CRIPTOVALUTE E HYPE

Come si è evoluto l'interesse delle persone?

Boller Mattia, De Rosso Daniel - Università degli studi di Milano Bicocca dipartimento DISCO, Data science

---

Sommario

I primi mesi del 2021 sono stati caratterizzati da una generale attenzione nei riguardi delle criptovalute, questo ha portato l'interesse di molte persone prima estranee a questo fenomeno, causando forti movimentazioni e volatilità nel mercato.

Quando il grande pubblico entra a far parte di un nuovo ambiente, spesso causa in esso reazioni inaspettate, guidate talvolta dall'irrazionalità delle folle. Un caso simile è stato quello della moneta chiamata Dogecoin, resa famosa tramite meme sui social e supportata da diversi influencer che ne hanno favorito la crescita.

Tramite questo progetto il nostro obiettivo è stato quello di cercare delle evidenze di causalità tra le variazioni di prezzo di Dogecoin e diversi eventi misurabili nei social.

Il nostro approccio data driven ci ha portato all'estrazione di diversi dataset riguardanti tweets storici e indici del prezzo della criptovaluta, con l'aggiunta di dati in tempo reale estratti da Twitter. Per fare ciò sono stati utilizzati strumenti quali: NiFi per la gestione del flusso di dati, Kafka per la gestione dello streaming di dati veloci, MongoDB per l'integrazione e la gestione della varietà dei dati e il tutto poi coordinato da diversi script python per le operazioni più specifiche.

Dopo aver reso i dati fruibili ed interpretabili, attraverso il software Tableau si è voluto ottenere delle visualizzazioni che provassero il nostro point permettendo di rendere i dati più facilmente osservabili.

Keywords

Cryptocurrency - Dogecoin - Twitter

Indice

Introduzione

Durante gli ultimi anni (2020/2021) si è sentito parlare sempre di più dell'argomento "Criptovalute". Oltre a Bitcoin, la più famosa tra le valute digitali, si sono fatte spazio molte altre monete, con meno valore, alcune legate a progetti vari, altre nate senza nessuno scopo se non quello di cercare di guadagnarci sperando in una crescita del prezzo. Quest'ultime in particolare, nei primi mesi del 2021, sono state soggette ad un forte aumento dell'interesse verso di loro dato dall'estrema volatilità che le caratterizza. La più famosa tra queste è sicuramente Dogecoin.

Dogecoin (DOGE) nasce come criptovaluta ironica ispirandosi al cane Shiba Inu, diventato virale nei social. Il progetto open-source è stato sviluppato da Billy Markus e nasce come fork di Litecoin nel dicembre del 2013. La crescita di dogecoin si basa esclusivamente sulla audience che è riuscita ad attirare nel corso degli anni. A favorire la sua espansione è stato anche il supporto ricevuto dal CEO di Tesla Elon Musk, il quale l'ha definita la sua criptovaluta preferita.

Attraverso questo progetto abbiamo quindi cercato di capire quanto fossero correlati fra loro l'interesse delle persone, il loro entusiasmo (hype in inglese) e le fluttuazioni del prezzo di Dogecoin, sia nel breve che nel lungo termine. Abbiamo provato a comprendere quanto potessero essere incidenti le idee e le dichiarazioni di una persona così influente come Elon Musk rispetto al valore di una moneta basata così tanto sul hype.

1. Descrizione dei dati

1.1 Dati veloci

Per analizzare l'interesse delle persone e il loro stato d'animo in correlazione alle fluttuazioni del prezzo di Dogecoin, si è deciso di fare uso di alcuni dati veloci provenienti da due diverse fonti.
I dati riguardanti il prezzo della criptovaluta sono stati ricavati da Coingecko, piattaforma nata nel 2014 e dedicata interamente al mondo delle criptovalute che tiene traccia di diverse informazioni riguardanti più di 8000 token diversi. Il sito web mette a disposizione degli sviluppatori in modo totalmente gratuito le proprie API, utilizzabili sia per ricavare dati storici, sia per ricevere informazioni varie riguardanti diverse criptovalute in tempo reale. [1]
In particolare, noi abbiamo impostato la nostra pipeline in modo che eseguisse una chiamata ogni 60 secondi e che ricevesse le seguenti informazioni riguardanti Dogecoin:

- Prezzo corrente (euro)
- Valore del market cap corrente (euro)
- Volume delle ultime 24 ore (euro)
- Timestamp dell'ultimo aggiornamento del prezzo

Ogni risposta arriva in formato json e racconta la situazione della criptovaluta al momento della richiesta.
Per misurare l'interesse delle persone invece si è deciso di basarsi su Twitter. Sono state utilizzate le API fornite dal social network per poter ricevere ogni tweet o retweet che contenesse all'interno del testo o degli hashtag collegati al post la parola "Dogecoin". Ogni tweet postato arriva in formato json, contenente moltissime informazioni riguardanti il post e l'utente creatore, solo alcuni di questi dati sono stati mantenuti come sarà successivamente spiegato nel paragrafo 2.3.

