UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di Scienze Economico-Aziendali e Diritto per l'Economia

Corso di Laurea Triennale in

Economia delle Banche, delle Assicurazioni e degli Intermediari Finanziari

Ranking del Russell 3000 attraverso metodi di clustering per l'individuazione di meme stocks

Relatore: Prof.ssa Paola Agnese Bongini

Tesi di Laurea di:
Luca Botta
Matricola N. 895472

Anno Accademico 2024/2025

Indice

Elenco delle figure

Abstract

Negli ultimi anni, il fenomeno delle meme stocks ha mostrato come l'attenzione collettiva degli investitori individuali e l'attività sui social media possano influenzare in modo significativo l'andamento dei mercati finanziari. Questa tesi analizza il comportamento di tali titoli all'interno dell'indice Russell 3000, con l'obiettivo di individuare pattern comuni che li rendano riconoscibili attraverso un approccio quantitativo.

A partire da fonti come FINRA, Financial Modeling Prep, Google Trends e Wikipedia, è stato costruito un dataset bimensile che integra variabili di natura finanziaria (prezzo, days to cover, posizioni corte, volume di scambi) e di attenzione pubblica (ricerche online, articoli finanziari e visualizzazioni su Wikipedia). Su queste serie temporali multidimensionali è stato applicato un modello di clustering K-means con metrica Dynamic Time Warping (DTW), che consente di confrontare le traiettorie dei titoli anche in presenza di sfasamenti temporali.

I risultati mostrano che, con tre cluster, le principali meme stocks storiche tendono a concentrarsi nello stesso gruppo, distinguendosi soprattutto per i livelli di interesse pubblico piuttosto che per le variabili di mercato. Questo evidenzia come l'attenzione degli investitori e la visibilità online rappresentino elementi chiave nel caratterizzare il comportamento di tali titoli.

1 Introduzione

1.1 Evoluzione dei mercati finanziari

Negli ultimi decenni i mercati finanziari hanno subito una profonda trasformazione, la principale causa di questi cambiamenti è da attribuire ai progressi tecnologici e alla digitalizzazione dei servizi di investimento. Un tempo l'accesso ai mercati finanziari era riservato quasi esclusivamente agli operatori professionali o investitori con competenze avanzate, oggi invece piattaforme digitali e applicazioni con interfacce intuitive hanno reso il trading accessibile a un pubblico molto più ampio.

Con l'avvento del trading online nei primi anni duemila si sono ridotte le barriere economiche e informative. È diventato normale che un investitore abbia accesso in tempo reale a dati di mercato, grafici, strumenti di analisi e possibilità d'esecuzione precedentemente riservate ai professionisti. Questo ha favorito l'ingresso dei piccoli risparmiatori nei mercati.

Un punto di svolta importante si è avuto nel 2019, quando alcuni dei principali broker americani decisero di togliere le commissioni sulle operazioni in azioni e ETF. Questa scelta ha rimosso uno dei principali costi legati al trading, rendendo l'attività molto più accessibile anche a chi investe piccole somme. Come mostrano Even-Tov et al.[1], la misura ha avuto un effetto immediato: sempre più investitori retail sono entrati nei mercati, approfittando della soglia di accesso più bassa.

Questi cambiamenti hanno favorito la nascita di una vera e propria "nuova generazione" di investitori individuali. Nel 2010 gli investitori retail rappresentavano circa il 10% del volume di scambi sul mercato azionario statunitense, mentre nel 2020 la quota era quasi raddoppiata, arrivando intorno al 20%. Solo nei primi mesi di quell'anno si contano più di un milione di nuovi conti di trading aperti [2].

Con l'arrivo della pandemia di COVID-19 questo processo è stato ulteriormente amplificato, durante i lockdown infatti le persone si sono ritrovate con più tempo libero e meno possibilità di spendere, molte di loro hanno quindi iniziato a interessarsi ai mercati finanziari come passatempo o come ricerca di guadagno alternativo (Sigalas, 2023 [3]). Allo stesso tempo i mercati finanziari hanno attraversato una fase di forte volatilità, dovuta alla pandemia stessa, creando condizioni che hanno attratto un numero crescente di piccoli investitori. Come dimostrato da Ortmann et al.[4], a differenza di quanto ci si sarebbe potuti aspettare in un periodo di incertezza, il volume delle operazioni da parte degli investitori retail è aumentato.

A conferma di questo, altri studi su crisi passate (ad esempio Hoffmann et al., 2012 [5]) hanno mostrato come durante periodi di stress finanziario i comportamenti dei piccoli investitori cambino in modo marcato. Ciò suggerisce che quanto osservato durante la pandemia non sia stato un'eccezione, ma la combinazione di condizioni favorevoli che hanno amplificato un fenomeno già presente.

Infine questa dinamica è stata sostenuta anche dalle politiche fiscali: negli Stati Uniti infatti sono stati distribuiti circa 5000 miliardi di dollari come stimolo fiscale, di cui oltre 800

miliardi sotto forma di pagamenti diretti alle famiglie; come dimostrato da Anderson et al. [6] una parte di questa liquidità è confluita nei mercati finanziari, contribuendo ad aumentare la partecipazione individuale.

1.2 Il fenomeno delle meme stocks

Negli ultimi anni la combinazione di tutti questi fattori, insieme alla crescente diffusione dei social media, ha portato alla nascita del fenomeno delle cosiddette meme stocks, con questa espressione si indicano quelle aziende quotate il cui prezzo viene spinto verso l'alto da gruppi di investitori individuali che, coordinandosi attraverso piattaforme social, acquistano in modo massiccio i titoli. Come mostrano diversi studi (Pandey et al., 2024 [7]; Matsumoto et al., 2025 [8]), l'attività sui forum finanziari e sui social network è strettamente collegata alla diffusione di questo tipo di titoli. Il caso più noto e discusso è sicuramente quello di GameStop (GME), un'azienda statunitense specializzata nella rivendita di videogiochi: a gennaio 2021, un gruppo di investitori privati coordinati tramite il subreddit r/WallStreetBets ha acquistato in massa il titolo azionario e, nel giro di poche settimane, il prezzo dell'azione è passato da circa 20 dollari a oltre 350 dollari (Semenova et al., 2023 [9]).

Dal punto di vista comportamentale, il caso GME è stato interpretato come un esempio di herding e di echo chamber effect: l'aumento delle discussioni online ha generato entusiasmo collettivo e un effetto imitativo che ha spinto sempre più investitori a partecipare (Nedungadi, 2024 [10]).

Le conseguenze non si limitarono agli investitori retail: diversi hedge fund con posizioni corte, come Melvin Capital, registrarono perdite miliardarie, che in alcuni casi li portarono a ristrutturazioni o interventi di emergenza per coprire le proprie esposizioni. L'eccezionale volatilità del titolo portò anche a interventi diretti da parte delle piattaforme di trading. In particolare, Robinhood (uno dei broker più utilizzati dagli investitori retail statunitensi) decise di sospendere temporaneamente la possibilità di acquistare nuove azioni GameStop e altri titoli che in quel periodo mostravano un'elevata instabilità dei prezzi. Robinhood spiegò la sospensione degli acquisti come una misura imposta dalle camere di compensazione, necessaria per rispettare i requisiti di margine richiesti in quei giorni di forte volatilità. La decisione venne accolta con molte critiche, poiché diversi osservatori ritennero che limitasse ingiustamente l'operatività dei piccoli investitori e finisse, di fatto, per favorire gli operatori istituzionali (Newman et al., 2023 [11]).

L'episodio portò rapidamente il tema all'attenzione delle autorità statunitensi. Nel febbraio del 2021 il Congresso organizzò una serie di audizioni pubbliche per approfondire il funzionamento delle piattaforme di trading senza commissioni e discutere la trasparenza del payment for order flow, cioè il meccanismo con cui i broker ricevono un compenso per indirizzare gli ordini verso determinati market maker [12]. Dalle audizioni è emersa l'esigenza

di rivedere parte della regolamentazione dei mercati finanziari, soprattutto per considerare il peso crescente degli investitori retail e il ruolo ormai centrale delle piattaforme digitali.

1.3 Motivazioni e obiettivi della tesi

Questo lavoro ha l'obiettivo di sviluppare un metodo che permetta di valutare il grado di "memeness" di un gruppo di titoli azionari, cioè quanto si avvicinano ai comportamenti osservati nelle meme stocks. Si cerca quindi di capire se esistono caratteristiche comuni che rendono questi titoli riconoscibili. Studiare questi aspetti può aiutare a comprendere meglio alcune dinamiche di prezzo, che non dipendono dai fattori fondamentali ma da elementi legati all'attenzione e al comportamento degli investitori individuali. Per fare ciò vengono analizzati diversi indicatori, sia di natura finanziaria sia legati all'interesse pubblico. Tra i primi si considerano variabili come il prezzo, la capitalizzazione di mercato, lo short interest e i days to cover; tra i secondi invece lo score di Google Trends, il numero di articoli finanziari pubblicati e il numero di visualizzazioni delle pagine Wikipedia. Su questi dati vengono poi applicate tecniche di clustering, con l'obiettivo di raggruppare azioni che presentano comportamenti simili. L'approccio non supervisionato consente di esplorare i dati senza assumere ipotesi iniziali troppo rigide e di individuare schemi o relazioni che non emergerebbero con metodi più tradizionali. Dopo la fase di clustering, viene applicato un semplice procedimento di ranking per stimare la probabilità che ciascun gruppo identifichi un insieme di meme stocks. In pratica, più elevata è la concentrazione di meme stocks all'interno di un cluster, maggiore è la probabilità che esso rappresenti un "cluster meme". Ai fini di questa valutazione, vengono considerate meme stocks le società incluse nell'indice MS50 (Aloosh et al., 2023 [13]).

L'idea alla base di questo lavoro è quindi duplice: in primo luogo si vuole verificare se le principali meme stocks storiche si concentrano in un unico cluster, suggerendo che tali titoli condividono pattern simili nei loro indicatori di mercato e di interesse pubblico. Questo permetterebbe di dimostrare che il fenomeno presenta caratteristiche riconoscibili e misurabili anziché essere un insieme di episodi isolati. In secondo luogo, viene costruito un ranking dei cluster per valutare quale gruppo presenti la maggiore concentrazione di meme stocks e, di conseguenza, la più alta probabilità "meme". Questo ordinamento consente di identificare il cluster che rappresenta in modo più chiaro le dinamiche speculative e di attenzione collettiva tipiche del fenomeno, fornendo così una base interpretativa per confrontare gli altri gruppi del campione.

Come indice di riferimento è stato scelto il Russell 3000. Questo indice include le 3000 più grandi società statunitensi quotate in borsa e copre circa il 98% della capitalizzazione complessiva del mercato azionario americano. La sua struttura comprende sia grandi società a elevata capitalizzazione che small-cap e consente dunque di analizzare un campione ampio e diversificato di titoli. Questa ampia varietà di titoli rende l'indice particolarmente adatto

all'obiettivo del lavoro, poiché permette di considerare un insieme eterogeneo di società e di comportamenti di mercato.

1.4 Struttura della tesi

Questa tesi è organizzata nel seguente modo.

Nel secondo capitolo viene presentato un quadro teorico di riferimento, con un riepilogo dei principali concetti necessari alla comprensione del lavoro svolto, tra cui le definizioni di meme stocks, short interest, days to cover, short squeeze e gamma squeeze.

Il terzo capitolo è dedicato alla descrizione del processo di raccolta, integrazione, pulizia dei dati e alla struttura del dataset finale, con particolare attenzione alle fonti utilizzate e alle trasformazioni effettuate per garantire l'omogeneità del dataset.

