REPORT
The MemeStock Mania

Gwydyon Marchelli, Olga Cozzolino, Alessandro Ciocchetti,
Davide Bottinelli, Lorena Volpini

Luglio 2022

1 Introduzione

L'obiettivo del progetto consiste nello studio e nella successiva applicazione di tecniche di analisi avanzate al fenomeno MemeStock, esploso alla fine del Gennaio 2021 con il caso GameStop. In questo report descriviamo le tecniche applicate. Per i relativi risultati ottenuti ed il contesto del progetto si faccia riferimento all'articolo correlato.

2 Recupero dei Dati

In questa sezione verrano spiegate le tecniche di information retrieval e web surfing applicate per l'ottenimento dei dati. Cento ticker (il ticker è il codice utilizzato per identificare in modo univoco le azioni di un determinata azienda quotata in borsa) sono stati selezionati partendo dalla subreddit Wall Street Bets, ed i dati loro riguardanti sono stati estratti da social networks, aggregatori di notizie e database di dati finaziari per il 2021.

2.1 Reddit

Il primo step è stato analizzare le menzioni nell'anno 2021 sul social network Reddit, che funziona come agglomeratore di diversi siti (forum) nello stesso cappello, accedibili con lo stesso user. Ognuna di queste subreddit tratta di un argomento differente. Si è rivolta l'attenzione in particolare a 10 di questi:

- r/WallStreetBets (la più grossa e famosa subreddit riguardante il mercato azionario),
- r/Cryptocurrency,
- r/stocks,
- r/investing,
- r/pennystocks,
- r/spacs,
- r/stockmarket,
- r/options

- r/GME,
- r/Superstonk.

Non è risultato necessario alcuno scraping o crawling, poichè dal sito yolostocks.live è possibile scaricare le menzioni storiche giornaliere relative al 2021 per tutte e dieci le subreddit. Dopo un'analisi relativa al numero di iscritti e alla presenza dei ticker nelle singole subreddit, siamo giunti alla conclusione che le informazioni su r/WallStreetBets fossero sufficenti per individuare i ticker più importanti da cercare su Twitter e sui giornali.

2.2 Twitter

Il crawling della piattaforma social network Twitter è stato eseguito tramite la libreria tweepy utilizzando due diversi profili developer di tipo accademico. Sono stati raccolti tutti i tweet di una lista di hashtag selezionata per ogni giorno del 2021. Come orario delle giornate si è scelto un periodo di tempo compreso dalle 9.30 ora di New York (UTC+4) fino alle 9.30 del giorno successivo. In totale sono stati scaricati c.a. 15 milioni di tweet per una lista di 75 hashtag selezionati partendo dalle menzioni reddit di ticker finanziari dello stesso anno da diversi subreddit e la disponibilità dei dati finanziari per i suddetti ticker da database pubblici (partendo da una lista di 100, ed escludendo i ticker per cui non si sono trovati i dati finanziari oppure ticker il cui hashtag è palesemente usato con altro significato sul social Twitter).
Il formato con cui i tweet sono stati scaricati è il formato json, con all'interno le informazioni importanti riportate nei campi "data" ed "includes". Nel primo, vi è un elenco dei tweet individuati con la nostra ricerca, con campi quali id del tweet, id del user, se il tweet è un retweet, un reply o un quote, id del tweet parent, geolocalizzazione (se disponibile). In includes si trovano le stesse informazioni per i tweet parent, e le informazioni sugli user (id, descrizione, numero di followers e following, geolocalizzazione personale). Infine, si trovano anche i dati sulle menzioni, hashtag e cashtag[1] della lista di tweet scaricata. Queste informazioni sono state trasformate successivamente in sette file csv diversi per hashtag per giorno, e poi riadattati in un unico database per hashtag per giorno.