1.2 Dati storici

Utilizzando nuovamente le API crypto di Coingecko è stato possibile prelevare la serie storica del prezzo di Dogecoin nel periodo di riferimento (settembre 2020 - gennaio 2021). I dati ottenuti contengono diverse informazioni utili, come il numero di followers su diversi social in un determinato giorno e indici specifici di Dogecoin (volume di scambio giornaliero e market cap).
In aggiunta è stato affiancato un dataset che raccoglie i diversi tweets di Elon Musk. Questo con l'obiettivo di studiare la variazione del prezzo in relazione ai comportamenti social del CEO di Tesla.
Prima di essere utilizzati, i dataset sono stati puliti e alleggeriti da eventuali attributi superflui per le nostre analisi. Inoltre, dal registro di tutti i tweets di Elon Musk sono stati filtrati solo i record utili per le nostre indagini, si è pertanto ricercato solamente le stringhe di testo che contenessero specifiche parole chiave ("Dogecoin", "crypto", "doge", etc.) nell'arco di tempo preso in esame.
Tutti questi passaggi sono stati svolti attraverso script python che, una volta finita la parte di data processing, caricano i risultanti dataset in un database MongoDB.

2. Architettura

2.1 Apache NiFi

Per la gestione del flusso dei dati, si è utilizzato Apache Nifi.
Apache NiFi è una piattaforma integrata di logistica dei dati per l'automazione dello spostamento dei dati tra sistemi diversi. Assicura il controllo in tempo reale che semplifica la gestione dello spostamento dei dati tra qualsiasi origine e qualsiasi destinazione. È indipendente dall'origine dei dati, supporta fonti disparate e distribuite di diversi formati, schemi, protocolli, velocità e dimensioni come macchine, dispositivi di geolocalizzazione, flussi di clic, file, social feed, file di registro e video e altro ancora. [2]
Attraverso la piattaforma si è gestita tutta la parte di estrazione, preprocessing e storage dei dati raw, in particolare i nodi utilizzati più di interesse sono stati:

- **GetTwitter, InvokeHTTP:** nodi utilizzati per contattare le API di Twitter e Coingecko
- **PublishKafka_2_6, ConsumeKafka_2_6:** nodi utilizzati per pubblicare e consumare messaggi su Kafka
- **JoltTransformJSON, RouteOnAttribute:** nodi utilizzati per la manipolazione dei json e di cui si parlerà in modo più approfondito nel paragrafo 2.3.
- **PutMongo:** nodo utilizzato per il salvataggio dei documenti in MongoDB

Di seguito è possibile vedere una rappresentazione semplificata dell'architettura utilizzata.

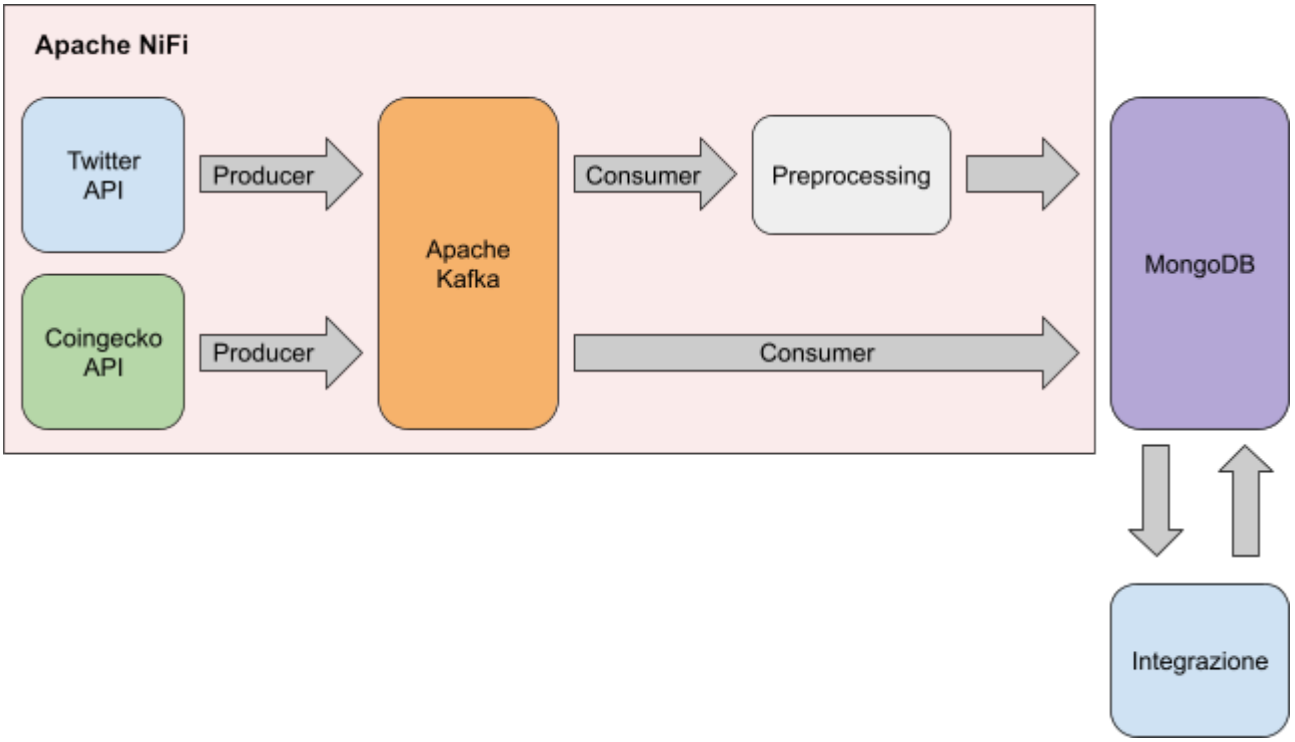

Immagine 2.1: schema dell'architettura adottata

2.2 Apache Kafka

Per gestire lo streaming di dati proveniente da Twitter e da Coingecko si è fatto uso di Apache Kafka. Kafka è una piattaforma per il data streaming distribuita che permette di pubblicare, sottoscrivere, archiviare ed elaborare flussi di record in tempo reale. È progettata per gestire flussi di dati provenienti da più fonti distribuendoli a più consumatori. In breve, consente di spostare grandi quantità di dati da un punto qualsiasi a un altro nello stesso momento. [3]

Sono quindi stati creati due producer per immettere nel cluster Kafka le risposte ricevute dalle API di Twitter e Coingecko e due consumer che si occupano di estrarre i messaggi inseriti dai producer nello storage temporaneo di Kafka, in modo che poi questi dati possano essere sottoposti a preprocessing e successivamente salvati in uno storage adeguato.

