Il caso GameStop
Un approccio data-driven

Marta la Franca[1] 866590, Alessandra Maggipinto[2] 872562, Alessandro Risaro[3] 825113

[1]Università degli studi di Milano-Bicocca, Data Science

[2]Università degli studi di Milano-Bicocca, Data Science

[3]Università degli studi di Milano-Bicocca, Data Science

Abstract

Nel 2021 si è assistito all' ascesa delle *meme stock*, ovvero azioni di società con scarse prospettive future ed oggetto di un elevato short selling, che negli ultimi mesi hanno registrato rendimenti straordinari per merito di gruppi di investitori retail che interagendo in determinati *subreddit* hanno deciso di acquistare in massa tali azioni, causando non pochi problemi ai grandi *Hedge Fund*, che stavano scommettendo sul loro ribasso. Tra tutte spicca *GameStop*, la quale è stata la prima ad avere un forte impatto mediatico. L'obiettivo di questo lavoro è quello di capire come si sia evoluto il prezzo dello stock *GameStop* (GME) tra l' 11 *Gennaio* 2021 e il 14 *Febbraio* 2021, confrontandolo con il numero di post e il sentiment degli utenti nel subreddit *"wallstreetbets"* e vedere se vi sia una connessione tra le operazioni di scambio compiute su GameStop e il suddetto gruppo Reddit. Inoltre si vuole fare luce su quali fossero le parole maggiormente usate dagli utenti del gruppo, sia in momenti estremamente rialzisti, che in momenti estremamente ribassisti.

Keyword: Gamestop, GME, Reddit, wallstreetbets, short selling, short squeeze, Robinhood

Introduzione

Nel mese di Gennaio 2021 tutti parlavano di GameStop, quello che è successo è stato un vero e proprio caso mediatico. Quella organizzata sul canale Reddit *r/wallstreetbets* viene descritta da molti come una battaglia di Davide contro Golia, in cui l'unione di piccoli investitori retail è riuscita a far tremare *Wall Street*.
Gamestop è un' azienda con un core business molto precario, in quanto si basa sul commercio videogiochi in store fisici, in un mondo in cui gli acquisti di videogiochi in formato digitale direttamente dalle console è sempre in maggiore espansione. La fragilità dell' azienda aveva portato molti *Hedge fund* a scommettere sul suo fallimento, attraverso l'apertura di posizioni ribassiste (*short*). Si cerca ora di far chiarezza su quali siano stati i principali avvenimenti registrati tra l' 11 *Gennaio* 2021 (giorno in cui il *Wall Street Journal* attribuisce l'inizio del rally del prezzo) fino al 4 *febbraio* 2021:

- 11/01/2021: Vengono nominati 3 **nuovi membri del consiglio d' amministrazione**, tra cui *Ryhan Choen* fondatore della piattaforma di e-commerce *Chewy*. Questo annuncio ha portato ad alcuni primi post in cui si parlava di Gamestop su *r/wallstreetbets*;
- 13/01/2021: Si raggiunge un aumento del prezzo delle azioni del 50% rispetto al 11 gennaio;
- 19/01/2021: *Citron Research*, una società finanziaria che si concentra sullo **shortselling** (ovvero sulle vendite allo scoperto), pubblica un tweet apostrofando coloro che detengono azioni *GME* come 'suckers', dicendo che il prezzo tornerà a 20$ molto rapidamente;
- 26/01/2021: **Elon Musk** twitta ripostando il link al subreddit *r/wallstreetbets* aggiungendo un commento: "Gamestonk!!", accrescendo sempre di più l'entusiasmo degli investitori;
- 27/01/2021: Il prezzo aumenta nel pre-market del 140%, aprendo a 354.83$. *Citron* e *Melvin* capital, due Hedge fund che stavano vendendo allo scoperto *GameStop*, chiudono le loro posizioni ribassiste aprendo nuove posizione a rialzo per fare arbitraggio (il così detto **short squeeze**), riportando comunque ingenti perdite.

- 28/01/2021: *Robinhood* e altre piattaforme di trading bloccano le transazioni di *GME*, causando molte polemiche. In questo giorno il prezzo è molto volatile, raggiungendo un massimo di 483$ e un minimo di 112.25$ per azione;
- 29/01/2021: la maggior parte delle piattaforme di trading riabilitano le transazioni delle azioni di *GameStop*.
- 02/02/2021: Dopo il grande rialzo dei giorni precedenti *GME* apre in ribasso del 50% a 140.76$ ad azione;
- 04/02/2021: La segretaria del tesoro degli Stati Uniti D'America **Janett Yellen** ospite ad un noto talk-show televisivo *"Good Morning America"* conferma di aver organizzato un incontro con i regolaratori della *SEC*, della *Commodity Futures Trading Commission* e anche della *FED*, in cui discuteranno se gli eventi recenti meritino ulteriori azioni per far si che i mercati finanziari funzionino correttamente, in modo efficente e gli investitori siano protetti. Questo annuncio causa un ulteriore ribasso delle azioni di *GameStop*, che chiude a $53,33$$ per azione.

---

1 Data Management

1.1 Strumenti e tecnologie utilizzate

Per raggiungere gli obiettivi di ricerca prefissati si è deciso di concentrarsi sulla **Velocità** e sulla **Varietà**. L' acquisizione e la memorizzazione di dati che vengono generati velocemente, come le submission di Reddit, viene effettuata tramite **Apache Kafka**, ovvero una tecnologia middleware che si pone tra il flusso di dati e l' utente attraverso un meccanismo a coda. Nello specifico caso applicativo la velocità è stata solamente simulata, infatti dovendo fare un' analisi a posteriori del fenomeno i post di interesse sono stati acquisiti attraverso l' **API Pushfit** di *Reddit*, portati ad un formato *csv* e in seguito è stata simulata l'acquisizione e la memorizzazione attraverso Kafka. I dati riguardanti prezzo e volume dello stock ad una granularità di 15 minuti sono stati scaricati tramite accesso al **terminale Bloomberg**, un software integrato nel settore finanziario utilizzato per accedere ad informazioni e dati finanziari in real time, che quest'anno a causa dell'emergenza sanitaria è stato reso accessibile agli studenti anche da remoto. I due dataset una volta integrati sono stati caricati su **MongoDB**, mediante un apposito script in **Python**. Inoltre è stato utilizzato **GitHub** per permettere lo scambio e tener traccia dei dataset e degli script utili al lavoro svolto.

1.2 Raccolta dei dati

La raccolta dati si è basata principalmente su due fonti: l' API di Reddit e il terminale Bloomberg.
Si è seguito il seguente schema logico:

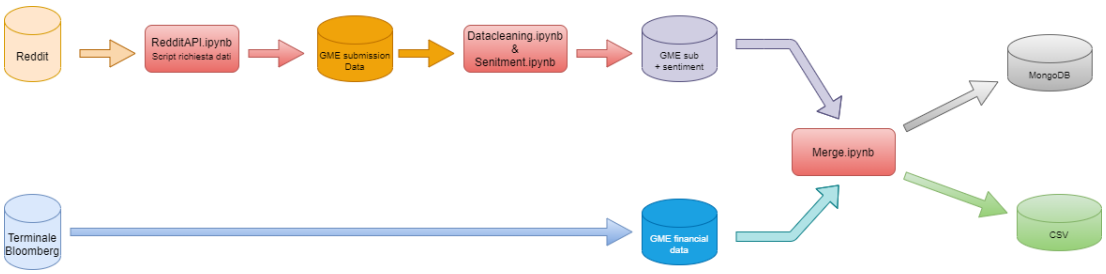

Figura 1: Pipeline progetto Data Management

1.2.1 Reddit API

L' acquisizione dei post pubblicati su *'wallstreetbets'* è stata effettuata utilizzando l' API Pushfit di Reddit, la quale permette attraverso l'end point *https://api.pushshift.io*, i parametri *after* e *before*, inserendo uno specifico subreddit d'interesse in *subreddit* e specificando una o più parole chiave in *q* di accedere ai post e ai commenti di un determinato gruppo reddit, in un arco temporale definito, contententi una o più keywords d'interesse. La raccolta dei dati di

Reddit è stata eseguita tramite il seguente algoritmo presente nello script *redditAPI.ipynb*:

Algorithm 1: redditAPI.ipynb

Data: after & before = periodo temporale di interesse; query = parola chiave per estrarre le submission; sub = in quale subreddit bisogna andare a cercare

Def `getData(`*query, after,before,subreddit*`):`
- url = PushfitRequestURL
- r = requests.get(url)
- data = json.loads(r.text)
- **return** data['data'] ; // ritorna una variabile data che contiene tutte le submission in formato json

Def `collectData(`*submission*`):`
- subData = list()
- r = requests.get(url)
- data = json.loads(r.text)
- title = subm['title']
- url = subm['url']
- author = subm['author']
- sub_id = subm['id']
- score = subm['score']
- created = datetime.datetime.fromtimestamp(subm['created_utc'])
- numComms = subm['num_comments']
- subData.append((sub_id,title,url,author,score,created,numComms))
- subStats[sub_id] = subData ; // inserisco i dati di ogni submission in un dizionario

Inserire i parametri d'interesse: query, before, after, subreddit
Inizializzare countsub = 0 e subStats = {}
data = getData(query,after,before,subreddit)
1 **while** *len(data)>0* **do**
2 - **for** *ogni submission in data* **do**
  - collectData(submission)
  - countsub+=1
- after= l'ultima data delle submission
- data = getData(query,after,before,subreddit)

N.B per ogni richiesta API vengono scaricate un massimo di 100 submission, inoltre è stato usato il seguente PushfitRequestURL:

```
"https://api.pushshift.io/reddit/search/submission/?title=GameStop|GME&size=100&after=1611676800&before=1611756000&subreddit=wallstreetbets"
```

Infine i dati relativi ad ogni submission sono salvati in formato csv seguendo il seguente schema:

Attributo	**Descrizione**
PostID	identificativo univoco delle submission
Url	url della submission
Title	titolo della submission
Author	autore della submission
Score	voto dato dagli utenti alla submission
PublishDate	data, ora e minuto pubblicazione post
Totalnofcomments	numero totale di commenti ad una submission

Tabella 1: schema degli attributi dei dati Reddit

A scopo meramente didattico per simulare la velocità è stata costruita un'architettura *Kafka* che permette l'acquisizione dei dati in tempo reale. Il dataset su cui è stata applicata tale architettura è quello ottenuto tramite lo script *redditAPI.ipynb*.