2.3 Gdelt

Per ottenere i dati sulle menzioni delle aziende legate ai ticker da analizzare, si è interrogato Gdelt (Global Knowledge Graph). Per ognuno di essi è stato scaricato un csv con dati quali la data, l'id dell'articolo, la fonte, il numero di mezioni, il tono medio, positivo e negativo, la polarità, l'activness, i temi, le persone e le organizzazioni. Da questa mole di dati sono stati generati, per ogni ticker, un file csv riportante: data, menzioni, tono medio giornaliero, polarità media giornaliera e activness media giornaliera. È stato inoltre generato un altro csv con solo la data ed il numero di menzioni. Per quanto riguarda il significato dei campi significativi:

- Tono: è il tono medio di ogni articolo (tono positivo meno il tono negativo) con un range di valori compreso tra -100 e 100. Valori comuni sono tra -10 e +10, con valore 0 indicante tono neutro. Da tenere conto che un documento con un tono vicino a 0 può avere una bassa emotività oppure può avere tono positivo e tono negativo quasi equivalenti tra loro. Per questo motivo è utile considerante anche il positive e negative score oppure guardare la variabile "polarità".
- Positive Score, percentuale delle parole con una connotazione positiva.

[1] hashtag preceduto dal simbolo del dollaro utilizzato esclusivamente per titoli azionari

- Negative Score, percentuale delle parole con una connotazione negativa.
- Polarità, la percentuale delle parole che compaiono nel dizionario tonale di gdelt. È un indicatore di quanto emotivamente polarizzato sia il testo. Se la polarità è alta ma il tono è neutrale significa che il testo ha valori del tono positivo e negativo quasi equivalenti.
- Activity Reference Density, percentuale di parole 'attive'. Quest'indice rende la "activeness" generale di un testo. Un testo 'attivo' si distingue da un testo 'clinicamente descrittivo'.

2.4 Dati finaziari

Sono stati raccolti due diversi tipi di dati finanziari: sia i dati del mercato azionario che i dati specifici degli short selling.

2.4.1 Yahoo finance

I dati finanziari sono stati ottenuti tramite API con la libreria python yfinance che si interfaccia con il database di Yahoo Finance. In particolare sono stati scaricati dati (poi inseriti in csv per ticker) riguardanti il prezzo di chiusura a fine giornata (AdJ close), il volume di transazioni (volume) e un segnale binario (0 o 1) nel caso anomalo.

2.4.2 Short selling

I dati riguardanti gli short selling (azioni vendute allo scoperto) sono stati ottenuti scaricando i database dal sito www.finra.org. Come per yahoo finance sono stati recuperati i dati relativi ai ticker selezionati considerando l'intervallo prescelto che va dal 1/1/2021 al 31/12/2021. Si è proceduto alla pulizia del dato eliminando le colonne poco rilevanti che indicavano il valore di Short Volume e di Short Exempt Volume, mantenendo il solo Total Volume. Inoltre i volumi giornalieri degli stessi titoli ottenuti su facility differenti sono stati raggruppati (lo stesso titolo, scambiato su mercati differenti nel medesimo giorno).[2]

3 Data processing

Una volta ottenuti i dati, tramite visualizzazione dei dati preliminari, sono stati selezionati 16 ticker per il training set e 2 per il test set, i cui dati da processare e preparare per il dataset finale.

3.1 Aggregazione dei dati

Una volta collezionati i dati richiesti, si è operato per aggregare le informazioni contenute nei vari documenti in dei database unici per ticker. In particolare, in questa fase si è scelto di aggregare per mese e per ticker, unendo le informazioni contenute dalle estrazioni giornaliere di Twitter, le menzioni di Reddit da yolostocks.live, i dati finanziari e il sentiment riguardo ai ticker dai giornali, con le modalità descritte precedentemente. Questa aggregazione è stata eseguita con il codice Python usando pandas dataframe poi storati su un account universitario di Google Drive. I dati Twitter hanno richiesto una considerevole pulizia; si è infatti deciso di eliminare i tweet senza inerenza al tema finanziario tramine esclusione per cashtag e/o per sentiment analisi con tema "finance" e "work".[3]

[2] Questo passaggio è stato eseguito successivamente alla selezione dei ticker per il training set ed il test set, per cui sono stati scaricati i dati degli short selling solo per la lista dei 18 ticker finali

[3] Questa esclusione per sentiment ha seguito la stessa procedura descritta nel successivo paragrafo 3.2

3.2 Sentiment analysis

Il testo dei tweets relativi ad ogni ticker è stato preprocessato sostituendo menzioni di users e hyperlink con placeholders. In seguito è stato applicato il modello roBERTa-base già allenato su 124 milioni di tweets, che sfrutta TweetEval benchmark per la sentiment analysis. Si sono ottenuti sentiment scores ('positive' e 'negative') per ogni tweet e per ogni quotation. Per quanto riguarda i retweet, invece, si è optato per attribuire lo stesso sentiment del tweet originale. Importante sottolineare come il modello sia stato scaricato e fatto girare in locale, evitando così problemi legati a second-usage di dati sensibili.