2.3 Preprocessing

Al momento dell'arrivo dei dati da Twitter, essi vengono sottoposti ad una procedura di preprocessing. I documenti json che descrivono i tweet forniti dalle API Twitter sono particolarmente complessi, dato il gran numero di informazioni che contengono e dato come cambia la loro struttura in base ad alcune caratteristiche del post. In particolare, si ottengono strutture del json diverse stiamo trattando un tweet o un retweet e se il testo del post è stato troncato per lunghezza eccessiva.
Per riuscire ad ottenere dei documenti omogenei sono state applicate svariate trasformazioni ai dati originali attraverso i nodi forniti da Apache NiFi JoltTransformJSON e RouteOnAttribute.

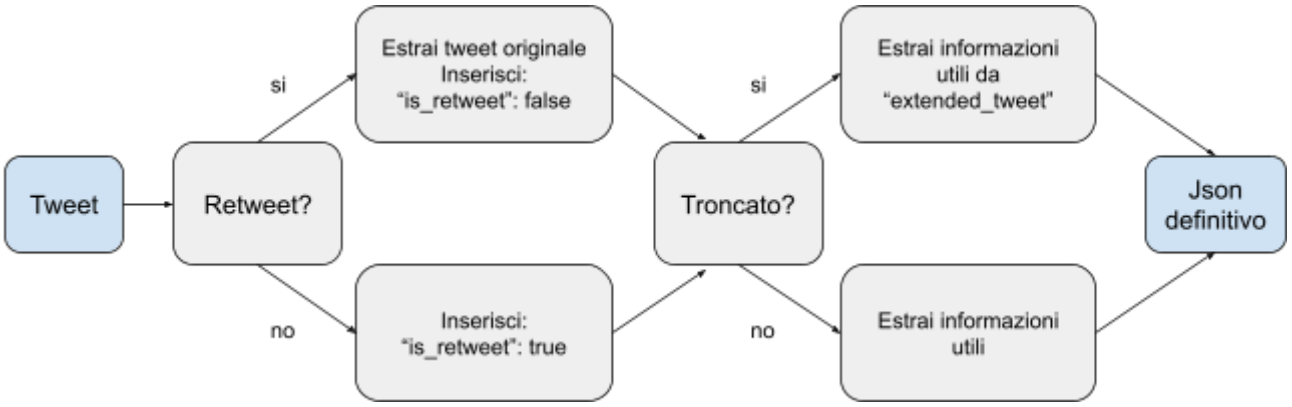

Immagine 2.2: schema delle trasformazioni applicate ai documenti json

Nell'immagine 2.2 è possibile vedere il percorso seguito per effettuare il preprocessing dei dati di Twitter. Innanzitutto viene effettuato il controllo sul campo "isretweet" e in caso di sua presenza se ne estrae il contenuto rappresentante il tweet originale soggetto del retweet. A fine controllo viene inserito un nuovo campo "is_retweet" con possibili valori "true" o "false" per poter sapere in seguito se il documento si riferisce ad un post originale o ad un repost. In seguito a questa prima operazione, si passa all'estrazione delle singole informazioni necessarie nel caso del nostro progetto. Anche in questo caso viene effettuato un controllo su un campo, "istruncated" e nel caso abbia valore "true" sappiamo che le informazioni a noi necessarie sono contenute nel campo "extended_tweet".
I controlli sono stati eseguiti sfruttando i nodi RouteOnAttribute che permettono di far percorrere strade diverse al flusso di dati in base al valore di uno o più attributi, mentre le trasformazioni sono state eseguite attraverso i nodi JoltTranformJSON che consentono, attraverso specifici comandi, di effettuare diversi tipi di trasformazioni.
I documenti json finali riguardanti i tweet sono così strutturati:

```
{
        "timestamp_ms" : long ,
        "user_id" : long ,
        "text" : "text" ,
        "hashtags" : [text] ,
        "is_retweet" : boolean
}
```

2.4 Storage

Per lo storage dei dati si è scelto di utilizzare un database MongoDB, database non relazionale che mantiene al suo interno i dati in forma di documento in stile json. Per noi è stata la scelta più ovvia data la natura dei nostri dati iniziali, completamente in formato json e quindi facili da trattare e salvare attraverso MongoDB.
Sono state create 5 collezioni per lo storage dei documenti:

- DogeHistorical: dati storici riguardanti i prezzi di Dogecoin
- ElonTweets: dati storici rigurdanti i tweets di Elon Musk
- Coingecko: dati veloci sul prezzo corrente di Dogecoin
- Twitter: dati veloci sui tweet riguardanti Dogecoin
- CoinTweetData: dati di Twitter e Coingecko integrati, aggregati e arricchiti (dettagli nel paragrafo 2.5)