Sono stati così creati un *Producer* e un *Consumer* utilizzando la libreria *Kafka-Python*, su un *topic* chiamato *reddit*, sono state caricate le submission tramite un ciclo *for* in cui 100 submission per volta vengono trasformate in formato *JSON*, caricate sul broker e dopo aver eseguito queste operazioni si aspetta un periodo di tempo pseudo-casuale tra 5 e 10 secondi. In contemporanea il Consumer consuma i dati che da *JSON* vengono inseriti in un dataframe Pandas.

Il topic è stato creato utilizzando il comando da terminale:
kafka-topics.bat –create –topic reddit –bootstrap-server localhost:9092

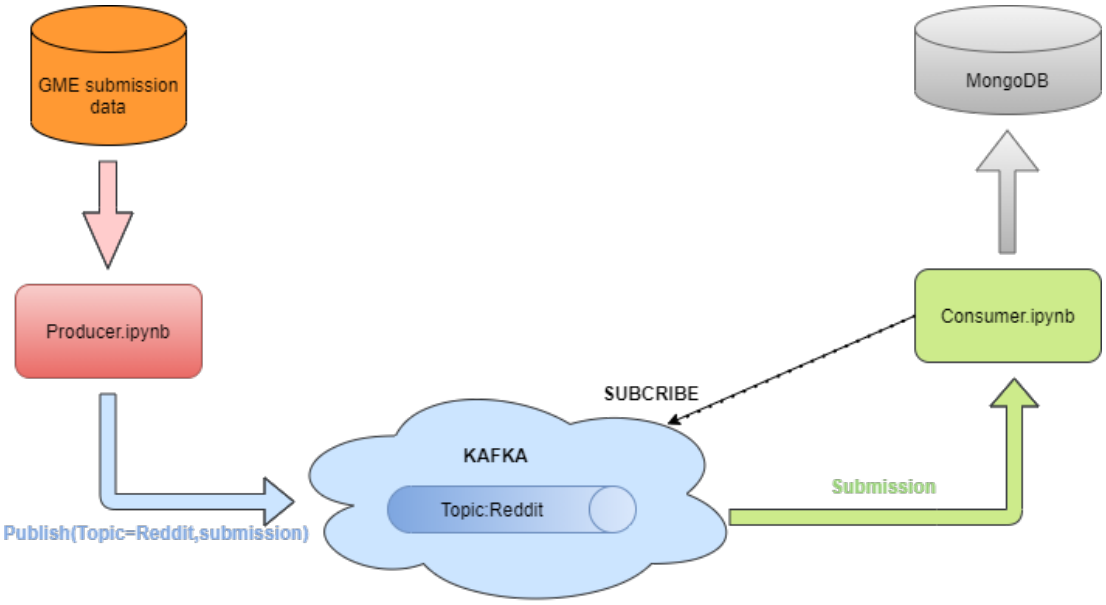

Figura 2: Pipeline architettura Kafka

Di seguito si può vedere nel dettaglio il procedimento dei due algoritmi:

Algorithm 2: Producer.ipynb

Data: submission - insieme delle submissions presenti in "wallstreetbets";
KafkaProducer(data, channel) - sends data to channel kafka

```
for i in range(0, numrow/100) do
    foreach submission do
        submission= reddit(max(100 submissions))
        KafkaProducer(submission, reddit)
        while "GME" submission are finished do
            submission= reddit(max(100 submission))
            KafkaProducer(submission, reddit)
            time.sleep(random(1,5))
```

Algorithm 3: Consumer.ipynb

Data: KafkaConsumer(channel) - receives data from channel kafka

```
consumer.subscribe(['reddit']) - subscribe to topic: submission=empty pandas
 DataFrame
for message in consumers do
    message = message.value
    temp=saveToJson(message)
    submission=concat([submission,temp])
    saveToMongoDB(submission)
```

N.B: I dati sono stati salvati in JSON per avere una maggior facilità nel caricare i dati su MongoDB.

Lo schema logico dei documenti JSON è sostanzialmente invariato.

1.2.2 Acquisizione dati finanziari

I dati riguardanti prezzo e volume delle azioni *GME*, ad una granularità di 15 minuti, sono stati scaricati tramite accesso al *terminale Bloomberg*, un software integrato nel settore finanziario utilizzato per accedere ad informazioni e dati finanziari in real time, che quest'anno a causa dell'emergenza sanitaria è stato reso accessibile agli studenti anche da remoto. Tramite il terminale si è riusciti ad accedere anche ai dati degli scambi pre e post market, così da avere una

visione più ampia di prezzo e volume quando si vogliono indagare specifici giorni. Avere accesso al *terminale Bloomberg* porta a problemi di replicabilità nel tempo, quindi si è pensato ad un modo alternativo per poter acquisire dati finanziari e poter ripetere il lavoro svolto anche nel futuro. Attraverso l'uso dell' **API *yfinance*** che permette di scaricare con un semplice script *Python* dati finanziari a diversi livelli di granularità (1m, 2m, 5m, 15m, 30m, 60m, 90m, 1h, 1d, 5d, 1wk, 1mo, 3mo) dal sito *Yahoo! finance* e restituisce dati in un dataframe o una serie *pandas*.
Il processo di acquisizione dei dati relativi a *GME* viene svolto attraverso lo script yFinance.ipynb in cui devono essere definiti i parametri *Ticker*, *Start*, *End*, *Interval*. I dati vengono salvati in formato csv.
N.B: I dati a 1 minuto sono disponibili solo per i precedenti 7 giorni da quando si lancia lo script, mentre tutti i restanti dati intraday (1m<interval < 1d) sono disponibili solo per i precedenti 60 giorni. I dati acquisiti tramite il *terminale Bloomberg* sono stati scaricati in un formato csv secondo il seguente schema:

Attributo	**Descrizione**
DATE	data, ora e minuto della rilevazione
PRICE	prezzo di *GME*
VOLUME	volume scambiato di *GME*
SMAVG(15)	media mobile a 15 lag temporali

Tabella 2: schema degli attributi dei dati finanziari

1.3 Qualità dei dati

Mentre per i dati finanziari non si sono riscontrati problemi di qualità del dato, non si può dire lo stesso per quanto riguarda le submission di *Reddit*. Infatti tra il 26 e il 27 Gennaio vi sono state 10 ore in cui l'API non ha risposto e i dati sono andati persi, attraverso alcune ricerche è emerso che il problema d'incompletezza è stato causato dalla moltitudine di post e commenti pubblicati su *'wallstreetbets'* dopo il tweet di *Elon Musk* nella notte del 26 gennaio, che hanno portato addirittura alla chiusura per qualche ora del subreddit da parte dei moderatori il giorno seguente.

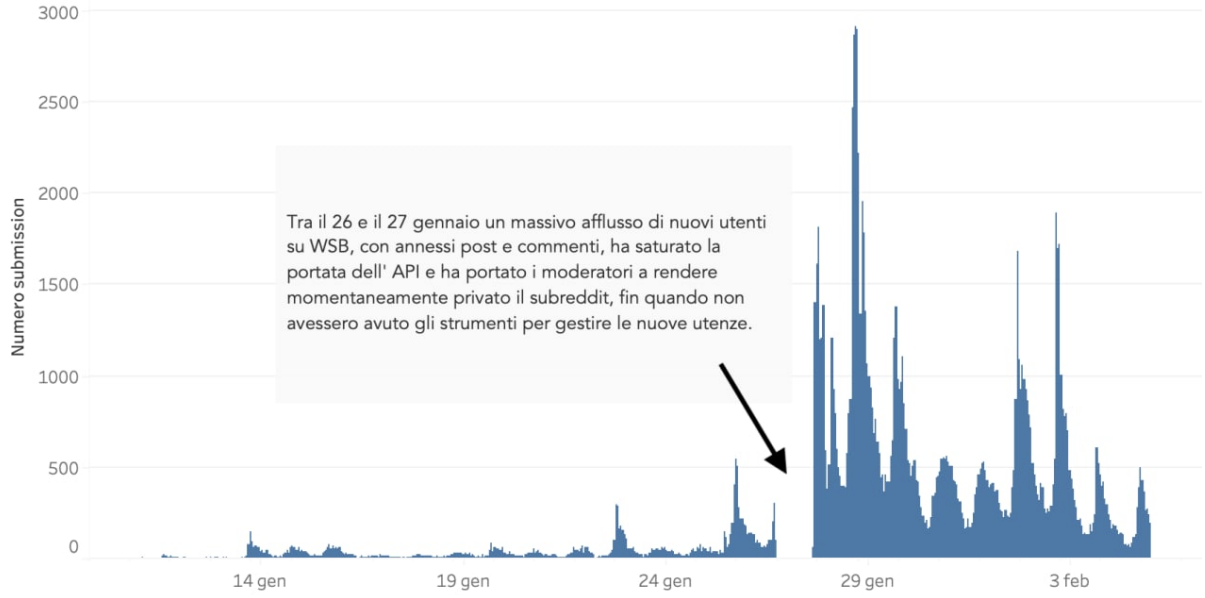

Figura 3: Incompletezza dei dati

1.4 Sentiment Analysis

Sui titoli delle submission di Reddit scaricate attraverso lo script *redditAPI.ipynb* è stato svolto prima una operazione di preprocessing e di pulizia dei dati testuali ed in seguito è stata svolta la sentiment analysis per riuscire a capire come è cambiato l'umore degli utenti di *'wallstreetbets'* durante il periodo temporale d'interesse.