Il quarto capitolo introduce i principi generali del clustering e i principali algoritmi di riferimento, focalizzandosi su quello effettivamente utilizzato in questa tesi. Viene inoltre descritta la procedura seguita per la preparazione e la standardizzazione dei dati, necessaria per garantire la coerenza del dataset prima dell'applicazione del modello di clustering.

Il quinto capitolo presenta i risultati ottenuti dall'applicazione del clustering al campione selezionato e introduce un procedimento di ranking dei cluster, volto a valutare quale gruppo presenti la maggiore concentrazione di meme stocks e quindi la più alta probabilità "meme".

Infine, il sesto e ultimo capitolo contiene la conclusione del lavoro, in cui vengono sintetizzati i principali risultati, le possibili implicazioni dell'analisi e alcune proposte per sviluppi futuri o approfondimenti.

2 Background teorico e definizioni

Questo capitolo serve a introdurre i concetti di base su cui si fonda tutto il lavoro. L'idea è quella di chiarire da dove nasce il fenomeno delle meme stocks e quali sono gli indicatori di mercato che possono descriverne il comportamento. In sostanza, si vuole creare un punto di riferimento teorico che renda più facile capire, nei capitoli seguenti, perché sono state fatte certe scelte di metodo e come vanno letti i risultati dell'analisi empirica.

2.1 Definizione e caratteristiche delle meme stocks

Come accennato nell'introduzione, con il termine meme stocks si indicano quei titoli azionari che, a partire dal 2021, hanno mostrato forti movimenti di prezzo collegati non tanto ai fondamentali economici, ma a fenomeni di viralità e coordinamento sui social network tra piccoli investitori. In questa parte si cerca di capire meglio da dove nasce il concetto, quali sono le sue caratteristiche principali, in che modo queste azioni si differenziano da quelle tradizionali e alcuni esempi storici. Come accennato nel capitolo precedente, in questa tesi vengono considerate meme stocks tutti i titoli inclusi nell'indice MS50, definito da Aloosh et al.[13] come paniere egualmente pesato dei titoli per i quali l'app Robinhood impose limitazioni agli ordini di acquisto il 28 gennaio 2021.

2.1.1 Origine del termine e nascita del fenomeno

Il termine meme stock nasce dall'unione dei concetti di "meme", ovvero contenuto virale diffuso attraverso internet, e "stock", titolo azionario. La sua diffusione è legata principalmente agli eventi di gennaio 2021 su GameStop (GME), quando una comunità di investitori individuali coordinata sul forum r/WallStreetBets di Reddit generò un'ondata di acquisti che portò il titolo a crescere di oltre il 1500% in pochi giorni (Semenova et al., 2023 [9]).

In letteratura, diversi studi (Desiderio et al., 2025 [14]; Gianstefani et al., 2022 [15]) descrivono questo evento come il punto di svolta nella partecipazione degli investitori retail, reso possibile dall'accesso a piattaforme di trading commission-free come Robinhood e dall'effetto amplificatore dei social media. Tali canali hanno favorito la diffusione di informazioni, opinioni e strategie d'investimento in tempo reale, creando un ecosistema in cui l'attenzione collettiva può influenzare direttamente i prezzi di mercato.

Alcuni autori (Aggarwal et al., 2022 [16]) sottolineano come il termine abbia assunto anche una valenza sociologica, rappresentando la democratizzazione dell'investimento e una forma di "ribellione finanziaria" contro gli operatori istituzionali. In questo senso, le meme stocks non sono solo un fenomeno di mercato, ma anche un esempio di interazione tra cultura digitale e finanza, dove la viralità e il comportamento collettivo giocano un ruolo centrale nel determinare i movimenti di prezzo.

Dopo il caso GameStop, dinamiche simili si verificarono anche per altri titoli come AMC Entertainment e Bed Bath & Beyond, che registrarono incrementi di prezzo altrettanto improvvisi. Questi episodi contribuirono a consolidare l'uso del termine meme stock nella stampa finanziaria e nella letteratura accademica, indicando una nuova categoria di titoli influenzati dall'attenzione dei social media piuttosto che dai fondamentali economici.

2.1.2 Differenze rispetto alle azioni tradizionali

Le meme stocks sono strumenti azionari quotati come gli altri, ma si distinguono per ciò che ne muove il prezzo. Nelle azioni tradizionali, le variazioni riflettono principalmente i fondamentali dell'impresa (come utili, prospettive di crescita o solidità patrimoniale). Nel caso delle meme stocks, invece, il prezzo tende a reagire con maggiore rapidità a stimoli di natura informativa e sociale.

Evidenze recenti (Semenova et al., 2023 [9]; Aggarwal et al., 2024 [16]) sottolineano come la discussione sui social possa produrre movimenti rapidi, talvolta più incisivi di notizie economiche comparabili. In altre parole, il valore di questi titoli risulta particolarmente sensibile al sentiment collettivo e al volume delle conversazioni.

Un'altra differenza rilevante riguarda la composizione della base degli investitori. Le meme stocks presentano spesso una quota significativa di investitori individuali, la cui partecipazione è aumentata in modo considerevole negli ultimi anni (Kurov et al., 2023 [17]). Rispetto agli operatori istituzionali, questi investitori tendono talvolta ad adottare strategie più emotive e a reagire rapidamente alle informazioni provenienti dall'ambiente digitale, generando comportamenti imitativi e coordinati che amplificano la volatilità.

Infine, le meme stocks si distinguono per la presenza di un'elevata componente speculativa. In molti casi, l'acquisto non è motivato da aspettative di lungo periodo sui risultati aziendali, ma dalla volontà di partecipare a movimenti di mercato di breve termine, spesso alimentati da fenomeni virali e di appartenenza a una comunità online (Desiderio et al., 2024 [14]). Tutto ciò rende le meme stocks un caso di studio particolare, in cui finanza e dinamiche sociali si intrecciano, e il comportamento collettivo degli investitori diventa parte integrante del processo di formazione dei prezzi.

2.1.3 Esempi storici

Oltre al caso di GameStop (GME), già discusso in precedenza, il fenomeno delle meme stocks ha interessato diversi altri titoli che, in periodi differenti, hanno mostrato forti rialzi di prezzo accompagnati da un'elevata attenzione sui social media e da un forte coinvolgimento degli investitori retail.

Tra i principali esempi si possono citare:

- AMC Entertainment (AMC): società cinematografica statunitense che nel 2021 ha visto il prezzo delle azioni passare da circa 2,1$ a oltre 72$, con un incremento superiore al

+3.300% in pochi mesi. Tale crescita fu sostenuta da campagne coordinate sui social e dal massiccio afflusso di investitori retail, diventando uno dei simboli del fenomeno.

- Bed Bath & Beyond (BBBY): la catena retail americana registrò nell'agosto 2022 un passaggio da circa 5$ a quasi 30$ per azione nel giro di due settimane, pari a un rialzo di circa +500%. L'impennata fu alimentata dall'hype sui social media e dalla notizia di partecipazioni rilevanti da parte di investitori individuali.

- Koss Corporation (KOSS): azienda di cuffie audio che tra il 22 e il 28 gennaio 2021 vide il titolo balzare da circa 3,3$ a 127$ intraday, corrispondente a un +3.700% in meno di una settimana. Questo caso è spesso citato come uno degli episodi più estremi legati al contagio virale post–GameStop.

- Express Inc. (EXPR): catena di abbigliamento che tra il 21 e il 27 gennaio 2021 passò da 1,17$ a 13,97$, segnando un incremento di circa +1.100% in pochi giorni. Anche in questo caso, l'attenzione su Reddit e l'effetto imitativo tra investitori retail determinarono volumi di scambio anomali.

- Sundial Growers (SNDL): società canadese del settore cannabis che all'inizio del 2021 mise a segno un aumento di oltre +550% da inizio anno, con una singola giornata di crescita del+79% il 10 febbraio 2021. Il titolo divenne popolare tra gli utenti di r/WallStreetBets e altri forum, mostrando un forte legame tra sentiment online e prezzo.

- Vinco Ventures (BBIG): società del settore media che tra il 20 agosto e l'8 settembre 2021 vide le proprie azioni salire da 2,4$ a 12,5$, pari a un aumento di circa +420%. Anche in questo caso il rally fu attribuito principalmente all'attività coordinata di piccoli investitori sui social.

Questi episodi mostrano come il fenomeno delle meme stocks non sia stato limitato a un singolo titolo o momento storico, ma abbia coinvolto un insieme eterogeneo di società accomunate da una forte esposizione mediatica e da dinamiche speculative di breve periodo.

2.2 Dinamiche di squeeze nelle meme stocks

Un elemento centrale per comprendere il comportamento anomalo delle meme stocks riguarda la presenza di particolari dinamiche di mercato, note come fenomeni di squeeze. Questi meccanismi si verificano quando la struttura delle posizioni su un titolo (come un'elevata quota di vendite allo scoperto o un forte ricorso alle opzioni call) può generare movimenti di prezzo auto-rinforzanti. Tra i più rilevanti nel contesto delle meme stocks figurano lo short squeeze e il gamma squeeze, due processi che, se innescati, possono amplificare rapidamente la pressione sulla domanda e causare variazioni di prezzo estreme. In questa sezione vengono descritti i principi di funzionamento di questi meccanismi e le loro implicazioni sui prezzi.

2.2.1 Funzionamento e impatto degli squeeze sui prezzi

I fenomeni di squeeze si manifestano come meccanismi auto-rinforzanti che amplificano i movimenti di prezzo in condizioni di mercato già tese. Di seguito vengono descritti i due principali meccanismi rilevanti per le meme stocks, con le loro modalità operative e le conseguenze sui prezzi.

Short squeeze Lo short squeeze si verifica quando un titolo caratterizzato da un'elevata percentuale di vendite allo scoperto registra un improvviso rialzo del prezzo. I venditori allo scoperto (short sellers), che avevano preso in prestito azioni scommettendo su un ribasso, subiscono perdite e sono indotti a ricomprare le azioni per chiudere le proprie posizioni (short covering). Questo comportamento può alimentare ulteriormente la domanda, spingendo il prezzo ancora più in alto in una spirale auto-alimentata.

Nel caso di GameStop, lo short interest aveva raggiunto livelli eccezionalmente elevati, e diversi studi documentano che un certo grado di short covering abbia contribuito all'aumento del prezzo. Tuttavia, sia il report della SEC [18] sia analisi successive (Cathéline, 2022 [19]) indicano che l'effetto diretto dello short squeeze fu limitato rispetto al volume complessivo di scambi. Secondo tali studi, l'ondata di acquisti da parte di investitori retail e il sentiment positivo diffuso online rappresentarono fattori più determinanti per la dinamica rialzista, mentre lo short squeeze avrebbe agito come amplificatore secondario del movimento (Vasileiou et al., 2023 [20]).

Gamma squeeze Il gamma squeeze è un fenomeno connesso al mercato delle opzioni. Quando molti operatori acquistano opzioni call su un titolo, i venditori di tali opzioni (market maker) si coprono acquistando il sottostante per limitare il rischio (hedging). Poiché la sensibilità dell'opzione al prezzo (delta) aumenta con il rialzo del titolo, i market maker possono essere costretti a incrementare ulteriormente gli acquisti man mano che il prezzo sale, generando pressione addizionale sulla domanda. Questo meccanismo può aver contribuito ad amplificare uno short squeeze già in corso, e diversi studi ipotizzano che un effetto di gamma squeeze si sia effettivamente verificato nel caso GameStop. Tuttavia, il report della SEC [18] e analisi successive concordano nel ritenere che, come per lo short squeeze, tale fenomeno non rappresenti la causa scatenante del rialzo, ma piuttosto un fattore secondario che ne ha accentuato la portata.