2.5 Integrazione e arricchimento

I dati derivanti da Twitter e da Coingecko sono stati integrati fra loro, aggregati e arricchiti per permetterci di studiarne le relazioni e di ricavare informazioni. Queste operazioni sono state automatizzate attraverso uno script ad-hoc in Python programmato per essere eseguito una volta al giorno, andando così ad integrare tutti i dati del giorno prima.
Lo script si occupa di creare un nuovo documento che integra al suo interno tutte le informazioni aggregate di uno specifico giorno. Ogni documento contiene quindi un primo campo “day” che indica a quale giorno i dati contenuti nel json si riferiscono. Subito dopo si presenta una array chiamato “data” che contiene 288 oggetti, ognuno rappresentante uno slice da 5 minuti del giorno descritto dal documento. Ogni slice contiene gli ultimi ottenuti in quel lasso di tempo per quanto riguarda prezzo, market cap e volume di Dogecoin e svariate informazioni riguardanti i tweet registrati negli scorsi 5 minuti. In particolare viene contato il numero di tweet e retweet intercettati e viene inserito un array contenente tutti gli hashtag presenti nei post. Vengono mantenuti solo gli hashtag contenenti lettere e numeri.
Le informazioni riguardo ai tweet vengono arricchite effettuando una sentiment analysis sui testi dei post attraverso la libreria TextBlob. La libreria, dopo aver analizzato il testo del tweet, restituisce un valore di polarità, il quale viene utilizzato per inserire nell’oggetto json anche il conteggio dei tweet riconosciuti come positivi, neutrali o negativi.
Di seguito viene riportata la struttura di un documento integrato:

```
{       “day” : date,
        “data” :  [{    “date” : date ,
                        “price” : double ,
                        “market_cap” : double,
                        “volume” : double ,
                        “number_of_tweet_tot” : int ,
                        “number_of_tweet” : int ,
                        “number_of_retweet” : int ,
                        “hashtags” : [text] ,
                        “posTweets” : int ,
                        “negTweets” : int ,
                        “neutTweets” : int      }]
}
```

3. Visualizzazioni

Tutte le infografiche sono state sviluppate con lo strumento Tableau [5] ed è possibile visionarle nella versione interattiva al link di Tableau Pubblic:
https://public.tableau.com/app/profile/daniel8426/viz/ProgettoDataViz_16227353929240/dogecoin

