteristics is the use of the "priority on order" principle, which is described in [24]. When an n-gram is matched, all matches of any other n-grams that are components of the matched n-gram are ignored. This is called precedence granted for n-grams with higher "n" in the implementation particular to Aigents. If the tetragram "not","a","bad","thing" is matched, for example, both the bigram "bad","thing" and the unigram "bad" previous are ignored and dismissed. Similar to the example, the matching bigram "no", "good" ignores and undervalues both of its constituent unigrams, "no" and "good." Furthermore, the model features a setting that allows for the logarithmic scaling of the counted frequencies, and our investigations have shown that activating this setting improves performance.

Bert based models. In [21] the usage of 15 BERT-based models were trained using different data.

Distilbert-base-uncased.

It is a streamlined version of BERT that retains 97% of BERT's language understanding while being 40% smaller and 60% faster, trained on the same dataset as BERT, English Wikipedia and Toronto Book Corpus[25]

finiteautomata/bertweet-base-sentiment-analysis.

The base is built in BerTweet, which is a RoBERTa model. And trained on SemEval 2017 corpus (around ~40k tweets)[26]

cardiffnlp/twitter-roberta-base-sentiment.

The base model used was RoBERTa trained on ~58M tweets and with the TweetEval benchmark is is finetuned for sentiment analysis. It's based on the Google's Bert Model that released in 2018, and was proposed by inhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov[17]

ProsusAI/finBERT.

The BERT basic model, which was developed to analyze financial texts, was in heavy use. Financial PhraseBank, which contains 4845 sentences from finance news, the TRC2-financial dataset, which contains 400K sentences, and the FiQA Sentiment dataset were used to train it.[27]

moussaKam/barthez-sentiment-classification.

A seq2se2 model for French, built using BERT as a base[28]

textattack/bert-base-uncased-imdb.

TextAttack is a Python framework developed by the researchers for adversarial training and data augmentation in NLP. BERT, a base model that was trained using IMDB data - taset[29]

initeautomata/beto-sentiment-analysis.

TASS 2020 corpus was used to train a transformer-based library for sentiment analysis, emotion analysis, and hate speech identification. (around ~5k tweets)

siebert/sentiment-roberta-large-english. This model, which is a refinement of RoB-ERTalarge, was tested using data from 15 different datasets.[30]

sagorsarker/codeswitch-spaeng-sentiment-analysis-lince.

a BERT-based model for sentiment analysis, name entity recognition, and language identification. It may be utilized with mixed languages, including English, Spanish, Hindi, and Nepali, and was trained using the LinCE dataset.[31]

aychang/roberta-base-imdb.

The IMDB dataset was used to train the roBERTa base model.[30]

rohanrajpal/bert-base-multilingual-codemixed-cased-sentiment.

Based on the SAIL 2017 dataset, the bert-base-multilingual-cased base model was tuned.[32]

abhishek/autonlp-imdb_sentiment_classification-31154.

IMDB dataset used to train a BERT-based model.

VictorSanh/roberta-base-finetuned-yelp-polarity.

Based on RoBERTa, this model was refined using the Yelp polarity.

severo/autonlp-sentiment_detection-1781580.

IMDB dataset used to train a BERT-based model. Model precision: 0.930 and accuracy: 0.9426.

mrm8488/distilroberta-finetuned-tweets-hate-speech

distilroberta-base. Hate speech detection has been fine-tuned using the tweets hate speech detection dataset.

2.2.6 Model Comparison

In [21], the performance of the models was assessed using the Pearson correlation coefficient for each of the four metrics (Positive, Neural, Negative and Contradictive which is SQRT(positive * ABS(negative))) between the values "predicted" by the model and the "ground truth" values over 490 reference tweets and posts. The ground Truth, simply refers to a manual classification that has been done by two independent reviewers of the test dataset.

The study shows that the correlation of the "out-of-the-box" Aigents model, "aigents," is about 0.33, and that after fine-tuning, The correlation for "aigents+" is 0.57. The terms "ensemble(all)" and "ensemble(top 3)" refer to the average metrics for all models (aigents+, aigents, and finBERT), respectively). It's important to acknowledge that different models can provide different results based on the dataset. RoBERTa's performance has no significant difference from other models such as the VADER.

2.3 Text Summarization

Automatic text summarization involves condensing a document while maintaining the essential information's meaning and content.[33] The two main ways that currently exist are the abstractive and extractive. The first technique suggests the creation of sentences,

which comes after key contents and ideas have been summarized, with the strength of paraphrasing the original content. Extractive, on the other hand, is choosing a group of words, sentences and phrases and creates, as the name suggests, an extraction. This often has the disadvantage of sounding less like a human, while on the other hand abstractive has a more natural tone.

Sequence-to-sequence, a paradigm for abstractive summarization that relies on encoder-decoder architectures based on and more recently Transformers, has gained popularity. The desire to gather additional summarizing datasets with more abstract summaries has grown over the past few years. Recent research has expanded the success of text production, including abstractive summarization, by utilizing such pre-training for Transformer-based sequence-to-sequence models. Study introduces the Gap Sentence Generation (GSG), which is a self-supervised task and is used to pretrain a transformer encoder-decoder on huge documents, known as Pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence models, or PEGASUS.[18]

Considering their 568M parameter model this model equals or outperforms other state of art models in text summarization. In addition to that, this model will be also used in the methodology below, because it further pushed forward the state of art by using a text corpus specialized in HugeNews.[18]

2.3.1 Text Summarization - Prior work

MASS[34] masked sequence-to-sequence generation has been suggested, which reconstructs a sentence fragment from the rest of the phrase. One sentence fragment was chosen at random.

UniLM [35] proposed training on three different language modeling tasks: sequence-to-sequence (word-level mask), bidirectional (word-level mask with next sentence prediction), and unidirectional (left-to-right and right-to-left).

T5[36] introduced C4, a sizable text corpus obtained from Common Crawl, which we also utilize in some of our models, and demonstrated the benefit of scaling up model size (to 11 billion parameters) and pre-training corpus. They also adapted the text-to-text framework to a range of NLP applications. With randomly distorted text spans of various span sizes and mask ratios, T5 was pre-trained.

BART[37] pre-trained sequence-to-sequence models using a denoising autoencoder. BART learnt to rebuild the original text by corrupting it with an arbitrary noise function. The text infilling noising tool for generation tasks uses single mask tokens to mask randomly sampled text spans.

In comparison to the models above, in PEGASUS [18] multiple whole sentences that sort ones are being masks. The same research successfully demonstrates the objective, that sentences are chosen based on importance, rather than randomly.

2.4 Limitations

It is quite interesting that over time, more and more research papers were written analyzing a few cryptocurrencies, basically for price prediction with various ways. For Sentiment analysis on the other hand, the most common case is the analysis of bitcoin. From my perspective, as the market has great fluctuations, it would be extremely important to consider that X predictions applied on Y conditions using Z rules, and take another step by showing, what is up with Streaming Data. How do these conditions evolve, and since we are analyzing the Voice of the Crowd how does this Voice change by the time? Was it always the case that the X Crypto affected the public, or it all started from the public audience, when someone advertised or announced the release of a new coin? What is the average affect and for how long does the public, or Twitter, talk about a significant event, and for how long does this public in X weight affect the price, if there is such a case.

This is something that is missing from the market, and in order for anyone to observe and to end up on a result, they would need to rerun a similar analysis from scratch, losing time on analysis historical data, without being able to see maybe the presence and to share this knowledge.

This tool that was created, has the potential to be the new Google for researchers, Traders, cryptocurrency analysts and so on and so on. Of course, this could be just the beginning, as further enhancements will be mentioned on the next chapters.

In regards to the sentiment analysis, users might need to take into consideration the sarcasm, the idioms, the negations and non-text data.

2.5 Real-Time analysis

In times of crisis, there might be a need to measure the effect of the crisis in real time. We might need to draw a strategy by checking both historical and real time data, but one is for sure; the more you learn the more you earn. Combing this quote of Warren Buffet with the speed of grabbing this news, the more affective can be any type of strategy, creating a loop of observe – orient – decide – act – observe.

In the business world, operational decision support systems now include interactive analytical dashboards. Such features' analytical components resolve optimization issues that are invisible to the human eye, and their interactive components actively include users in the optimization process using graphical user interfaces (GUIs). Although interactive analytical dashboards in operational DSS are supposed to be valuable to enterprises, little is known about their efficacy or how they affect people's cognitive capacities. By investigating and empirically assessing the effects of interactive analytical dashboard features on situation awareness (SA) and task performance in operational DSS, this research [38] helps closing this gap, proving how useful the dashboards are especially in real time decision making.

3 Methodology

3.1 Research Model

On this chapter, we analyse the steps that were taken in order to give some body to our research. Many approaches were tried during this development, though only some of these will be mentioned, just for their limitations. Starting from Step 1, it is always the data.

- The source,
- the size,
- the quality
- the refresh rate in our case had to be analyzed before anything else.

Since we are talking about real time data, we are talking about performance, speed, big data, technologies behind, system Specifications. To take it one step further, we use also historical data for our model. So by combining historical and real time data, we understand that this is one problem, that an action had to be taken quickly, otherwise the data size would hit the system specifications, causing slow performance and unstable results.

Starting from sources, for the voice of the crowd it's used Twitter Data, Yahoo finance News and Newspaper3k API. For CryptoCurrencies Price, we are scraping the prices from CoinGecko in real time.

The Data Retrieval, Preprocessing, Analyzing and Storing is being done with The Python Programming Language. Python then stores this data, using SQL Server Management Studio. In order to further exploit the analysis and to make this public, SQL Server is connected with Power Bi which later is published to the web and embeded in a 3rd party site, allowing the users to join without even using any type of credentials.

For automating this procedure, Python scripts are in the loop feeding and cleaning the database almost every minute. Power BI detects these changes and gets updated in real time. All these are included in Figure 4.

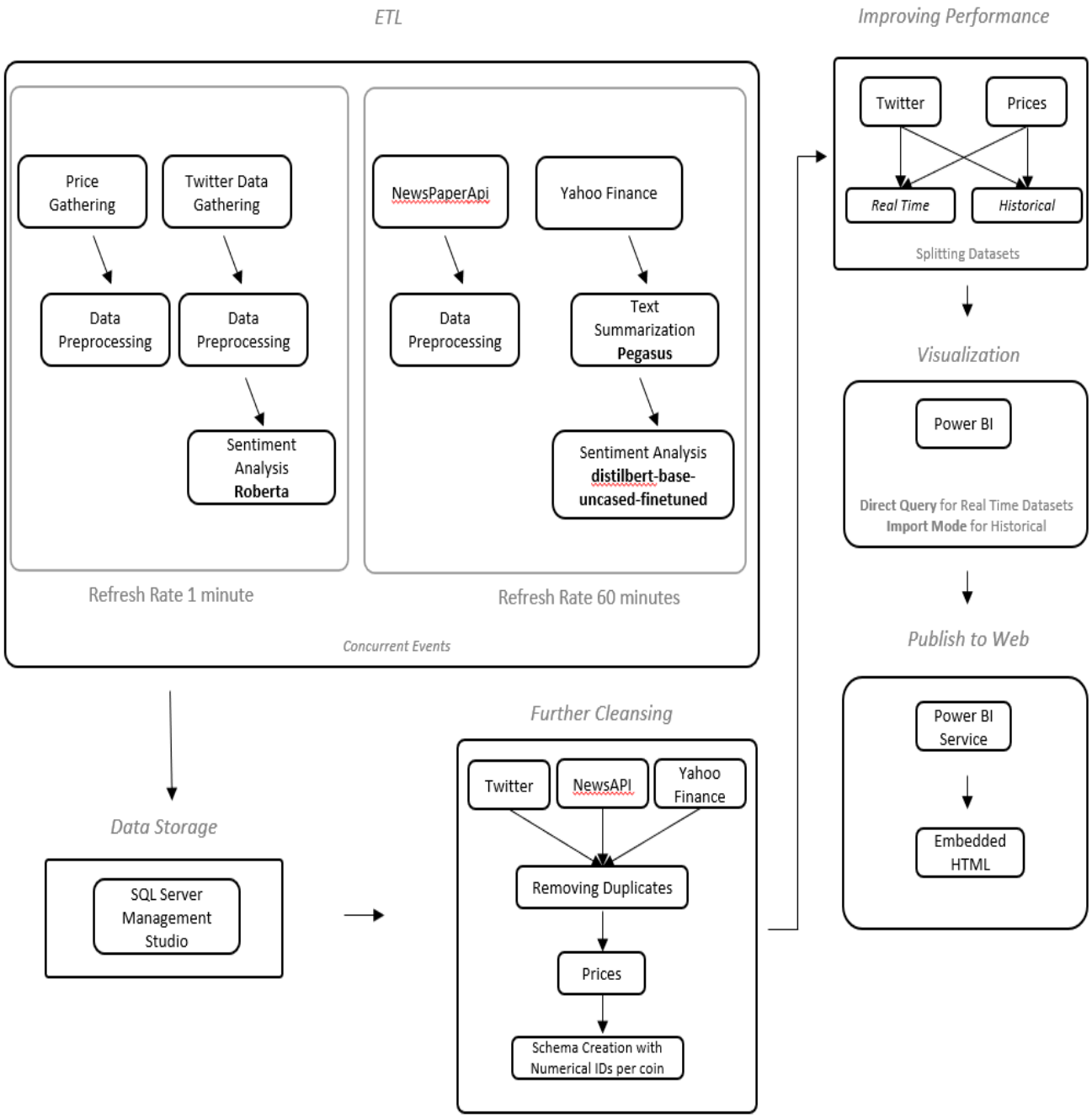


Figure 4: A general overview of The Methodology of the Research

4 Data Collection

4.1 Twitter Data Collection

In order collect Data from Twitter, a few ways were tried. First way, was the Tweepy Library for which there is a need of Developer's Twitter account to be created. In addition to that, there were a few limitations on this library such as, the query could not expand beyond 7 days from the current date and the number of rows was limited to 100 per query. Since, there was need to collect also historical data, this way was easily rejected. Second way, was the Library Twint but the cons of Snscrape Library were greater in terms of data interoperability, if anyone would consider that the latter can be connected with

- Twitter
- Instagram
- Reddit
- Facebook
- Weibo
- Telegram
- Mastodon

In addition to that, the way of structuring the query was a simple "copy paste" from twitter's query, so it makes it a little bit easier for any researcher to write the query in Python, the same way that they would be filtering their Twitter's feed in real life. Taking all the above into consideration, the query that was used, included the following parameters, where "until_date" and "since_date" were dynamic.

"(btcusd OR bitcoin OR ethereum OR ethusd OR Litecoin OR ltcusd OR bchusd OR etcusd OR zrxusd OR xrpusd OR xlmusd OR eosusd OR repusd OR solana or Crypto or Ctyptocurrency) min_replies:1 min_faves:10 min_retweets:1 until:" + str(until_date) + "since:" + str(since_date)

More keywords could be included in the query, and this is something that could change on the fly, as the script is running every minute automatically. In order to speed up the performance a little bit, the number of tweets that are being collected every minute is set to 100, out of which almost 30% are dropped due to low engagement score. It is worth to mention, that this limit can be adjusted, and any researcher can adjust this, allowing to search for tweets even from 2012.

Engagement score that was mentioned above, was used only for filtering out tweets in the data retrieval using a simple rule: If Count of likes, retweets and Replies equals 0 then this would be dropped.

The remaining tweets were stored into SQL Server Management Studio 2019. An example of the code is displayed below in Figure 5.

```
limit: int = 100
master_df = pd.DataFrame()
for company,query in queries:
    tweets = []

    o = 0

    for tweet in sntwitter.TwitterSearchScraper(query).get_items():

        if len(tweets) == limit:
            break
        else:
            query_field = query.split("until")[0]
            tweets.append(
                [company, query_field, tweet.date, tweet.username, tweet.content, tweet.likeCount, tweet.retweetCount, tweet.replyCount, tweet.coordinates,
                 tweet.hashtags, tweet.user.followersCount, tweet.user.favouritesCount, tweet.sourceLabel, tweet.quotedTweet, tweet.user.friendsCount, tweet.user.location, 0])
        o = o+1
        print(o)
    df = pd.DataFrame(tweets, columns=['Company','Query','Date', 'User', 'Tweet', 'likeCount', 'retweetCount', 'ReplyCount','Coordinates', 'Hashtags','Followers', 'Favorites', 'sourcelabel','Quotes','friends','location','Score'])
    df['Date2']= df['Date'].apply(lambda a: pd.to_datetime(a).date())
    df = df.loc[df['Followers'] > 0]
    df['Weight'] = (df['likeCount'] + df['retweetCount'] + df['ReplyCount'] + df['Followers'])
    df = df.loc[df['Weight'] >0]
    df['Score'] = ((df['likeCount'] + df['retweetCount'] + df['ReplyCount']) / df['Followers'])*100
    df = df.loc[df['Score'] > 0]
    master_df= master_df.append(df, ignore_index=False)
```

Figure 5: Code Snippet for gathering Twitter Data with Python

Considering a few system limitations, such as unstable internet connection and old hardware, the average number of collected tweets per 30 minutes was from 200 to 800. This number consists of tweets that has already been preprocessed, sent for sentiment analysis and stored in SQL Server, while at the same time, other sources that we will review below where scrapped, summarized and used for our analysis

An example of the tweets that were collected in a specific day, in November 2022 is displayed in Figure 6.