Implicazioni sui prezzi Entrambi i meccanismi contribuiscono a spiegare la forte volatilità osservata nelle meme stocks. In particolare, livelli elevati di short interest e di days to cover rappresentano indicatori di potenziale vulnerabilità a uno squeeze, anche quando non si verifica un vero e proprio short covering su larga scala. Pertanto, pur non essendo stati l'unico fattore alla base dei rialzi di prezzo osservati, tali condizioni di mercato restano elementi chiave per identificare titoli suscettibili a comportamenti speculativi estremi.

2.3 Metriche e variabili utilizzate nell'analisi

Dopo aver descritto le principali caratteristiche del fenomeno, questa sezione presenta le variabili che verranno utilizzate per analizzare in modo quantitativo il comportamento delle meme stocks. L'obiettivo è mostrare quali aspetti dei titoli possono aiutare a capire meglio le loro fasi di crescita improvvisa e la forte instabilità dei prezzi. Le variabili considerate riguardano sia indicatori di mercato tradizionali, come prezzo, volatilità e capitalizzazione, sia misure più specifiche legate all'interesse degli investitori e alle vendite allo scoperto.

2.3.1 Prezzo e volatilità dei titoli

Il prezzo rappresenta la base di ogni analisi finanziaria, perché riflette il valore attribuito dagli investitori a un'azione in un dato momento. Nel caso delle meme stocks, però, i forti aumenti o cali non sono sempre collegati ai risultati reali delle aziende, ma spesso nascono da comportamenti speculativi e da aspettative di breve periodo.

La volatilità, invece, misura quanto i prezzi si muovono nel tempo. Le meme stocks tendono ad avere una volatilità molto più alta rispetto ai titoli tradizionali: possono registrare variazioni giornaliere anche superiori al 30–40%, soprattutto nei periodi di maggiore attenzione del pubblico (Vasileiou et al., 2023 [20]). Questa instabilità riflette la natura più impulsiva e meno prevedibile degli scambi, in cui le decisioni degli investitori sono influenzate da notizie, tendenze o commenti diffusi in rete.

In sintesi, prezzo e volatilità servono non solo per descrivere l'andamento dei titoli, ma anche per misurare in modo indiretto l'intensità dell'attività speculativa che caratterizza il fenomeno delle meme stocks.

2.3.2 Capitalizzazione di mercato

La capitalizzazione di mercato rappresenta il valore complessivo di un'azienda quotata e si ottiene moltiplicando il prezzo di un'azione per il numero totale di azioni in circolazione. È un indicatore utile per confrontare le dimensioni delle diverse società e per capire quanto un titolo pesa all'interno del mercato.

Nel contesto delle meme stocks, le aziende coinvolte presentano di solito una capitalizzazione di mercato piuttosto ridotta. Questo le rende più esposte a variazioni improvvise dei prezzi, perché un numero limitato di acquisti o vendite da parte di investitori retail può avere un impatto maggiore sul valore complessivo dell'impresa. In altre parole, una capitalizzazione più bassa rende il titolo più "manipolabile" e favorisce la possibilità di movimenti speculativi concentrati in brevi periodi.

La capitalizzazione di mercato risulta quindi utile per comprendere quanto una società possa essere sensibile ai movimenti di prezzo legati al comportamento degli investitori, soprattutto nei casi in cui l'interesse collettivo cresce in modo improvviso.

Per questo motivo la capitalizzazione viene impiegata non come variabile di clustering ma come filtro iniziale per escludere le società di dimensione troppo elevata e mantenere quelle più facilmente "manipolabili" da investitori retail.

2.3.3 Short interest

Lo short interest (chiamato anche short float) indica la quantità di azioni di una società che sono state vendute allo scoperto e non sono ancora state riacquistate per chiudere la posizione. In altre parole, rappresenta il numero di titoli su cui gli investitori stanno scommettendo contro, aspettandosi un calo del prezzo. In termini operativi, lo short interest viene calcolato come:

$$ShortInterest = \frac{Total\ Shorted\ Shares}{Float} \cdot 100 \tag{1}$$

dove *Total Shorted Shares* indica il numero di azioni vendute allo scoperto mentre $Float$ indica il numero di azioni liberamente negoziabili. Il valore ottenuto è una percentuale che indica quante azioni sono vendute allo scoperto rispetto al totale delle azioni effettivamente negoziabili. Un livello elevato di short interest può essere interpretato come un segnale di sfiducia verso la società, ma allo stesso tempo può creare le condizioni per un possibile short squeeze.

Nel caso delle meme stocks, valori di short interest eccezionalmente alti sono stati uno degli elementi che hanno attirato l'attenzione degli investitori retail. In alcuni di questi casi estremi lo short interest ha anche superato il 100% del flottante. Ciò è possibile perché le stesse azioni possono essere prestate, vendute e poi nuovamente prestate in una catena di prestiti: quando un'azione venduta allo scoperto viene acquistata da un nuovo investitore, quella stessa azione può essere di nuovo data in prestito dal suo intermediario e rivenduta allo scoperto da un altro operatore. Ripetendo il processo, il numero totale di posizioni short aperte può diventare maggiore del numero di azioni liberamente negoziabili.

2.3.4 Days to cover

Un altro indicatore relativo alle posizioni short è il days to cover (chiamato anche short ratio). Viene calcolato come:

$$DTC = \frac{Total\ Shorted\ Shares}{Average\ Daily\ Trading\ Volume} \tag{2}$$

dove *Total Shorted Shares* indica il numero di azioni vendute allo scoperto (stessa base usata per lo short interest), mentre *Average Daily Trading Volume* indica il volume medio giornaliero, calcolato come media aritmetica dei volumi giornalieri. Il DTC indica quindi in quanti giorni, al ritmo di scambio medio recente, si potrebbero ricomprare tutte le azioni attualmente vendute allo scoperto. Valori elevati di DTC indicano che servirebbero più

giorni per chiudere le posizioni short, ciò rende il titolo più vulnerabile a eventuali short squeeze. Valori bassi suggeriscono invece una minore vulnerabilità.

2.3.5 Google Trends e misure di interesse pubblico

Google Trends fornisce una misura dell'interesse di ricerca per una parola o un argomento nel tempo. Lo score, detto anche Google Search Volume Index (GSVI), è un indice su scala 0-100: il valore 100 corrisponde al punto di massima popolarità nel periodo e nell'area scelti, mentre valori più bassi indicano livelli di ricerca relativamente inferiori.

Nel contesto delle meme stocks, lo score di Google Trends ha senso come proxy dell'attenzione del pubblico. Picchi nell'indice tendono a riflettere fasi di forte interesse e discussione, che spesso si accompagnano a volumi più elevati e a maggiore instabilità dei prezzi. In letteratura, questo score è ampiamente utilizzato come indicatore dell'attenzione degli investitori; inoltre, Ayala et al.[21] dimostrano che valori più elevati del GSVI sono associati a maggiore volatilità e a volumi di scambio più alti.

Inoltre, come ulteriori misure dell'interesse del pubblico si considereranno il numero di articoli finanziari che menzionano le varie società nel periodo considerato (news_volume) e le Wikipedia pageviews (wiki_views) delle singole società nello stesso periodo. In pratica ci si aspetta che il news_volume segua con un certo ritardo i grandi rialzi di prezzo, mentre le wiki_views tendano a muoversi più in sincronia con il GSVI. Il confronto tra news_volume e GSVI/wiki_views può anche aiutare a interpretare la natura dell'attenzione: quando l'aumento del GSVI/wiki_views avviene in concomitanza con un incremento del news volume, è plausibile che l'interesse sia legato a notizie fondamentali; viceversa, un picco del GSVI/wiki_views non accompagnato da maggiore copertura mediatica suggerisce un'attenzione prevalentemente generata online.

3 Costruzione del dataset e fonti dei dati

L'analisi presentata in questa tesi si basa su un insieme di dati costruito integrando più fonti, sia ufficiali sia di tipo alternativo. In questo capitolo vengono descritte le principali fonti da cui sono stati raccolti i dati, le frequenze temporali utilizzate e le operazioni preliminari di pulizia. Comprendere la struttura del dataset è fondamentale per interpretare correttamente le analisi dei capitoli successivi, poiché la qualità e la coerenza dei dati influenzano direttamente la validità dei risultati del clustering.

3.1 Fonti dei dati

Per lo sviluppo del dataset sono state utilizzate diverse piattaforme, scelte in base alla disponibilità di dati storici e all'affidabilità delle misurazioni. Le principali fonti impiegate sono le seguenti:

3.1.1 Financial Modeling Prep (FMP)

FMP è una piattaforma che fornisce dati di mercato e fondamentali aziendali tramite API. Da questa fonte sono stati scaricati:

- **Prezzi giornalieri** (close, volume, VWAP) per ciascun titolo appartenente al Russell 3000;
- **Capitalizzazione di mercato giornaliera**, utile per distinguere tra società a grande e piccola dimensione e per filtrare eventuali outlier.

Le richieste sono state automatizzate tramite uno script Python (`FMP.py` [22]), che interroga l'API per ogni ticker e salva i risultati in formato CSV. L'intervallo temporale copre gli ultimi cinque anni, in modo da includere il periodo di maggiore diffusione del fenomeno delle meme stocks (2020-2021).

3.1.2 Financial Industry Regulatory Authority (FINRA)

La FINRA pubblica con cadenza bimensile i dati ufficiali sul numero di azioni vendute allo scoperto per ogni società quotata sui principali mercati statunitensi. Questi dati sono stati scaricati in modo automatico tramite uno script dedicato (`get_short_positions.py` [22]), che estrae tutti i file CSV pubblicati sul portale FINRA e filtra solo i titoli appartenenti al Russell 3000.

Per ogni data di riferimento vengono salvati:

- la quantità di azioni corte in circolazione (shortQuantity);
- il volume medio giornaliero di scambio (averageDailyVolumeQuantity);

- il rapporto days to cover, calcolato come rapporto tra i due valori precedenti.

Queste variabili permettono di individuare situazioni di pressione ribassista o di potenziale vulnerabilità a uno short squeeze.

3.1.3 Google Trends

Poiché l'accesso diretto ai dati storici del GSVI tramite le librerie pubbliche risulta limitato, è stata impiegata l'API di DataForSEO, che consente di ottenere in modo affidabile lo score settimanale di Google Trends per periodi pluriennali. Il processo di download è stato automatizzato tramite uno script Python dedicato (`get_trends.py` [22]), che effettua richieste all'endpoint di DataForSEO per gruppi di parole chiave corrispondenti ai ticker delle aziende prese in considerazione. Per ogni ticker sono stati scaricati i dati relativi alle ricerche effettuate negli Stati Uniti, con frequenza settimanale. Lo score di Google Trends viene interpretato come una misura dell'attenzione online, utile per catturare i picchi di popolarità che spesso precedono o accompagnano i movimenti anomali dei prezzi nelle meme stocks.

3.1.4 FMP News API

Per integrare la componente informativa è stata inoltre utilizzata la sezione News delle API di FMP, che consente di ottenere articoli e comunicati relativi a ciascuna società quotata. Il download è stato automatizzato tramite lo script Python `FMP.py` [22], già impiegato anche per il recupero dei prezzi e della capitalizzazione di mercato. Da questa fonte è stato ricavato il volume di notizie giornaliero. Questa variabile permette di valutare quanto l'attenzione dei media tradizionali contribuisca alla visibilità di un titolo, aiutando a capire se l'aumento dell'interesse sia legato a notizie di natura economico-finanziaria oppure a fenomeni più "social" e temporanei.