3.1 Infografica serie storica e tweets Elon Musk

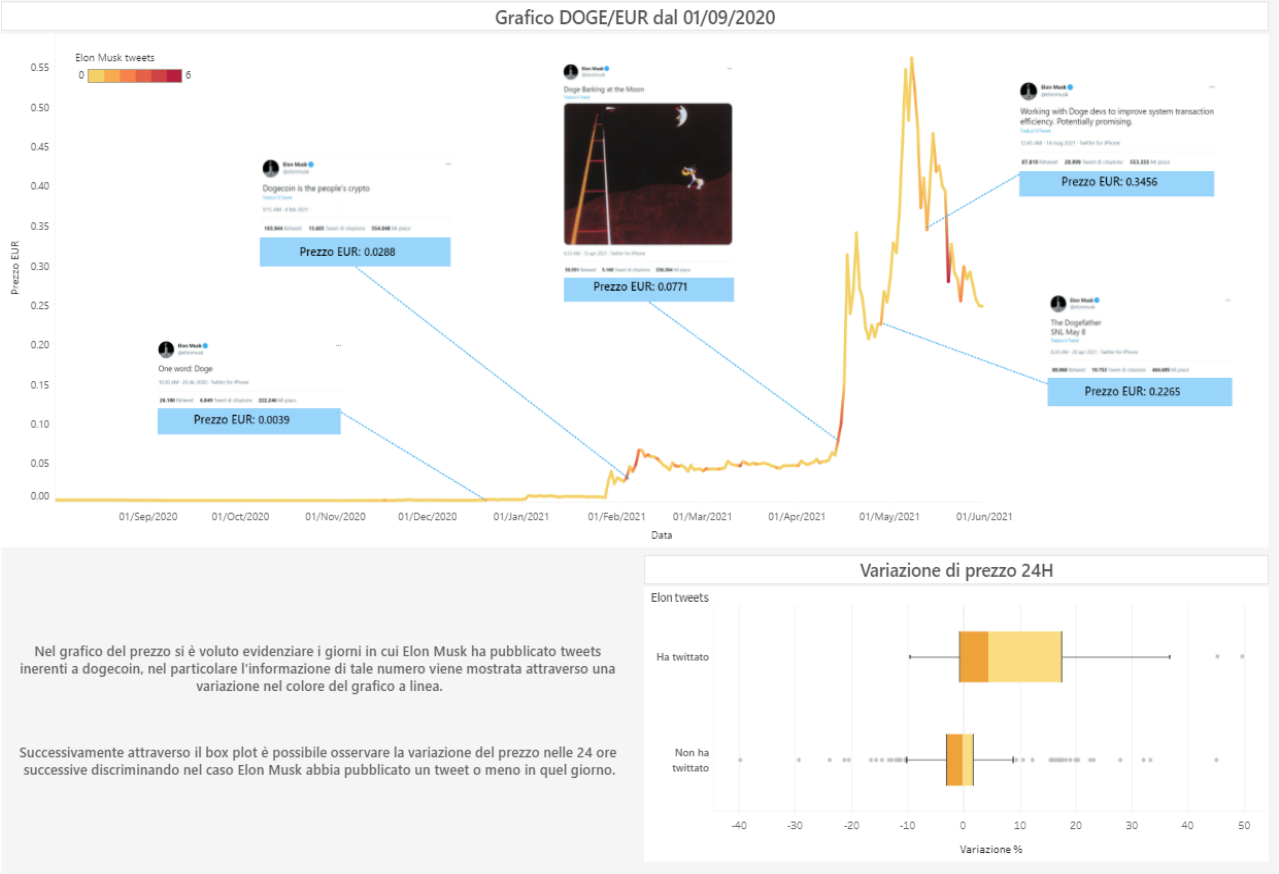

Immagine 3.1: infografica serie storica del prezzo e tweets Elon Musk

Nella visualizzazione relativa all'immagine 3.1 sono stati analizzati i tweets di Elon Musk e la loro influenza sulla variazione del prezzo di Dogecoin dal giorno 01/09/2020. Nella prima visualizzazione si è voluto mettere in risalto i giorni in cui Elon Musk ha pubblicato dei tweets riguardanti la criptovaluta. Ciò è osservabile nelle zone del grafico dove il colore tende al rosso ed eventualmente reso più esplicito dalle immagini allegate che riportano i tweets più importanti prima di una forte movimentazione del prezzo.
Nell'infografica sottostante si può apprezzare tramite un box plot l'effettiva variazione del prezzo delle 24H successive ad un tweet di Elon Musk rispetto al resto dei giorni.

3.2 Infografica interesse social nel tempo

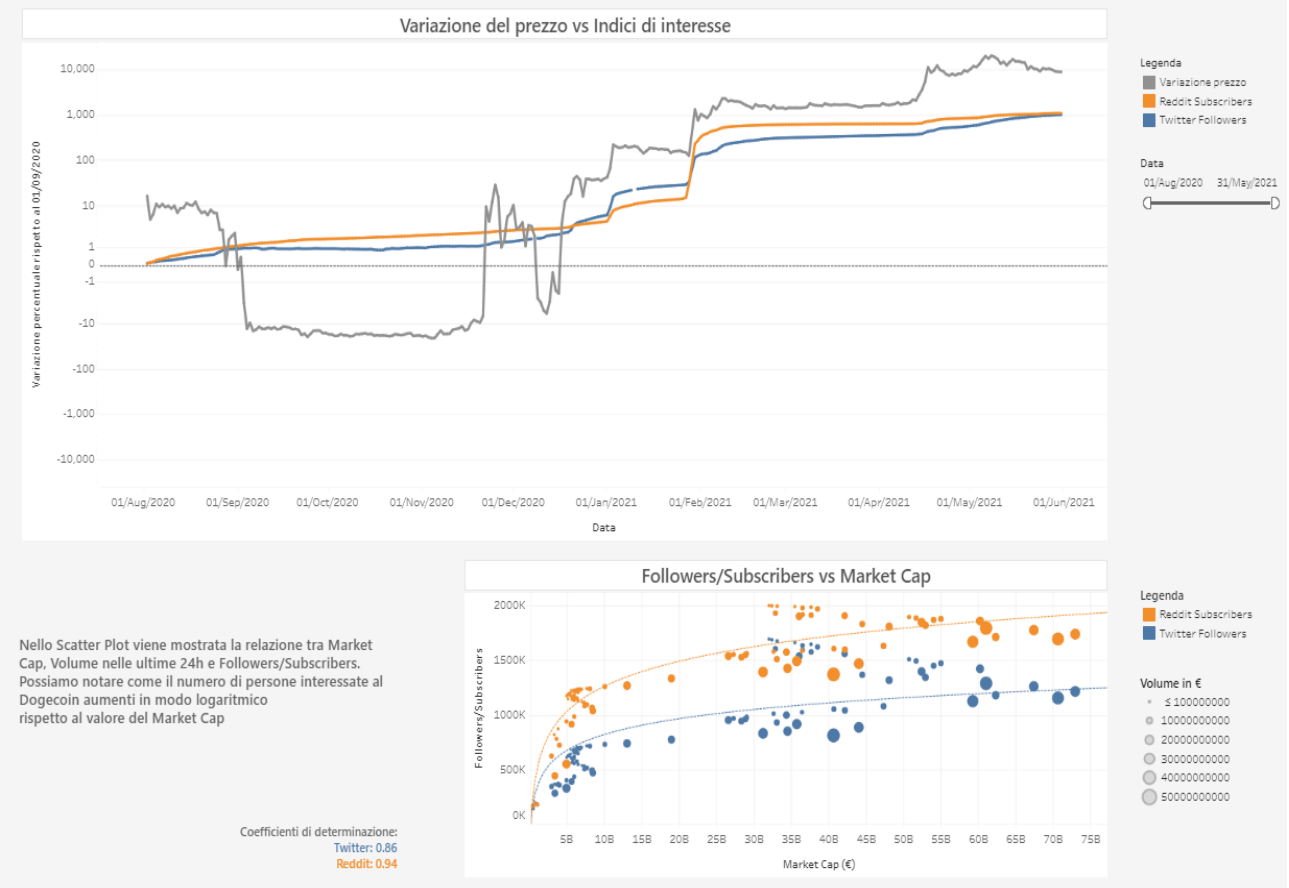

Immagine 3.2: infografica dell'interesse social al variare del tempo

Nella visualizzazione relativa all'immagine 3.2 si è cercato di valutare come l'interesse delle persone riguardo a Dogecoin cambia in relazione alla variazione del prezzo della valuta. Si sono tenuti in considerazione il numero di followers su Twitter e il numero di subscriber su Reddit (forum di riferimento e che vanta una community molto ampia di utenti interessati al mondo delle criptovalute).

