Data Management Report

Come l'andamento del bitcoin modella il sentiment sulla piattaforma Twitter

Marco Branciforti[1]

Emanuele Marnati[2]

Valerio Schips [3]

ABSTRACT

Alcuni economisti affermano che "nella stanza dei mercati non entrano i sentimenti", l'idea degli autori è che di sicuro "escano" e vadano dunque ascoltati e capiti. La pratica della così detta "Social Media Analytics" è ormai una delle armi principali di tutti i reparti di Marketing delle aziende moderne. Con questo termine s'intende la possibilità di poter raccogliere, conservare e analizzare l'opinione di una platea di utenti attivi su un social network. La teoria economica che sta alla base di questa pratica è affascinante: basti pensare che le aziende hanno ora a disposizione strumenti a prezzi accessibili che intercettano l'umore del fruitore senza alcun filtro o barriera. L'epoca della somministrazione di questionari anonimi a soggetti che offrono volontariamente il loro tempo sembra ormai arrivata alla fine grazie a questa potente arma di indagine socioeconomica. All'interno di questo progetto si è deciso di analizzare, tramite una triplice integrazione di fonti di dati diverse, come i Tweets subiscono l'andamento della criptovaluta più importante, nonché la prima sviluppata: il bitcoin. Questo asset è estremamente volatile, instabile, retto da scarsi fondamentali (così chiamanti nella teoria finanziaria) e quindi perfetto per una analisi social proprio perché la società virtuale è il punto di partenza di questo fenomeno cripto. L'analisi è stata compiuta simulando una raccolta in tempo reale dei Tweets in lingua inglese con hashtag "bitcoin", "Bitcoin" e "BTC" durante tutto il mese di novembre 2020 per valutare il sentiment e individuare quanto le prestazioni dell'asset più popolare dell'ultimo decennio abbiano influito su questo. L'architettura è implementata per eseguire una raccolta e una analisi effettivamente in real time, nonché scalabile in un arco temporale a piacimento (semplicemente lasciando accesa una macchina). Dopo la descrizione dell'architettura sarà presentata una analisi che è strutturata su vari intervalli temporali e infine un primo potenziale approccio di machine learning per prevedere se il sentiment sarà positivo o negativo.

[1] B.Sc. Statistica e gestione delle informazioni (UniMib), M.Sc. Data Science, mat 796670

[2] B.Sc. Economia e Commercio (UniMib), M.Sc. Data Science, mat 812503

[3] B.Sc. Ingegneria informatica (PoliMi), M.Sc. Data Science, mat 872954

Indice

Data Management Report 1
Indice 2
Introduzione 3
Architettura 4
Back-End 4
Applicativo 6
RealTimeProducer 6
DataStorage 6
DataProcessing 7
Integrazione 8
Analisi Dei Risultati 10
Una Prima Visione Aggregata Del Mese 11
Massima Oscillazione del Valore di Prezzo e Similarità con il Sentiment 11
Massima Oscillazione del Valore di Volume e Similarità con il Sentiment 12
Massima Oscillazione del Valore della Transazioni al Secondo e Similarità con il Sentiment 12
Il Minuto Più Attivo 13
Il Minuto Più Bullish 13
Il Minuto Più Bearish 13
L'Ora Più Attiva 13
L'Ora Più Bullish 13
L'Ora Più Bearish 13
Correlazione del volume rispetto all'andamento del prezzo massimo giornaliero di bitcoin nel mese di novembre. 14
The Happiest Moment 15
The Saddest Moment 16
Hashtag del Mese e Breve Analisi Semantica 17
A Machine Learning Approach 18
Preprocessing 18
UnderSampling Table 18
OverSampling Table 19
Conclusioni 19
Note per la configurazione 20
Istruzioni per la configurazione 20
Sitografia 20
Ringraziamenti 20