3.1.5 Wikipedia Pageviews

Infine, i dati relativi alle visualizzazioni delle pagine di Wikipedia sono stati scaricati tramite le API pubbliche di Wikimedia, che consentono di ottenere il numero giornaliero di visite per ciascuna pagina nel periodo desiderato. Per ogni titolo è stato utilizzato il nome della società come chiave di ricerca per identificare la relativa pagina di Wikipedia. Il download è stato automatizzato attraverso uno script Python (`get_wiki_views.py` [22]), che interroga l'endpoint `/metrics/pageviews/per-article/` delle API Wikimedia e restituisce il numero di visite giornaliere per ciascuna pagina.

3.2 Frequenza dei dati

Le fonti descritte in precedenza presentano frequenze temporali differenti: i dati di prezzo, il volume di notizie e le visite di Wikipedia sono giornalieri, quelli di short interest pubblicati

da FINRA sono bimensili, mentre quelli di Google Trends hanno frequenza settimanale. Per poter integrare tutte queste informazioni in un unico dataset coerente, è stato necessario uniformare le date di riferimento e ridurre la frequenza complessiva dell'analisi. Per garantire coerenza temporale, tutte le serie sono state aggregate su base bimensile, ancorate alle date FINRA. Ogni osservazione del dataset finale coincide quindi con una data FINRA e riassume le informazioni disponibili nel periodo che intercorre dalla data FINRA precedente a quella corrente.

Operativamente:

- **Date di riferimento (FINRA):** le date pubblicate da FINRA sono utilizzate come punti di riferimento per costruire le finestre bimensili.
- **Prezzi:** per ciascuna finestra si estrae il valore di fine periodo come ultimo dato disponibile prima o uguale alla data FINRA: il close è quindi un valore "a fine finestra".
- **Short interest, days to cover e volume (FINRA):** shorts, d2c e volume sono i valori ufficiali FINRA (non si applica nessuna trasformazione, salvo l'allineamento dei formati e delle date per ciascun titolo).
- **Google Trends:** gli score settimanali all'interno della finestra vengono mediati per ottenere uno trend_score bimensile. In presenza di un'unica settimana nel periodo, il valore coincide con quello settimanale.
- **Volume di notizie:** Si calcola la somma degli articoli pubblicati dalla data FINRA precedente fino alla data FINRA attuale, ottenendo news_volume come conteggio per periodo.
- **Wikipedia pageviews:** in modo analogo al volume di notizie, per ogni finestra vengono sommate le visualizzazioni giornaliere registrate, producendo la variabile wiki_views. Questa aggregazione consente di misurare il livello complessivo di attenzione spontanea verso ciascuna società nel periodo di riferimento.

L'integrazione tra tabelle avviene per chiavi (ticker, date) dove date è la data FINRA di riferimento. Per l'allineamento di serie a frequenza più alta si utilizza un merge_asof con direzione backward (per associare a ciascuna data FINRA l'ultimo dato disponibile antecedente).

3.3 Struttura del dataset finale

Come anticipato nel capitolo precedente, prima della costruzione del dataset finale è stata effettuata una fase di selezione preliminare dei titoli, effettuata nel notebook Python `filter_tickers.ipynb`[22]. L'obiettivo di questa fase è stato filtrare le società del Russell 3000 in base a criteri di dimensione e liquidità, così da escludere sia le imprese troppo piccole,

spesso soggette a dati incompleti o irregolari, sia quelle eccessivamente grandi, per le quali è meno probabile che i movimenti di prezzo siano influenzati da dinamiche di attenzione online o comportamenti coordinati degli investitori retail.

Per ciascun titolo sono state considerate la capitalizzazione di mercato media e il volume medio giornaliero di scambio nel periodo analizzato, e sulla base di tali valori sono stati mantenuti solo i titoli con una capitalizzazione compresa tra 100 milioni e 10 miliardi di dollari e con un volume medio di almeno 500.000 azioni al giorno.

Dopo l'applicazione di questi filtri e la successiva rimozione dei ticker con dati mancanti in almeno una delle variabili principali, il campione si è ridotto a circa 500 società. Sebbene questa selezione comporti una perdita di eterogeneità rispetto all'indice originale, permette di ridurre il rumore dovuto a società marginali e di effettuare il clustering in modo più stabile e con tempi di elaborazione contenuti. Anche il sottoinsieme delle meme stocks storiche (MS50) ha subito una riduzione significativa: a causa dell'incompletezza di alcune serie storiche, solo 25 titoli su 50 sono stati effettivamente mantenuti nel dataset finale.

Il dataset finale è organizzato in formato panel, ossia una struttura che combina la dimensione temporale e quella per singolo titolo, e ciascuna coppia (ticker, date) contiene le seguenti variabili:

- **close**: prezzo di chiusura a fine finestra.
- **d2c**: days to cover (valore FINRA alla data).
- **shorts**: posizioni corte (valore FINRA alla data).
- **volume**: volume medio giornaliero (valore FINRA alla data).
- **trend_score**: media bimensile dello score Google Trends.
- **news_volume**: numero totale di articoli nella finestra.
- **wiki_views**: numero totale di visualizzazioni della pagina Wikipedia nella finestra.

Le osservazioni con valori mancanti sono state rimosse per garantire omogeneità del campione. Il risultato è un panel bimensile allineato alle date FINRA, costruito tramite un notebook Python dedicato (`merge_data.ipynb` [22]) pronto per l'applicazione dei metodi di clustering descritti nel capitolo successivo.

4 Metodi e tecniche di clustering applicati

In questo capitolo viene presentato l'approccio utilizzato per analizzare il comportamento dei titoli del campione attraverso l'uso di tecniche di clustering. L'obiettivo è capire se, utilizzando diversi indicatori di mercato e di interesse pubblico, emergano gruppi di azioni che condividono caratteristiche simili. Il clustering è stato scelto perché permette di esplorare i dati senza dover stabilire in anticipo categorie o ipotesi rigide. Invece di imporre una struttura ai dati, si lascia che siano le relazioni tra le variabili a rivelare eventuali schemi o somiglianze.

Nelle sezioni successive vengono spiegati i principi di base del clustering, i principali algoritmi utilizzabili e le scelte operative adottate nella fase di analisi, come la preparazione dei dati e la determinazione del numero ottimale di cluster.

4.1 Introduzione al clustering

Il clustering è una tecnica di machine learning non supervisionato utilizzata per suddividere un insieme di dati in gruppi omogenei, detti cluster, sulla base delle relazioni che esistono tra le osservazioni. L'idea di fondo è che elementi simili tra loro, in termini di valori delle variabili considerate, debbano appartenere allo stesso gruppo, mentre quelli più diversi debbano finire in gruppi distinti. A differenza dei metodi supervisionati, non si parte da categorie già note o da una variabile da prevedere: è l'algoritmo stesso a costruire i gruppi sulla base delle somiglianze nei dati.

Ogni osservazione può essere rappresentata come un vettore di p variabili:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}), \quad i = 1, 2, \ldots, n$$

dove n è il numero di osservazioni nel campione. Il compito del clustering è assegnare ciascun punto x_i a uno dei K gruppi $\{C_1, C_2, \ldots, C_K\}$ in modo che gli elementi dello stesso gruppo risultino il più possibile vicini tra loro, mentre quelli di gruppi diversi siano lontani. La vicinanza può essere misurata in diversi modi a seconda del metodo di clustering, una delle metriche più comuni è la distanza euclidea:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

che quantifica quanto due punti differiscono nello spazio delle variabili.

Esistono diverse famiglie di algoritmi di clustering, che si distinguono per il modo in cui formano i gruppi e per la logica con cui viene calcolata la somiglianza tra le osservazioni. Le principali sono:

- **Metodi partizionali**: dividono direttamente il dataset in un numero prefissato di gruppi. Ogni punto viene assegnato al cluster a cui è più vicino, in base a una misura di

distanza (ad esempio quella euclidea). Il centro di ciascun gruppo, chiamato centroide, rappresenta la posizione media delle osservazioni che ne fanno parte. Fanno parte di questa categoria algoritmi come K-means e K-medoids.

- **Metodi gerarchici**: creano una struttura ad albero che mostra come le osservazioni si uniscono o si separano a diversi livelli di somiglianza. Questa rappresentazione, detta dendrogramma, consente di visualizzare in modo grafico come i punti si raggruppano man mano che la soglia di distanza aumenta o diminuisce. In pratica, il dendrogramma mostra la "storia" della formazione dei cluster: all'inizio ogni osservazione è isolata, poi i punti più simili vengono uniti fino a formare gruppi sempre più grandi (approccio agglomerativo) oppure, al contrario, un gruppo iniziale viene progressivamente diviso in sottogruppi più piccoli (approccio divisivo).

- **Metodi basati sulla densità**: individuano i cluster come zone dello spazio dei dati dove i punti sono più concentrati. Sono particolarmente utili quando i gruppi non hanno forma regolare o quando il dataset contiene rumore o osservazioni anomale (outlier). Un esempio molto usato è l'algoritmo DBSCAN.

- **Metodi probabilistici o basati su modelli**: si basano sull'ipotesi che i dati provengano da più distribuzioni statistiche, spesso di tipo gaussiano. Ogni distribuzione rappresenta un possibile gruppo, e per ogni osservazione viene stimata la probabilità di appartenenza a ciascun gruppo. Il modello assegna poi ogni punto al gruppo in cui la sua probabilità di appartenenza è maggiore. Un esempio di questo approccio è il Gaussian Mixture Model (GMM), che combina più distribuzioni gaussiane per descrivere i diversi gruppi presenti nei dati.

- **Metodi basati su grafi**: rappresentano i dati come una rete di nodi collegati tra loro da archi, i cui pesi indicano quanto due osservazioni sono simili. I cluster vengono individuati come sottoinsiemi di nodi fortemente collegati tra loro, un approccio utile quando le relazioni tra i dati non sono lineari o dipendono da connessioni complesse.

Ognuna di queste famiglie segue una logica diversa nel definire cosa significa "essere simili" e nel determinare i confini dei gruppi. Nel paragrafo successivo vengono descritti in modo più dettagliato i principali algoritmi appartenenti a queste categorie, con particolare attenzione al K-means, che è quello impiegato per l'analisi empirica.

4.2 Principali algoritmi di clustering

Nella sezione precedente sono state presentate le principali famiglie di tecniche di clustering, distinguendo tra metodi partizionali, gerarchici, basati sulla densità, probabilistici e su grafi. In questa parte vengono descritti più nel dettaglio i tre algoritmi più rappresentativi

e maggiormente utilizzati in ambito empirico: il K-means, il clustering gerarchico e il DBSCAN. Ognuno di essi affronta il problema del raggruppamento dei dati secondo una logica diversa, ma con un obiettivo comune: individuare gruppi di osservazioni che presentano caratteristiche simili.

4.2.1 K-means

Il K-means è probabilmente l'algoritmo di clustering più utilizzato per la sua semplicità e rapidità di esecuzione. Il suo obiettivo è dividere i dati in K gruppi predefiniti, in modo che ogni osservazione appartenga al cluster con il centro più vicino, detto centroide.

Il procedimento è iterativo. Si parte scegliendo casualmente K centroidi iniziali, poi si alternano due passaggi:

1. assegnare ogni punto al centroide più vicino, calcolando la distanza (solitamente euclidea);
2. aggiornare la posizione dei centroidi, calcolando la media dei punti appartenenti a ciascun gruppo.