1.4.1 Text Cleaning

Al fine di rendere la Sentiment Analysis il più efficiente possibile sono state applicate alcune strategie di pulizia del testo presentate all'interno dello script *'Data_ Cleaning.ipynb'*, in particolare uno dei problemi principali riscontratosi è stata la presenza delle emoticon e delle emoji, molto utilizzate nel contesto di Reddit. Si è ritenuto che le emoticon siano dati da tenere in considerazione per determinare il sentiment di un post, per questo motivo inizialmente si è svolto un processo di data cleaning in cui sono stati eliminati gli spazi iniziali, gli spazi multipli e sono stati inseriti spazi tra le emoticon, che spesso sono inserite una dietro l'altra, successivamente attraverso l'uso delle librerie *'emot'* e *'emoji'* si è proceduto trasformando le emoji e le emoticon nel loro corrispettivo verbale (ad esempio trasformando l'emoji del razzo nella parola "Rocket"). Infine l'operazione è stata completata attuando altre strategie di cleaning, in particolare: rimozione degli URLs presenti nei titoli; rimozione della punteggiatura; trasformazione di tutte le lettere in minuscolo; rimozione di tutti gli HTML tags; rimozione di tutte le stop word attraverso l'uso della libreria NLTK.

1.4.2 Get sentiment

Il sentiment è stato ottenuto attraverso la libreria *NLTK* (Natural Language Toolkit) di *Python*, in particolare usando un tool specifico chiamato *Vader*, il quale viene utilizzato per determinare il sentiment di testi provenienti da post pubblicati sui Social Network.

L'output che si ottiene una volta eseguita la sentiment con *Vader* consiste in 4 metriche:

- Positive: proporzione del testo che viene classificato come sentiment positivo;
- Neutral: proporzione del testo che viene classificato come sentiment neutro;
- Negative: proporzione del testo che viene classificato come sentiment negativo;
- Compound: Singola metrica normalizzata per la valutazione del sentiment, assume valori compresi tra -1 (sentiment estremamente negativo) e +1 (sentiment estremamente positivo). Viene calcolata sommando i punteggi del sentiment di ogni parola nel testo e poi normalizzando il risultato tra -1 e 1 .

1.5 Data integration

Per poter rispondere alle nostre domande di ricerca abbiamo dovuto effettuare un'integrazione tra i dati delle submission effettuate su *wallstreetbets* durante l'intervallo di tempo di interesse e i dati relativi all'andamento delle azioni GME. I due dataset ottenuti sono stati integrati secondo una logica di *left outer join* (dati reddit-dati finanziari) in un unico dataset tramite il linguaggio di programmazione Python, in particolare l'integrazione è avvenuta prima del caricamento dei dati su *MongoDB*, lo script che mostra l'operazione è *merge.py*. Per ricavare le informazioni relative all'andamento delle azioni nella giornata considerata abbiamo effettuato un'**integrazione temporale**, dove la chiave considerata è stata **DATE**, cioè la data comprendente anche le ore e i minuti. Per quanto riguarda le submission è stata considerata la chiave **Publish Date**, cioè la data di pubblicazione. Solo a questo punto i documenti vengono caricati su *MongoDB*.

Alcune date di pubblicazione presentano un orario che non ha alcun riscontro nel dataset finanziaro, a causa della differente granularità dell'orario di acquisizione dei dati finanziari e degli orari di pubblicazione delle submission con cui vengono raccolti i dati. Per favorire l'integrazione dei dati, si è deciso di affrontare tale problema arrotondando tutti i dati relativi alle submission al quarto d'ora più vicino. Questa correzione è stata effettuata direttamente all'interno dello script *merge.ipynb*, in un'ottica *process-driven*.

1.6 Storage su MongoDB

Si è deciso di utilizzare come *DBMS MongoDB*, pertanto i dati integrati sono stati caricati in formato *JSON* sul software *MongoDB* tramite il linguaggio *Python*, ed in particolare la libreria Pymongo, in un database chiamato **DataMan** ed inseriti in una collezione denominata **gmereddit**. Questo approccio è pensato anche nell'ottica in cui si abbia un maggiore volume di dati da gestire, in quanto *MongoDB* attraverso il sistema di sharding permette sia di replicare i dati su più *cluster* per avere una buona tolleranza al fallimento, sia il ribilanciamento automatico

da un cluster sovraccarico di dati ad un *cluster* scarico, permettendo così una buona scalabilità orizzontale. In seguito si sono eseguite alcune *query*, per testare il funzionamento dei dati uniti attraverso lo script *merge.py*. In particolare:

- *query 1*: trova un documento qualsiasi, viene estratto in particolare il primo documento.

  ```
  gmereddit_collection.find_one();
  ```

- *query 2*: trova un documento in cui vi sia una submission che presenta uno score pari a 8.

  ```
  gmereddit_collection.find_one( {"Score" : 8} );
  ```

---

2 Data Visualization

La scelta delle visualizzazioni è stata fatta per rispondere alle domande di ricerca di questo lavoro:

1. Esiste una relazione tra prezzo delle azioni *GameStop* e numero di post pubblicati su *"wallstreetbets"*?
2. Esiste una relazione tra prezzo delle azioni *GameStop* e l'umore degli utenti di *"wallstreetbets"* e quest'ultimo influenza il volume di scambio?
3. Come cambiano le parole utilizzate dagli utenti di *"wallstreetbets"* per descrivere *GameStop* e qual è il sentiment legato ad esse in diversi archi temporali?

2.1 Le visualizzazioni utilizzate

Per rispondere alle nostre domande di ricerca abbiamo utilizzato 5 infografiche diverse, le prime 3 per avere una visione più ampia del fenomeno di interesse, mentre nelle ultime due si osserva il fenomeno ad una granularità più elevata in alcuni specifici giorni.

N.B i nomi utilizzati per descrivere le visualizzazioni sono stati ricavati dal seguente catalogo *https://www.data-to-viz.com/*

Prima infografica La prima infografica è stata ottenuta dalla combinazione di due diverse visualizzazioni:

- Una **Treemap** interattiva che permette all'utente di indagare in un arco temporale selezionato quali sono le parole maggiormente utilizzate dagli utenti di "wallstreetbets". L'utente può inoltre decidere quante parole visualizzare all'interno della *treemap*, che si presentano in base alla loro frequenza, di default vengono visualizzate 15 parole;
- Un **Percent stacked bar chart** interattivo che permette all'utente di indagare in un arco temporale selezionato quale sia la percentuale di submission positive, negative e neutre.

La combinazione di queste due visualizzazioni permette all'utente di rispondere alla terza domanda di ricerca ed è anche possibile utilizzare questa infografica in combinazione alle altre infografiche per avere una visione più completa del fenomeno d'interesse.
Di seguito viene riportata una rappresentazione sommaria dell' infografica in cui si considera l'arco temporale di osservazione dell'intero fenomeno:

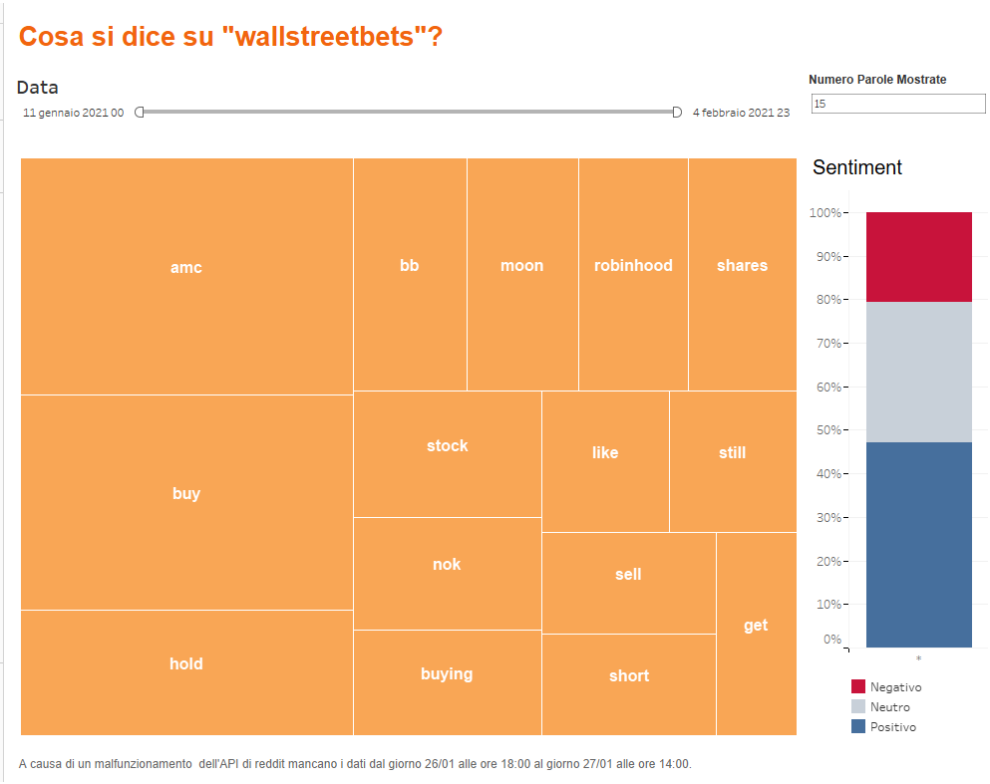

Figura 4: Prima infografica

Seconda infografica la seconda infografica è stata ottenuta dalla combinazione di due diverse visualizzazioni:

- Uno **Scatterplot** che presenta sull' asse delle ascisse il numero di post pubblicati giornalmente su *"wallstreetbets"* e sull' asse delle ordinate il prezzo di chiusura giornaliero delle azioni *GameStop*. Ogni osservazione rappresentata all'interno dello scatterplot corrisponde ad un determinato giorno. Data la presenza di outliers per favorire la leggibilità del grafico ad entrambi gli assi è stata applicata la trasformazione logaritmica.
- Due **Box plot** che rappresentano la distribuzione delle osservazioni rispetto alle variabili degli assi dello scatterplot. Sono stati aggiunti con l'idea di rendere più chiara al fruitore dell'infografica la reale distribuzione delle variabili, che all' interno dello scatterplot era stata "distorta" dall' applicazione della trasformazione logaritmica ad entrambi gli assi.

La combinazione di queste due visualizzazioni permette all' utente di rispondere alla prima domanda di ricerca precedentemente posta.

Viene in seguito riportata la seconda infografica:

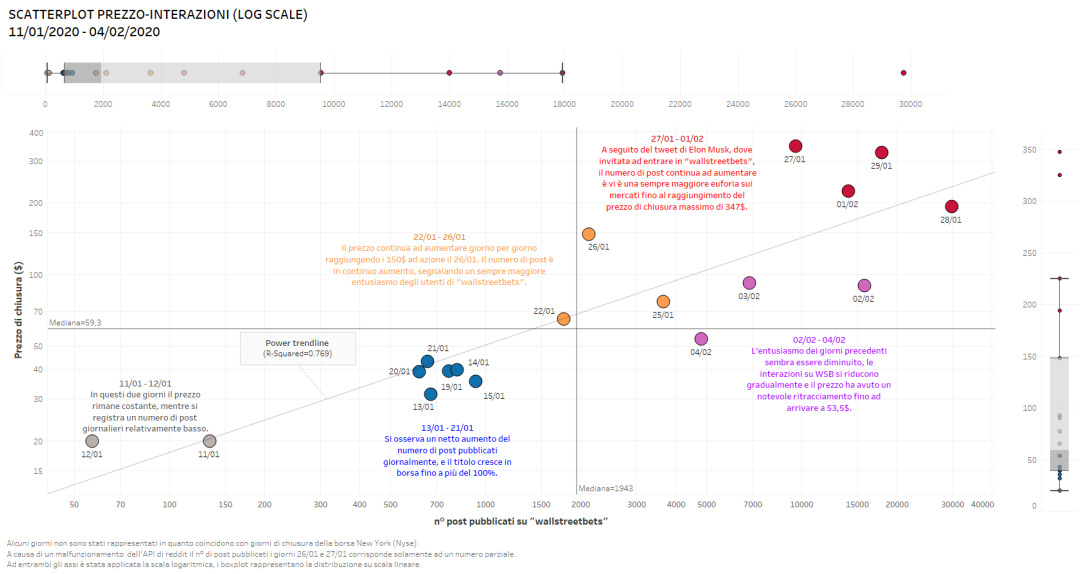

Figura 5: Seconda infografica

Terza infografica anche la terza infografica è stata ottenuta dalla combinazione di due diverse visualizzazioni:

- Un **Candle stick** che permette, attraverso l'uso delle così dette *candele giapponesi*, di analizzare tramite il corpo della candela la variazione tra prezzo di apertura e prezzo di chiusura giornaliero di *GameStop* e anche di capire quanto il prezzo giornaliero sia stato volatile tramite l'ombra superiore e l'ombra inferiore che determinano rispettivamente il massimo ed il minimo raggiunto dal prezzo durante la giornata di borsa. Alla candela viene dato un colore blu quando il prezzo di chiusura è maggiore del prezzo di apertura (ovvero quando vi è stato un rialzo del prezzo), mentre in caso di ribasso del prezzo le è assegnato un colore rosso. La scelta di rappresentazione del prezzo tramite un grafico a candela è stata preferita alla rappresentazione del prezzo tramite un *line chart*, in quanto questa avrebbe tratto in inganno l'utente facendo emergere anche dei valori di prezzo nei giorni in cui la borsa è stata chiusa.

- Un **Lollipop Chart** che rappresenta l'andamento del sentiment medio giornaliero nel periodo di nostro interesse. E' stato inoltre aggiunto il valore medio del sentiment registrato nell'arco temporale considerato con la relativa banda di normalità, così da poter capire la presenza di giorni che hanno registrato un sentiment non normale.
  Viene di seguito riportata la terza infografica:

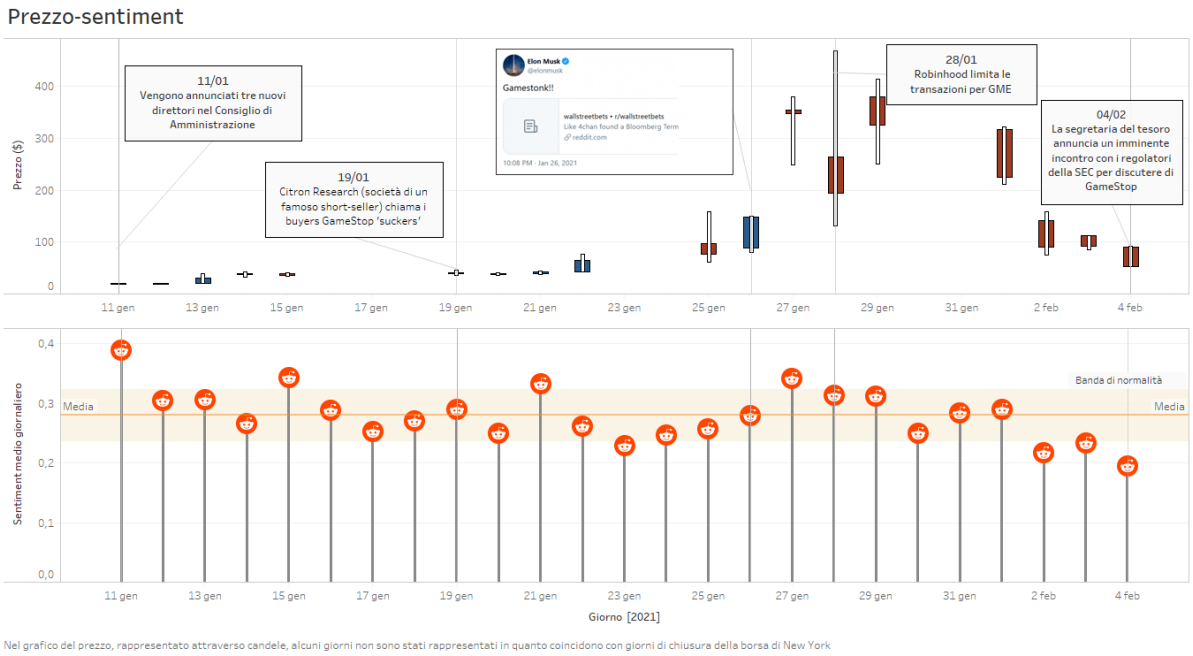

Figura 6: Terza infografica

La combinazione di queste due visualizzazioni permette di rispondere alla seconda domanda di ricerca, anche se questa viene indagata ad un basso livello di granularità.

Quarta e **quinta infografica** queste due infografiche sono state sviluppate con l'idea di andare ad osservare il fenomeno ad un maggiore livello di granularità, si concentrano su due giorni di particolare interesse, rispettivamente il 28 Gennaio (il giorno di maggiore volatilità) e il 4 Febbraio (il giorno dei regolatori). La quarta e la quinta infografica derivano dalla combinazione di tre visualizzazioni:

- Un **Line chart** che rappresenta l'andamento del prezzo delle azioni di Gamestop in fase di *pre-market*, *regular market* e *post-market* nel giorno d'interesse;
- Un **Area chart** che rappresenta l'andamento del volume di scambio delle azioni di GameStop in fase di *pre-market*, *regular market* e *post-market* nel giorno d'interesse;
- Un **Line chart** che rappresenta l'andamento del sentiment degli utenti di *"wallstreetbets"* in fase di *pre-market*, *regular market* e *post-market* nel giorno d'interesse. E' stato inoltre aggiunto il valore medio del sentiment registrato durante il giorno considerato con la relativa banda di normalità

La combinazione di queste tre visualizzazioni permette di osservare da più vicino il fenomeno d'interesse, così da poter dare una "lente di ingrandimento" su determinati eventi al fruitore delle infografiche.