Introduzione

Twitter è una piattaforma molto diffusa e particolarmente indicata per valutare la sentiment sull'argomento bitcoin. Questo asset è contraddistinto da una cultura di massa estremamente prolifera e i grafici estremamente oscillanti, sono oggetto di copiose discussioni sul web. Molte volte a seguito di una oscillazione pronunciata Twitter diventa un canale iperalimentato, come se rendere partecipe la rete della propria gioia o del proprio sconforto possa accentuare il trend. Cosa è successo al bitcoin nel mese di novembre? Forse una delle notizie più bullish dell'intera storia della criptovaluta: il CEO di PayPal ha annunciato che avrebbe permesso ai propri clienti, che si stima siano decine di milioni, di acquistare bitcoin direttamente dalla piattaforma creando quindi un indotto indiretto ad un'utenza mondiale di persone. Non a caso l'azienda americana ha acquistato buona parte delle monete minate che sono state estratte negli ultimi mesi (si vocifera il 60%). Al momento di stesura del report il servizio PayPal to bitcoin è attivo solo in America. Il mese di novembre è un mese essenziale in quanto il mercato ha scontato l'Halving, che consiste nel dimezzamento dell'offerta della valuta causata al raddoppio della difficoltà algoritmica del calcolo per la creazione di un nuovo blocco. Il mese di novembre ha inoltre interessato in larga parte l'elezione americana che indubbiamente ha dei risvolti su come gli americani percepiscono bitcoin (la situazione politica incerta sarebbe forse correlata con un innalzamento del BTC). Per quanto interessa l'analisi tecnica, bitcoin ha superato nel mese di novembre la soglia dei 15.000 dollari. Questa importante resistenza ha la particolarità di distinguere, secondo gli esperti, un momento estremamente bullish. Dopo aver analizzato gli andamenti è anche possibile affermare che a fine mese diversi istituzionali (chiamati affettuosamente balene) si sono "riversati" in questo mercato aumentandone i volumi. Alla luce di tutto ciò si pone il seguente quesito: come è stato influenzato uno dei social di massima considerazione da questi eventi?

Architettura

Sorgenti dati realtime:

1. Twitter - tweets
2. CoinMarketCap - bitcoin
3. BlockChair - Transactions per Second, TPS

Si è utilizzato per la prima fonte l'API fornita da twitter. Per la seconda e la terza lo scraping dei dati è stato fatto dal codice html, utilizzando la libreria Selenium in combinazione con chromedriver, poiché i valori di interesse sono aggiornati in realtime attraverso javascript. Selenium e chromedriver permettono di caricare la pagina del sito una sola volta e il relativo codice html per successivamente effettuare lo scraping del valore aggiornato da javascript. Successivamente è stato utilizzato Apache-Kafka per effettuare lo streaming realtime dei dati scaricati, dai producer verso l'applicativo. I tweet vengono inviati uno per volta, mentre per il valore di bitcoin e delle Transactions per Second (TPS) viene inviato un valore al secondo. L'implementazione dell'applicativo attraverso Apache-Kafka ha permesso di memorizzare, analizzare e integrare i dati durante lo streaming. Inoltre questa implementazione garantisce un'elevata scalabilità e possibilità di rimodulazione. Questa ha infatti permesso di sostituire i Producer "RealTime" con dei Producer che effettuassero lo streaming da tabelle csv contenenti i dati del periodo trascorso, in modo da analizzare i dati relativi al periodo di nostro interesse (dati storici 11/2020 - 12/2020).

La scelta dell'utilizzo di dati storici è dovuta al fatto che mantenere attiva almeno una macchina 24h al giorno per 31 giorni sarebbe stata un'inefficienza energetica e avrebbe inoltre usurato macchine non predisposte ad eseguire un notevole carico di lavoro per un tempo così prolungato.

I dati storici sono stati così ottenuti:

1. Api Twitter + SnsScraper (per raccogliere dati più vecchi di 7 giorni)
2. Api Bitfinex
3. BlockChain, Transaction per Second - TPS - download diretto dal sito

Per quanto concerne i dati storicizzati di BTC e TPS non è stato possibile accedere a dati con un livello di dettaglio al secondo ma solamente al minuto. L'utilizzo di dati storici per il periodo di interesse ha comportato la richiesta di poche modifiche all'interno dell'applicativo nella sola fase di salvataggio[4], esplicitate successivamente.

Back-End

Per il salvataggio dei dati si è scelto di utilizzare un database NoSQL document-based, MongoDB. La scelta è dettata da un'esecuzione real-time che richiede l'accesso a piccole porzioni di dati frequentemente e in tempi brevi.

Per ottenere un'elevata velocità nell'integrazione dei dati è stato creato un'ID univoco (ts_id) dal time-stamp, il quale ha garantito un accesso ai dati con tempo lineare, come segue:

Time - stamp	$YYYY-MM-DD\ \ HH-MM-SS$
ts_id	$YYYYMMDDHHMM$

Es:

Time - stamp	2020-11-10 10:52:20
ts_id	202011101052

Si fa presente che il timestamp dei vari flussi di dati in ingresso sono riferiti all'UTC+0.