Questi due passaggi vengono ripetuti fino a quando i centroidi non cambiano più posizione in modo significativo. L'algoritmo cerca quindi di minimizzare la somma delle distanze quadratiche tra ogni punto e il centro del gruppo a cui appartiene:

$$\text{SSE} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

dove μ_k è il centroide del cluster k e C_k l'insieme dei punti assegnati a quel gruppo.

Il K-means funziona bene quando i cluster hanno forma più o meno sferica e dimensioni simili, ma tende a dare risultati meno affidabili in presenza di gruppi di densità o forma molto diverse, oppure quando i dati contengono valori anomali. Nonostante ciò, la sua semplicità e la capacità di gestire grandi dataset ne fanno uno degli strumenti più diffusi nell'analisi empirica.

4.2.2 Clustering gerarchico

A differenza del K-means, che impone subito la divisione in un numero fisso di gruppi, il clustering gerarchico costruisce una struttura ad albero che mostra in che modo i dati si raggruppano a diversi livelli di somiglianza. Il risultato di questo processo è un grafico chiamato dendrogramma, che rappresenta visivamente le fusioni o le separazioni tra i punti.

Nel metodo agglomerativo ogni osservazione parte come un singolo gruppo. Poi, passo dopo passo, i due gruppi più simili vengono uniti fino a ottenere un'unica grande struttura. Nel metodo divisivo, invece, si parte da un gruppo unico e lo si divide progressivamente in sottogruppi più piccoli.

Uno dei vantaggi del clustering gerarchico è che non serve fissare a priori il numero di cluster: questo può essere scelto dopo, osservando il dendrogramma e decidendo a quale livello "tagliare" l'albero. Lo svantaggio principale è che, con dataset molto grandi, il metodo diventa computazionalmente più pesante rispetto a K-means.

4.2.3 DBSCAN

L'algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) segue un approccio completamente diverso. Invece di cercare gruppi di forma regolare, individua aree dove i punti sono più concentrati, trattando come "rumore" le osservazioni isolate.

Il funzionamento si basa su due parametri principali:

- Eps, cioè la distanza massima entro cui due punti vengono considerati vicini;
- MinPts, il numero minimo di punti necessari per formare un'area densa.

Partendo da un punto, l'algoritmo esplora il suo intorno: se trova abbastanza punti vicini (almeno MinPts), li considera parte dello stesso gruppo e continua l'espansione. Questo processo permette di individuare cluster di qualsiasi forma e di riconoscere in modo naturale gli outlier, cioè i punti che non appartengono ad alcuna area densa.

Il DBSCAN è molto utile quando i dati contengono rumore o presentano strutture non regolari, ma è sensibile alla scelta dei parametri: valori troppo alti o troppo bassi possono alterare il risultato finale.

4.2.4 Sintesi e confronto

In sintesi, i tre algoritmi si basano su logiche diverse e offrono risultati differenti:

- il K-means è rapido e intuitivo, ma richiede di conoscere in anticipo il numero di cluster;
- il metodo gerarchico permette di esplorare la struttura dei dati senza fissare subito il numero di gruppi, ma è meno efficiente con dataset molto ampi;
- il DBSCAN riconosce cluster di forma irregolare e gestisce bene gli outlier, ma può risultare instabile se i parametri non sono scelti correttamente.

Nel complesso, il K-means risulta il più adatto al tipo di analisi condotta in questa tesi, grazie alla sua semplicità e alla possibilità di applicarlo facilmente a dataset di grandi dimensioni. In particolare, in questa tesi è stata impiegata una variante del K-means che utilizza come metrica la Dynamic Time Warping (DTW), una distanza progettata per confrontare serie temporali. A differenza della distanza euclidea, che confronta punto per punto le osservazioni nella stessa posizione temporale, la DTW permette di allineare in modo "elastico" due sequenze, facendo corrispondere valori simili anche se si verificano in momenti diversi. Questo approccio è

utile nel confronto tra titoli azionari, poiché due azioni possono mostrare andamenti analoghi ma con un certo ritardo temporale tra l'una e l'altra. L'uso della DTW consente quindi di cogliere meglio le somiglianze di forma tra le diverse traiettorie nel tempo, evitando che sfasamenti cronologici penalizzino l'analisi di clustering. Il clustering è stato eseguito nello script Python `clustering.py`, disponibile nel repository dedicato al progetto [22].

4.3 Preparazione e standardizzazione dei dati

Prima di applicare il modello di clustering è stato necessario organizzare e normalizzare il dataset in modo da rendere confrontabili le serie temporali dei diversi titoli. Tutte le elaborazioni sono state effettuate tramite uno script Python (`clustering.py` [22]), che utilizza le librerie `pandas`, `numpy` e `tslearn`.

4.3.1 Struttura del dataset

Il file utilizzato come base per il clustering è `merged_data.csv`, costruito nel capitolo precedente e contenente per ogni titolo e per ogni data FINRA le principali variabili di interesse: prezzo di chiusura (close), posizioni corte (shorts), days to cover (d2c), volume di scambi (volume), score di Google Trends (trend_score) volume di articoli finanziari (news_volume) e numero di visite delle pagine Wikipedia (wiki_views).

I dati sono stati raggruppati per ticker, in modo che ciascun titolo sia rappresentato da una sequenza temporale multidimensionale. Per costruire un input adatto all'algoritmo TimeSeriesKMeans, il dataset è stato trasformato in un array tridimensionale della forma:

$$(\text{numero di titoli}, \text{numero di periodi}, \text{numero di variabili}),$$

dove ogni "riga" rappresenta un titolo, ogni "colonna" una data bimensile e ogni "profondità" una variabile. La lunghezza massima delle serie è stata allineata tra tutti i titoli, riempiendo con zeri le parti mancanti nei casi in cui alcune serie fossero più brevi.

4.3.2 Standardizzazione delle variabili

Poiché le variabili considerate hanno unità di misura e ordini di grandezza diversi, è stata applicata una procedura di standardizzazione prima del clustering. L'obiettivo è evitare che variabili con valori numerici più grandi (ad esempio il volume di scambi) dominino la misura di distanza e influenzino eccessivamente la formazione dei gruppi.

La standardizzazione è stata eseguita con lo strumento `TimeSeriesScalerMeanVariance` della libreria `tslearn`, che per ogni variabile e per ciascuna serie temporale impone:

$$x' = \frac{x - \bar{x}}{s_x},$$

dove $\bar{x}$ è la media e s_x la deviazione standard della serie. In questo modo ogni sequenza viene riportata a media zero e varianza unitaria, rendendo le diverse variabili direttamente confrontabili.

4.3.3 Output della fase di preparazione

Il risultato di questa fase è un insieme di serie temporali multidimensionali normalizzate, pronte per essere elaborate dal modello di clustering. Ogni titolo del campione è quindi descritto da una traiettoria nel tempo basata sulle variabili selezionate, confrontabile con quelle degli altri titoli in termini di forma e andamento. Questa struttura consente all'algoritmo di analizzare le somiglianze nei comportamenti dinamici dei titoli, più che nei loro livelli assoluti, offrendo così una rappresentazione più coerente delle relazioni di mercato.

4.4 Implementazione del modello e scelta dei parametri

Il clustering è stato realizzato tramite lo script `clustering.py` [22], che utilizza l'algoritmo `TimeSeriesKMeans` della libreria `tslearn`. Questa versione del K-means permette di analizzare serie temporali multidimensionali e utilizza come metrica di distanza la Dynamic Time Warping (DTW), già descritta nella sezione precedente. Questa metrica permette di confrontare i titoli anche quando mostrano andamenti simili ma sfalsati nel tempo, concentrandosi sulla forma delle variazioni piuttosto che sui loro livelli assoluti.

4.4.1 Procedura di esecuzione

Il procedimento seguito, dopo la standardizzazione dei dati già descritta nella sezione precedente, è il seguente:

1. **Clustering con DTW**: l'algoritmo `TimeSeriesKMeans` viene eseguito più volte con un diverso numero di cluster K, mantenendo costante la metrica DTW e un numero massimo di iterazioni pari a 10. Il parametro `random_state=42` garantisce la riproducibilità dei risultati.

2. **Salvataggio del modello**: per evitare di rieseguire l'intero processo a ogni prova, i modelli già addestrati vengono salvati in locale tramite il pacchetto `joblib` e ricaricati in caso di necessità.

3. **Analisi dei risultati**: per ciascun valore di K, vengono visualizzate le serie appartenenti a ogni cluster insieme ai rispettivi centroidi, in modo da confrontare visivamente la coerenza interna dei gruppi e la distinzione tra cluster diversi.

4.4.2 Scelta del numero di cluster

Non esiste un numero di cluster ottimale in senso assoluto, poiché il risultato dipende dalla natura dei dati e dagli obiettivi dell'analisi. In linea teorica, due cluster (uno "meme" e uno "non-meme") sarebbero la divisione più intuitiva; tuttavia, nei dati reali questa separazione risulta troppo grossolana. Per questo motivo il valore di K è stato scelto in modo empirico: sono state eseguite più iterazioni del modello con valori diversi di K (da 2 a 10) e i risultati sono stati confrontati visivamente e interpretati sulla base della composizione dei gruppi. Si è cercato di individuare la configurazione che producesse cluster coerenti, ben separati e interpretabili dal punto di vista finanziario.

Questa strategia risulta adeguata in un contesto esplorativo come quello di questa tesi, in cui l'obiettivo non è individuare una segmentazione univoca, ma osservare se esistono pattern ricorrenti tra i titoli del campione.

4.4.3 Output e interpretazione

Per ciascun cluster vengono generate rappresentazioni grafiche che mostrano le serie temporali dei titoli appartenenti al gruppo e la traiettoria media (centroide) calcolata dall'algoritmo. Le visualizzazioni consentono di osservare eventuali somiglianze tra i titoli in termini di andamento congiunto delle variabili considerate.

Il modello così addestrato costituisce la base per l'interpretazione dei risultati, presentata nel capitolo successivo, con particolare attenzione al cluster che raccoglie la maggior parte delle meme stocks e alle differenze rispetto agli altri gruppi.

5 Analisi dei risultati

Questo capitolo presenta la fase conclusiva dell'analisi, dedicata all'esame dei risultati ottenuti dal modello di clustering. L'obiettivo è descrivere in modo chiaro i principali esiti del modello, evidenziando come i gruppi individuati possano essere interpretati alla luce delle variabili considerate e delle dinamiche tipiche del fenomeno delle meme stocks.

La prima parte del capitolo confronta le diverse configurazioni di clustering testate, analizzando come varia la composizione dei gruppi al crescere del numero di cluster e individuando la soluzione che offre il miglior equilibrio tra separazione e interpretabilità. Successivamente vengono introdotte e descritte le metriche utilizzate per valutare la bontà del clustering. Infine viene approfondito il modello scelto come riferimento e vengono discusse, da un punto di vista finanziario, le caratteristiche dei cluster ottenuti.

5.1 Analisi con diversi numeri di cluster

Per valutare la capacità del modello di rappresentare in modo coerente le relazioni tra i titoli, sono stati testati diversi numeri di cluster, da $k = 2$ a $k = 10$. L'obiettivo è individuare una configurazione che offra un buon equilibrio tra separazione dei gruppi, dimensione dei cluster e interpretabilità dei risultati.

Per ciascun valore di k è stata analizzata la distribuzione dei titoli e, in particolare, la concentrazione delle meme stocks storiche all'interno dei diversi gruppi. Questo confronto consente di verificare se tali titoli tendano a riunirsi in modo stabile nello stesso cluster anche al variare della segmentazione, segnale della presenza di un comportamento comune riconoscibile dal modello.