Nella prima infografica è possibile apprezzare come i numeri social siano cresciuti proporzionalmente alla crescita del valore della moneta. La scelta di utilizzare una scala logaritmica è legata alla crescita esponenziale che la moneta ha avuto nei mesi di riferimento, rendendo altrimenti di difficile comprensione i dati su una scala lineare. Inoltre lavorare in termini di variazione percentuale ci ha permesso di confrontare diversi valori senza ricorrere a visualizzazioni multiasse, le quali possono essere usate per manipolare le informazioni.

Nell'infografica sottostante si è poi tracciato l'andamento dei valori social (sempre con asse logaritmica) ottenendo un trend con coefficienti di determinazione elevati (rispettivamente di 0.86 e 0.94 per followers Twitter e subscribers Reddit).

3.3 Infografica dati veloci Twitter

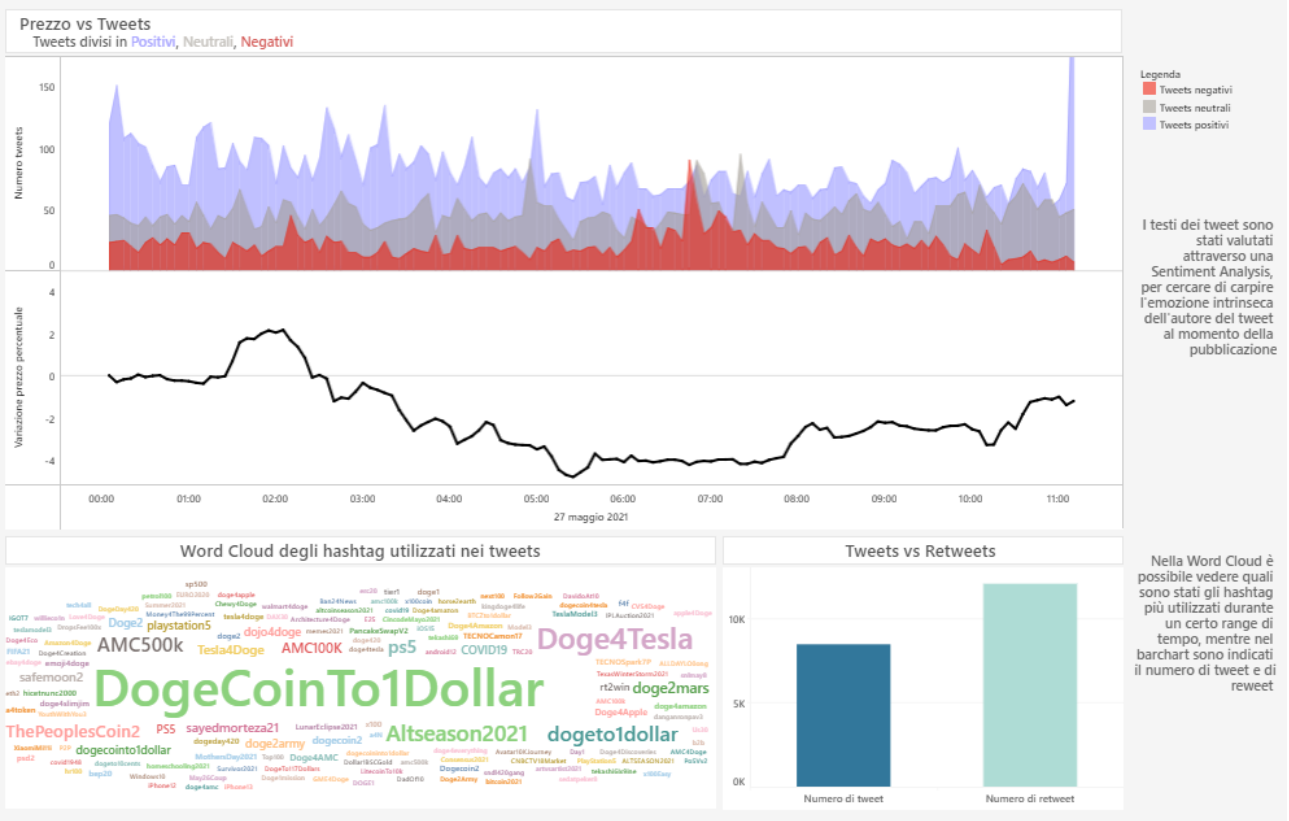

Immagine 3.3: infografica dati veloci twitter ottenuti in tempo reale

Nella visualizzazione relativa all'immagine 3.3 sono stati rappresentati i dati raccolti riguardanti l'andamento del prezzo di Dogecoin nell'arco di alcune ore e le informazioni riguardanti i tweets catturati durante lo stesso arco di tempo.
Nell'infografica principale viene tracciata la variazione di prezzo in percentuale dall'inizio della registrazione, con associati anche i diversi tweets divisi secondo la classificazione effettuata dal sentiment analysis (positivi, negativi o neutri).
Attraverso la word cloud sottostante vengono anche mostrati gli hashtag contenuti nei tweets, la cui dimensione è proporzionale al numero totale di tweets che lo contengono.
Nella versione interattiva dell'infografica è possibile filtrare i dati per un determinato arco di tempo e osservare la conseguente variazione della word cloud e "Tweets vs Retweets" relativamente al filtro selezionato.