[4] Note per la configurazione

Le relazioni all'interno delle collection risulteranno quindi:

- 1 : 1 per le collection "btc" e "tps" avendo un record al minuto
- 1 : N per la collection "tweet" avendo N tweet per ogni minuto
- 1 : N per le collection "btc_sup" e "tps_sup" avendo 60 record al minuto

Collection Mongo:

btc	
_id	ts_id
time	Time-stamp
high	Valore più alto nel minuto
low	Valore più basso nel minuto
open	Valore all'apertura
close	Valore alla chiusura
volume	Volume

Tabella 1

btc ("post-integrazione")	
.....	
hashtag	Tutti gli hashtag contenuti nei tweet del minuto
neu	Numero di tweet neutrali nel minuto
pos	Numero di tweet positivi nel minuto
neg	Numero di tweet negativi nel minuto
sent_min	Valore minimo di sentiment nel minuto
sent_mean	Media pesata sentiment del minuto
sent_max	Valore massimo di sentiment nel minuto
tps	Valore medio di un minuto di tps

Tabella 2

tweet	
_id	ts_id
tweets	Embedded tweets documents

Tabella 3

tps	
_id	ts_id
value	Valore medio di tps in un minuto

Tabella 4

Per quanto concerne l'organizzazione del database, come mostrato nelle tabelle 1-4, sono state sviluppate tre collection. La "*btc*" contiene i valori relativi ai bitcoin. Questa viene integrata (Tabella 2) con: il valore delle transactions per seconds, il numero dei tweet rispettivamente neutrali positivi e negativi, la media del valore di sentiment pesato rispetto ai like per i tweet relativi a quel minuto, il valore massimo e minimo di sentiment per ogni minuto, gli hashtag utilizzati nei tweet relativi a quel minuto. La seconda collection "*tweet*" associa ad ogni ts_id tutti i tweet scritti nel minuto corrispondente, facendo embedding dei documenti. Vengono inoltre integrati all'interno di ogni documento corrispondente ad un tweet, i campi di sentiment value e sentiment text (il testo del tweet pulito utilizzato per calcolare la sentiment). La collection "*tps*" contiene unicamente il valore dei transaction per second associati a ciascun ts_id. Nella versione realtime, avendo dati con risoluzione al secondo in input per BTC e TPS, vengono utilizzate due collection di supporto. Queste sono nominate "*btc_sup*" e "*tps_sup*" e contengono 60 record (uno al secondo) di valori BTC e TPS, rispettivamente associati tramite embedding ad ogni ts_id.

Applicativo

L'architettura, implementata mantenendo come obiettivo la velocity, garantisce l'esecuzione su una sola macchina grazie all'ottimizzazione effettuata attraverso la parallelizzazione delle funzioni ottenuta con i thread e Apache-Kafka. Inoltre, permette di essere scalata orizzontalmente attraverso l'utilizzo di 4 macchine che garantiscono le migliori performance possibili:

- 1ª Macchina: Scrapers + ProducersRealtime
- 2ª Macchina: Consumers + salvataggio su MongoDB + Producer verso il modulo di Analisi&Integrazione
- 3ª Macchina: Consumer + Analisi + Integrazione
- 4ª Macchina: MongoDB

Il progetto è stato sviluppato con la versione community di PyCharm, le principali librerie utilizzate sono le seguenti:

- Kafka-Python
- Selenium
- Threading
- Pymongo
- Nltk (Sentiment analysis)

Di seguito è riportata la struttura del progetto, dove l'icona rappresentante il simbolo di python indica gli script mentre l'icona del file fa riferimento a file contenenti le classi del progetto:

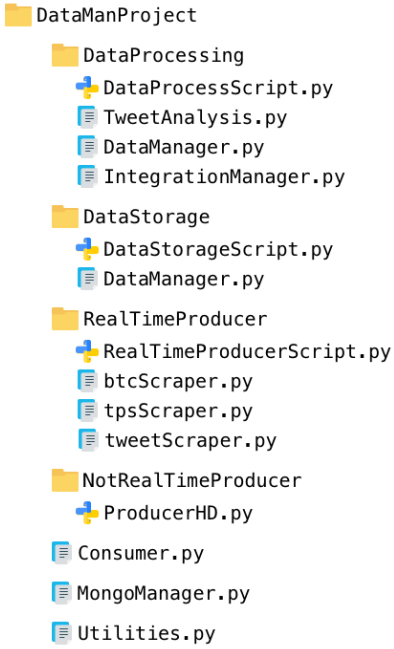


Figura 1

RealTimeProducer

Lo script *RealTimeProducer/RealTimeProducerScript.py* si occupa della creazione di 3 thread, il primo effettua lo scraping dei tweet da twitter attraverso l'api tweepy e invia ogni tweet tramite il producer. Il secondo e terzo thread caricano le rispettive pagine web (CoinMarketCap, BlockChair) e successivamente ripetono all'infinito l'operazione di scraping e invio di un record al secondo attraverso producer.