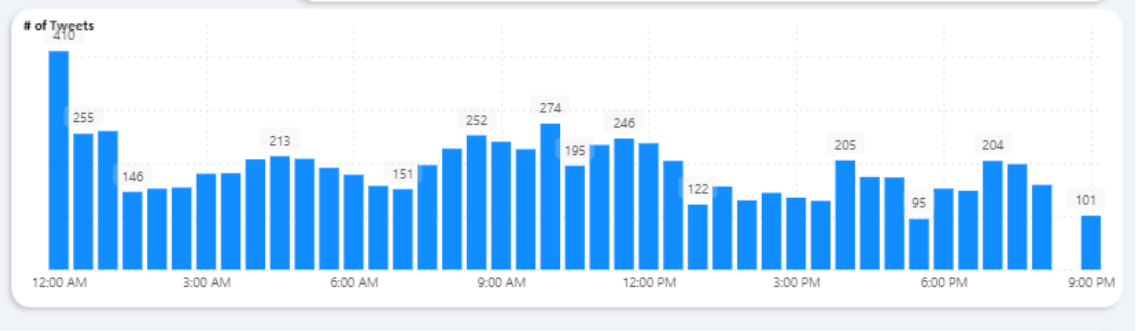


Figure 6: Count of Tweets, with sentiment every 30 minutes. Date: 25/11/2022

The information that was gathered with Snscrape from Twitter is displayed in the table 1 that is attached below:

	Renamed	**Status**	**Short Comments**
Company	Company	In use	Used to dynamically select queries
Query	Query	In use	
Tweet.date	Date	In use	Timestamp
Tweet.username	User	In use	
Tweet.Content	Tweet	In use	
Tweet.likecount	likeCount	In use	
Tweet.retweetCount	retweetCount	In use	
Tweet.replyCount	ReplyCount	In use	
Tweet.coordinates	Coordinates	Dropped	No significant info
Tweet.hashtags	Hashtags	Dropped	
Tweet.user.followersCount	Followers	In use	
Tweet.user.favouritesCount	Favorites	In use	
Tweet.sourceLabel	Source	Dropped	
Tweet.quotedTweet	Quotes	Dropped	
Tweet.user.friendsCount	Friends	In use	
Tweet.user.location	Location	Not in use	No significant info
Weight	**Weight**	Dropped	Calculated Column
Positive		In use	Merged on 2nd stage
Neutral		In use	Merged on 2nd stage
Negative		In use	Merged on 2nd stage

Table 1: Columns that retrieved from Twitter, calculated columns and columns from Sentiment Analysis Script merged.

4.2 Crypto Price Data Collection

In order to gather Price Data, various ways were used such as Binance APIs, CoinGecko API, but in order to avoid any future limitations the usage of BeautifulSoup and Requests library was a dealbreaker. These two libraries are scrapping data every minute from CoinGecko, the world's largest independent cryptocurrency data aggregator that has over 200 million monthly page views, and 10 million + Monthly users.

A piece of code is also attached below, in the Figure 7:

```
#CRYPTO API

r = requests.get('https://www.coingecko.com/?locale=en&page=1')
df4 = pd.read_html(r.text)[0]
df4 = df4[['Coin', 'Price', 'Mkt Cap']]
df4['Coin'] = df4['Coin'].replace(r'\s+', ' ', regex=True)
df4['Price'] = df4['Price'].apply(lambda x: x.replace(",", "").replace("$", "")).astype(float)
df4['Mkt Cap'] = df4['Mkt Cap'].apply(lambda x: x.replace(",", "").replace("$", "")).astype(float)
df5 = pd.DataFrame(df4)
df4['UpdateTime'] = pd.to_datetime('now') + pd.Timedelta(hours=2)
# df['UpdateTime'].map(lambda x: pd.to_datetime(x, yearfirst=True).tz_convert('Europe/Athens'))
df4['Date'] = df4['UpdateTime'].dt.date
df4['Hour'] = df4['UpdateTime'].dt.hour
df4['Minute'] = df4['UpdateTime'].dt.minute
df4.to_sql(name='API_Crypto3', con=engine, index=False, if_exists='append')
```

Figure 7: Scrapping Data from CoinGecko Page 1. More Pages can be included.

Considering that this part of code could run every a few seconds on a different server, only a few columns where stored in the database. Of course, other columns such as open, high, low, close can be easily calculated using a simple aggregation. On the other hand, more than 1 page could be tracked and stored in the same database using a for each loop. With a similar way, data can be retrieved from various sites overcoming some future problems, such as hacked sites, which can be a case in cryptocurrency world from time to time.

Example of the table that is being stored in our database is being displayed in Figure 8:

	Coin	Price	Mkt Cap	UpdateTime	Date	Hour	Minute
1	Bitcoin BTC	19363.43	371105587372	2022-10-24 22:04:31.910	2022-10-24	22	4
2	Ethereum ETH	1344.13	161460527518	2022-10-24 22:04:31.910	2022-10-24	22	4
3	Tether USDT	0.999709	68433219160	2022-10-24 22:04:31.910	2022-10-24	22	4
4	BNB BNB	273.55	44641515973	2022-10-24 22:04:31.910	2022-10-24	22	4
5	USD Coin USDC	1	43809428228	2022-10-24 22:04:31.910	2022-10-24	22	4
6	XRP XRP	0.456853	22735974874	2022-10-24 22:04:31.910	2022-10-24	22	4
7	Binance USD BUSD	1	21798247498	2022-10-24 22:04:31.910	2022-10-24	22	4
8	Cardano ADA	0.361287	12636505603	2022-10-24 22:04:31.910	2022-10-24	22	4
9	Solana SOL	28.45	10181009947	2022-10-24 22:04:31.910	2022-10-24	22	4
10	Dogecoin DOGE	0.059477698184	8116624480	2022-10-24 22:04:31.910	2022-10-24	22	4
11	Polkadot DOT	5.9	6851128601	2022-10-24 22:04:31.910	2022-10-24	22	4
12	Polygon MATIC	0.898229	6672380363	2022-10-24 22:04:31.910	2022-10-24	22	4
13	Lido Staked Ether STETH	1340.01	5991648742	2022-10-24 22:04:31.910	2022-10-24	22	4
14	Shiba Inu SHIB	9.965027E-06	5864264475	2022-10-24 22:04:31.910	2022-10-24	22	4
15	Dai DAI	0.999327	5784963119	2022-10-24 22:04:31.910	2022-10-24	22	4
16	TRON TRX	0.061817737409	5694345681	2022-10-24 22:04:31.910	2022-10-24	22	4
17	Wrapped Bitcoin WBTC	19394.13	4737744349	2022-10-24 22:04:31.910	2022-10-24	22	4
18	Avalanche AVAX	15.79	4681077753	2022-10-24 22:04:31.910	2022-10-24	22	4
19	Uniswap UNI	6.18	4642326985	2022-10-24 22:04:31.910	2022-10-24	22	4
20	OKB OKB	15.95	4013035933	2022-10-24 22:04:31.910	2022-10-24	22	4

Figure 8: Stored Data in SQL Server. UpdateTime is being calculated

4.3 Yahoo Finance News Data Collection

This part of data retrieval is the most complex and insightful, as the usage of it could benefit any researcher, trader and so on. This piece of code simply:

1) opens yahoo Finance at Google every 60 minutes.
2) then it searches for some specific tickers (more than 1), such as Bitcoin, Ethereum, Crypto,
3) then it collects and cleans the URL's that are given by Yahoo Finance News
4) then it opens these URL's one by one, summarizing their texts using abstractive state of art model,
5) next it finds the sentiment score using the sentiment Pipeline that is a distilbert-base-uncased-finetuned model,
6) and finally, it stores this data to SQL server as well.

It is worth to be mentioned that this is running every 1 hour in order to get the first page of Yahoo Finance. The procedure is fast enough, if anyone would consider that all news related to the field of interest can be summarized in less than one minute. This dataframe includes also the sentiment analysis calculation of each given URL.

It goes without saying that, the time of data retrieval could be adjusted along with the page number of Yahoo Finance, but for our case the first page contains always the hottest news, so this is the reason that only that was retrieved. The 1 hour was set as default time of refresh, because it was noticed that the news are not changing in the same refresh rate as the tweets for example.

The Model that was used, will be further analyzed on the next chapters. Last but not least, this part of code is running sequentially to the other scripts that are used for scrapping prices, tweets and since this is running on the backend part, the end-user of the tool won't have to run anything from their end. Instead, the look and feel of the Frontend in Power Bi, is guiding the user to filter the results that have been already gathered in the SQL-Server and published to Power BI, giving the option to analyze both Historical and Real Time queries. The longer this procedure runs, the greater insights will gather and provide to any researcher.

The Yahoo Finance News Data collection could be easily separated as a unique piece of code, that helps the researchers or any user to collect the news of their favorite topic and summarize them in a short timeframe. It could be also applied to different topics not related to cryptocurrency

To better understand this procedure, a flowchart is being displayed in Figure 8 below:

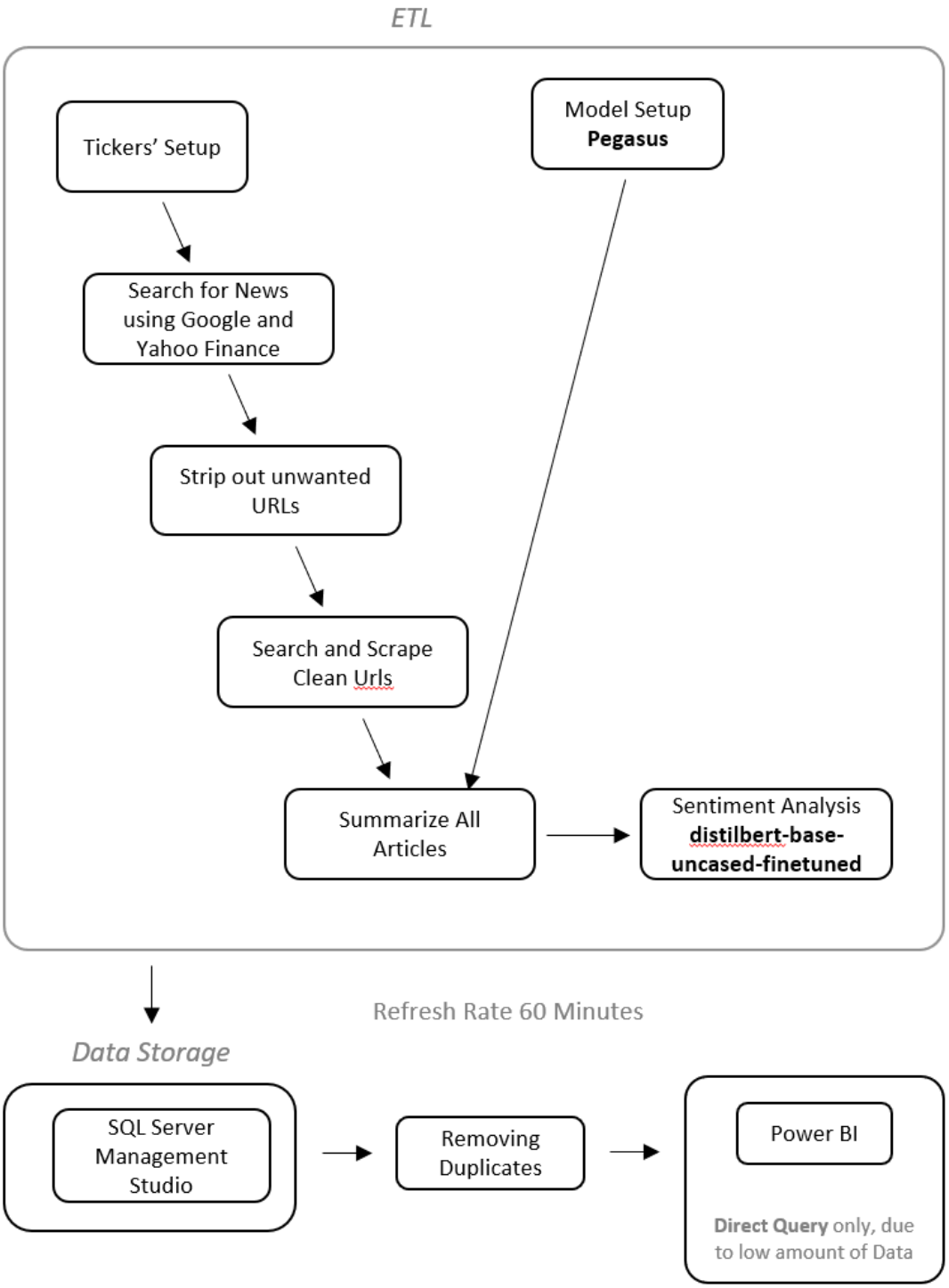


Figure 8: Yahoo Finance Scrapping, Summarizing, Applying Sentiment Score

The final Outcome of this script is displayed in Figure 9.

Summary	Publised	URL	Sentiment
Asia's Matrixport is raising $100 million at $1.5 billion valuation. Firm has no risk of insolvency with respect to FTX-linked products	2022-11-25T02:10:28Z	https://finance.yahoo.com/news/crypto-lender-seeks-1-5-021028564.html	POSITIVE
FTX, a cryptocurrency exchange, has filed for bankruptcy. FTX was one of the largest cryptocurrency exchanges in the world	2022-11-24T23:44:05Z	https://finance.yahoo.com/news/15-biggest-cryptocurrency-exchanges-world-234405631.html	NEGATIVE
Cathie Wood still holds her $1 million per coin forecast. Grayscale Bitcoin Trust is another fund to enter the market	2022-11-24T23:00:00Z	https://finance.yahoo.com/news/crises-cathie-wood-says-crypto-230000231.html	NEGATIVE
Exchange plans to invest $1 billion in recovery fund. FTX collapsed last month amid liquidity crisis in crypto sector	2022-11-24T16:48:23Z	https://finance.yahoo.com/news/1-binance-commit-1-bln-164823854.html	NEGATIVE
About 100 people lost their jobs, CEO Cavazzoli said. Exchange put plans to expand into Latin America on hold	2022-11-24T16:10:17Z	https://finance.yahoo.com/news/argentine-crypto-exchange-lemon-sheds-161017812.html	NEGATIVE
Digital asset has rallied 43% in the past month. Bitcoin, ether and Ethereum all fall in the past week	2022-11-24T14:47:05Z	https://uk.finance.yahoo.com/news/crypto-token-crash-ftx-collapse-litecoin-doge-solana-144705733.html	NEGATIVE
Financial advisors need to understand best practices. FTX is not the first time a crypto exchange has failed	2022-11-24T13:45:00Z	https://finance.yahoo.com/news/crypto-self-custodying-offers-protection-134500250.html	NEGATIVE

Figure 9: Final Outcome and visualization of Yahoo Finance Summarization and SA

4.4 NewsPaper API Data Collection

Newspaper3k API was firstly included in the analysis in order track any correlation with the Prices, though, because of the limitations (low number of API call) this was just retrieved and placed into the frontend, for a quick news update removing any significance from its content. Users can have a quick overview over the top headlines with the following query:

q='crypto OR BTC OR Bitcoin OR btcusd OR cryptocurrency OR Ethereum OR ETH OR Tether OR "Binance Coin" OR BNB OR XRP OR OR FTX OR Solana OR Dogecoin OR Polkadot',

4.5 Order Book Data Collection

Order Book Analysis was one of the latest improvements in the report, and the results could be used in place of the selected crypto-price. In general, the order book shows the volumes and prices that people would place into an order. As many orders could have been not taken place, this could provide another point of view to the end user, checking possible correlation with Sentiment analysis and Weighted bids and asks price per update time. For our case, this type of data includes for now only the BTCUSTD and it's connected with Binance using the requests library.

Example of this Data Collection is attached in Figure 10

```
symbol = "BTCUSDT"
url = "https://api.binance.com/api/v3/depth"

interval = Decimal('0.1')

params = {
    "symbol": symbol,
    "limit": 100,
}

data = requests.get(url, params).json()
bids = pd.DataFrame(data["bids"], columns=['price', 'quantity'], dtype=float)
bids['type'] = "bids"

asks = pd.DataFrame(data["asks"], columns=['price', 'quantity'], dtype=float)
asks['type'] = "asks"
```

Figure 10: Order Book Data Retrieval

4.6 Data Cleaning and schema Creation

Starting from the Prices in CoinGecko, since each site, might include some unwanted symbols and/or spaces the cleaning of unwanted symbols was performed while retrieving this data, as shown in Figure 7. As almost 100 Rows are imported into the Database every minute, repeating some strings, and especially the Coin Names is affecting the performance. For this reason, a simple procedure is storing all Distinct Coins from the Staging Table into a new Dimension Table which is called Dim_Crypto, providing a unique ID

for them. On the next step, the staging Table is joined with the Dim_Crypto and this generates a new Table which contains the ID from the Dim_Crypto and the Price, Update Date from the STG_Table.

Procedure is called through python, using SQL Statements. The procedure and script are attached in the Figures 11 and 12 respectively.