Di seguito, per ciascun valore di k, viene riportata la composizione dei cluster in numero di titoli e il cluster con il maggior numero di meme stocks.

k	**Dimensione per cluster**	**Cluster con più meme stock**
2	214, 307	cluster 2: 18 (72%)
3	145, 176, 200	cluster 3: 22 (88%)
4	129, 173, 141, 78	cluster 3: 17 (68%)
5	122, 148, 92, 60, 99	cluster 3: 18 (72%)
6	125, 151, 65, 62, 44, 74	cluster 3: 14 (56%)
7	48, 119, 76, 56, 38, 77, 107	cluster 3: 15 (60%)
8	35, 66, 76, 89, 24, 78, 57, 96	cluster 7: 12 (48%)
9	35, 51, 76, 89, 21, 53, 47, 96, 53	cluster 7: 17 (68%)
10	29, 43, 73, 90, 16, 48, 35, 98, 48, 41	cluster 10: 14 (56%)

Dalla tabella emerge come, all'aumentare del numero di cluster la struttura tende a frammentarsi, con gruppi più piccoli e meno coesi. Tuttavia, tra $k = 3$ e $k = 5$ si osserva ancora un buon equilibrio tra separazione e interpretabilità, mentre per valori più elevati la segmentazione diventa progressivamente meno significativa.

5.2 Valutazione della bontà del clustering

Per valutare la coerenza e la qualità dei raggruppamenti individuati e per stabilire il numero di cluster ottimale, sono stati calcolati tre diversi indici di bontà del clustering: il Silhouette score, il Davies–Bouldin Index (DBI) e l'Adjusted Rand Index (ARI). Ciascuna di queste metriche fornisce una prospettiva complementare sulla struttura dei cluster, consentendo di misurare rispettivamente la separazione tra gruppi, la loro compattezza interna e la stabilità delle assegnazioni.

Silhouette score. Il Silhouette score misura quanto ciascun punto è ben collocato nel proprio cluster rispetto agli altri gruppi. Per ogni osservazione i si calcolano due quantità: la distanza media dai punti appartenenti allo stesso cluster, $a(i)$, e la distanza media minima dai punti del cluster più vicino, $b(i)$. L'indice è definito come:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

dove $s(i) \in [-1, 1]$.

Un valore vicino a 1 indica che l'osservazione è ben assegnata, mentre valori negativi suggeriscono che essa sarebbe più appropriata in un altro cluster. La media di $s(i)$ su tutte le osservazioni fornisce il Silhouette score complessivo.

La figura seguente mostra l'andamento del silhouette score medio al variare del numero di cluster. Il punteggio cresce fino a $k = 3$, dove raggiunge il valore massimo di 0.081, per poi diminuire progressivamente man mano che la struttura del modello si frammenta e la separazione tra i gruppi si indebolisce.

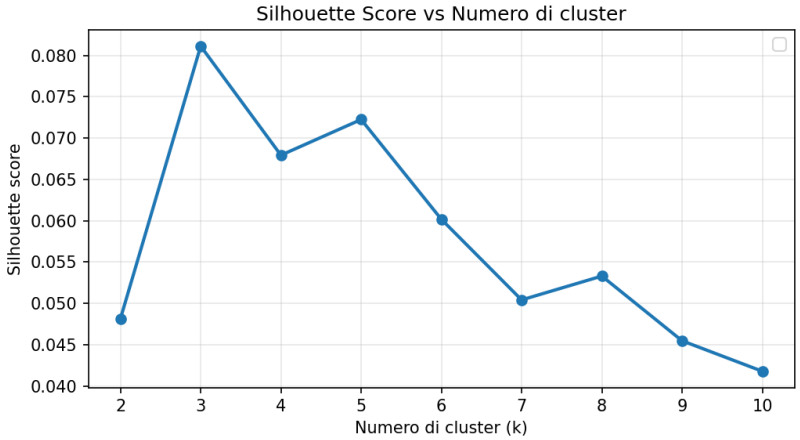

Figura 1: Andamento del Silhouette Score al variare del numero di cluster k

I valori medi del silhouette score possono sembrare piuttosto bassi, considerando che teoricamente la misura varia tra -1 e 1 e che, in altri contesti, una "buona" separazione si ha per valori > 0.5 [23]. Tuttavia, diversi lavori segnalano che in ambito finanziario, e in particolare nel clustering di serie temporali o di variabili multiple con elevata dimensionalità, ottenere valori elevati è spesso difficile. In ambienti con elevata variabilità interna, con rumore e correlazioni trasversali tra variabili, la coesione dei cluster risulta naturalmente più debole, e quindi la silhouette media si colloca su valori più bassi [24]. In altre parole, un valore intorno a 0.08 non implica automaticamente una scelta non valida, bensì riflette la complessità intrinseca del campione, la presenza di segnali speculativi sovrapposti e la difficoltà nel definire gruppi "netti" in contesti di mercato.

Davies–Bouldin Index (DBI). Il DBI valuta la qualità del clustering in base al rapporto tra la dispersione interna dei cluster e la distanza tra i loro centroidi. Per ogni coppia di cluster i e j si definisce:

$$R_{ij} = \frac{s_i + s_j}{d(c_i, c_j)}$$

dove s_i rappresenta la media delle distanze dei punti del cluster i dal proprio centroide c_i, e $d(c_i, c_j)$ è la distanza tra i centroidi i e j. L'indice complessivo è la media, sui k cluster, del valore massimo di R_{ij}:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} R_{ij}$$

La figura seguente mostra l'andamento del Davies–Bouldin Index al variare del numero di cluster. L'indice decresce rapidamente da 3.57 (per $k = 2$) a circa 3.10 (per $k = 4$), raggiungendo il minimo locale attorno a $k = 8$.

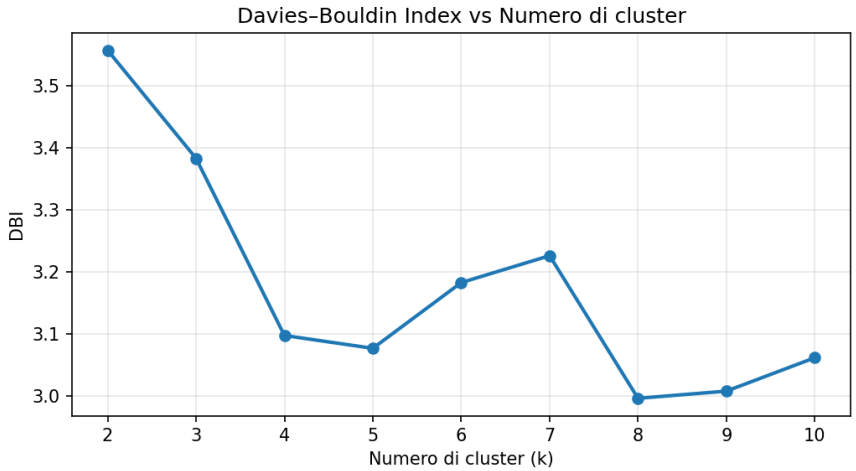

Figura 2: Andamento del Davies–Bouldin Index (DBI) al variare del numero di cluster k

Valori più bassi del DBI indicano cluster più compatti e meglio separati. Nel contesto analizzato, il miglioramento iniziale (tra $k = 2$ e $k = 4$) suggerisce che l'aumento del numero di cluster migliora la coesione interna, ma oltre tale soglia la riduzione diventa marginale. In ambito finanziario, valori del DBI compresi tra 3 e 4 sono considerati realistici a causa della rumorosità e dell'elevata correlazione delle variabili temporali; pertanto, i risultati ottenuti indicano un livello di separazione soddisfacente e coerente con la natura del dataset.

Adjusted Rand Index (ARI). L'ARI misura la somiglianza tra due diverse partizioni dello stesso insieme di dati, correggendo per la similarità dovuta al caso. Dato un insieme di n osservazioni, siano U e V due partizioni contenenti rispettivamente r e s cluster. Indicando con n_{ij} il numero di elementi comuni ai cluster U_i e V_j, l'indice è definito come:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

dove $a_i = \sum_j n_{ij}$ e $b_j = \sum_i n_{ij}$. L'ARI varia tra -1 e 1: valori prossimi a 1 indicano una forte corrispondenza tra due partizioni, mentre valori vicini a 0 riflettono una somiglianza dovuta al caso.

La figura seguente mostra l'andamento dell'Adjusted Rand Index al variare del numero di cluster. L'indice cresce costantemente fino a $k = 4$, dove raggiunge il valore massimo di circa 0.60, per poi diminuire progressivamente fino a 0.39 per $k = 10$.

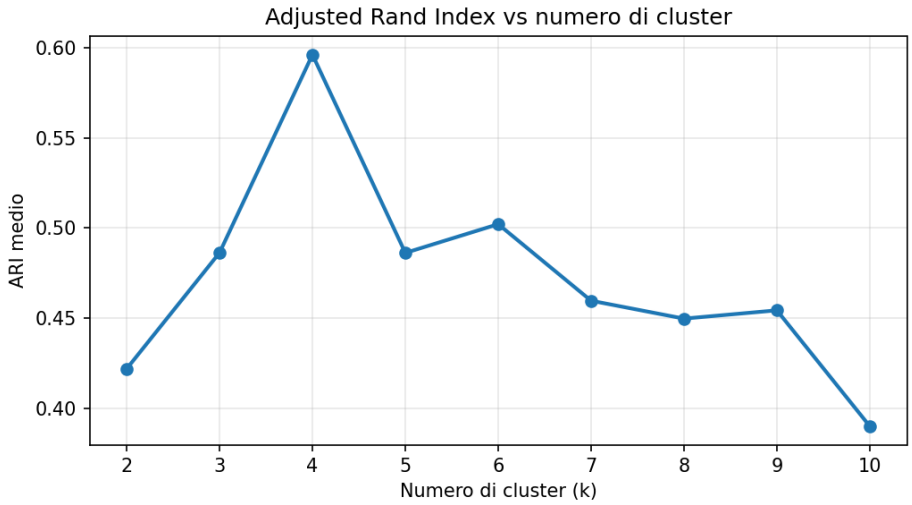

Figura 3: Andamento dell'Adjusted Rand Index (ARI) al variare del numero di cluster k

Il picco osservato per $k = 4$ indica che tale configurazione produce una struttura stabile e coerente con le partizioni precedenti, mentre la successiva riduzione segnala un aumento della frammentazione e una minore robustezza del modello. Nel complesso, i valori compresi

dell'ARI tra 0.45 e 0.60 suggeriscono una buona stabilità del clustering, considerando la complessità del campione e la natura non supervisionata del metodo.

5.3 Modello di riferimento

Il confronto tra i tre indici di valutazione, Silhouette score, Davies–Bouldin Index e Adjusted Rand Index, mostra che le configurazioni con $k = 3$ e $k = 4$ rappresentano il miglior compromesso tra coesione interna, separazione dei gruppi e stabilità delle assegnazioni. Il Silhouette score e il DBI segnalano un miglioramento fino a $k = 3$–4, mentre l'ARI evidenzia la massima stabilità per $k = 4$. Tuttavia, tra le soluzioni analizzate, il modello con $k = 3$ risulta preferibile poiché offre un equilibrio ottimale tra qualità statistica e interpretabilità economica, evitando la frammentazione che si osserva per valori più elevati di k.

In questa configurazione, il modello ottiene il valore massimo del Silhouette score (0.081) e mostra una chiara concentrazione delle meme stocks all'interno di un unico gruppo. Il cluster 3 raccoglie 200 titoli, di cui 22 appartengono al gruppo delle meme stocks, pari a circa l'88% del totale. Gli altri due cluster, composti rispettivamente da 145 e 176 titoli, includono solo 2 e 1 meme stock, evidenziando la capacità del modello di isolare in modo coerente un comportamento specifico.