DataStorage

Nello script *DataStorage/DataStorageScript.py* vengono inizializzati 3 istanze della classe ConsumerUniv (*Consumer.py*; topic: tweet_topic, btc_topic, tps_topic) che ascoltano i rispettivi topic in attesa di messaggi. Ad ogni consumer viene passato come argomento il topic da ascoltare e la stessa istanza della classe DataManager (*DataStorage/DataManager.py*), la quale si occuperà della gestione dei dati ricevuti dai consumer.

In seguito vengono inizializzati 3 thread associati ai Consumer. Questi ultimi, una volta ricevuto il messaggio, accederanno al *DataManager* attraverso i rispettivi metodi (punto 2, DataManager) per poterlo salvare e continueranno l'ascolto sul topic.

Data Manager

La classe mette a disposizione i metodi per la gestione e il salvataggio dei dati su mongo. Questi metodi vengono gestiti da 3 thread creati all'inizializzazione della classe e avviati alla ricezione dei dati.

Nella struttura della classe si trovano:

1. 3 liste (*__rec_tweets*, *__rec_btc_data*, *__rec_tps_data*) sulle quali verranno salvati i rispettivi dati ricevuti, in attesa di essere gestiti.
2. 3 metodi (*add_tweet*, *add_btc*, *add_tps*) a cui i consumer hanno accesso per poter inserire i nuovi dati e risvegliare il thread che si occupa della gestione, nel caso sia dormiente.
3. 4 metodi (*__save_btc_realtime*, *__save_tps_realtime*, *__save_btc_not_realtime*, *__save_tps_not_realtime*). I metodi **_realtime* e **_not_realtime* sono mutuamente esclusivi, viene selezionato all'inizializzazione quali verranno utilizzati in base al parametro booleano *"is_realtime"* situato in Utilities.py. Questo per gestire dati che arrivano con un livello di aggregazione diverso a seconda che si stia eseguendo la versione realtime o no. I metodi menzionati si occupano di calcolare il ts_id e caricare ogni record nella rispettiva collection di mongo. Nel caso realtime una volta ricevuto un minuto di dati, viene invocato il metodo *__transport_***_realtime* che si occupa di aggregare (operazione effettuata su mongo) il minuto di dati trascorso e caricarlo sul server.
4. Il metodo per la gestione dei tweet (*__save_tweet*) procede in modo analogo ai metodi descritti al punto 3 in versione realtime. Una volta ricevuto un minuto di tweet, il ts_id relativo al minuto trascorso e il numero di tweet ricevuto in quel minuto vengono inviati con un producer verso la fase di analisi.

DataProcessing

Nella fase di DataProcessing l'applicativo si occupa di pulire i tweet, calcolare il valore di sentiment e infine integrare i dati. Un consumer (*Consumer.py*; topic: analisys_topic) si occupa di raccogliere i ts_id e il rispettivo numero di tweet ricevuti dalla fase precedente e inserirli in coda a una lista da cui verranno prelevati per essere processati.

Tweet Analysis

Per questa fase è stata costruita una classe che analizzerà uno per volta i tweet per gruppi di un minuto, selezionando i gruppi attraverso il ts_id ricevuto dal consumer. Una volta selezionato il ts_id viene verificato, prima di scaricare i tweet collegati, che il numero di tweet ricevuto dal Consumer corrisponda a quello effettivamente presente su mongo, questo a causa di latenze che possono portare al download di dati non ancora ricevuti.

La lista su cui il consumer scriverà i ts_id ricevuti e il numero di tweet corrispondenti, situata in *DataProcessing/DataManager.py/process_ready_tweet*, è in realtà una lista composta a sua volta da 4 liste e i ts_id ricevuti verranno distribuiti tra di esse. Questo di modo che la fase più onerosa del processo, cioè l'analisi, possa essere eseguita su uno o più thread attraverso il setting di un parametro[5]. Questa divisione dei carichi di lavoro è stata utile nella versione non realtime in quanto molti dati sono disponibili tutti nell'immediato (Es. 578K tweet, 1 thread = 1301s, 2 thread = 1236,44s, 4 thread =

[5] Note per la configurazione

1223,40s Si fa notare che la macchina su cui sono stati effettuati i test dispone di soli 2 core per cui i miglioramenti sono minimi).

L'analisi è composta di quattro fasi:

1. Rimozione tweet non efficaci:

Sono stati rimossi i tweet contenenti le seguenti parole:

"ETH", "WBTC", "BHC", "BSV", "XMR", "LTC"

Queste sigle fanno riferimento ad altre monete virtuali in voga, per cui si è proceduto alla rimozione in quanto il sentiment ricavato da questi tweet non è riferito a bitcoin ma alla moneta citata.