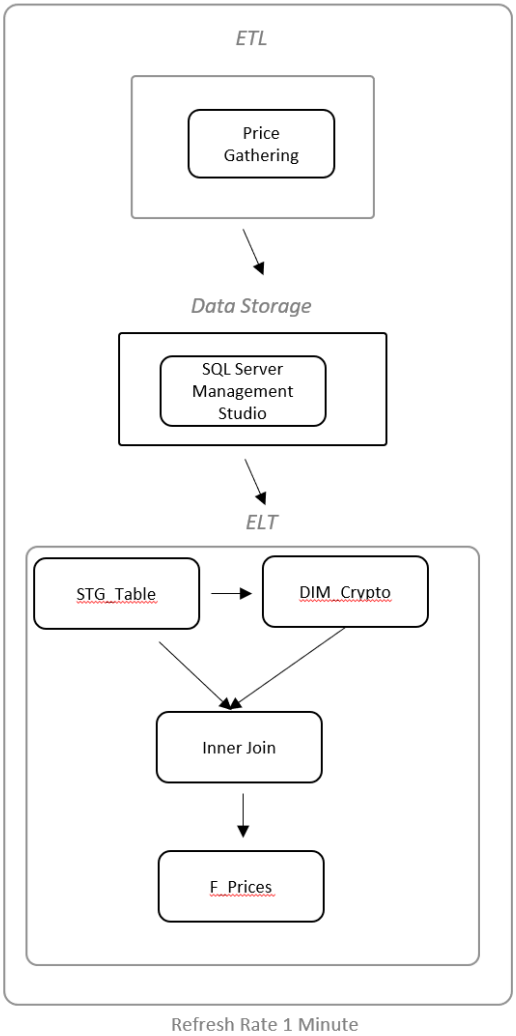


Figure 11: Schema Creation

```
begin try
    drop table DIM_Crypto
end try
Begin Catch
    --No action
End Catch

begin try
    drop table F_Crypto_Prices
end try
Begin Catch
    --No action
End Catch
```

```
SELECT distinct coin, identity(int, 1,1) as ID into DIM_Crypto FROM [THESIS].[dbo].[API_Crypto3]

select b.ID,Price,[Mkt Cap], [UpdateTime],[Date],[Hour],[Minute]
into F_Crypto_Prices from API_Crypto3 a join DIM_Crypto b on a.Coin = b.coin
```

Figure 12: Snippet of SQL Statements used to replace Coins with IDs

Cleaning the tweets was an entirely different procedure, as the better preprocessing the more accurate the model would work. For this reason, two major functions were created, checking one by one each word that is included in every tweet. The scenarios that had to be addressed along with the removal of spaces, hashtags, extra lines and symbols were to identify if the word is also the username or a URL. After joining all clean tweet words into one sentence, this sentence was ready to be encoded and the imported into the model for sentiment analysis.

Figure 13 describes the step of preprocessing what mentioned above.

```
def process_tweet(tweet):
    tweet_words = []

    for word in tweet.split(' '):
        re.sub('@[A-Za-z0-9_]+', '', word)
        re.sub('#', '', word)  # removes hastag '#' symbol
        re.sub('RT[\s]+', '', word)
        re.sub('https?:\/\/\S+', '', word)
        re.sub('\n', ' ', word)

        if word.startswith('@') and len(word) > 1:
            word = '@user'

        elif word.startswith('http'):
            word = "http"
        tweet_words.append(word)
    tweet_proc = " ".join(tweet_words)



    #sentiment analysis

    encoded_tweet = tokenizer(tweet_proc,return_tensors = 'pt')
    output = model(**encoded_tweet)

    scores = output[0][0].detach().numpy()
    scores = softmax(scores)
    score_dict ={}
    for i in range(len(scores)):
        l = labels[i]
        s = scores[i]
        score_dict[l] = s


    return  score_dict
```

Figure 13: Preprocessing Tweets

As preprocessing, finding sentiment score and storing these tweets into the Database were sequential, one more SQL statement was included in the Python Script in order to possible duplicates. All steps that mentioned until now in this chapter are running with a refresh rate 1 minute. SQL Statement is included in the Figure 14.

```
;WITH CTE3 AS(
   SELECT
    Company
   ,Query
   ,[Date]
   ,[User]
   ,[Tweet]
   ,[Negative]
   ,[Neural]
   ,[Positive]
    ,RN = ROW_NUMBER()OVER(PARTITION BY  Company
   ,Query
   ,[Date]
   ,[User]
   ,[Tweet]
   ,[Negative]
   ,[Neural]
   ,[Positive] ORDER BY [Tweet] )
   FROM [THESIS].[dbo].[Tweetó]
)
DELETE FROM CTE3 WHERE RN > 1
```

Figure 14: Removing Duplicates from Tweets Table

4.7 Automated Data Retrieval

The basic concept of this tool, is to analyze the sentiment analysis in real time. With that said, the process of collecting data, analyzing them and visualizing them had to be as closer to the real time as possible. All scripts from above on the first stage were running while someone had to push to execute button, though, this didn't create any data stream. For this reason, data retrieval was later included into a function that was scheduled to run every a few seconds.

Due to the nature of each data-source, some datasets had to be retrieved with a different refresh rate. For this reason, two different functions were created including different tasks. The functions that will be analyzed below are the Def API_runner2 and Def API_runner.

Starting for the API_runner2, this function was built in order to retrieve data from sources, where the refresh rate could be more flexible. It runs the parts that were mentioned above and are related to

- NewsPaper API
    - SQL Server Feed

- Yahoo Finance News
    - Yahoo Finance Sentiment Analysis
    - SQL Server Feed

The remaining function, API_runner, was built in order run everything in a strict refresh rate. This function starts from

- Crypto Price Data Collection
    - Data Cleansing
    - SQL Server Feed

- Twitter Data Collection
    - Twitter Data Cleansing
    - Twitter Sentiment Analysis
    - SQL Server Feed

- Order Book Data Collection
    - Data Cleansing
    - SQL Server Feed

- Data Cleaning and Schema Creation
    - Splitting Datasets in Real Time and Historical

Figure 15 Describes how these two functions are automated

```
import os
from time import time
from time import sleep

for y in range(440):
    API_runner2()
    for i in range(30):
        api_runner()
        print('API Runner completed')
        sleep(60)  # sleep for x seconds

    print('API Runner2 completed')
    sleep(10) #sleep for x seconds
exit()
```

Figure 15: Automating the refresh Rate with different limits. Sleep time is adjusted in a different limit.

4.8 Hybrid Connection to Power BI

In this chapter, we will review the process that was followed in order to create a flow from the sources to the end user as well as the observations that helped on improving the performance.

Power BI is a Business Intelligence Tool, which is great for visualizing the findings providing useful insights to the end user. The product itself is free and can be easily downloaded from Microsoft. The downloaded version that runs on the PC or Mac is called Power BI desktop (current version: December 2022) and users can have a complete Data Analysis, report creation, transformation, and visualization. It can be connected with multiple sources (in our case it's connected with SQL Server) and can be distributed to other users via the Power BI service, where data governance can be applied, setting rules on which kind of Data is visible to each enrolled user to this report.

Figure 16 shows the flow that was described above

Figure 16: An overview of Power BI, Source: AnyonColsulting.com

Having an overview of what this tool can do in order to help the end-users get their results, it is interesting to deep dive on how to connect our sources to the model. Considering also that the database itself contains both Historical and Real-Time datasets, we might need to use different techniques in order to update the Power BI.

In the Power BI world, there are two major types of importing the data into the Power BI: The import mode and the direct query. The first way captures a snapshot of the data and stores this to the disk. When it's time to refresh or to run a query, the data are loaded into the memory of the Power BI. After that stage all query results are very fast.

More details in regards to the compression and optimization of this mode is that, every time that we refresh the data set, VertiPaq Storage Engine is compressing, optimizing and storing the data to the disk. It is being noticed that the 10 GB of data can be compressed to 1 GB in size. The reason that this mode cannot be used for real time is that there is a limit on the refresh rate, about 8-48 times per day depending on the Power BI capacity.

For this reason, the second mode which is the DirectQuery is used providing the advantage of:

- Unlimited Refresh Rate
- Real Time Reports using Automatic Page Refresh Feature and more.

A third type that is the Push Mode that is another way of streaming data especially from Rest-APIs which can be implemented as an alternative to the direct query, though it

requires more development knowledge as it uses the Powershell in order to push changes automatically to the Power BI.

Figure 17 shows ways to connect and refresh the Power BI

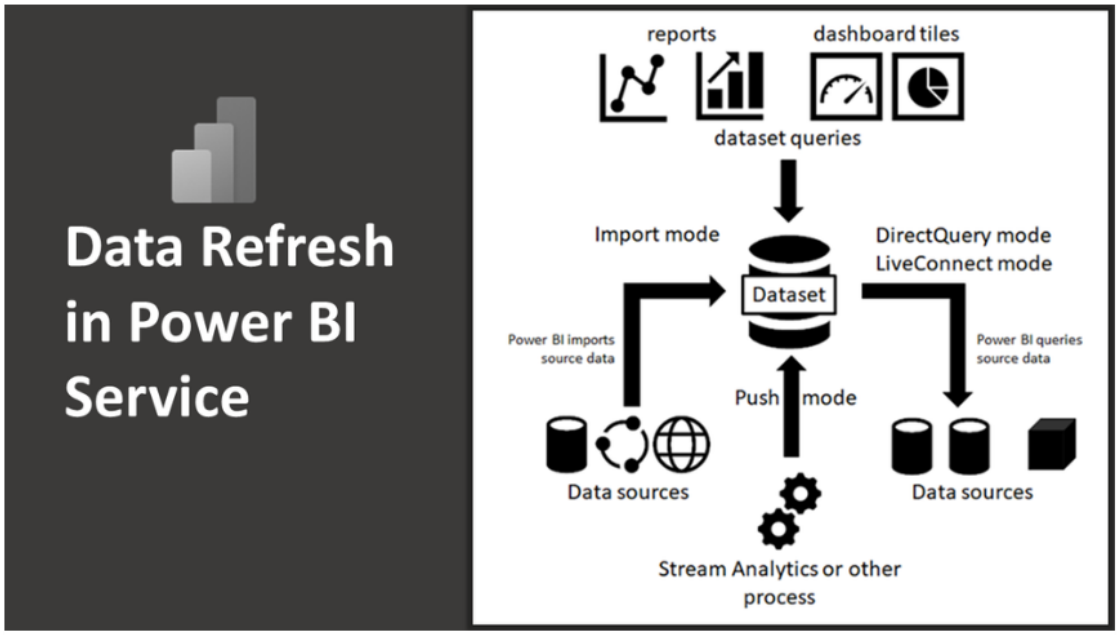


Figure 17: Power BI Refresh and Queries, Source: techcommunity.microsoft.com

Even though the concept of this project, might prompt you that DirectQuery is the way to go, that was also the first mistake that was noticed. Considering that every hour, thousands of rows are being added to the tables, and that Power BI is running queries on the same tables for every filter that you may apply on the Frontend, then performance is getting slower day by day. For instance, when this project started, each visual could have taken less than 5 seconds to be refreshed, but after 1 week of data, the same visual could have taken 10x more to be refreshed.

For this reason, all datasets, especially the heavy ones, had to be partitioned in Historical and Real Time. The partition of the same data will be further analyzed on the next sections, though for our scope, we need to consider that Historical Data will be stored in Power BI using the import mode, refreshing a few times per day. Real Time Data on the other hand are still using the DirectQuery mode.

With this way, performance was still maintained close to the first stage, allowing the users to better explore real time or historical trends with just a few clicks.

Figure 18 shows ways to hybrid mode that was followed

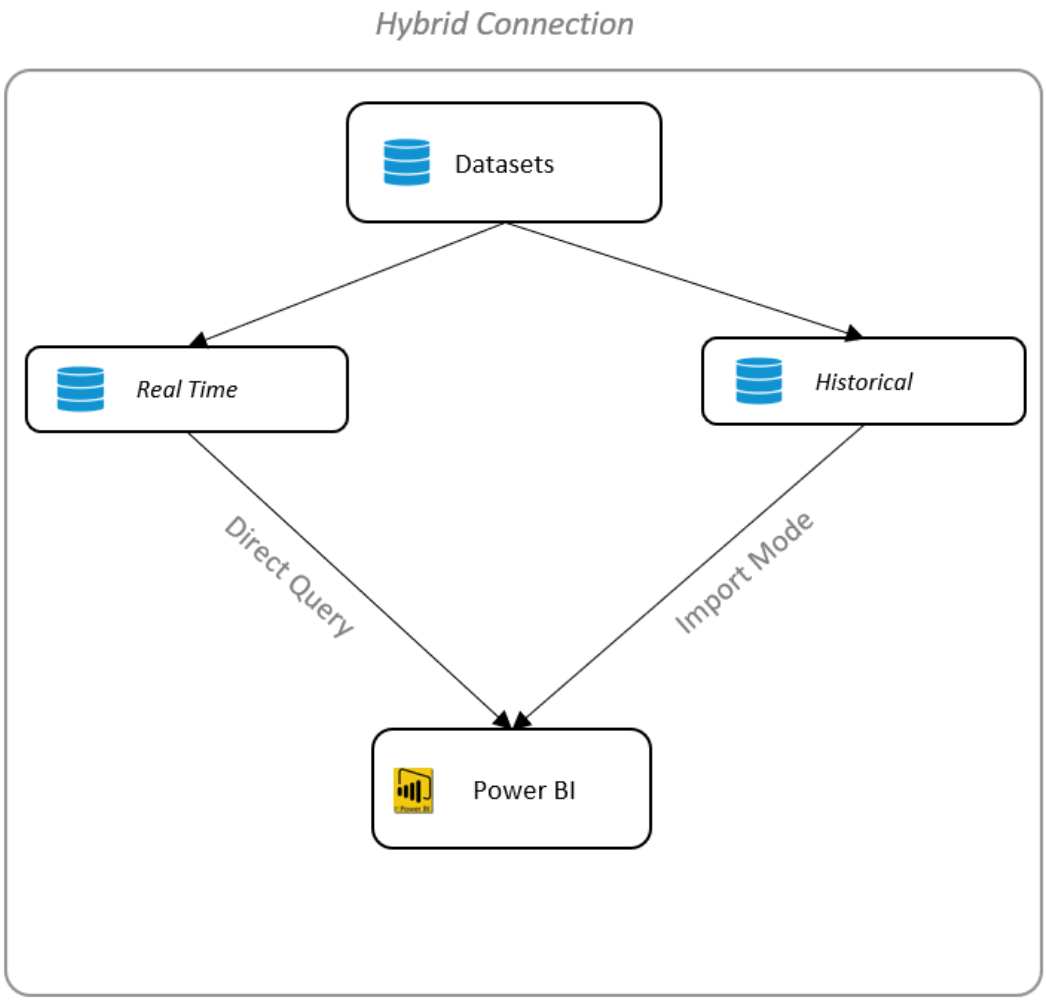


Figure 18: Power BI connection to sources

Taking a short timestamp from our dataset, the F_Crypto_Prices Table, from the 24th of October until the 3rd of December contains more that 3 million rows, while the F_Crypto_Prices_Real cannot exceed the 100 thousand per day, with the applied refresh rate. Similarly, for the same period almost 530 thousand tweets have been stored to the historical datasets, containing also the content itself, while the Real Time Dataset usually

contains less than 30 thousand tweets, taking into consideration also the cleaning that is applied on the original dataset.

In order to optimize the report, it is worth to be mentioned that all data cleaning is performed prior to the partitioning.

4.9 Real Time BI integration to the web

As the data model of Power BI has been already finalized, a few reports were created and published to the Power BI service. The latter offers some options to the report owner, such as, what is the refresh rate for the imported datasets, who has access to this report, and what is the Row Level Security that is applied. Since the purpose of this project is to be public, the part of credentials had to skipped. In order to achieve this, Power BI has the choice to publish the report to the web, using an iframe code that can be used for HTML embed wherever it is possible.

As a next step, and since the reports were published to the web, there was a need to gather all these into one webpage. For this reason, a draft domain was provided from wix.com in which users can visit and discover their findings, in real time, without passwords 24/7. The Project name that describes this webpage has been selected to be the Cryptofeel.eu, a name that combines the cryptocurrency with the sentiment analysis. The sections of this webpage are the following:

- Crypto
- News Summarization
- Sentiment Analysis
- Blog

An overview of the site is being displayed in Figure 19

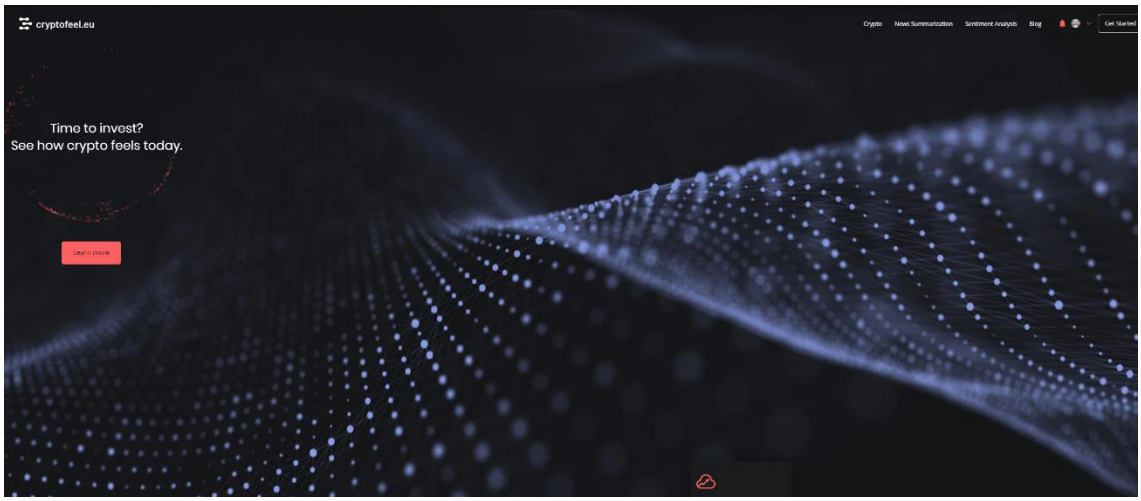


Figure 19: Landing Page and Context of cryptofeel.eu

The first section which is the crypto, is providing some real time and historical data for more than 100 Cryptocurrencies that have been scrapped so far. Some of the Key Point Indicators are being reflected in CoinGecko as well, but for our case some others have been included or modified.

Users, first, will get some information related to the CryptoFeel.eu, for instance the number of cryptocurrencies that have been tracked through the backend process, the number of Tweets, News and Articles that have been stored and analyzed.

Next, some trends and latest news are sitting at the top of the page, that can provide some useful insights on what the users need to focus on. At the middle part, which is the main focus of this report, users can select their coin of choice and check how this progresses at the last 24 hours.