3.3 Grafi di Twitter

Per ogni giorno del 2021, è stata considerata una finestra temporale dei dieci giorni precedenti. Per questa finestra, si è creato un edge di grafo considerando la menzione su Twitter di un nodo (user) da parte di un altro nodo (altro user). La connessione è pesata su quante volte quella menzione compare nella finestra. Per ognuna di queste finestre, si sono contati il numero di nodi, di edges, il degree massimo (numero massimo di vicini), il degree medio e la presenza di nodi influencer (calcolando sull'anno i nodi con la closeness più alta). Questi dati sono stati poi trasferiti nel database finale ed utilizzati insieme agli altri dati per la successiva clusterizzazione e classificazione.

4 Clusterizzazione e Classificazione

A questo punto del progetto, si è reso necessario costruire un singolo dataset comprendente i 16 ticker (per il training set) e 2 ticker (per il test set) da poi successivamente clusterizzare e classificare.

4.1 Creazione del dataset per classificazione e clusterizzazione

Per mantenere la serializzazione dei dati, si è reso necessario trasformare le feature nei valori medi di finestre temporali con rolling (con la stessa logica con cui sono stati costruiti i grafi di Twitter descritti precedentemente), e quindi estrarre da ogni finestra per ogni giorni i valori di media, mediana, standard deviation di ogni feature (eccetto per i dati dei grafi, essendo questi già costruiti su tale finestre). Complessivamente si sono così generate 104 features per 352 giorni per 16 ticker per il training set(352 per due per il test set), contando anche una feature per il nome del ticker, e due feature target comprese tra 0 e 1 rappresentanti l'andamento del prezzo e la volatilità. Le feature numeriche sono state poi normalizzate con minmaxscaler della libreria sklearn. Importante notare come la normalizzazione non sia stata realizzata sull'intero dataset ma per ticker, con lo scopo di mantenere le caratteristiche costanti nel ticker stesso, evitando che un ticker specifico (come GME) potesse basare con valori alti gli altri. Inoltre la normalizzazione di massimo e average degree è stata processata insieme, per mantenere la coerenza tra le due feature. Infine si è analizzata la correlazione tra le 104 features, individuandone 29 non necessarie per la successiva analisi.

4.2 Clusterizzazione

Essendo il dataset un insieme di timeseries multivariate, un approccio classico alla clusterizzazione non è realizzabile. Per questo motivo, come primo step è stata applicata una PCA (Principal Component Analysis) su tre componenti. La successiva analisi dei loadings e degli scores dei componenti ha validato il nostro approccio. In particolare, l'analisi dei loadings ha mostrato come la componente principale PC1 descrivesse la presenza nei subreddit, mentre la PC2 descrive i dati finanziari e le

news, e la PC3 descrive twitter ed i corrispettivi grafi. Successivamente, si è trasformato il dataset da avere giorni per ticker come records, ad un dataset con solo i ticker, fornendo come feature i tre valori dei componenti principali per giorno (in totale, 1056 feature). Su questo dataset è stato possibile applicare sia un approccio unsupervised kmeans che clustering gerarchico (entrambe con distanza euclidea). Per entrambi i clustering, sono stati selezionati 3 clusters, per rappresentare i casi anomali memestock, i casi memestock più instabili e i casi memestock legati ad aziende con maggiore valore sul mercato finanziario. Si è poi applicata la stessa clusterizzazione al dataset di test, e la clusterizzazione ha confermato le nostre ipotesi in merito alla classificazione dei casi memestock.

4.3 Classificazione

Per la successiva classificazione, è stata selezionata la variabile target cumulative price return (con defizione booleana 1 per positivo e 0 per negativo). Sono stati tentati diversi modelli: regressore logistico, Ridge classifier, LDA, Linear SVC, SVC non linear e random forest. A tutti i modelli è stato applicato una parameter tuning con cross validation. Si è preferito puntare su una buona precision sacrificando la recall, per evidenziare con sicurezza le anomalie. Il modello migliore risultante è stato random forest. Considerando la feature importance, si è notato come i grafi e le news, quindi gli eventi che accadono in contemporanea o successivamente, non aiutato alla predizione.