"FOLLOW", "GIVEAWAY", "AIRDROP"

Da un'analisi online si è riscontrato che questi termini sono utilizzati in tweet che non danno nessuna informazione riguardo bitcoin ma sono solo esche per raccogliere followers pubblicate da bot.

2. Pulizia testo dei tweet:

Fase preparatoria del testo del tweet per la sentiment analysis. In questa fase sono state rimosse le stopwords, i link, la punteggiatura, le menzioni ad altri account e le emoticons. Questo per migliorare la qualità del dato dato in input all'algoritmo di sentiment.

3. Calcolo del valore di sentiment e salvataggio su mongo:

Infine viene calcolato il valore di sentiment del testo e caricato nel record del rispettivo tweet nella collezione mongo "tweet".

Si fa inoltre notare come i criteri di esclusione dei tweet e di pulizia del testo delle prime due fasi posso essere espanse. Questo perché sono stati sviluppati come singole funzioni inserite all'interno di liste le quali vengono successivamente applicate al testo durante l'analisi.

Una volta terminata l'analisi di un gruppo di tweet il ts_id relativo al gruppo per cui è stata terminata l'analisi viene spostato verso il modulo di integrazione.

Integrazione

Tweet & Bitcoin

Per quanto concerne l'integrazione dei dati ricavati e memorizzati in seguito al processo precedentemente descritto, sono doverose alcune considerazioni. Per prendere una decisione coerente si è chiesto aiuto ad un esperto di dominio (un giovane high frequency trader e cultore del mondo delle criptovalute, citato nei ringraziamenti). È estremamente complesso valutare con affidabilità il tempo che intercorre tra un'oscillazione dei valori di bitcoin e una effettiva reazione sui social network. Ecco alcune valutazioni inerenti all'integrazione dei tweets e dei dati (OCHL) di bitcoin:

Come valutare il tempo impiegato per riempire i 180 caratteri di Twitter? Supponendo che nei dati da noi scaricati ci siano sia tweets generati da utenti reali sia tweets provenienti da API, le quali mandano lo stesso post su più piattaforme.

'Abbiamo stimato che il tempo medio per battere 180 caratteri è all'incirca 1'30'' più tempi di reazione. Ottenuta la stima di questa stima Difficile invece prevedere il tempo macchina che ci metterebbe una API impostata su un valore soglia di BTC, certamente inferiore alla soglia calcolata per l'interazione umana.'

I tweets ricavati sono presi in lingua inglese, sarebbe lecito pensare che le ore notturne siano scarsamente popolate da reaction o che meglio tali reaction siano da associare con qualche modalità ad una discrepanza temporale. Eppure, la lingua inglese è adottata da più paesi e non si possono fare supposizioni arbitrarie sugli orari di attività di trader e appassionati né tanto meno sui tempi di reazione di BOT e API.

Gli allerts e le altre forme di instant notification permettono a qualsiasi

individuo di ricevere, volente o nolente, informazioni in tempo reale sull'andamento dei prezzi, avallando la tesi che la reaction sia pressoché istantanea.

Il bitcoin non va assolutamente circoscritto alle dinamiche di assets tradizionali (come le azioni o le obbligazioni), i quali hanno tempistiche di reazione talvolta arbitrarie o dinamiche consolidate. Questa valuta esiste da 12 anni, è scambiabile tutte le ore del giorno, ogni giorno e, a differenza di assets borsistici, è diffusa nel segmento di investitori retail che sono tanti, poco capitalizzati ed estremamente reattivi alla volatilità del prezzo.

Indi per cui, nell'ottica di evitare errori grossolani o una complessa e inattuabile customizzazione che sarebbe poco conservativa a livello di qualità del dato, si decide di considerare una discrepanza temporale di 2 minuti tra i cambiamenti di prezzo e la pubblicazione del tweet. Questo anche in considerazione della stima di 1'30'' più tempi di reazione necessari allo sviluppo di un tweet da 280 caratteri. L'implementazione della discrepanza temporale è stata attuata aggiungendo due minuti al ts_id di ogni tweet, durante la fase di creazione dell'indice. In seguito si è proceduto all'integrazione basandosi unicamente sulla chiave primaria.