The KPIs that are introduced on this section are the

- Price
- -24h Price *(Price at maximum UpdateTime of previous day)*
- EI *(Evolution Index = Crypto Growth divided by the Total Market Growth)*
- Change *(Crypto Growth)*
- Today's Lowest *(Price related)*
- Today's Highest *(Price related)*
- Mkt Cap
- Last 7 Days Trend *(Price related)*

An overview of this section is being displayed below:

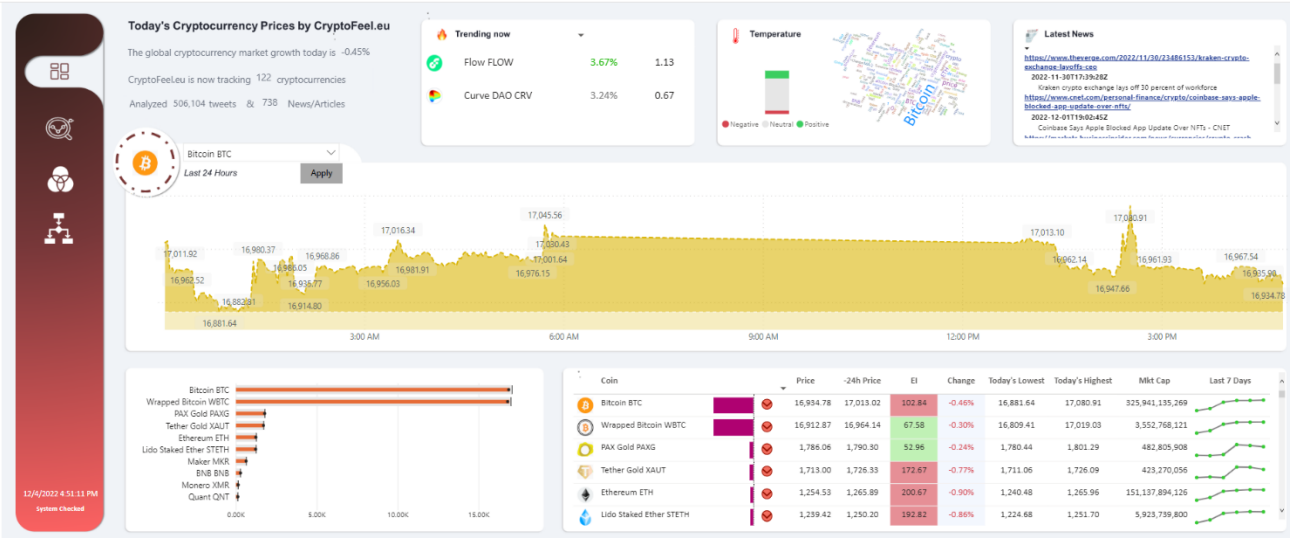


Figure 20: Crypto Section

Other subsections of Crypto include coin comparisons with Historical Data, where users can select different coins for Group A and Different for Group B and compare their historical Performance.

An overview of this section is being displayed below:

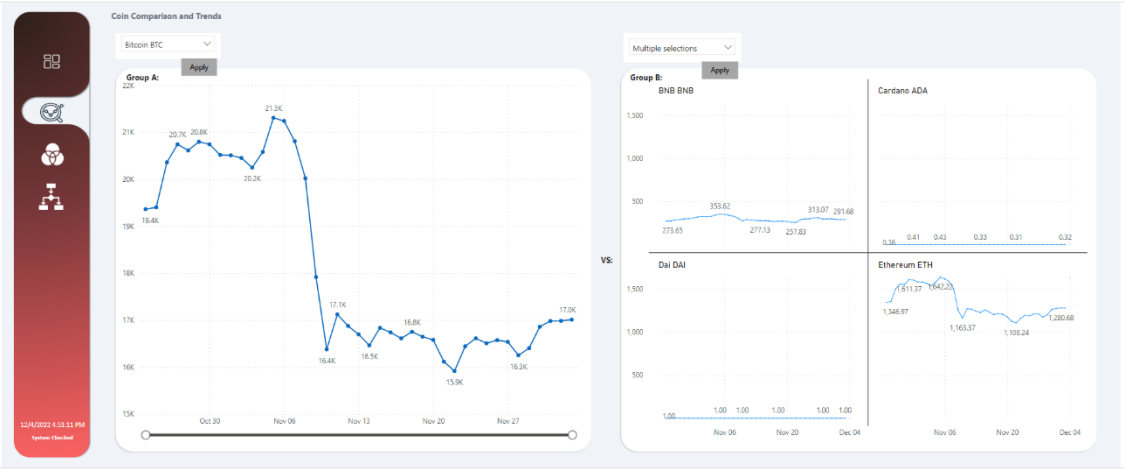


Figure 21: Coin Comparison

Another Subsection of this crypto analytics, is analyzing the

- Twitter Index versus Market Average price
- Average Time per Timeslot and selected coin

- Overall Performance

Figure 22 captures this analysis related to this subsection

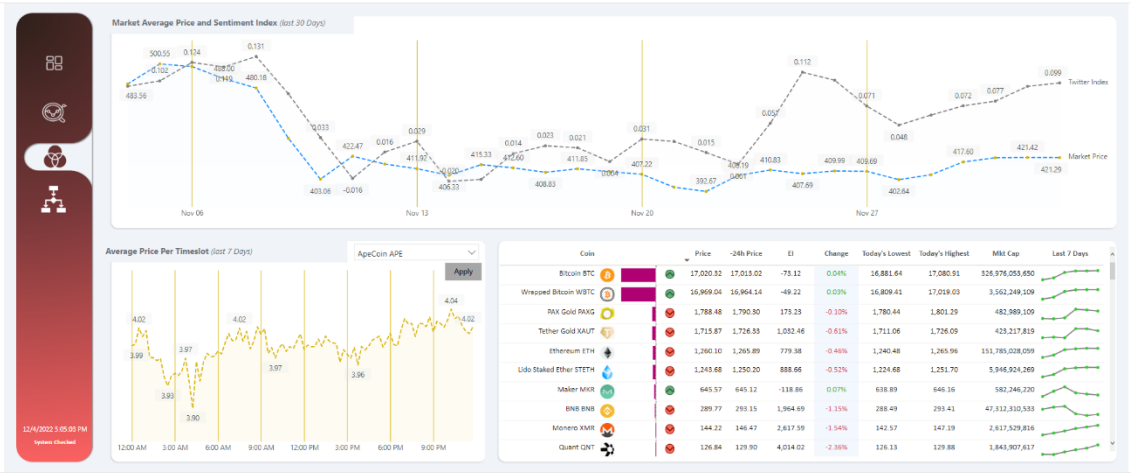


Figure 22: Overall Twitter Sentiment and Market Price. Overall Picture shows that there is some correlation with prices and Twitter Index of the total market.

Last subsection of this crypto analytics, is the good and bad news, where coins are been separated based on their Evolution Index. Evolution index can help the users understand how a specific coin performs compared to the total market. Since the result of this KPI can be a little tricky to understand some scenarios were created evaluation the Evolution Index.

The scenarios are numbered from 1 to 6 and are the following:

- Scenario 1 (good): Crypto Crowth <0 & Crypto Growth > Market Growth
- Scenario 2 (good): Market Crowth >0 & Crypto Growth > Market Growth
- Scenario 3 (good): Market Growth <0 & Crypto Growth > Market Growth
- Scenario 4 (bad): Crypto Crowth <0 & Crypto Growth < Market Growth
- Scenario 5 (bad): Crypto Crowth < Market Growth
- Scenario 6 (bad): Market Crowth >0 & Crypto Growth < Market Growth

An overview of this analysis is displayed below in Figure 23

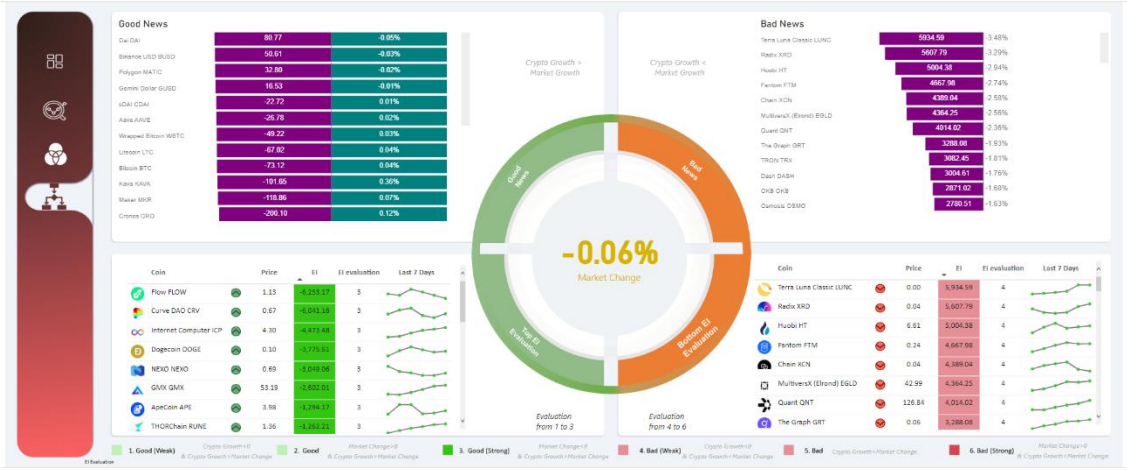


Figure 23: Good and Bad News

Next section is the News summarization. In this section, users can use the Text Summarization Tool which both abstractive and extractive summarization ways are in use. The tool works as follows:

- The user is pasting a link in the search bar,
- by clicking the “in a nutshell” button,
- the context is getting summarized in the boxes below.

The way of this summarization will be further analyzed in the Text Summarization section.

The benefits of this tool can be many, as it could work with any URL that includes text inside. In addition to that, key points of the extraction can be searched on the next section of sentiment analysis, checking how this topic is correlated with the total market or the price of a specific coin.

Figure 24 shows the result of a URL that was analyzed in real time.

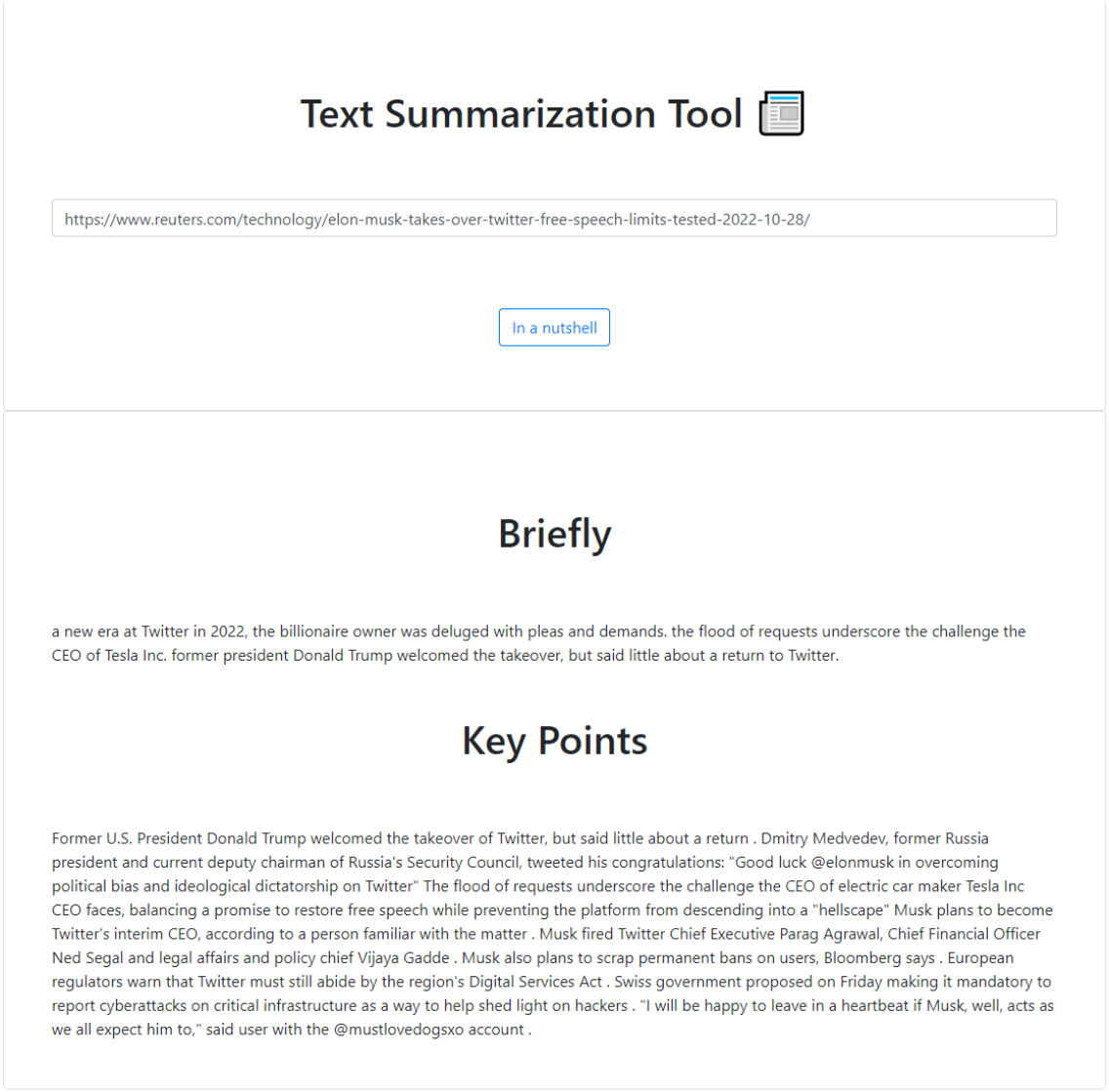


Figure 24: Text Summarization Tool

Another analysis that is being provided here is the Yahoo Finance Data Collection Section. This simply shows the latest news regarding cryptocurrencies, and can notify the users that something BIG has occurred. The URL of this summary is also provided, in case the user would like to get more details in regards to a specific topic.

This analysis is used also in the next section in which we are able to check how the markets corresponds to the news.

Figure 25 displays the section of News Summarization:

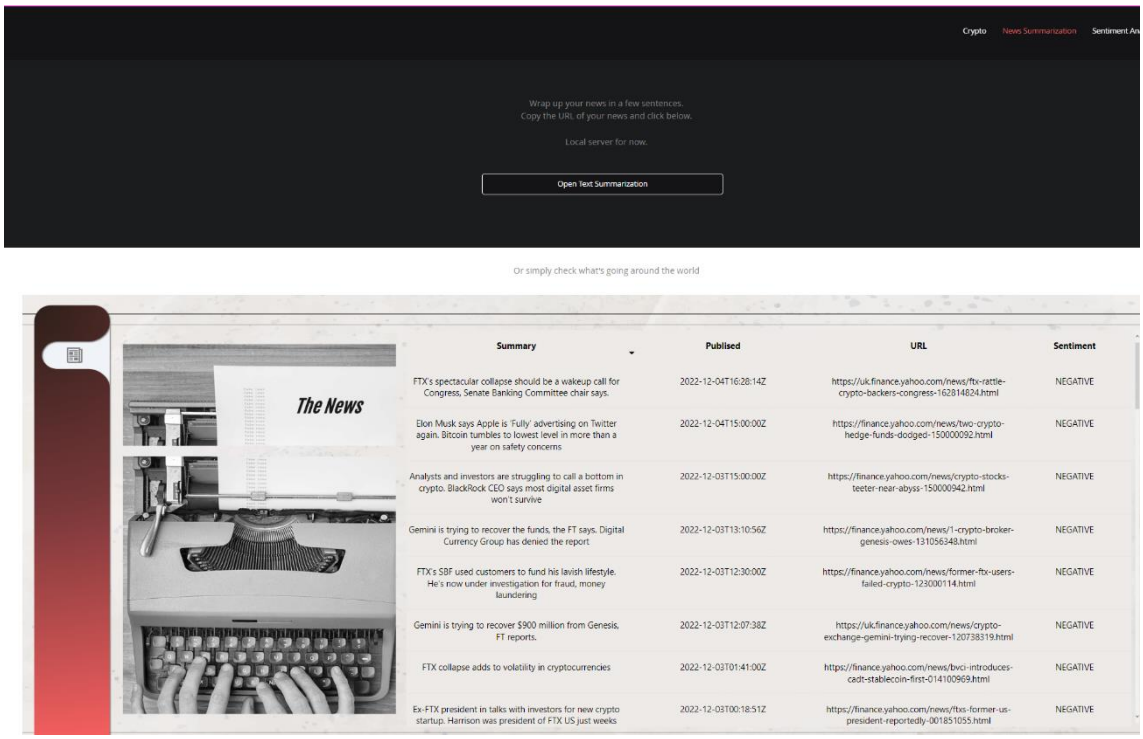


Figure 25: News Summarization Section

Next section of the CryptoFeel.eu is the Sentiment Analysis. This Report splits in two sections related to the Text Sources: Twitter & Yahoo Finance.

Twitter part, allows to the user to search in real time or the historical data, and get decisions based on some keywords on Twitter. Since the selected keyword is filtering the RealTime dataset, the Twitter Index in the chart is getting filtered as well. Then the users can also select the coin of the choice in order to see how Twitter is getting affected by the Price or vice versa. This section is quite interesting, if anyone could consider that, this dataset can be exported for additional analysis, finding causality, correlation and so on. Additional filters can be applied, filtering out posts with low engagement, number of friends, followers, replies and so on.

Figure 26 displays, how bitcoin relates to the Twitter Index, the amount of tweets that included the world “Bitcoin”, a word cloud and a part of the posted tweets.