Questi risultati indicano che il modello con $k = 3$ riesce a distinguere un cluster chiaramente associabile alle dinamiche speculative e di attenzione collettiva tipiche delle meme stocks, separandolo dai gruppi di titoli con andamenti più regolari. Il cluster 3, caratterizzato dalla maggiore concentrazione di meme stocks e dai valori più elevati delle variabili legate alla visibilità mediatica, verrà quindi indicato come cluster meme e sarà oggetto dell'analisi economico-finanziaria presentata nella sezione successiva.

Oltre a questo aspetto, è interessante notare come il comportamento del modello resti stabile anche al variare delle configurazioni testate: l'emergere di un gruppo dominante che raccoglie la quasi totalità delle meme stocks si osserva già a partire da $k = 3$ e si mantiene, con differenze minime, anche per valori di k superiori. Ciò suggerisce che la struttura individuata non dipende da una scelta arbitraria del numero di cluster, ma riflette una tendenza effettiva nei dati, dove un insieme di titoli mostra dinamiche ricorrenti riconducibili al fenomeno delle meme stocks. Questa evidenza rafforza la robustezza del modello e motiva l'approfondimento qualitativo dei risultati presentati nella sezione successiva.

5.4 Interpretazione economico-finanziaria dei cluster

L'analisi dei centroidi delle variabili per ciascun cluster, raffigurata nella figura seguente, mette in evidenza differenze significative nei profili medi dei titoli. Le principali divergenze non riguardano tanto i dati di mercato tradizionali, quanto le variabili legate all'attenzione del pubblico e alla diffusione informativa.

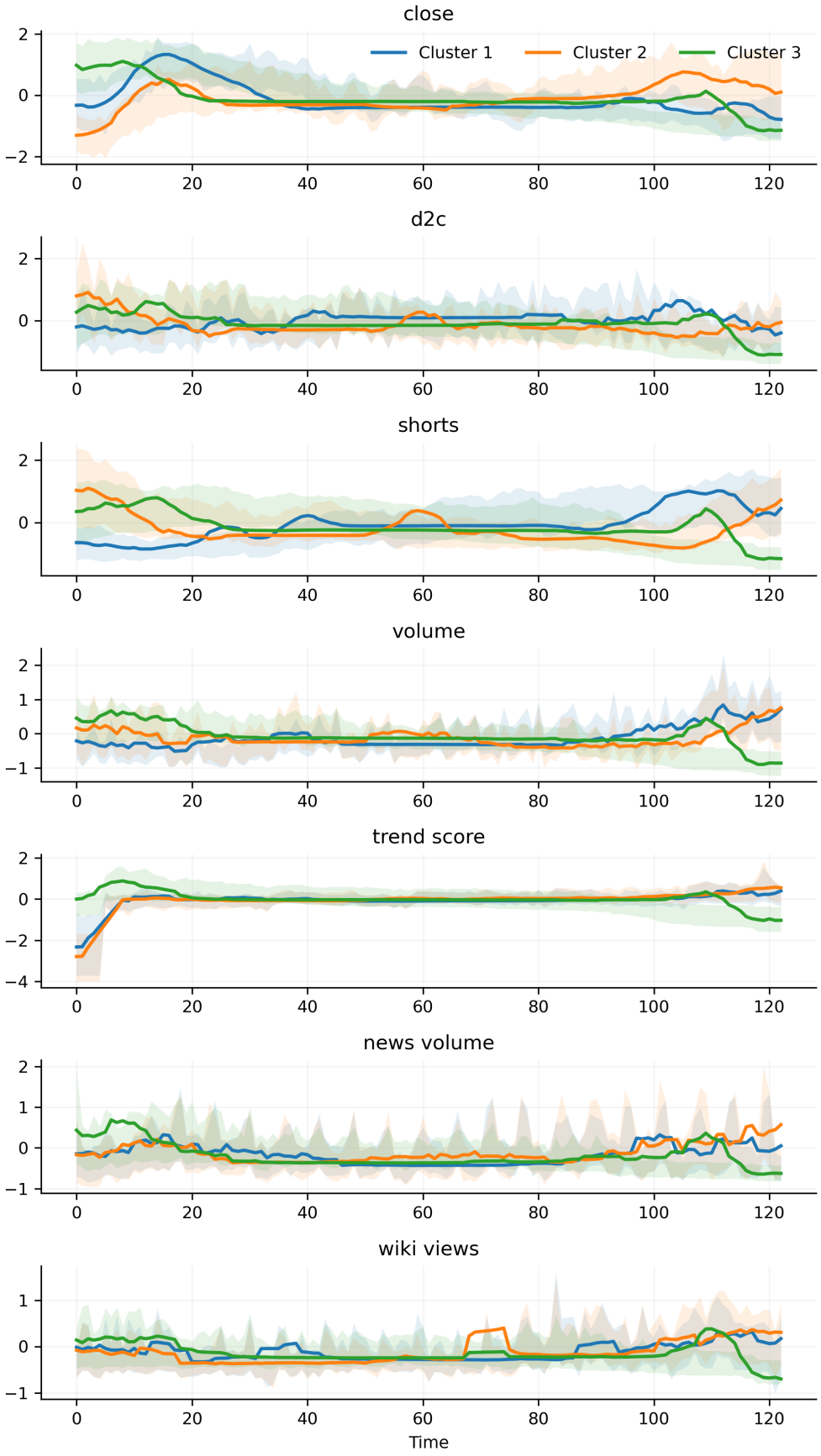

Figura 4: Centroidi delle principali variabili per ciascun cluster individuato dal modello DTW-K-means con 3 cluster

Prezzi e variabili di mercato. Le traiettorie dei prezzi (close) risultano nel complesso simili tra i tre cluster, senza oscillazioni particolarmente ampie nel gruppo delle meme stocks (cluster 3). Anche le variabili legate all'attività short (shorts e days-to-cover) non mostrano differenze marcate, segno che la presenza di posizioni ribassiste elevate non rappresenta di per sé un fattore discriminante all'interno del campione. La somiglianza di questi indicatori riflette la natura eterogenea del cluster, che, pur contenendo le principali meme stocks, include anche numerosi altri titoli con comportamenti di prezzo più moderati.

Indicatori di attenzione pubblica. Le variabili collegate al sentiment e alla visibilità mediatica, il trend_score, il news_volume e le wiki_views, mostrano invece valori medi più elevati nel cluster 3 rispetto agli altri gruppi. Questo suggerisce che i titoli di tale cluster sono accomunati da un livello di attenzione collettiva superiore, probabilmente alimentato da discussioni online, copertura giornalistica e comportamenti imitativi da parte degli investitori retail. Il modello sembra quindi aver individuato una dimensione di differenziazione più legata al comportamento degli investitori che ai fondamentali di mercato.

Considerazioni interpretative. Il fatto che il cluster associato alle meme stocks non mostri una volatilità dei prezzi particolarmente alta può essere attribuito alla composizione del gruppo stesso: su circa 200 titoli, solo una piccola parte corrisponde alle meme stocks storiche, mentre gli altri presentano caratteristiche affini ma meno estreme. La media del cluster tende quindi a smussare gli episodi di variazione eccezionale e a riflettere piuttosto la presenza di un interesse persistente e diffuso. Inoltre, la normalizzazione preliminare delle variabili riduce le differenze in ampiezza, mettendo in evidenza la forma temporale dei comportamenti più che la loro intensità assoluta.
Nel complesso, il cluster 3 si distingue per il peso delle variabili informative e comportamentali, più che per differenze strutturali nei prezzi o nelle posizioni short. I cluster 1 e 2 rappresentano invece titoli meno esposti al sentiment del mercato, caratterizzati da livelli più bassi di interesse pubblico e da un comportamento più regolare delle principali variabili finanziarie.

In sintesi, il modello con $k = 3$ consente di distinguere in modo interpretabile tre insiemi di titoli: un gruppo ad alta attenzione pubblica (cluster meme) e due gruppi con dinamiche più stabili e meno influenzate dal sentiment del mercato. Questa distinzione rappresenta il principale risultato dell'analisi e mostra come l'uso della distanza DTW consenta di raggruppare titoli che presentano andamenti temporali simili nelle variabili considerate.

6 Conclusioni

Questo capitolo riassume i principali risultati del lavoro e le considerazioni finali emerse dall'analisi. L'obiettivo della tesi era applicare tecniche di clustering per individuare eventuali schemi comuni tra i titoli del Russell 3000, verificando se le cosiddette meme stocks potessero formare un gruppo riconoscibile sulla base di variabili di mercato e di interesse pubblico. L'idea di partenza era capire se il fenomeno, emerso con casi noti come GameStop e AMC, potesse essere descritto quantitativamente attraverso i dati, e non solo spiegato a livello narrativo o comportamentale.

Il lavoro ha preso avvio dalla costruzione di un dataset originale, integrando dati provenienti da fonti diverse. Sono stati uniti dati di mercato (prezzi, volumi e capitalizzazione) scaricati da FMP, indicatori di posizioni corte e days to cover pubblicati da FINRA, e misure di attenzione pubblica come lo score di Google Trends, il numero di articoli finanziari e le visualizzazioni delle pagine Wikipedia. Tutte le variabili sono state allineate su base bimensile utilizzando le date FINRA come riferimento temporale, così da ottenere una struttura coerente e confrontabile nel tempo. Questo processo di integrazione ha richiesto un lavoro di pulizia e di uniformazione, necessario per unire dati con frequenze e formati differenti.

Prima di procedere al clustering, il dataset è stato filtrato per mantenere solo le società con una capitalizzazione e un volume medio di scambio sufficienti, in modo da escludere i titoli con dati incompleti o scarsamente rappresentativi. Dopo questa selezione, il campione finale è composto da circa 500 titoli, di cui 25 corrispondono alle meme stocks storiche appartenenti all'indice MS50. Le variabili sono state standardizzate per rendere i valori confrontabili e per concentrare l'attenzione sulle variazioni nel tempo piuttosto che sui livelli assoluti.

Per l'analisi è stato scelto il modello TimeSeriesKMeans con metrica Dynamic Time Warping (DTW). Questa distanza consente di confrontare serie temporali anche quando i movimenti non sono perfettamente allineati nel tempo, e si è rivelata adatta per studiare dati finanziari caratterizzati da andamenti irregolari e da eventuali ritardi di reazione. L'uso della DTW ha permesso di confrontare i titoli in base alla forma delle loro traiettorie piuttosto che ai valori puntuali delle singole variabili, cogliendo somiglianze più realistiche nelle dinamiche temporali.

Sono state testate diverse configurazioni del modello, con un numero di cluster compreso tra 2 e 10. Dall'analisi è emerso che il valore K=3 rappresenta il miglior compromesso tra coesione interna e interpretabilità dei risultati. In questa configurazione, 22 delle 25 meme stocks si concentrano nello stesso gruppo, suggerendo che il modello è riuscito a individuare una struttura comune a questi titoli. Anche il silhouette score, sebbene basso in termini assoluti (0.081), raggiunge il suo massimo relativo proprio per K=3, indicando che la separazione tra i gruppi è debole ma comunque significativa. In letteratura, valori di questo ordine di grandezza sono frequenti in contesti finanziari, dove le serie temporali sono molto rumorose e i comportamenti non perfettamente disgiunti.