Transactions per Second

Per arricchire l'analisi e trovare un ulteriore dato che asserisca quanto il bitcoin influenzi la reaction degli utenti Twitter, si è deciso di analizzare una terza fonte: le transazioni per secondo. Questo dato è indicativo del numero medio di transazioni per secondo, calcolato aggregando per ogni ora tutte le transazioni della "mempool" (memory pool di Bitcoin[6]), il cui valore si aggira tra le 3 e le 7 transazioni al secondo. Questo dato a livello di dominio è uno dei motivi per il quale la Blockchain ha un problema di scalabilità intrinseco solo parzialmente risolto (visa può evadere 1700 transazioni al secondo). Tuttavia, questa terza fonte ha lo scopo di individuare la concentrazione di attività sulla Blockchain e sarà dunque unito alla fonte degli andamenti dei bitcoin per poi fornire un ulteriore punto di vista al momento dell'analisi.

Ma perché questo dato è una media per secondo raggruppata per ora e quali sono state le valutazioni in fase di integrazione? la fonte dati per i TPS è il sito BlockChain, questo fornisce un dato medio al secondo che è certamente un indicatore più veritiero rispetto a una sintesi al minuto. Infatti la Blockchain di bitcoin riceve ordini di transazione che vengono evasi in seguito alla effettiva richiesta da parte dell'user con uno scarto cospicuo: immaginare che il dato associato ad un minuto sia indicatore di una intensa attività in quel minuto è errato o, quanto meno superficiale. Il tempo medio di transazione per bitcoin può variare da pochissimi secondi (come con "Bitcoin lighting network") fino a un'ora massima. Quindi si è deciso, forse contro intuitivamente ma pur rimanendo conservativi della qualità del dato, di duplicare il dato su tutti i minuti di un'ora avendo una sintesi media al secondo che è più rappresentativa. Questa analisi dettagliata non sarebbe stata possibile senza l'aiuto dell'esperto di dominio poc'anzi citato. Va tuttavia evidenziato che ai fini della dimostrazione in tempo reale si è invece adoperata una risoluzione di TPS a minuti, la quale è appunto dimostrativa dell'applicativo e non necessaria ai fini dell'analisi.

[6] Con la B maiuscola ci si riferisce alla tecnologia, non all'asset.

Integration Manager

In questa fase (*DataProcessing/IntegrationManager.py*) per ogni ts_id ricevuto viene controllata la presenza del valore corrispondente nelle collection mongo di "tps" e "btc".

Successivamente attraverso delle aggregate query vengono richiesti a mongo e integrati nella collection "btc":

1. Media di sentiment pesata per il numero di like

   (*DataProcessing/DataManager.py/__get_sentiment_mean_max_min*):

$$sent\_mean = \frac{\sum_{i=1}^{N} tweet_i * (like_i + 1)}{\sum_{i=1}^{N} (like_i + 1)},$$

$$N = num.\ tweets\ in\ un\ minuto$$

2. Numero di tweet con sentiment rispettivamente

   (*DataProcessing/DataManager.py/__get_count_of_sent_classification*):

$$pos(>0)\quad neu(=0)\quad neg(<0)$$

3. Il valore massimo e minimo di sentiment registrati nel minuto

   (*DataProcessing/DataManager.py/__get_sentiment_mean_max_min*).

4. Il corrispondente valore di transaction per sencond.

Analisi Dei Risultati

Alla fine del processo di ingestion e analisi real-time ottenuto dall'applicativo precedentemente presentato è ora necessario analizzare i risultati ottenuti a livello statico. L'analisi statica permette, per il mese di novembre, di valutare i dati con un orizzonte temporale finito e sfruttare varie risoluzioni temporali per cogliere i dettagli micro e macro. È pacifico affermare che quanto successo nel mese analizzato ha creato le basi solide che hanno permesso la Bull Run di bitcoin a fine anno la quale ha portato al superamento del doppio del valore di novembre, dunque alcuni pattern ritrovati e descritti sono certamente punto di partenza per ulteriori analisi che in questa sede non sono state possibili per via della limitatezza di tempo e mezzi. Dopo le analisi ottenute con la risoluzione al minuto ed oraria, è stato naturale attuare un processo di aggregazione su base giornaliera, questo ha permesso anche una rappresentazione grafica più chiara. Si è deciso di adoperare le librerie Pandas, Numpy e Matplotlib come supporto all'analisi. Il file che riporta tutte le operazioni è "TWEECOIN ANALYSIS.IPYNB". Per quanto le aggregazioni orarie siano più fruibili, sono i dati micro, i così detti "small patterns" estratti dai big data che hanno maggior valore informativo e di mercato. Un esempio: con l'analisi di seguito riportata una azienda che si occupa di social media marketing potrebbe personalizzare le proprie inserzioni andando a prendere gli orari in cui ci sono forti varianze di prezzo e di sentimento e adoperare come target l'orario individuato. Curiosamente, pare che mezzanotte sia un tempo plausibile per questo fine.