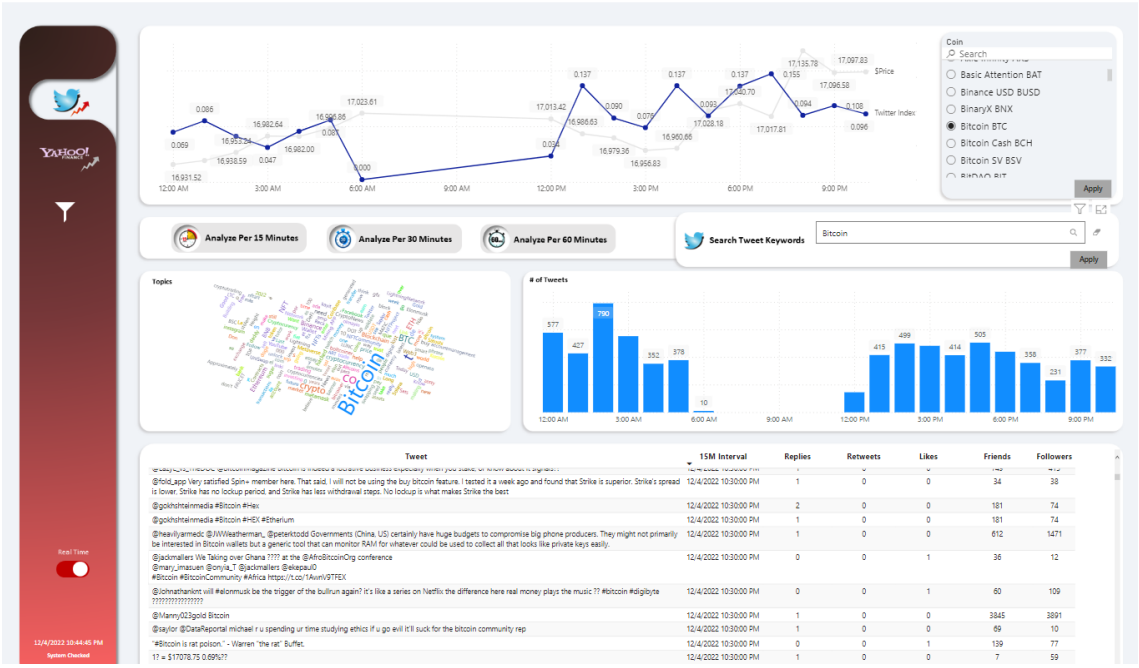


Figure 26: Twitter Sentiment Analysis with specific keyword search (Real Time)

Historical Data of this section includes also a date picker, which can be used in order see when a specific trend started, or what is the causality between the news using a specific keyword for the Tweets and a Coin Picker for the price, as shown in figure 27.

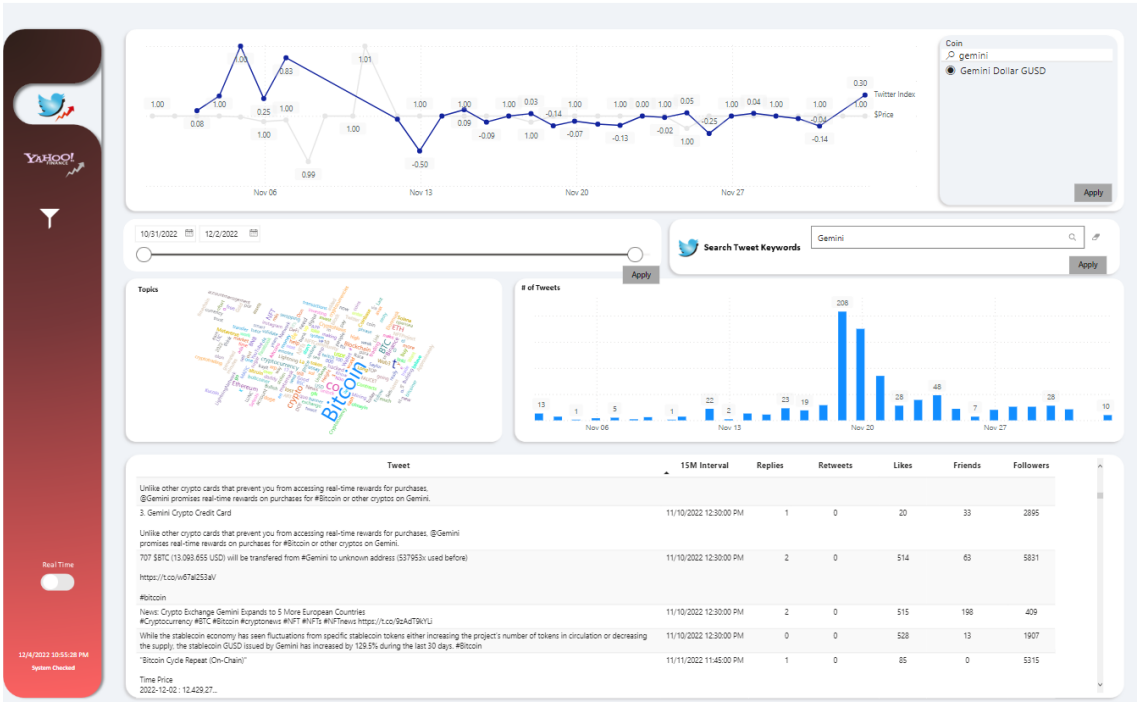


Figure 27: Twitter Sentiment Analysis with specific keyword search (Historical)

Last part of this section is the Yahoo Finance analysis in which, end users can see – for now – how the world news are progressing along with the price of the total market. Again,

the Yahoo Index can be filtered by some specific News that might be of the users' interest, such as "FTX". Figure 28, shows the analysis that be extracted from the tool, related to the Yahoo Finance source.

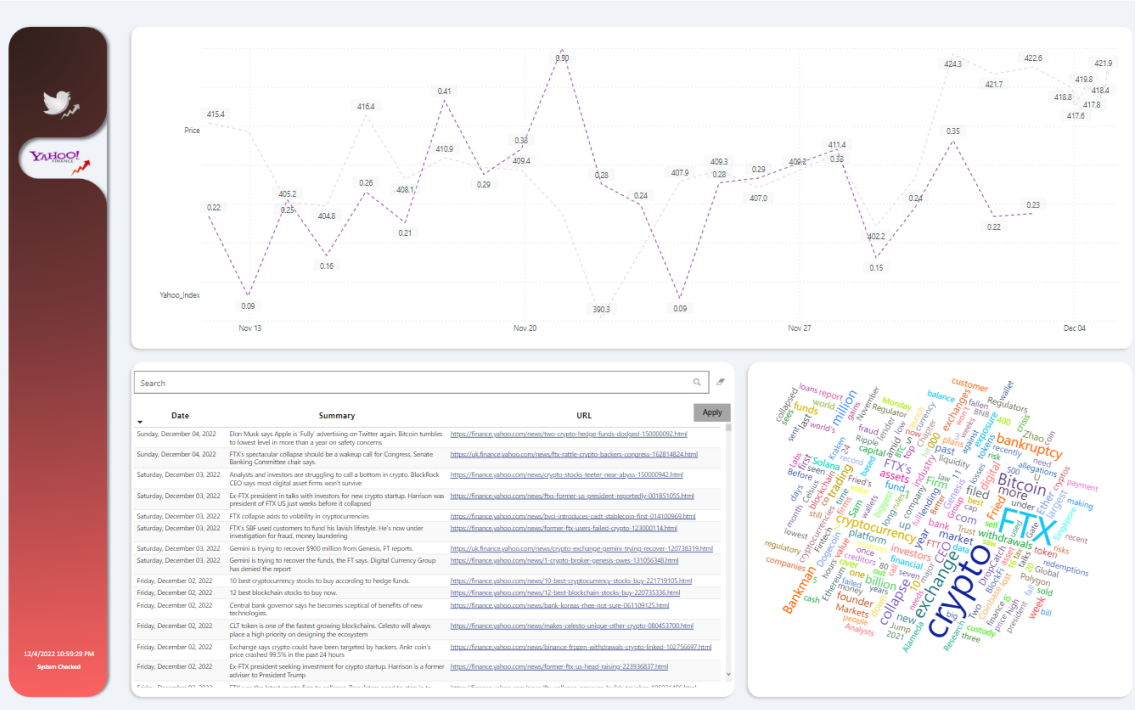


Figure 28: Yahoo Sentiment Analysis and Total Market Price. No Filters applied.

Last Section of the web page is the blog, where all site members can post various questions, like, share etc. The main scope of this section though is to include further static analysis that could be added at any time. Such as causality tests for specific keywords and coins. This blog can be used as a great inventory, adding more knowledge to this research.

5 Sentiment Analysis in practice

As discribed also in 2.2, SA is an automatic process that is trying to classify text data based on their polarity. In our case, the polarity has three values: positive, negative and neutral. The usage of this knowledge can help companies understand if their customers

are talking positive, negative or neutral about their products, campaigns, actions and so on. Older days would have a person manually browsing through relevant tweets, reading them, and categorizing them according to their sentiment. As you might expect, not only this does not scale, it is expensive, takes a lot of time, and is subject to human error.

Fortunately, recent developments in AI have made it possible for businesses to deploy machine learning models for tweet sentiment analysis that are just as accurate as people. Companies can use machine learning to analyze tweets in real-time, around-the-clock, at scale, and to quickly evaluate thousands of tweets. More significantly, they can receive the insights they need at the right time.

Why analyze Twitter sentiment? The two most popular use cases are assessing user feedback and monitoring comments to identify potential problems early on. Companies utilize this for a wide range of use cases.

5.1 Roberta

This research, uses the Twitter-roBERTa-base (cardiffnlp/twitter-roberta-base-sentiment) for Sentiment Analysis, which is a BERT model and belongs also to the State of the Art models, trained on ~58M tweets and with the TweetEval benchmark is is finetuned for sentiment analysis. It's based on the Google's Bert Model that released in 2018, and was proposed by inhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. [17]

Roberta is being used in the API_runner that is described above, while the new data are getting scrapped through twitter. Each tweet is encoded and provides a score for the positive, neutral, or negative labels.

These labels are joinning the dataframe that contains the tweet information that is described in the Tweet Data Collection section. The Figure 29 includes a part of the tweets that were retrieved using the Tweet Data Collection procedure, along with their sentiment. Even though, the tweets were cleaned, removing any kind of hashtags, symbols etc, the whole context is being stored in the SQL Server. With this way, more than one models can be used to analyze the polarity of the same dataset providing great information for future researchers of this tool.

Date	User	Tweet	likeCount	retweetCount	ReplyCount	Followers	Favorites	friends	location	Score	Negative	Neutral	Positive
2022-12-03 23:09:16.000	BahadirKilic11	Where are you @cz_binan...	1	0	1	274	2299	710	Burbank, CA	0.72992700729927	0.6389502	0.3287187	0.032331
2022-12-03 22:05:36.000	evoskuil	@HAYA_META @EdanYa...	0	0	1	7864	6670	98		0.0127161749745676	0.08504605	0.7255057	0.1894482
2022-12-04 01:18:24.000	0xEthereumYoda	#Ethereum price update: ...	0	0	1	442	0	1		0.226244343891403	0.04516673	0.8876899	0.0671433
2022-12-04 13:57:00.000	rovercrc	I longed #Bitcoin on the ret...	12	4	11	400956	19119	374	Bitcoin ????	0.00673390596474426	0.006365613	0.2798104	0.713824
2022-12-03 22:37:28.000	westkripto	Tam spottan aldigim coin h...	2	0	0	48420	7969	158	Istanbul, Türk...	0.00413052457662123	0.1454685	0.8005162	0.05401524
2022-12-04 00:24:06.000	Rusmir_Rusko	It's ok to admit you were w...	0	0	1	83	9311	150		1.20481927710843	0.09804904	0.5318045	0.3701464
2022-12-04 16:01:53.000	Alon_Vigos	Compilation of beautiful NF...	3	0	2	62	101	68		8.06451612903226	0.003355684	0.1302366	0.8664078
2022-12-04 11:39:27.000	duquevalerivave	This zesty faucet from @_...	1	0	0	9	10	72	Venezuela	11.1111111111111	0.01092589	0.6140051	0.375069
2022-12-04 17:00:07.000	OrderflowES	Later today I will post my lo...	1	0	0	599	6482	2	Desk	0.16694490180301	0.02024292	0.7060041	0.2737529
2022-12-04 01:02:15.000	BitcoinConfs	Block 765786 Hash: 0x...0...	0	0	1	1256	4	0		0.0796178343949045	0.1018488	0.8520375	0.04611364
2022-12-03 23:30:25.000	btcsentimentCOM	?? #Crypto Whales vs. #St...	1	0	1	8765	1816	421		0.0228180262407302	0.06668628	0.8567038	0.0766098
2022-12-03 22:26:20.000	octopus_olaf	Bitcoin ist die Zukunft. #N...	0	0	1	751	10436	526		0.133155792276964	0.05125491	0.7794808	0.1692644
2022-12-04 11:06:28.000	0xEthereumYoda	#Ethereum price update: ...	1	0	0	441	0	1		0.226757369614512	0.04504472	0.8877029	0.06725251
2022-12-04 13:17:04.000	markwild_369	$1.99 for https://t.co/UBO...	0	0	1	720	3423	372	Bangkok, Th...	0.138888888888889	0.006777633	0.153578	0.8396443
2022-12-04 19:55:04.000	whalesradar_com	#YFIUSDT #YFI Signal #...	0	0	1	2084	2	1	Istanbul	0.0479846449136276	0.09367854	0.8384684	0.06785314
2022-12-04 00:43:01.000	0xEthereumYoda	#Ethereum price update: ...	0	0	1	442	0	1		0.226244343891403	0.04714694	0.8834265	0.06942648
2022-12-04 18:05:33.000	QuantumHodler	@CValorem @knutsvanhol...	1	0	0	213	16406	171	Buenos Aires,...	0.469483568075117	0.6443016	0.3411721	0.01452637
2022-12-03 22:35:05.000	CryptoRedhawks	We are all misfits!! Join our ...	1	1	0	115	404	544		1.73913043478261	0.00296309	0.09492159	0.9021154
2022-12-04 02:52:25.000	0xEthereumYoda	#Ethereum price update: ...	1	0	0	443	0	1		0.225733634311512	0.04317681	0.8859727	0.07085048
2022-12-03 22:13:48.000	fadyamroush	@Bitcoin_info_ar ????? ?...	0	0	1	1128	1757	446	Spain	0.0886524822695035	0.1942539	0.7636017	0.04214427

Figure 29: Retrieved Tweets and Sentiment Analysis

Since there is also a date available in our Dataset, this is providing us the opportunity to see how this works with the price, creating a nice field of study regarding correlation and causality.

5.2 Distil Bert Model

The classification of topics is possible using this paradigm. The raw model can be applied to next sentence prediction or masked language modeling, but its main purpose is to be improved upon in a subsequent job. On the development set, this model achieves an accuracy of 91.3. In the Yahoo Finance News Data collection part, this model was used after summarizing the text with Pegasus that will be further analyzed in the upcoming sections. Figure 30 includes a piece of code that was used in order to use this model for sentiment analysis and the matching with the summarized URL's. Of course, one of the purposes of CryptoFeel.eu is to give the chance to the users, to select the model of their choice. Having said that, more models can be combined in the near future.

```
# 5. Adding Sentiment Analysis
print('Calculating sentiment.')
sentiment = pipeline("sentiment-analysis")
scores = {ticker: sentiment(summaries[ticker]) for ticker in monitored_tickers}

# # 6. Exporting Results
print('Exporting results')

def create_output_array(summaries, scores, urls, dates):
    output = []
    for ticker in monitored_tickers:
        for counter in range(len(summaries[ticker])):
            output_this = [
                ticker,
                summaries[ticker][counter],
                scores[ticker][counter]['label'],
                scores[ticker][counter]['score'],
                urls[ticker][counter],
                dates[ticker][counter]
            ]
            output.append(output_this)
    return output
```

Figure 30: Distil Bert Snippet

6 Text Summarization in practice

In the news summarization section, it was mentioned that summarization techniques were tried either to allow the user to perform a real time summarization, or to get the latest news summarized with a sentiment score as well.