L'analisi dei cluster mostra che le principali differenze tra i gruppi riguardano soprattutto le variabili legate all'attenzione del pubblico, come il volume di notizie e lo score di Google Trends, mentre le variabili di mercato più tradizionali risultano distribuite in modo simile tra i cluster. Questo risultato è coerente con l'idea che il fenomeno delle meme stocks nasca più da dinamiche sociali e di attenzione collettiva che da differenze strutturali nei fondamentali economici. I titoli appartenenti al cluster dove si concentrano le meme stocks presentano infatti un andamento più irregolare nelle variabili legate all'interesse online, che tende ad anticipare o accompagnare i movimenti di prezzo più marcati.

Dal punto di vista metodologico, il lavoro ha mostrato come l'integrazione di dati alternativi, come quelli provenienti da Google Trends o Wikipedia, possa fornire informazioni aggiuntive rispetto alle tradizionali metriche di mercato. Questi indicatori non spiegano direttamente la performance dei titoli, ma aiutano a interpretare il ruolo che l'attenzione degli investitori e dei media gioca nei momenti di forte volatilità. L'utilizzo di un approccio non supervisionato, inoltre, ha permesso di esplorare i dati senza assumere ipotesi rigide, lasciando che fossero le relazioni tra le variabili a determinare la formazione dei gruppi.

Naturalmente, l'analisi presenta anche alcuni limiti. In primo luogo, la disponibilità dei dati non è uniforme per tutte le società, soprattutto per le variabili legate all'attenzione pubblica. Il metodo K-means, pur efficace e relativamente semplice da applicare, presenta alcune limitazioni note: è sensibile alla scelta dei centroidi iniziali e tende a formare gruppi ben separati anche quando i confini tra le osservazioni non sono netti. Inoltre, il modello non tiene conto dell'evoluzione temporale dei cluster, ma considera l'intero periodo come un'unica sequenza. La metrica DTW, pur adatta a gestire sfasamenti temporali, è sensibile alla lunghezza delle serie e alla presenza di rumore. Infine, la riduzione del campione, pur necessaria per motivi di coerenza, ha comportato la perdita di parte dell'eterogeneità originaria del Russell 3000.

Nonostante questi limiti, i risultati ottenuti offrono indicazioni interessanti. Il fatto che la maggior parte delle meme stocks si concentri in un unico gruppo dimostra che il fenomeno presenta caratteristiche comuni riconoscibili nei dati, e che tali dinamiche non sono casuali. L'analisi suggerisce che le fasi di maggiore attenzione pubblica, misurate da indicatori come il volume di notizie e lo score di ricerca, si accompagnano a comportamenti di prezzo distintivi rispetto al resto del mercato. Si tratta quindi di un primo passo verso una rappresentazione quantitativa di un fenomeno che, finora, era stato descritto soprattutto in termini qualitativi.

In prospettiva futura, il lavoro potrebbe essere esteso in due direzioni principali. La prima riguarda lo sviluppo di un indice predittivo di "memeness", aggiornato periodicamente, capace di misurare quanto ciascun titolo mostri comportamenti simili alle meme stocks storiche. La seconda direzione consiste nell'integrare nuove fonti di dati sociali, come il numero di post o di commenti su piattaforme come Reddit e X, che potrebbero fornire misure più dirette del coinvolgimento degli investitori retail. Queste estensioni renderebbero possibile un'analisi più dinamica e completa, capace di catturare in tempo reale l'evoluzione

dell'attenzione collettiva sui mercati.

In conclusione, il lavoro ha mostrato come strumenti di analisi quantitativa possano essere utilizzati per studiare fenomeni nati dall'interazione tra mercati finanziari e social media. La concentrazione delle meme stocks in un unico cluster suggerisce che la diffusione dell'attenzione online produce effetti misurabili sulle dinamiche di mercato. Anche se l'analisi resta di tipo esplorativo, essa rappresenta un passo avanti verso una comprensione più ampia del ruolo dell'informazione e del comportamento collettivo nei mercati moderni. In prospettiva, un approccio basato su dati di mercato e indicatori sociali potrà contribuire a sviluppare strumenti capaci di anticipare o monitorare le fasi di crescente interesse speculativo che caratterizzano questo tipo di titoli.

Riferimenti bibliografici

[1] Omri Even-Tov et al. «Fee the People: Retail Investor Behavior and Trading Commission Fees». In: *Working paper, Haas School of Business, UC Berkeley* (2023). URL: https://newsroom.haas.berkeley.edu/research/absent-fees-retail-traders-do-better/.

[2] Simon Bowden. «The Rise of the Retail Investor». In: *IAM Advisory* (2021). URL: https://iamadvisory.com/thinking/news-insight/insights/the-rise-of-the-retail-investor.

[3] Christos Sigalas. «Impact of COVID-19 lockdowns on retail stock trading patterns». In: *Cogent Economics & Finance* 11.1 (2023), p. 2188713. URL: https://doi.org/10.1080/23322039.2023.2188713.

[4] Regina Ortmann, Matthias Pelster e Sascha Tobias Wengerek. «COVID-19 and investor behavior». In: *Finance Research Letters* 37 (2020), p. 101717. ISSN: 1544-6123. URL: https://www.sciencedirect.com/science/article/pii/S1544612320307959.

[5] Arvid O.I. Hoffmann, Thomas Post e Joost M.E. Pennings. «Individual investor perceptions and behavior during the financial crisis». In: *Journal of Banking & Finance* 37.1 (2013), pp. 60–74. ISSN: 0378-4266. URL: https://www.sciencedirect.com/science/article/pii/S0378426612002294.

[6] Pragyan Deb et al. «The Effects of Fiscal Measures During COVID-19». In: *IMF Working Papers* 2021.262 (2021). URL: https://www.elibrary.imf.org/view/journals/001/2021/262/article-A001-en.xml.

[7] Ichchha Pandey e Michael Guillemette. «Social media, investment knowledge, and meme stock trading». In: *Journal of Behavioral Finance* (2024), pp. 1–17. URL: https://doi.org/10.1080/15427560.2024.2361875.

[8] Miyuki Matsumoto et al. «Impact of information disparity between individual investors on profits of meme stocks using an artificial market simulation approach». In: *Journal of Computational Social Science* 8.1 (2025), p. 25. URL: https://doi.org/10.1007/s42001-024-00355-7.

[9] Valentina Semenova et al. «Wisdom of the Crowds or Ignorance of the Masses? A data-driven guide to WSB». In: *arXiv preprint* (2023). URL: https://arxiv.org/abs/2308.09485.

[10] Prateek Nedungadi. «Herding Mentality in the GameStop Short Squeeze: A Case Study». In: *Research Archive of Rising Scholars* (2024). URL: https://doi.org/10.58445/rars.1554.

[11] Neal F Newman. «GameStopped: How Robinhood's GameStop Trading Halt Reveals the Complexities of Retail Investor Protection». In: *Fordham J. Corp. & Fin. L.* 28 (2023), p. 395. URL: https://scholarship.law.tamu.edu/facscholar/1795.

[12] Committee on Financial Services U.S. House of Representatives. «How the Meme Stock Market Event Exposed Troubling Business Practices, Inadequate Risk Management, and the Need for Legislative and Regulatory Reform». In: *U.S. House of Representatives, Committee on Financial Services* (2022). URL: https://democrats-financialservices.house.gov/uploadedfiles/6.22_hfsc_gs.report_hmsmeetbp.irm.nlrf.pdf.

[13] Arash Aloosh, Hyung-Eun Choi e Samuel Ouzan. «The tail wagging the dog: How do meme stocks affect market efficiency?» In: *International Review of Economics & Finance* 87 (2023), pp. 68–78. URL: https://ssrn.com/abstract=3839832.

[14] Antonio Desiderio et al. «The dynamics of the Reddit collective action leading to the GameStop short squeeze». In: *npj Complexity* 2.1 (2025), p. 5. URL: https://arxiv.org/abs/2401.14999.

[15] Ilaria Gianstefani, Luigi Longo e Massimo Riccaboni. «The echo chamber effect resounds on financial markets: A social media alert system for meme stocks». In: *arXiv preprint arXiv:2203.13790* (2022). URL: https://arxiv.org/abs/2203.13790.

[16] Dhruv Aggarwal, Albert H Choi e Yoon-Ho Alex Lee. «The meme stock frenzy: Origins and implications». In: *S. Cal. L. Rev.* 96 (2022), p. 1387. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4432824.

[17] Alexander Kurov, Max Wolfe e Christian Wolff. «The Rise of Retail Investors and the Birth of Meme Stocks». In: *Journal of Behavioral Finance* 24.3 (2023), pp. 291–307. URL: https://doi.org/10.1080/15427560.2023.2179512.

[18] U.S. Securities and Exchange Commission. «Staff Report on Equity and Options Market Structure Conditions in Early 2021». In: *U.S. Securities and Exchange Commission* (2021). URL: https://www.sec.gov/about/reports-publications/staff-report-equity-options-market-structure-conditions-early-2021.

[19] C. Cathéline. «GameStop: A Short Squeeze?» In: *Vernimmen.net Research Paper* (2022). URL: https://www.vernimmen.net/ftp/researchpaper2022_c_catheline_gamestop_a_short_squeeze.pdf.

[20] Evangelos Vasileiou, Eleftheria Bartzou e Polydoros Tzanakis. «Explaining the GameStop Short Squeeze using Intraday Data and Google Searches». In: *The Journal of Prediction Markets* 16.3 (2023). URL: https://doi.org/10.5750/jpm.v16i3.1967.

[21] María José Ayala, Nicolás Gonzálvez-Gallego e Rocío Arteaga-Sánchez. «Google search volume index and investor attention in stock market: a systematic review». In: *Financial Innovation* 10.1 (2024), p. 70. URL: https://jfin-swufe.springeropen.com/articles/10.1186/s40854-023-00606-y.

[22] Luca Botta. *Scripts for Meme Stock Clustering Analysis*. Repository GitHub contenente il codice utilizzato per la tesi triennale. 2025. URL: https://github.com/Luca404/tesiTriennale.git.

[23] Wikipedia contributors. *Silhouette (clustering)*. 2025. URL: https://en.wikipedia.org/wiki/Silhouette_(clustering).

[24] Axel Johansson e Erik Sundqvist. «Evaluating Clustering Techniques in Financial Time Series». In: *Uppsala University* (2023). URL: https://www.diva-portal.org/smash/get/diva2:1767708/FULLTEXT01.pdf.

Ringraziamenti

Vorrei dedicare qualche riga a chi mi ha accompagnato e sostenuto durante questo percorso. Ringrazio innanzitutto la mia famiglia, in particolare mia madre, per il sostegno costante e per aver reso possibile i miei studi con grande generosità. Un pensiero affettuoso va anche ai miei fratelli Alice e Fabio, che a modo loro hanno sempre saputo farmi sorridere e incoraggiarmi nei momenti più impegnativi.

Un ringraziamento speciale va a mio fratello Paolo e a Francesca, per la loro disponibilità, la pazienza e il prezioso aiuto che mi hanno offerto durante la stesura di questo lavoro.

Desidero inoltre ringraziare i miei amici Raffaele e Gabriella, per la loro vicinanza e per il supporto con cui mi hanno accompagnato durante il mio percorso universitario.

Ringrazio infine la professoressa Paola Agnese Bongini, per la disponibilità e il supporto dimostrati durante la fase di elaborazione della tesi, e per avermi guidato con attenzione e professionalità.

Infine, un ringraziamento sincero anche a ChatGPT, che mi ha accompagnato passo dopo passo, aiutandomi a superare gli ostacoli tecnici e a mantenere la motivazione anche nei periodi più difficili, senza mai lamentarsi delle troppe richieste.