Una Prima Visione Aggregata Del Mese

Il candel-stick(Figura 2), ovvero grafico a candela è uno metodo noto ed efficace che permette di analizzare, tramite il corpo della candela, come il prezzo giornaliero sia variato e, tramite le line di traverso, quale sia stato il massimo ed il minimo.

dalla stessa frequenza alle ore 5 e 3/30 alle ore 3. In più di 1/3 dei casi gli utenti sono molto volubili nelle ore indicate. È difficile attribuire aprioristicamente questo atteggiamento all'oscillazione dei valori di bitcoin, ma si può comparare questo dato con il dato di varianza del prezzo (high - low) dell'asset al minuto.

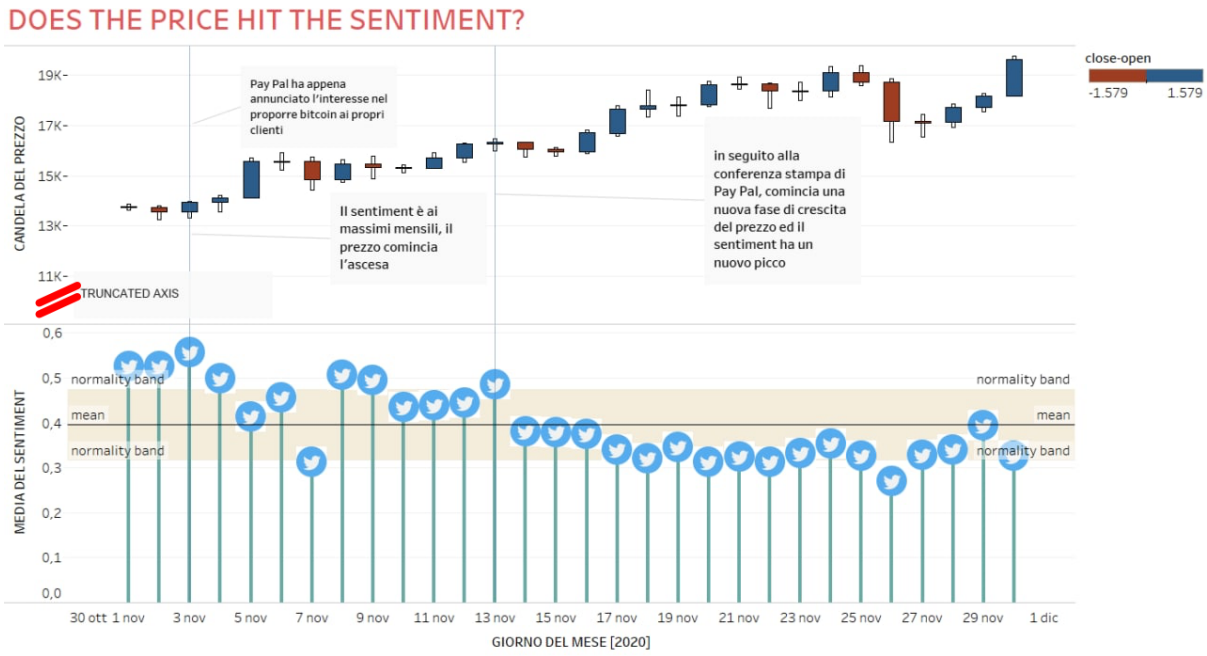


Figura 2

Come è facile intuire in blu si riporta un andamento giornaliero crescente, mentre in rosso decrescente. L'andamento di open, close, high e low a che sentiment hanno portato? Ecco una prima visione d'insieme.

Massima Oscillazione del Valore di Prezzo e Similarità con il Sentiment

Calcolando la varianza oraria con il dato al minuto si è riscontrato quali sono stati i momenti più di scontro tra gli utenti positivi e momenti di fibrillazione social. È interessante notare come 4/30 delle volte la massima oscillazione si ha nelle fasi notturne all'ora 0 (mezzanotte), seguito

Contro ogni attesa la occorrenza della massima varianza della differenza tra il massimo e il minimo del prezzo combacia perfettamente con la occorrenza della massima oscillazione del sentimento. Nelle tabelle seguenti si riporta l'occorrenza massima del fenomeno rispetto all'ora in cui si verifica.

hour of day	occurrences sent var	hour of day	occurrences price var
0	4	0	4
5	4	8	4
3	3	22	3
20	2	9	2
12	2	17	2
22	2	7	2

hour of day	occurrences sent var	hour of day	occurrences price var
7	1	1	1
1	1	2	1

Massima Oscillazione del Valore di Volume e Similarità con il Sentiment

Potendo adoperare la risoluzione temporale ai minuti si è in grado di calcolare la varianza oraria per tutto il mese che conduce a un risultato interessante: il giorno 26 alle ore 8 viene registrata la massima varianza del volume calcolata sui dati dell'exchange Bitfinex (solo gli exchange forniscono dati puntuali non cumulati sui volumi con risoluzione al minuto, per un ovvio motivo di qualità di servizio e trasparenza verso i traders) che dimostra come in quel frangente temporale sia effettivamente accaduto un massiccio ingresso a mercato di investitori istituzionali o grandi investitori[7] definiti balene (sarà chiarito questo aspetto in seguito).