6.1 Pegasus Abstractive

The model that was a nice fit to the financial news stands for Pegasus. A unique financial news dataset comprised of 2K articles from Bloomberg was used to fine-tune this algorithm. Figure below includes a snippet of calling the model with their tokenizer.

```
# 2. Setup Model
model_name = "human-centered-summarization/financial-summarization-pegasus"
tokenizer = PegasusTokenizer.from_pretrained(model_name)

model = PegasusForConditionalGeneration.from_pretrained(model_name)
```

Figure 31: PEGASUS model setup

For the case of the real time text summarization tool, since the text could be also related to non-financial data, the t5 model was used, providing very fast results. The usage of this model on this tool, could be combined with other models as well, allowing the user to select among a few given options.

```python
def summarizeTEXT(url):
    warnings.filterwarnings("ignore")
    article = Article(url)
    article.download()
    article.parse()
    article.nlp()

    ## preprocess the input text
    preprocessed_text = article.text.strip().replace('\n',' ')
    t5_input_text = 'summarize: ' + preprocessed_text

    model = T5ForConditionalGeneration.from_pretrained('t5-small')
    tokenizer = T5Tokenizer.from_pretrained('t5-small')

    device = torch.device('cpu')

    tokenized_text = tokenizer.encode(t5_input_text, return_tensors='pt', max_length=512).to(device)
    summary_ids = model.generate(tokenized_text, min_length=30, max_length=512)
    summary = tokenizer.decode(summary_ids[0], skip_special_tokens=True)

    return summary
```

Figure 32: Use of t5-small

6.2 Extractive

For the extractive part of this section, the text had to split in order to optimize performance. Then summarization pipeline which is a dependency from Transformers' Library was used in order to summarize these chunks of sentences that have been gathered from the whole content. It is interesting to say, that both t5-small and summarization pipeline are called in order to bring the results to the end user. T5-small is connected with the "Briefly" and summarization is called into the "Key points" part. Aim of this combination is to provide more insights to the user.

```python
r = requests.get(url)
soup = BeautifulSoup(r.text, 'html.parser')
results = soup.find_all(['h1', 'p'])
text = [result.text for result in results]
ARTICLE = ' '.join(text)

## preprocess the input text
#preprocessed_text = article.text.strip().replace('\n',' ')
ARTICLE = ARTICLE.replace('.', '.<eos>')
ARTICLE = ARTICLE.replace('?', '?<eos>')
ARTICLE = ARTICLE.replace('!', '!<eos>')
#t5_input_text = 'summarize: ' + ARTICLE

sentences = ARTICLE.split('<eos>')
# print(sentences)
current_chunk = 0
chunks = []

for sentence in sentences:
    if len(chunks) == current_chunk + 1:
        if len(chunks[current_chunk]) + len(sentence.split(' ')) <= max_chunk:
            chunks[current_chunk].extend(sentence.split(' '))
        else:
            current_chunk += 1
            chunks.append(sentence.split(' '))
    else:
        print(current_chunk)
        chunks.append(sentence.split(' '))

for chunk_id in range(len(chunks)):
    chunks[chunk_id] = ' '.join(chunks[chunk_id])
```

Figure 33: preprocessing for extractive summarization

6.3 Django Framework

A few limitations were noticed on the idea of letting a user running a python script. What is widely common, is that someone would download and install a Python Environment, and after installing the necessary libraries, some pieces of code would be tested. The idea of news summarization in this research is to provide a way, for someone to use this python script without downloading or copying anything to their computer.

For this reason, Django Framework was tested and used. Django is a free open-source web framework, that is used to create websites.

Along with some basic HTML bootstraps, functions created in python, results can be connected and transferred to the user. For this project, the server is local, and in order to run it requires from the user to execute the following command: python runserver manage.py

```
<!doctype html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1, shri
    <link rel="stylesheet" href="https://cdn.jsdelivr.net/npm/bootstrap@4.3.
    <title>In a nutshell</title>
  </head>
  <body>


    <div class='container mt-5'>

        <form type='post' method='post' class='card p-5'>
            <h1 class='mt-5  text-center'>Text Summarization Tool </h1>

            <h1 class='mt-5  text-center'></h1>
```

Figure 34: HTML bootstrap with Django

7 Discussion

The concept of this research is based on data that are created a stream; thus, the results can create interesting scenarios to be further analyzed. The benefit of using such a tool, is that anyone could filter out unwanted content based on their selections. The size of Tweets that is being gathered everyday could be adjusted providing even greater insights, though, this would have some computational costs. Despite that, roBERTa the main model for sentiment analysis in tweets was up to the task, handling the heavy work without any issue, though more finetuning and training can be achieved.

For the Yahoo Finance Data, Pegasus unties the hands of many, considering the extremely fast and accurate results that are being provided. It would be interesting to connect the key content of a URL with a key content in Twitter, providing an extra weight to the Sentiment Index of Yahoo news. This Tool also provides various intervals both for prices and news/tweets and be easily used to see for how long, some specific news affected a Price, and for how long this was affecting the opinion of the crowd. These datasets can be used also for Granger Causality and forecasting using Vector Autoregression.The more these APIs are running the more insights and filters can be provided for real time analysis.

From the beginning of this project some quite interesting facts have occurred and have been collected through CryptoFeel.eu.

Some results on the figure below can show a scenario with the ChatGPT [39], which was introduced on the 30th of November in 2022, and how this is interacting with crypto and the general sentiment analysis around this topic.

The query that was used in the 4.1 proves that along with cryptos, people went ahead and tweeted ChatGPT with cryptocurrency keywords altogether. Some of the key topics that were identified were the cybersecurity and predictions, topics that were already mentioned in first sections of this research. The period that was selected for these results were between the 1st and 28th of December 2022.

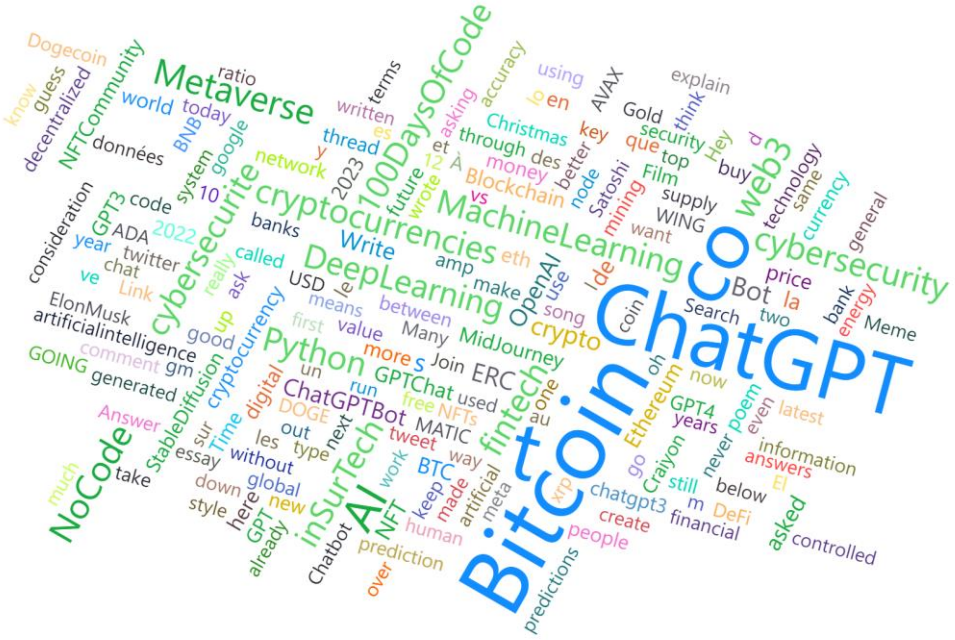


Figure 35: ChatGPT and Crypto topics

In addition to that, some of the first experiments that people started to do, while combining cryptocurrency and ChatGPT was to write poems and sonnets, as the figure below indicates.

ChatGPT Bitcoin sonnet:

In the depths below the azure sea
A force doth stir and boil the brine
Bitcoin's proof-of-work doth furiously
Mine for the digital gold so fine

ChatGPT prompt: can you me write a song about bitcoin in the style of a freestyle rap

result:
Bitcoin, bitcoin, oh how sweet
My money is mine, I can't be beat
No government control, no corporate feat
Bitcoin, bitcoin, I'm in the elite

I asked @OpenAI's chat bot ChatGPT to explain #Bitcoin's Proof of Work algorithm using a 5/10/15/20 year old's vocabulary.

Here are the results ????

"Ode to Bitcoin Influencer Dum Dums" by OpenAI ChatGPT & Cryptostonks

Dum dums, oh dum dums, you make us all roll our eyes
With your predictions of Bitcoin reaching the skies

You predicted it would rise to two-hundred and fifty thousand, a price never seen 1/3

"The Time Traveler's Bitcoin" Dr Who TV show #ChatGPT https://t.co/2lqxtlDPL2

#ChatGPT

Prompt: Write a tweet thread that explains why someone should run their own bitcoin node

??

Figure 36: First Impressions over using ChatGPT and crypto

Moreover, some additional graphs can prove that there should be a correlation between bitcoin (for example) and ChatGPT twitter index, in a way that when the people started using the latter, their emotions could have created a hype around the Bitcoin itself. However, the sample is still too low at the moment, and the results could be saved to see how this trend progresses.

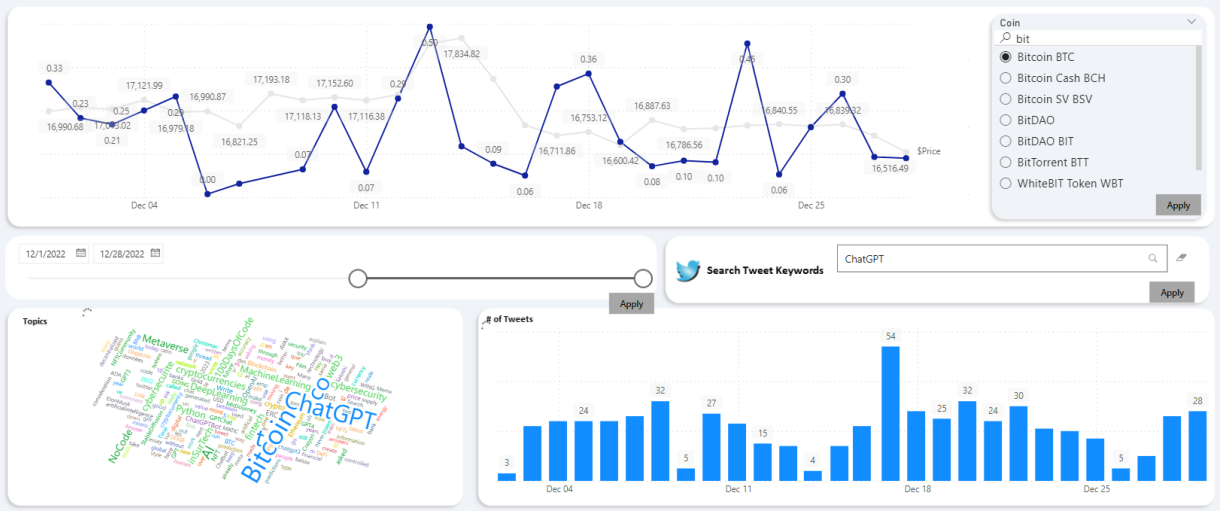


Figure 37: ChatGPT Twitter index and Bitcoin Prices. The sample is too low, but a small correlation can be observed.

By using State-of-the-Art Models for summarization and sentiment analysis, CryptoFeel.eu can promise to any end user, either this is a researcher or a trader, that this could work as their own eyes in the market of cryptocurrency, combining both real time and historical data for all needs.

An eye in the market could see for instance the impact of a specific event in the market, as it recently happened with FTX bankruptcy, where a huge number of tweets that were gathered in 4.1 was mentioning the FTX keyword. Even when the company stopped operating, please were still talking about this, but on a lower frequency. FTX was ranking in the first places of search in CryptoFeel.eu both in Twitter and in Yahoo Finance News.

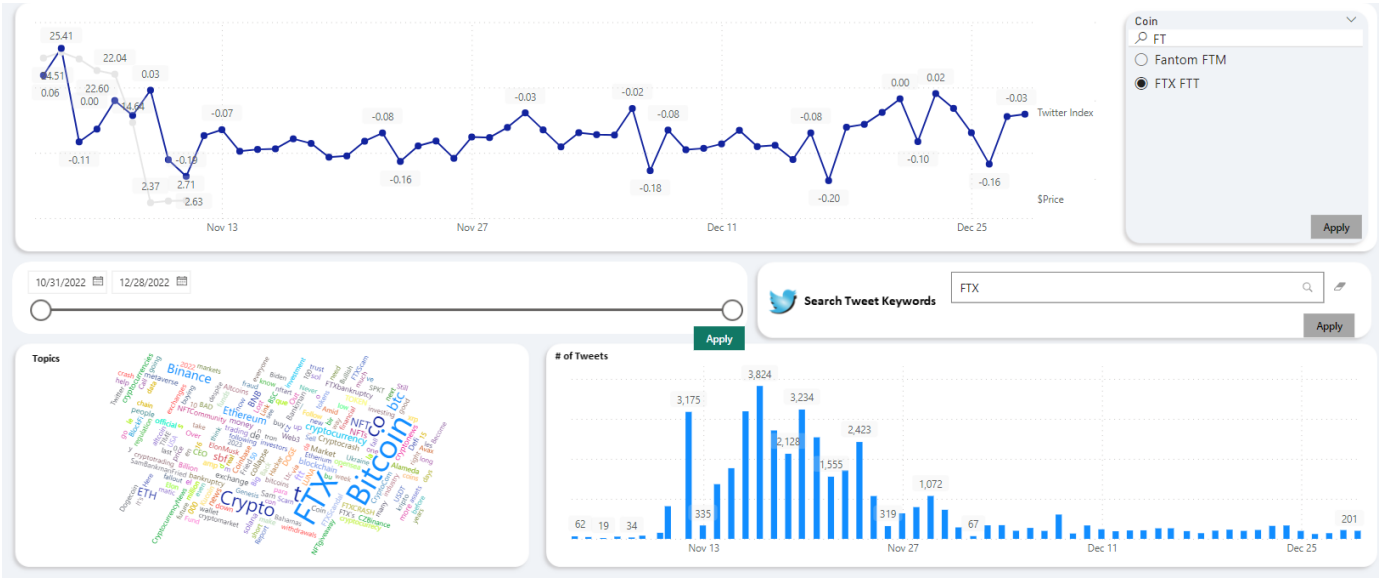


Figure 38: Keyword FTX Twitter Index, volume of Tweets and after bankruptcy effects. A negative trend while typing the FTX on Twitter along with the high column of Tweets can justify the event. The Word-cloud on the left window can easily help the end user understand what are the topics that are being addressed.

By combining the sentiment analysis with the summarization of Yahoo Finance news, the Keyword “FTX” can also indicate that something significant has happened, and that was a successful milestone in CryptoFeel.eu as it was able to catch this information from day one.

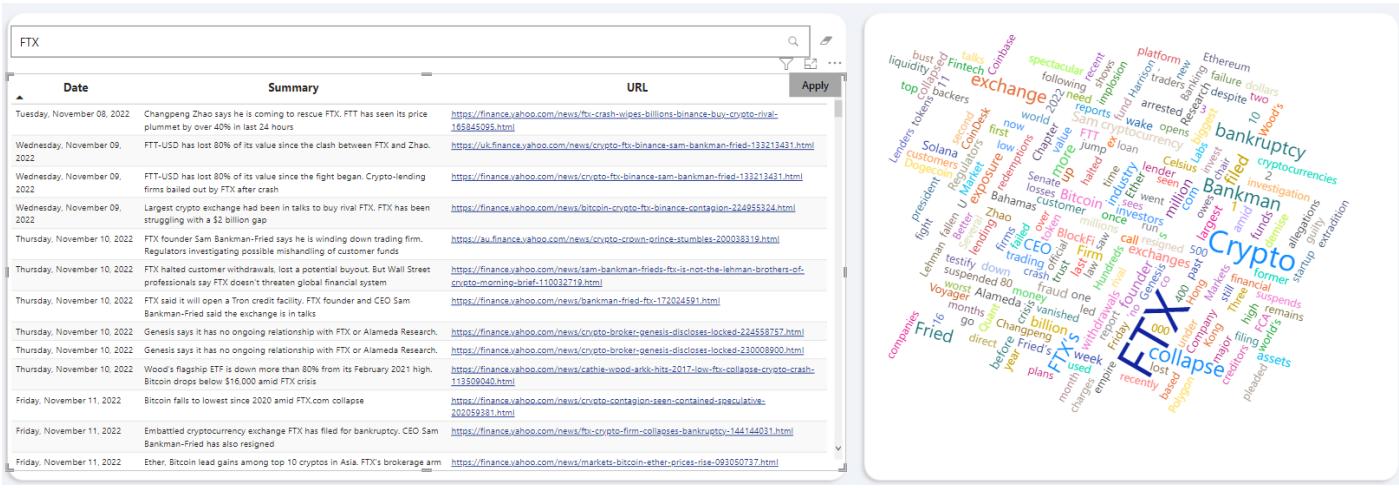


Figure 39: Yahoo Finance News and FTX. FTX, Collapse, Fried, CEO, Bankruptcy are some of the important keywords that are mentioned on the news.

8 Conclusion

The cryptocurrency market's characteristics as a young technology with enormous potential increase its allure to potential investors.

Successful user sentiment analysis and research into their predictive power are crucial in today's fast-paced, data-driven world. As a result, the combination of a sector that is extremely popular among social media users, like cryptocurrency, and the market's volatility present a special opportunity. By doing sentiment analysis, reviewing the literature, and investigating the relationship between social media data, finance news and prices, this work adds to the scant research on the subject combining the text summarization all along.

Furthermore, this research is using a creative way of having the topics of Real time, Sentiment analysis, Cryptocurrency prices and text summarization, working together, allowing the user to filter both in real time and historical data.

Additionally, these findings contribute to the body of earlier research about the influence of social media sentiment on the cryptocurrency market. The methodology used in this study could enhance more techniques for sentiment analysis, text summarization, allowing at the same time the dynamic calculation of possible price prediction now that everything is included in the same real time flow. Last but not least, in this study it was proved that big topics such as FTX, and ChatGPT are being searched on the social networks and their understanding can lead to exploring more areas such as “predictions”, “Root Cause Analysis” and so on. Last but not least, CryptoFeel.eu encourages the users to leave their feedback on the “Blog” section, allowing the tool to be improved. Future work would definitely use predictive models based on this project’s Data. Proving that these models can help traders on their strategies would be insightful.

Bibliography

[1] O. Sattarov, H. S. Jeon, R. Oh, and J. D. Lee, "Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis," in *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, Nov. 2020, pp. 1–4. doi: 10.1109/ICISCT50599.2020.9351527.

[2] P. V. Rane and S. N. Dhage, "Systematic Erudition of Bitcoin Price Prediction using Machine Learning Techniques," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, Mar. 2019, pp. 594–598. doi: 10.1109/ICACCS.2019.8728424.

[3] O. Kraaijeveld and J. de Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," *Journal of International Financial Markets, Institutions and Money*, vol. 65, p. 101188, Mar. 2020, doi: 10.1016/j.intfin.2020.101188.