N.B: Il line chart rappresentante il prezzo presenta in entrambe le infografiche un asse troncato, per evidenziarlo è stato utilizzato *Power Point*, quindi non compare in *Tableau Public*

Vengono di sotto riportate la quarta e la quinta infografica:

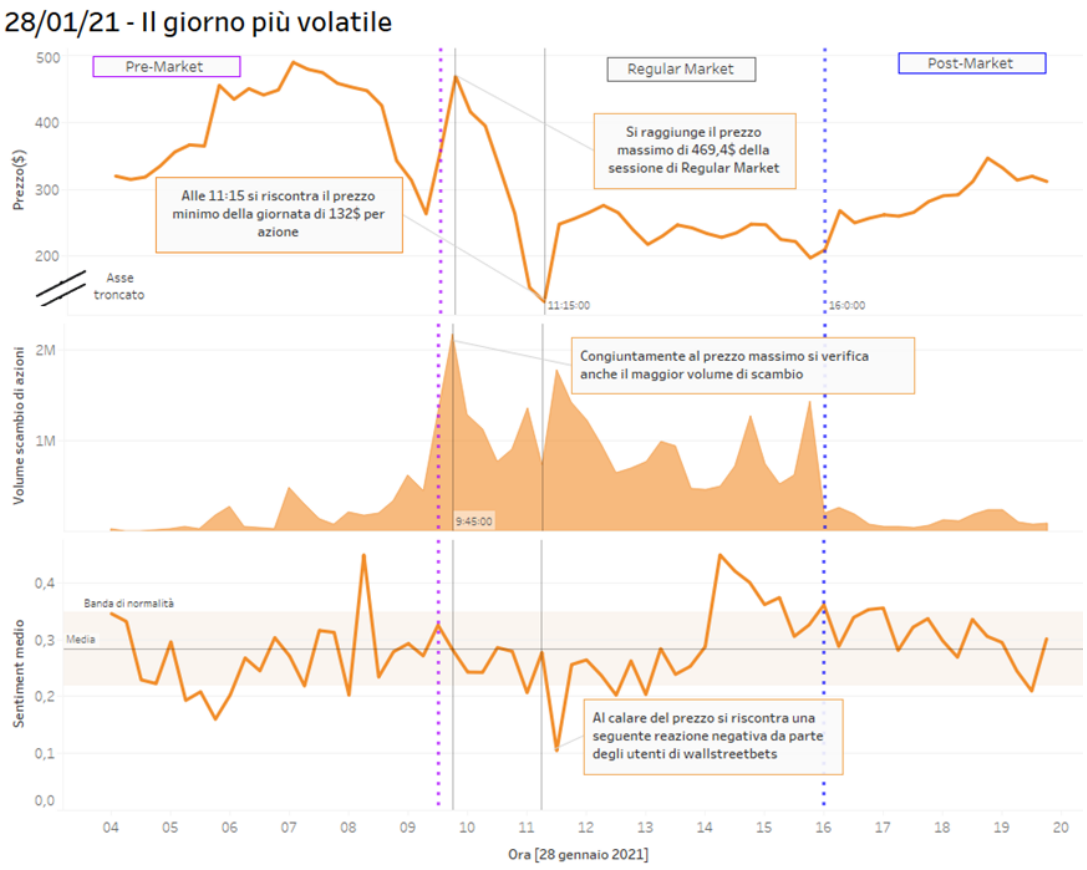

Figura 7: Quarta infografica

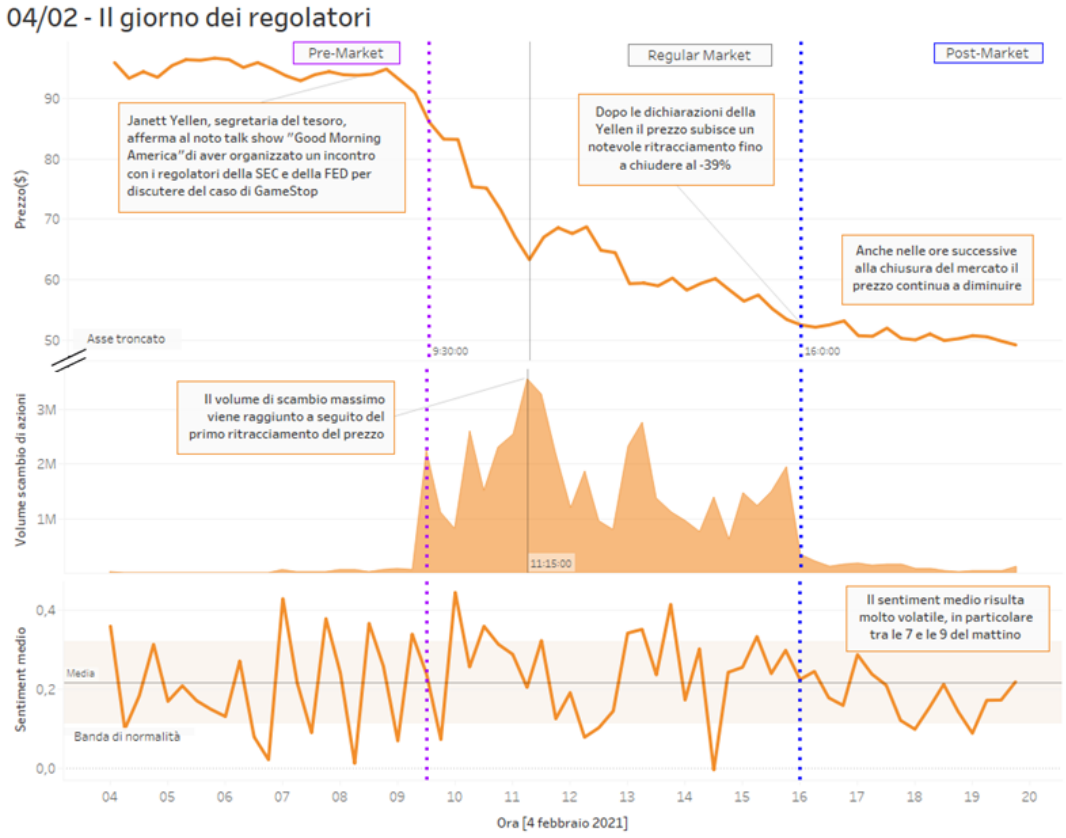

Figura 8: Quinta infografica

2.2 Valutazione qualità

Le iniziali infografiche prodotte sono state sottoposte a tre diverse tipologie di valutazione di qualità:

- Valutazione euristica
- Valutazione tramite utilizzo di user test
- Valutazione tramite questionario psicometrico

Le valutazioni hanno permesso il miglioramento delle visualizzazioni, fino al raggiungimento del prodotto finale precedentemente presentato.

2.2.1 Valutazione Euristica

Durante questa prima valutazione qualitativa le infografiche sono state sottoposte a sei persone, lasciandoli completa libertà di esplorazione e gli è stato chiesto di fare un *"think aloud"*, ovvero un ragionamento a voce alta su quello che stavano vedendo e che stavano capendo. Quando l'utente si bloccava o presentava difficoltà nell'interpretazione delle infografiche è stato registrato il problema presentatosi.

Vengono di seguito esposti i problemi che si sono presentati e le correzioni attuate:

- *Problema 1*: Nella seconda infografica l'utente cercava di trovare delle correlazioni tra punti dello scatterplot legati da colori simili (es. punti in blu e punti in azzurro)
  Soluzione 1: Si è deciso di cambiare i colori dei punti dello scatterplot usando colori molto diversi tra loro
- *Problema 2*: Sempre nella seconda infografica l'utente aveva problemi nella lettura dei valori presenti nei boxplot
  Soluzione 2: Si è deciso di aggiungere un asse proprio in scala lineare ad ognuno dei due boxplot
- *Problema 3*: Nella terza infografica l'utente aveva difficoltà a capire quale candela corrispondesse a quale barra del lollipop chart
  Soluzione 3: Si è deciso per i giorni di particolare interesse di aggiungere delle linee verticali che congiungono la candela alla barra del lollipop chart
- *Problema 4*: Sempre nella terza infografica molti utenti hanno riscontrato difficoltà nella lettura delle candele
  Soluzione 4: Dato che il candle stick è stato ritenuto essere il miglior modo per rappresentare i dati finanziari senza trarre l'utente in inganno, si è deciso di accettare che alcuni utenti non riescano a capire il significato delle candele
- *Problema 5*: Nella quarta e quinta infografica l'utente non riusciva a comprendere a cosa si riferissero le linee tratteggiate (indicanti l'apertura e la chiusura del regular market)
  Soluzione 5: Si è deciso di associare alle caption indicanti *pre-market* e *post-market* lo stesso colore delle linee che ne determinavano rispettivamente la fine e l'inizio

2.2.2 User Test

Durante questa fase della valutazione sono state presentate le infografiche a dodici persone a cui sono state sottoposte delle task da svolgere per ogni infografica, in queste task era richiesto all'utente di rispondere attraverso un approccio interattivo.