3.4 Valutazione empirica ed euristica delle infografiche

Per eseguire una valutazione sulle infografiche da noi create, sono stati effettuati 3 test per valutare i diversi aspetti delle visualizzazioni.
Il primo test si tratta di un test psicometrico, il quale consiste nel dare un punteggio da 1 a 6 alle varie caratteristiche delle infografiche. Il questionario è stato sottoposto a 28 persone e si sono ottenuti in generale ottimi risultati per le prime due infografiche. La visualizzazione dedicata invece ai dati veloci provenienti da Twitter e Coingecko è stata valutata in modo meno positivo rispetto alle altre, motivo per il quale potremmo apportare alcune modifiche durante sviluppi futuri.
Di seguito sono mostrati i 3 stacked barchart che rappresentano la varie risposte degli utenti al questionario in percentuale al totale.

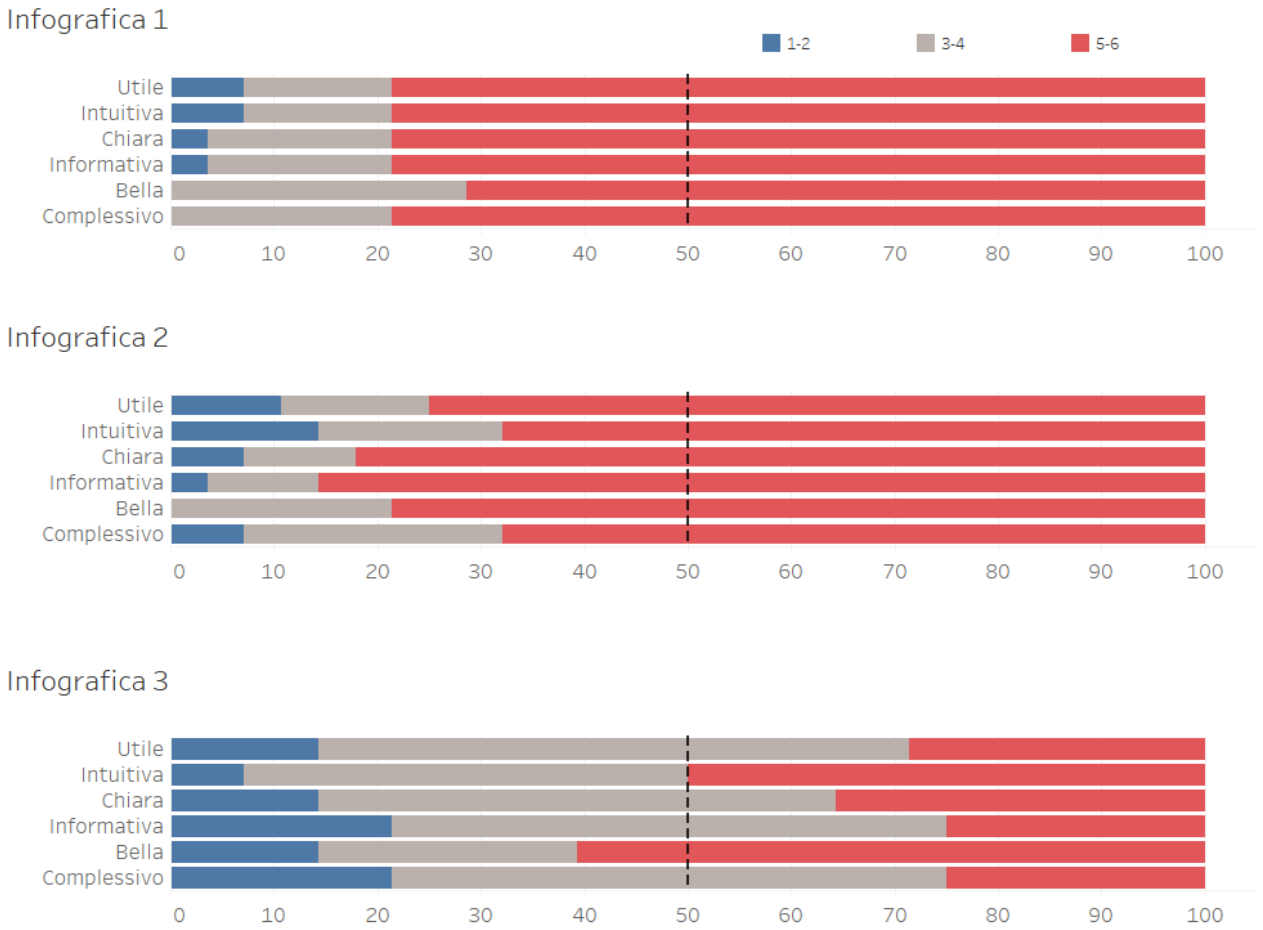

Immagine 3.4: infografica sulle risposte al test psicometrico

Sono successivamente stati sottoposti 7 utenti a delle task interattive riguardanti le infografiche, con l'obiettivo di studiare i risultati in termini di tasso di errore e tempo di risposta per il completamento del compito.
Le domande sono le seguenti:

- Relativamente alla infografica numero 1, qual è stato il giorno con il numero di tweets di Elon Musk più alto?
- Relativamente alla infografica numero 2, qual è la variazione percentuale dei tre indici il giorno 1 febbraio 2021?
- Relativamente alla infografica numero 3, qual è il momento un cui ci sono stati più tweets negativi e qual è l'hashtag più frequente a quell'ora?