[4] C. S. Wright, "Bitcoin: A Peer-to-Peer Electronic Cash System," *SSRN Electronic Journal*, 2008, doi: 10.2139/ssrn.3440802.

[5] S. Corbet, B. Lucey, A. Urquhart, and L. Yarovaya, "Cryptocurrencies as a financial asset: A systematic analysis," *International Review of Financial Analysis*, vol. 62, pp. 182–199, Mar. 2019, doi: 10.1016/j.irfa.2018.09.003.

[6] SHOBHIT SETH, "Explaining the Crypto in Cryptocurrency," *https://www.investopedia.com/tech/explaining-crypto-cryptocurrency/*. https://www.investopedia.com/tech/explaining-crypto-cryptocurrency/ (accessed Dec. 18, 2022).

[7] L. Tredinnick, "Cryptocurrencies and the blockchain," *Business Information Review*, vol. 36, no. 1, pp. 39–44, Mar. 2019, doi: 10.1177/0266382119836314.

[8] S. Kethineni and Y. Cao, "The Rise in Popularity of Cryptocurrency and Associated Criminal Activity," *Int Crim Justice Rev*, vol. 30, no. 3, pp. 325–344, Sep. 2020, doi: 10.1177/1057567719827051.

[9] J. M. Gavilanes, T. C. Flatten, and M. Brettel, "Content Strategies for Digital Consumer Engagement in Social Networks: Why Advertising Is an Antecedent of

Engagement," *J Advert*, vol. 47, no. 1, pp. 4–23, Jan. 2018, doi: 10.1080/00913367.2017.1405751.

[10] O. Kraaijeveld and J. de Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," *Journal of International Financial Markets, Institutions and Money*, vol. 65, p. 101188, Mar. 2020, doi: 10.1016/j.intfin.2020.101188.

[11] S. Wooley, A. Edmonds, A. Bagavathi, and S. Krishnan, "Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec. 2019, pp. 500–505. doi: 10.1109/ICMLA.2019.00093.

[12] N. Smuts, "What Drives Cryptocurrency Prices?," *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 3, pp. 131–134, Jan. 2019, doi: 10.1145/3308897.3308955.

[13] D. Shen, A. Urquhart, and P. Wang, "Does twitter predict Bitcoin?," *Econ Lett*, vol. 174, pp. 118–122, Jan. 2019, doi: 10.1016/j.econlet.2018.11.007.

[14] O. Sattarov, H. S. Jeon, R. Oh, and J. D. Lee, "Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis," in *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, Nov. 2020, pp. 1–4. doi: 10.1109/ICISCT50599.2020.9351527.

[15] X. Gao, W. Huang, and H. Wang, "Financial Twitter Sentiment on Bitcoin Return and High-Frequency Volatility," *Virtual Economics*, vol. 4, no. 1, pp. 7–18, Jan. 2021, doi: 10.34021/ve.2021.04.01(1).

[16] S. Corbet, C. Larkin, B. Lucey, and L. Yarovaya, "KODAKCoin: a blockchain revolution or exploiting a potential cryptocurrency bubble?," *Appl Econ Lett*, vol. 27, no. 7, pp. 518–524, Apr. 2020, doi: 10.1080/13504851.2019.1637512.

[17] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019.

[18] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," Dec. 2019.

[19] STAFFORD GLOBAL, "Finance 2.0: What is Cryptocurrency?," *https://www.staffordglobal.org/articles-and-blogs/finance-articles-blogs/cryptocurrency-bitcoin-finance-banking/*, Apr. 2021.

[20] E. Lansiaux, N. Tchagaspanian, and J. Forget, “Community Impact on a Cryptocurrency: Twitter Comparison Example Between Dogecoin and Litecoin,” *Frontiers in Blockchain*, vol. 5, Apr. 2022, doi: 10.3389/fbloc.2022.829865.

[21] A. Raheman, A. Kolonin, I. Fridkins, I. Ansari, and M. Vishwas, “Social Media Sentiment Analysis for Cryptocurrency Market Prediction,” Apr. 2022.

[22] F. Å. Nielsen, “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,” Mar. 2011.

[23] C. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/icwsm.v8i1.14550.

[24] A. Kolonin, “High-performance automatic categorization and attribution of inventory catalogs,” Feb. 2022.

[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” Oct. 2019.

[26] J. M. Pérez, J. C. Giudici, and F. Luque, “pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks,” Jun. 2021.

[27] D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” Aug. 2019.

[28] M. Kamal Eddine, A. Tixier, and M. Vazirgiannis, “BARThez: a Skilled Pretrained French Sequence-to-Sequence Model,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9369–9390. doi: 10.18653/v1/2021.emnlp-main.740.

[29] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, “TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP,” Apr. 2020.

[30] C. Siebert, J. Hartmann, M. Heitmann, and C. Schamp, “Accuracy of Automated Sentiment Analysis,” *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3489963.

[31] S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram, and M. Choudhury, “GLUECoS : An Evaluation Benchmark for Code-Switched NLP,” Apr. 2020.

[32] G. Aguilar, S. Kar, and T. Solorio, “LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation,” May 2020.

[33] J. Pilault, R. Li, S. Subramanian, and C. Pal, “On Extractive and Abstractive Neural Document Summarization with Transformer Language Models,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9308–9319. doi: 10.18653/v1/2020.emnlp-main.748.

[34] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “MASS: Masked Sequence to Sequence Pre-training for Language Generation,” May 2019.

[35] L. Dong *et al.*, “Unified Language Model Pre-training for Natural Language Understanding and Generation,” May 2019.

[36] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” Oct. 2019.

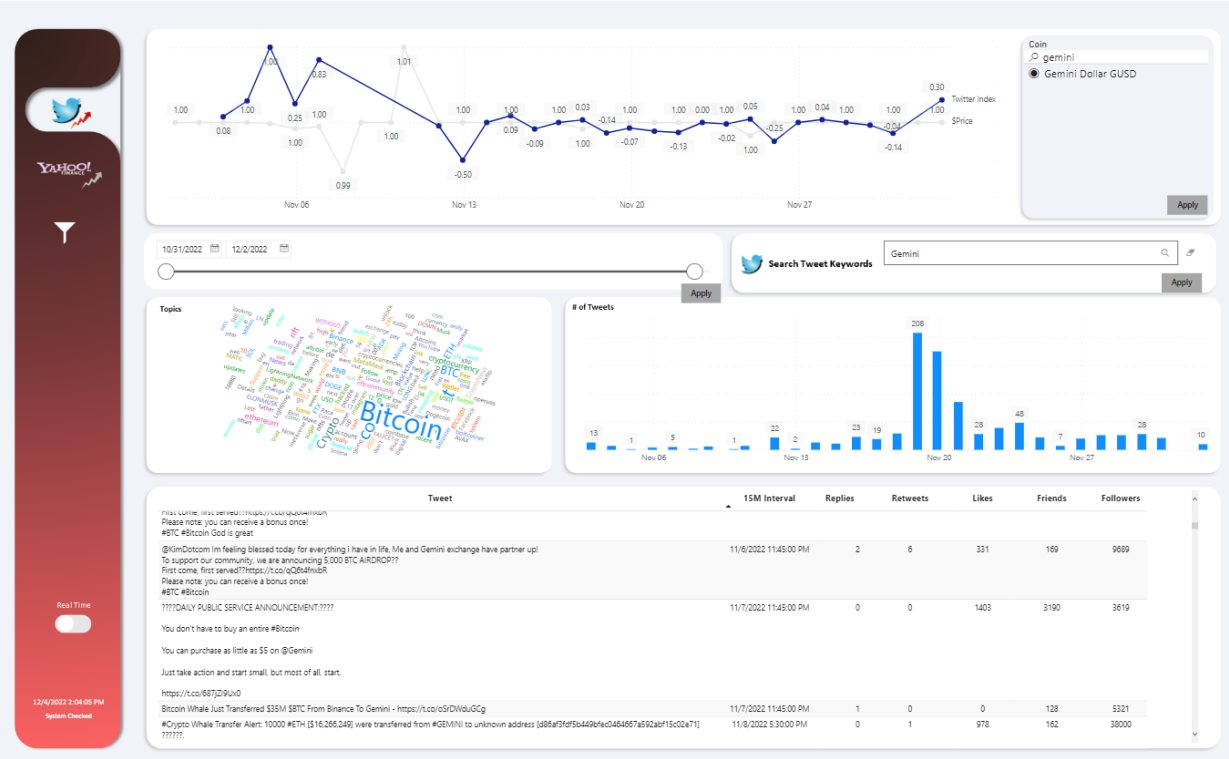
[37] M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” Oct. 2019.

[38] M. Nadj, A. Maedche, and C. Schieder, “The effect of interactive analytical dashboard features on situation awareness and task performance,” *Decis Support Syst*, vol. 135, p. 113322, Aug. 2020, doi: 10.1016/j.dss.2020.113322.

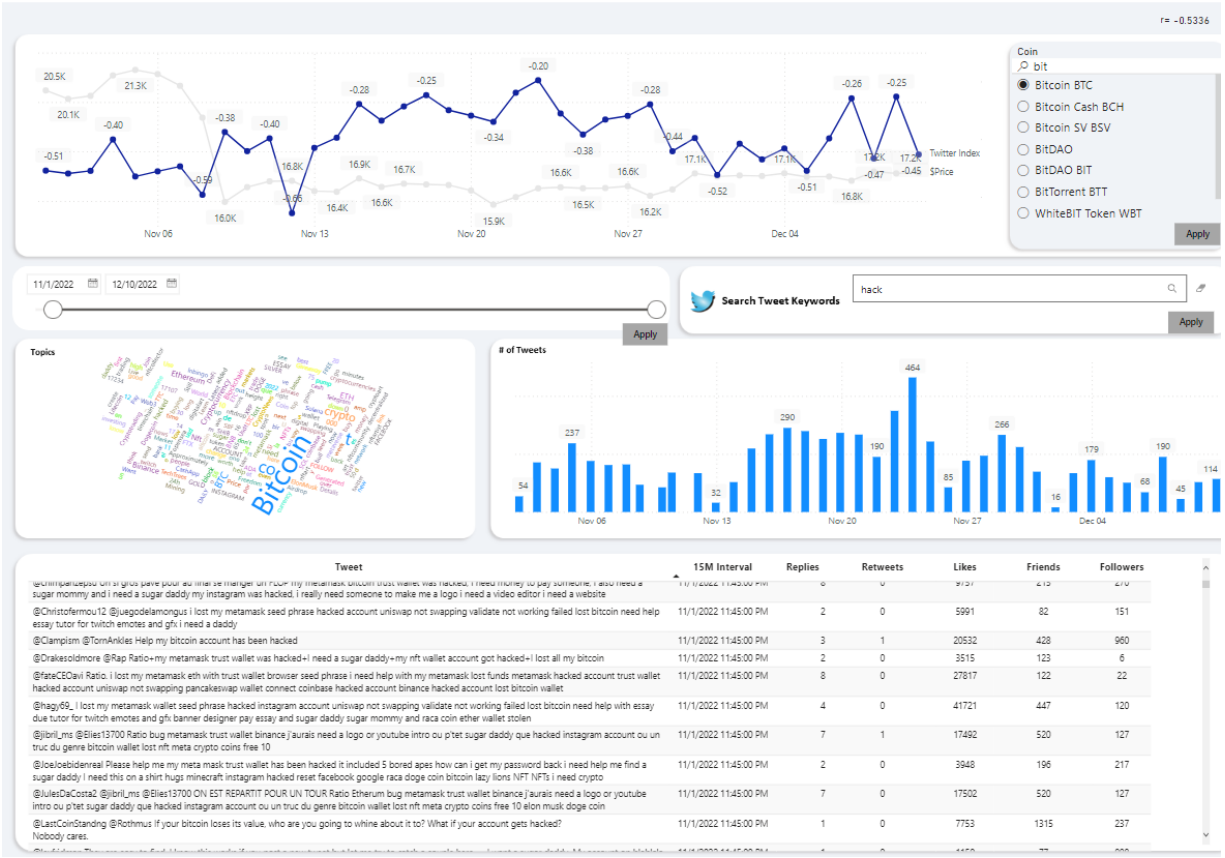
[39] OpenAI, “ChatGPT: Optimizing Language Models for Dialogue,” Nov. 30, 2022.

Appendix

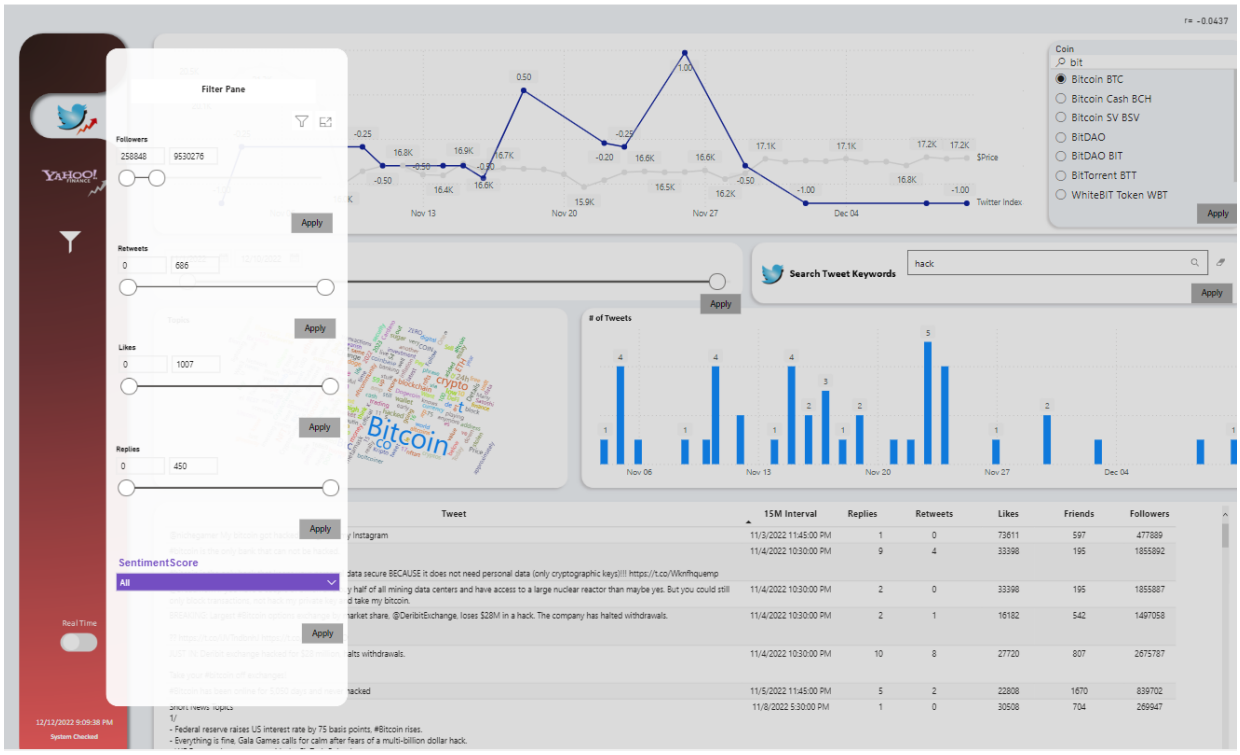
GEMINI coin results



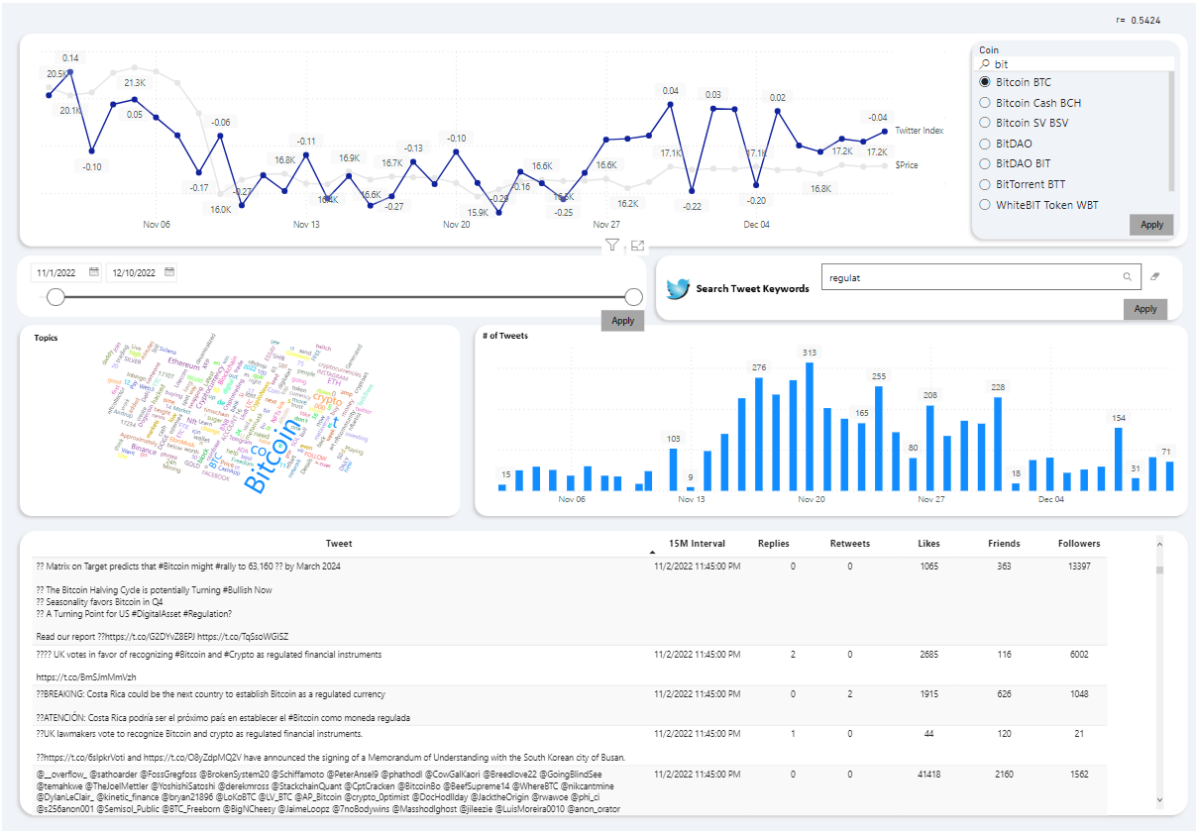
Correlation of “Hack” Sentiment Score on Twitter and Bitcoin Prices