1. Prima infografica
   - Qual è la parola più utilizzata all'interno di WSB il 29 gennaio alle ore 09:00? *amc*
   - Qual è la parola più utilizzata e quale la meno utilizzata nel periodo che va dal 19 gennaio 00:00 al 22 gennaio 00:00? *bb, hold*
   - In tutto il periodo considerato (dall'11/01 al 04/02), qual è la percentuale del sentiment neutro? $32,14\%$

I risultati delle task della prima infografica sono stati rappresentati attraverso i seguenti violin plot:

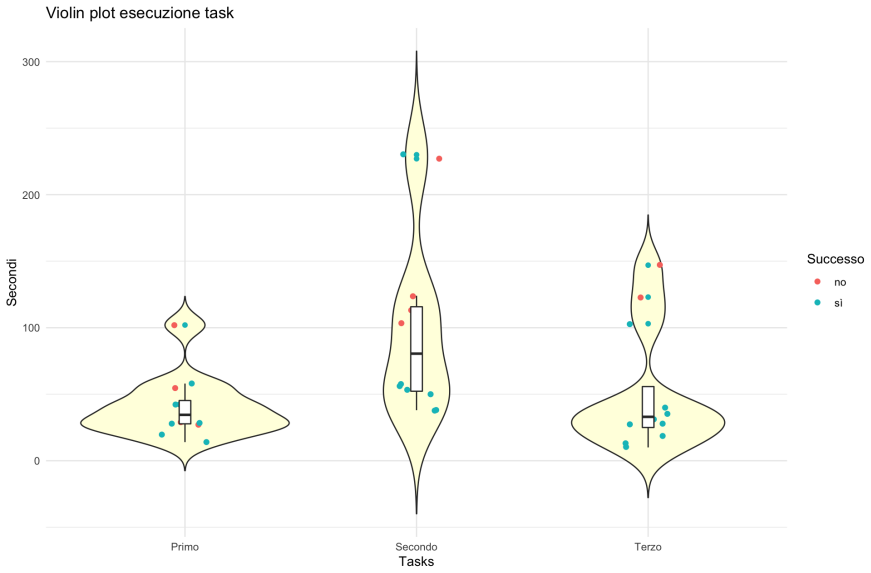

Figura 9: Risultati task prima infografica

Ottenuti questi risultati si è notato una maggiore difficoltà soprattutto per quanto riguarda l'esecuzione della seconda task, la quale richiedeva agli utenti di interagire in modo più complesso con la barra temporale. Si è deciso quindi di cambiare la struttura della barra temporale per rendere il suo utilizzo più semplice all'utente.

2. Seconda infografica

   - In che giorno si è raggiunto il massimo numero di post pubblicati? *28/01*
   - Nel giorno 29/01 qual è il numero effettivo di post pubblicati? *17922*
   - Dopo il giorno 22/01 quale giorno ha registrato un prezzo di chiusura minore del prezzo di chiusura mediano? *04/02*

I risultati delle task della seconda infografica sono stati rappresentati attraverso i seguenti violin plot:

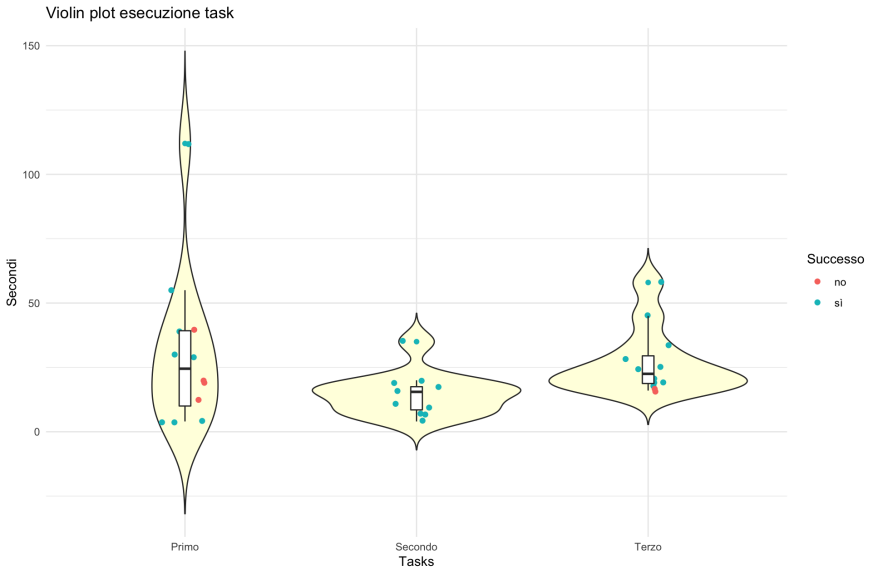

Figura 10: Risultati task seconda infografica

3. Terza infografica

   - Quale giorno presenta il sentiment maggiore? *11/01*
   - In che giorno viene raggiunto il prezzo massimo? *28/01*
   - Come valuti il trend del prezzo dal 01/02 al 04/02: Rialzista o Ribassista? *Ribassista*

   I risultati delle task della terza infografica sono stati rappresentati attraverso i seguenti violin plot:

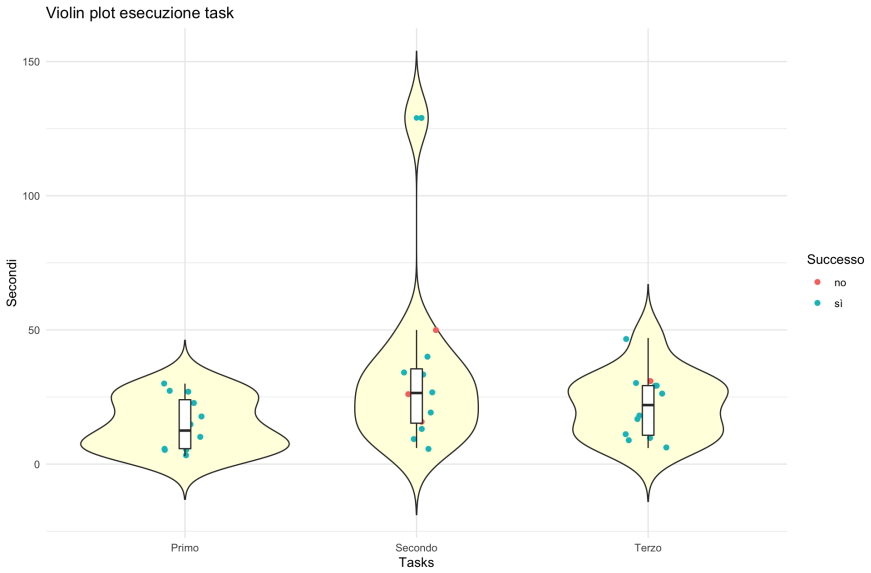

Figura 11: Risultati task terza infografica

4. Quarta infografica

   - Qual è il massimo volume di scambio raggiunto il giorno 28/01? *2178000*
   - A che ora si è raggiunto il sentiment minimo? *11:30*
   - Come valuti il prezzo Post-Market: rialzista o ribassista? *Rialzista*

   I risultati delle task per la quarta infografica sono stati rappresentati attraverso i seguenti violin plot:

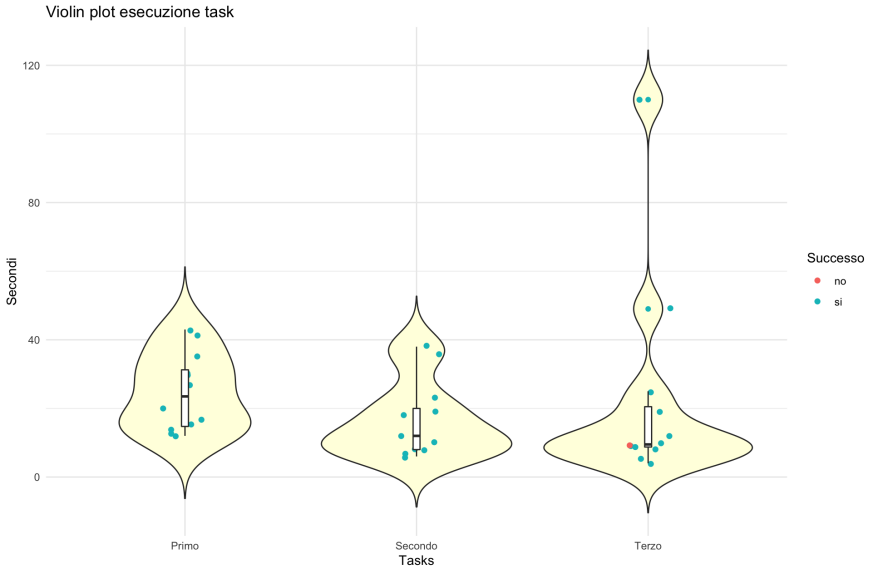

Figura 12: Risultati task quarta infografica

5. Quinta infografica

    - Tra l'orario di apertura e di chiusura del *Regular Market* il prezzo di Gamestop è aumentato? *Falso*
    - A che ora si è verificato il massimo volume di scambio? *11:15*

    I risultati delle task per la quinta infografica sono stati rappresentati attraverso i seguenti violin plot:

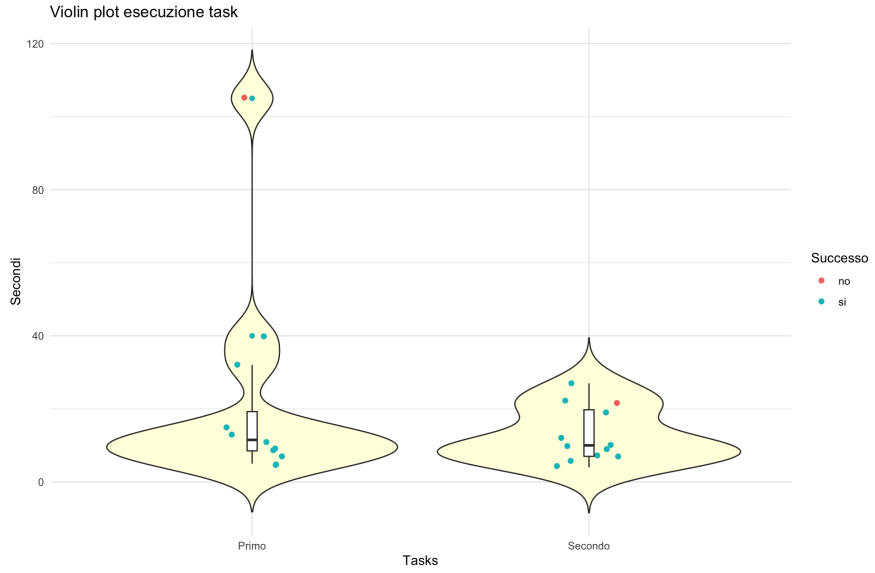

Figura 13: Risultati task quinta infografica

2.2.3 Questionario psicometrico

Nell'ultima fase di valutazione delle infografiche è stato presentato il questionario psicometrico Cabitza - Locoro a 24 persone articolato nelle seguenti domande:

- Come valuti la chiarezza dell'infografica?
- Come valuti l'utilità dell'infografica?
- Quanto valuti bella l'infografica?
- Come valuti l'intuitività dell'infografica?
- Quanto valuti informativa l'infografica?
- Come valuti complessivamente l'infografica?

Le risposte sono state acquisite attraverso lo strumento *"Moduli Google"*. I risultati sono stati rappresentati attraverso l'utilizzo di violin plot e correlogrammi per evidenziare le correlazioni tra le variabili utilizzate nei questionari:

Prima infografica:

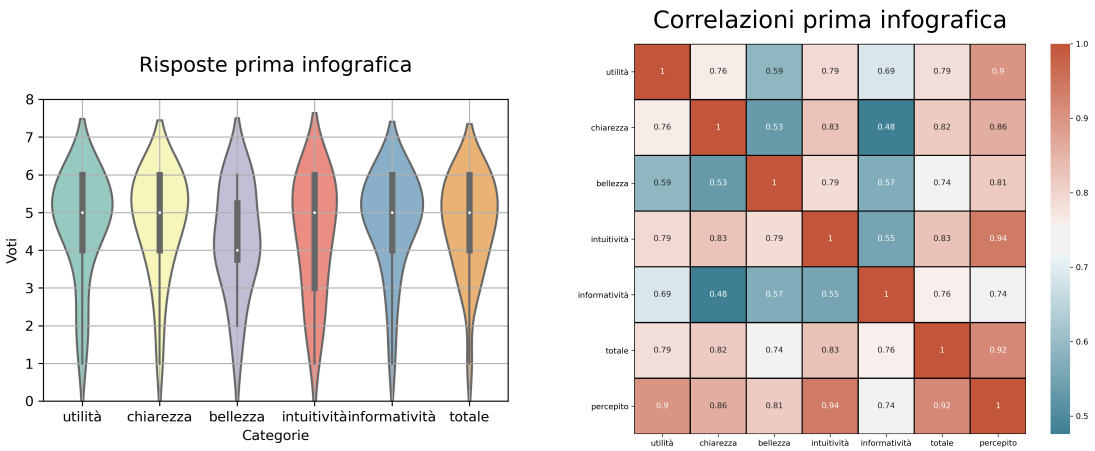

Figura 14: Risultati questionario psiconometrico prima infografica

Seconda infografica:

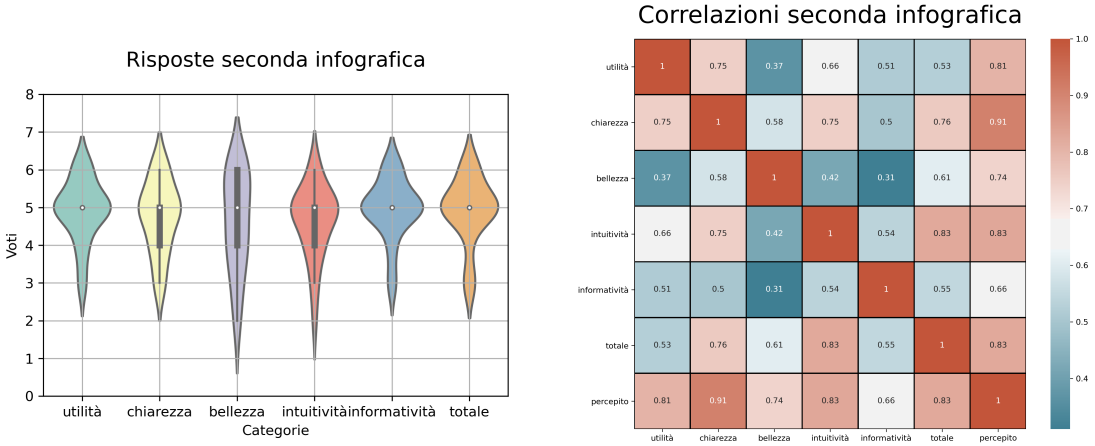

Figura 15: Risultati questionario psiconometrico seconda infografica

Terza infografica:

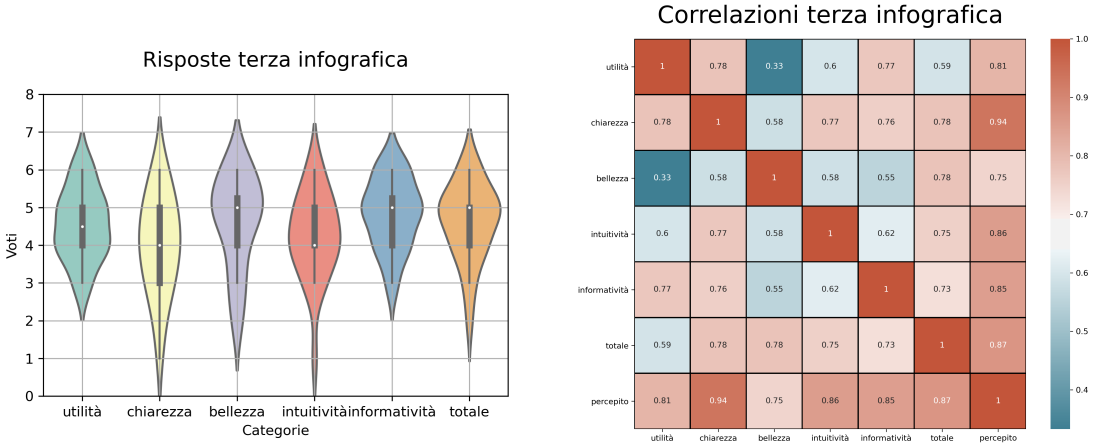

Figura 16: Risultati questionario psiconometrico terza infografica

Quarta infografica:

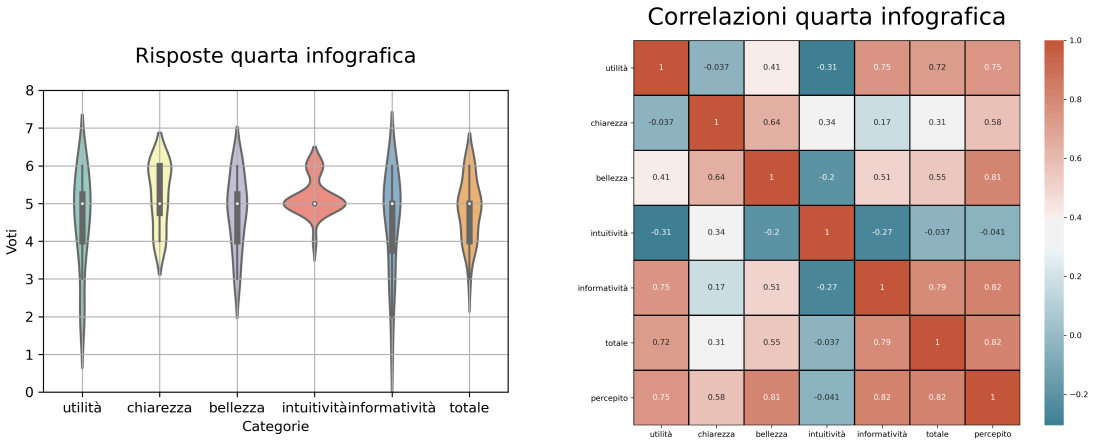

Figura 17: Risultati questionario psiconometrico quarta infografica

Quinta infografica:

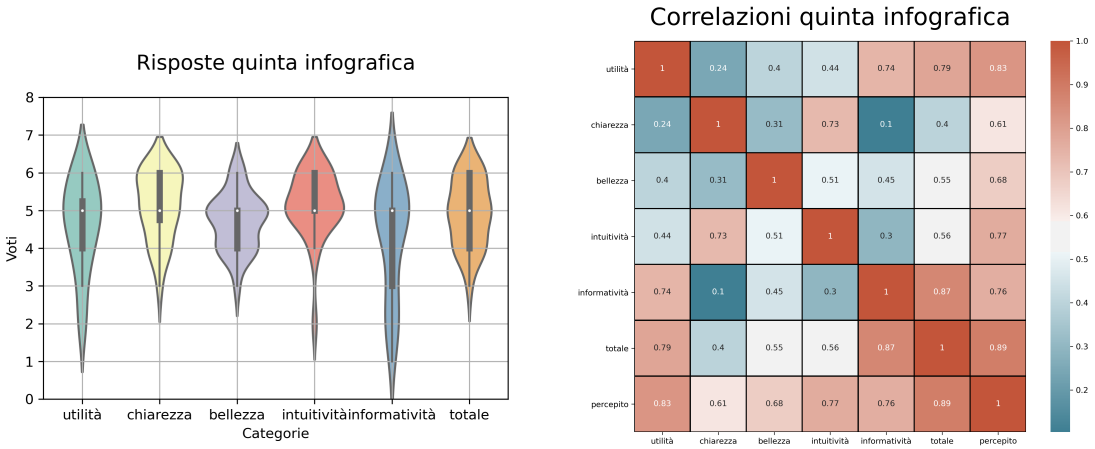

Figura 18: Risultati questionario psiconometrico quinta infografica

Dai risultati è possibile osservare come nella prima e nella seconda infografica l'informatività ha un peso minore sulla valutazione totale. Nella terza infografica ciò che ha peso minore sul totale è la bellezza. Invece nella quarta ad avere un peso minore sul totale è l'intuitività, nella quinta la bellezza. Nella prima infografica la mediana relativa alla bellezza risulta essere minore rispetto le altre, così come la mediana relativa all'intuività nella terza infografica.

---

3 Risultati e prospettive future

3.1 Risultati prima e seconda domanda di ricerca

I risultati di questo lavoro ci portano a pensare che il prezzo delle azioni *GameStop* abbia influenzato il sentiment e il numero di interazioni degli utenti di *"wallstreetbets"*, questo può essere osservato ad un livello macro nella seconda e nella terza infografica. In particolare nella seconda infografica possiamo notare come vi sia una correlazione positiva tra il prezzo di *GME* e il numero di post pubblicati su *WSB*, mentre nella terza infografica è più difficile osservare una correlazione sull'intero arco temporale, si osserva però un sentiment crescente nel periodo tra il 26/01 e il 29/01 che coincide con il maggiore rialzo osservato del prezzo, mentre tra l'01/02 e il 04/02 si osserva una riduzione del sentiment medio in corrispondenza di un importante ritracciamento del prezzo. Sebbene deve essere sottolineato il fatto che in tutto il periodo osservato il sentiment medio giornaliero su *"wallstreetbets"* sia sempre rimasto positivo, questo risultato è in linea con la nota euforia legata alle *"meme stock"* in questo particolare subreddit.
Si è voluto poi osservare il fenomeno con un maggiore livello di granularità concentrandosi in particolare sui giorni 28/01 e 04/02. Il 28 Gennaio è stato un giorno molto volatile per il prezzo,

si è quindi voluto osservare come gli utenti reagissero a questi forti shock, ponendo particolare attenzione al prezzo minimo raggiunto alle 11.15 AM, in quanto solo nei minuti successivi si è osservato un netto calo del sentiment medio, questo a confermare l'ipotesi che il prezzo influenzi il sentiment anche degli utenti di *WSB*. Inoltre non si è osservato una particolare relazione tra il volume di scambio e il sentiment, mentre il volume risulta essere molto legato al prezzo soprattutto ai massimi e ai minimi. Il 4 Gennaio a *"Good Morning America"* un noto talk-show televisivo americano, che va in onda dalle 7 alle 9 del mattino, la segretaria del tesoro degli Stati Uniti *Jannet Yellen* parla di un imminente incontro con i regolatori per discutere del caso *GameStop*. Si nota come l'annuncio della *Yellen* spaventi gli investitori e porti ad una forte vendita del titolo, causando un ribasso del prezzo che perde il 39%, mentre per il sentiment non si può dire molto se non che è stato molto volatile tutto il giorno in particolare durante la messa in onda del talk-show. In relazione al prezzo minimo locale delle 11.15 AM si è registrato il massimo picco di volume, ma non si è osservata una relazione tra sentiment degli utenti e volume di scambio. Possiamo quindi concludere che vi siano molti player esterni a "wallstreetbets" che compiono operazioni di scmabio su GameStop

3.2 Risultati terza domanda di ricerca

Per quanto riguarda le parole più usate sull' intero arco temporale studiato si osservano le 15 più utilizzate e la loro frequenza espressa in termini percentuali:

Parola	%
amc	17,45
buy	15,98
hold	9,29
bb	5,92
moon	5,77
robinhood	5,70
shares	5,63
stock	5,34
nok	4,71
buying	4,46
like	4,01
still	4,00
sell	3,97
short	3,92
get	3,66

Tabella 3: 15 parole più utilizzate 11/01/2021 - 04/02/2021

N.B: La percentuale è stata calcolata rispetto alle 15 parole più utilizzate e non rispetto a tutte le parole.

Si nota la preponderanza di parole prettamente rialziste come *buy*, *hold*, *buying* e *still*, mentre parole che caratterizzano un sentiment ribassista come sell e short vengono usate in modo più marginale all'interno del gruppo *Reddit*.

Si vuole ora andare ad indagare cosa dicevano gli utenti reddit in alcuni giorni in cui è stato studiato il fenomeno:

- 11/01 -13/01 :

Parola	%
gang	12,88
today	10,19
moon	9,48
shares	8,22
cohen	6,79
short	6,79
squeeze	6,61
buy	6,26
sell	5,90
gains	5,72
right	4,47
ryan	4,29
first	4,11
selling	4,11
still	4,11

Tabella 4: 15 parole più utilizzate 11/01 - 13/01

N.B: La percentuale è stata calcolata rispetto alle 15 parole più utilizzate e non rispetto a tutte le parole.

Notiamo che nei giorni precedenti lo scoppio del prezzo delle azioni di *GameStop* la parola *buy* non era molto usata all'interno del subreddit, mentre erano molto usate parole caratterizzanti un sentiment ribassista come *sell, selling, short.* Inoltre vediamo che viene molte volte nominato *Ryan Cohen*, nuovo membro del consiglio di amministrazione di *GamesStop* nominato proprio nel giorno 11/01, questo sottolinea come il rinnovo della dirigenza fosse uno dei temi più affrontati su *"wallstreetbets"* e uno dei fattore scatenanti della crescita del prezzo.

- 28/01:

Parola	%
amc	17,31
buy	16,98
robinhood	9,46
hold	8,11
bb	7,94
nok	6,67
trading	4,62
still	4,37
buying	4,16
stock	4,04
moon	3,96
sell	3,38
shares	3,16
us	2,94
go	2,84

Tabella 5: 15 parole più utilizzate 28/01/2021

N.B: La percentuale è stata calcolata rispetto alle 15 parole più utilizzate e non rispetto a tutte le parole.

Notiamo come *buy* e *hold* detengano anche in questo giorno un ruolo preponderante,

in particolare spiccano anche le parole *Robinhood* e *trading* in quanto proprio il 28 Gennaio *Robinhood* (nota piattaforma di trading) aveva bloccato le transazioni delle azioni *GameStop* sulla sua piattaforma. In questo giorno le parole che hanno un'accezione ribassista sembrano essere ancora meno utilizzate, ad esempio si vede che la parola *short* non compare tra le 15 più utilizzate dagli utenti di *"wallstreetbets"*

- 04/02:

Parola	%
amc	18,31
buy	12,29
hold	10,84
shares	6,87
moon	6,06
stock	5,89
still	5,87
short	5,83
today	4,65
like	4,51
buying	4,15
go	3,86
get	3,82
sell	3,82
new	3,18

Tabella 6: 15 parole più utilizzate 04/02/2021

N.B: La percentuale è stata calcolata rispetto alle 15 parole più utilizzate e non rispetto a tutte le parole.

Il 4 febbraio, sebbene le parole rialziste siano sempre molto utilizzate notiamo un incremento dell'utilizzo della parola*short*, questo particolare giorno è stato caratterizzato da un grande *sell off* delle azioni di *GameStop* causato dalle dichiarazioni delle istituzioni.

3.3 Prospettive future

Vedendo le parole più utilizzate dagli utenti di *WSB* emergono spesso *amc, nok, bb*, questi rappresentano i ticker di alcune società molto chiacchierate all'interno del subreddit, ovvero *AMC Entertainment Holdings* la più grande catena di cinema degli US, *Nokia* e *BlackBerry*, due società che producono principalmente cellulari e che da anni non sono più sulla cresta dell'onda. Sarebbe quindi interessante ampliare il progetto anche a queste "meme stock" e magari considerare un arco di tempo più recente, anche alla luce del notevole impatto mediatico che sta avendo in questi giorni AMC.