Di seguito vengono mostrati i risultati su grafico violin plot.

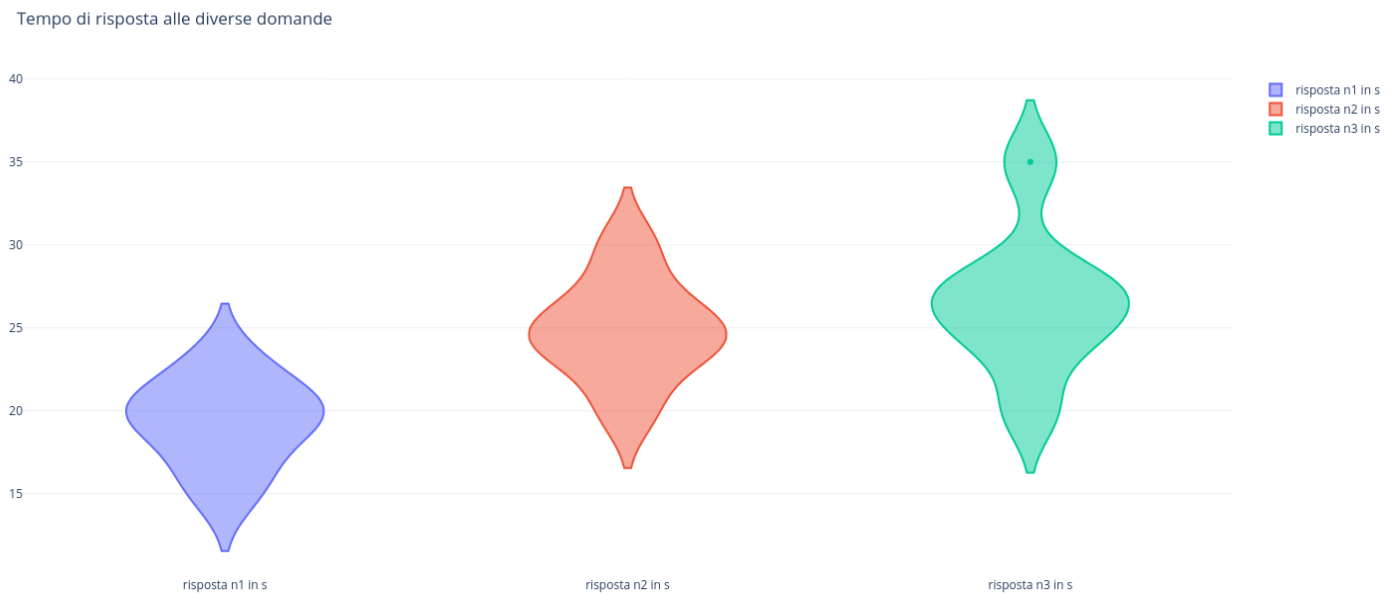

Immagine 3.5: infografica sul tempo di risposta alle domande sottoposte

Inoltre il tasso di risposte corrette relativamente alle tre domande è stato: ~1, ~0.86 e ~0.71.
Per concludere la parte della valutazione euristica è stato chiesto a tre utenti di effettuare un'analisi più in profondità, riportando problemi riscontrati e discutendone con noi ad alta voce.
Due di loro hanno riportato perplessità riguardo gli assi a variazione percentuale, definiti di difficile comprensione e non direttamente interpretabili. Un appunto è stato fatto anche riguardo ai dati veloci raccolti su Twitter, ci è stato suggerito di aumentare il numero di dati registrati così da rendere l'infografica più apprezzabile.
Complessivamente le infografiche sono state valutate curate e interessanti.

4. Conclusioni e sviluppi futuri

Attraverso questo progetto, si è cercato di capire se ci fossero correlazioni fra due degli argomenti di maggiore interesse negli ultime anni: social network e criptovalute.
Attraverso i dati estratti abbiamo potuto notare come il numero di followers di Dogecoin aumentasse in modo correlato al suo prezzo e abbiamo visto che effetti può avere un post di una persona influente come Elon Musk sul valore di una criptovaluta così volatile come Dogecoin.
Per concludere, il progetto potrebbe essere ulteriormente arricchito con un ottimizzazione strutturale per permettere l'archiviazione di dati di grande volume e raccolti in continuazione e in tempo reale. Sarebbe stato interessante aver potuto fare uso delle API avanzate di Twitter per accedere a tutti i tweets storici riguardanti Dogecoin, così da poter effettuare un'analisi più approfondita sul sentimento delle persone in corrispondenza ad eventi significativi, come improvvisi cali o aumenti di prezzo della criptovaluta.
Un altro miglioramento possibile nel futuro, sarebbe l'implementazione di un sistema di alert in real time basato su eventi descritti dai dati veloci estratti, sia riguardanti particolari tweet intercettati, sia riguardanti improvvisi movimenti del valore di Dogecoin.

Referenze

[1] Coingecko, documentazione API - https://www.coingecko.com/en/api

[2] Apache NiFi - https://it.cloudera.com/products/open-source/apache-hadoop/apache-nifi.html

[3] Come funziona Apache Kafka? - https://www.redhat.com/it/topics/integration/what-is-apache-kafka

[4] Jolt, JSON to JSON library - https://github.com/bazaarvoice/jolt

[5] Tableau, software per la visualizzazione di dati e creazione infografiche - https://www.tableau.com/