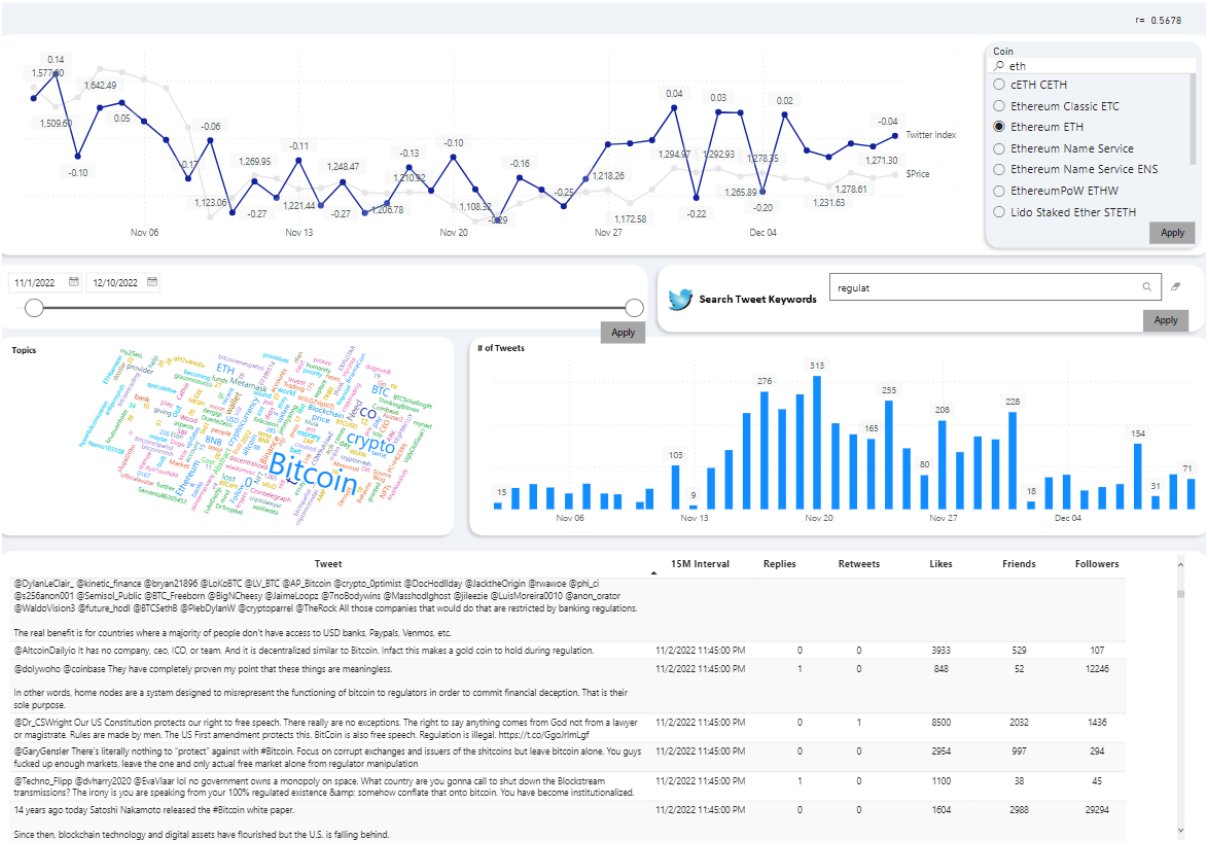
Same keywords’ Correlation, after filtering followers



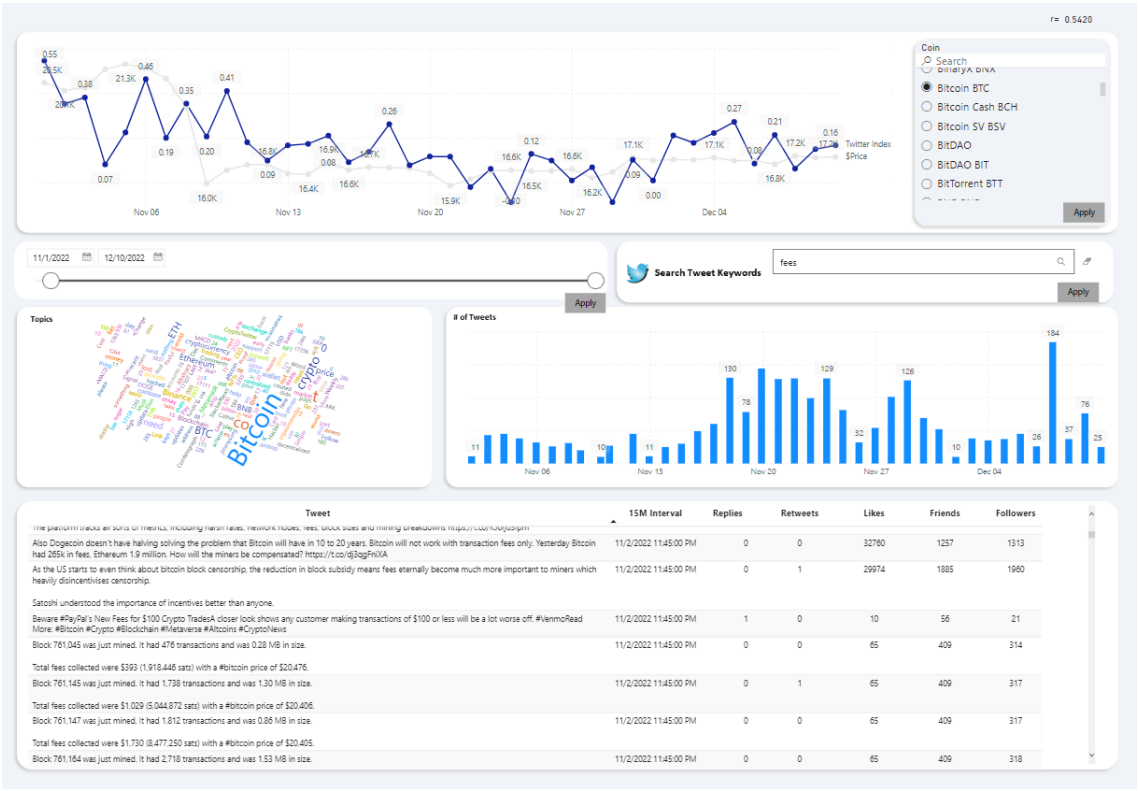
Correlation of “Regulat**” Sentiment Score on Twitter and Bitcoin Prices



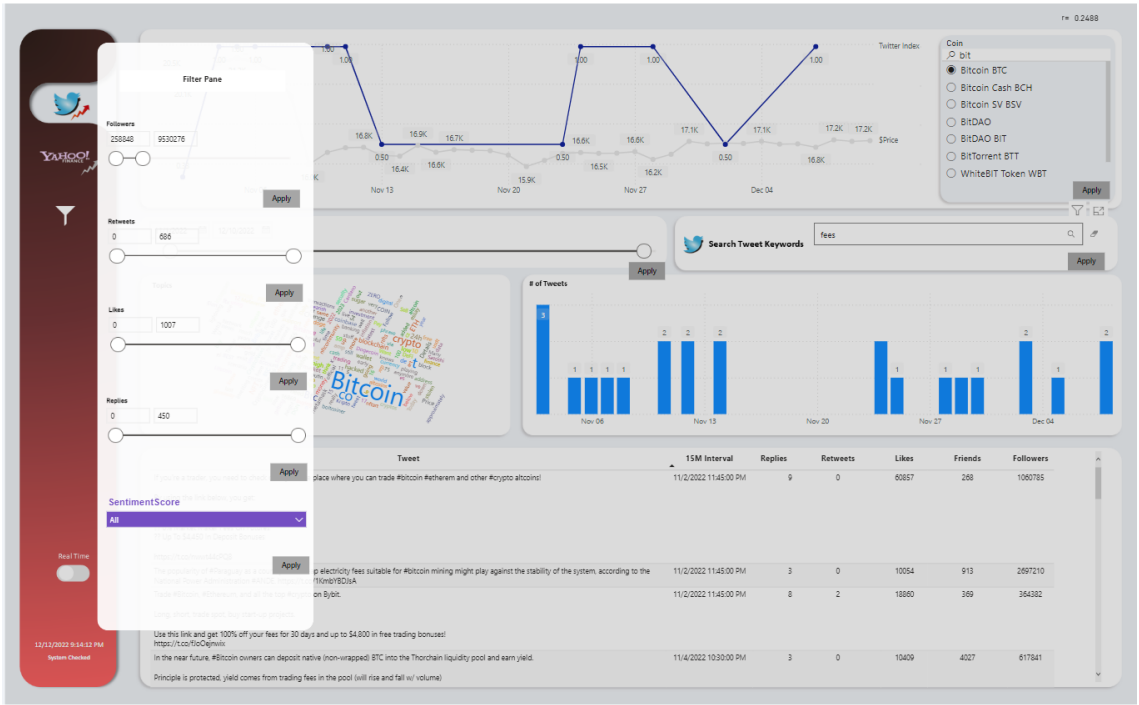
Correlation of “Regulat**” Sentiment Score on Twitter and Ethereum Prices



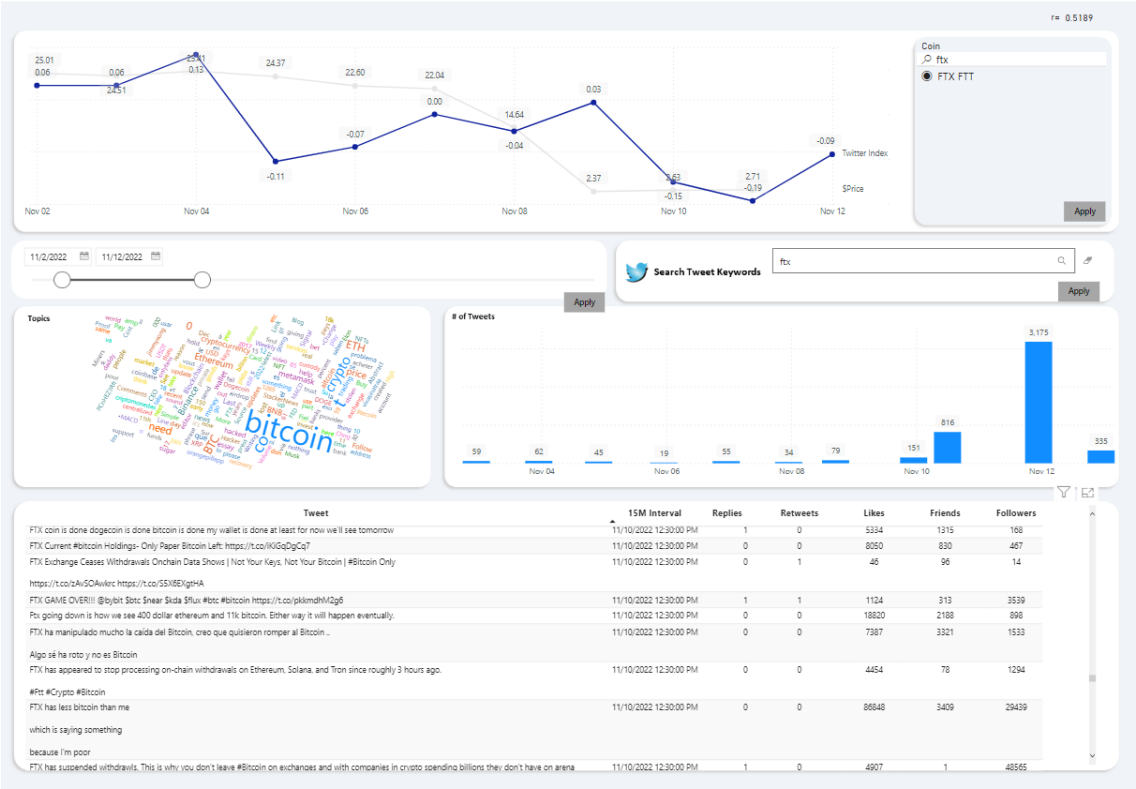
Correlation of “Fees” Sentiment Analysis on Twitter and Bitcoin Prices



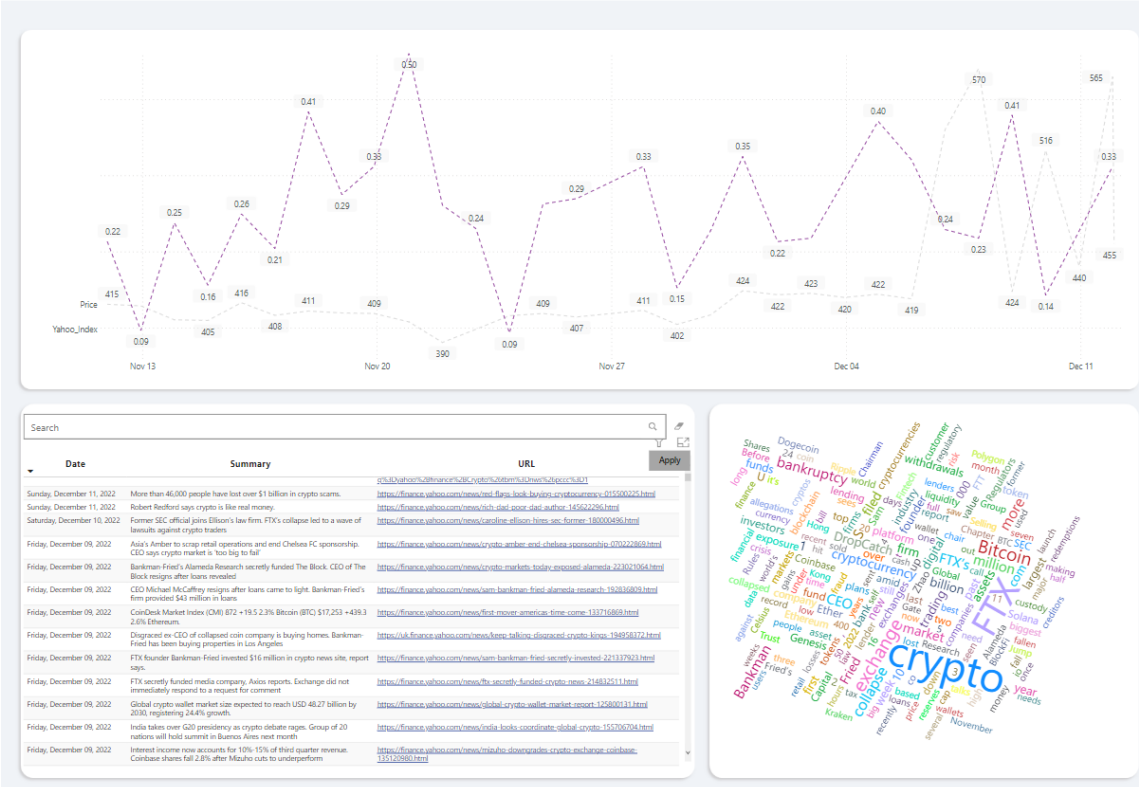
Same keywords’ Correlation, after filtering followers

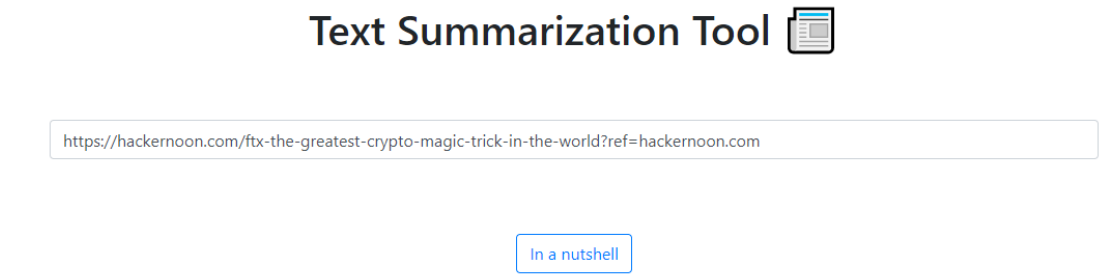


FTX Analysis, Sentiment Score and Price Correlation, Number of Tweets growth



Yahoo Finance News and Total Market Progress





Briefly

FTX is a crypto derivative exchange facilitator. it provides leverage to allow people to get more exposure in the markets using fewer deposits. if you were to deposit 100 USD, it would mean that a 20% market move could see you losing your entire deposit.

Key Points

FTX, a crypto derivative exchange facilitator, is an exchange that provides leverage to allow people to get more exposure in the markets using fewer deposits . The instruments that allow these operations are called “perpetual futures” and roll over month to month, without closing like traditional futures in the CME . The market participants on the other side of your trade who buy and sell the majority of these positions are “market makers” FTX was acquiring companies and sponsorships while providing a return for FTT through Alameda . Given that the FTT price is based on market fluctuations, it was a matter of time until one big actor (Binance) liquidated its position . Once the hat was gone, the rabbit could never be found while the company was left in Chapter 11 bankruptcy . A cryptocurrency can move in excess of 70% during a single trading session .