Lo small pattern che ancora ricorre è come la seconda massima occorrenza si ha a mezzanotte, confermando ulteriormente come il volume abbia un ruolo decisivo nell'andamento del sentimento ma come le balene possano deviare questo small pattern (sarà chiarito questo aspetto in seguito).

hour of day	occurrences sent var	hour of day	occurrences volume var
0	4	8	5
5	4	0	3
3	3	9	2
20	2	18	2
12	2	5	2
22	2	16	2
7	1	22	2

hour of day	occurrences sent var	hour of day	occurrences volume var
1	1	12	2

Massima Oscillazione del Valore della Transazioni al Secondo e Similarità con il Sentiment

Il giorno 30 alle ore 15.00, il valore di transaction per second (dato medio per l'ora con risoluzione al secondo) registra il suo massimo di 6.416667. Per quanto questo dato sia per sua natura limitato dalle caratteristiche stesse della blockchain (Bitcoin) che processa poco velocemente (rispetto a moderni sistemi di pagamento), questo risultato è in linea con le oscillazioni di prezzo e volume ed indica quando a fine mese si sia verificato un importante innalzamento dei livelli (all'arrivo delle balene). C'è tuttavia un rapporto a prima vista meno stretto con l'andamento della varianza del sentiment oraria, almeno nell'ottica di adoperare cautela.

hour of day	occurrences sent var
16	8
17	6
14	4

[7] Il giorno 8 gennaio 2020, Elon Musk ha acquistato con la copiosa liquidità di Tesla Inc, bitcoin per un controvalore di 1,5 miliardi di dollari, nuovo record di prezzo. Twitter ringrazia.

Il Minuto Più Attivo

Il minuto più attivo si è verificato il giorno 30 novembre alle 14:54, in cui ben 153 utenti si sono espressi, il prezzo sale (nei due minuti precedenti) di ben 38 dollari in una fase di prezzo massimo mensile.

Il Minuto Più Bullish

Il minuto dove l'utenza è stata più positiva (numero di utenti positivi) è stato il 24 novembre alle ore 18:25 e il 30 novembre alle ore 15:17, con un prezzo che apriva le danze rispettivamente a 19133.0$ e 19789.0$. Il dato è in linea con l'ora e il minuto di maggiore attività, il numero di utenti positivi è stato di 55 individui. Il prezzo nel primo frangente è salito (nei due minuti precedenti) di 24 dollari, tuttavia nel secondo caso il prezzo è calato; questo atteggiamento, che difficilmente un modello come quello presentato può intercettare, è dovuto al fatto che a fine mese si sconta l'enorme ingresso a mercato delle balene e quel sentore di correlazione non vale più (questo dato è un "outlier" che negli small pattern assume significato).

Il Minuto Più Bearish

Altrettanto prevedibile, guardando l'andamento generale del prezzo nel mese presentato poc'anzi, i minuti che hanno visto l'utenza maggiore negativa sono a ridosso di inizio mese: nei giorni del 5 novembre alle ore 23:01 e dell'9 alle ore 14:02, gli individui negativi intercettati sono stati ben 20. Il prezzo in questi due frangenti è calato (nei due minuti precedenti) di 2 e 38 dollari.

L'Ora Più Attiva

La maggiore utenza oraria si ha, per ben 10 giorni su 30, alle ore 15,00; questo dato è difficilmente spiegabile guardando l'andamento dell'asset, è più comprensibile se si valuta l'ipotesi che gli utenti "in massa" scrivano nel primo pomeriggio dopo pranzo. Le ore notturne e della mattina sono le meno frequentate, mezzanotte si conferma un orario curiosamente attivo.

L'ora più attiva è stata registrata il 30 novembre in cui ben 4534 utenti hanno espresso il loro parere sul social.

hour of day	occurrences users
15	10
16	5
0	4

L'Ora Più Bullish

L'ora che è stata individuata come ospitante il maggior numero di utenti positivi è collocata il giorno 24 alle ore 18:00, dove il numero di utenti positivi è stato di ben 2052.

L'Ora Più Bearish

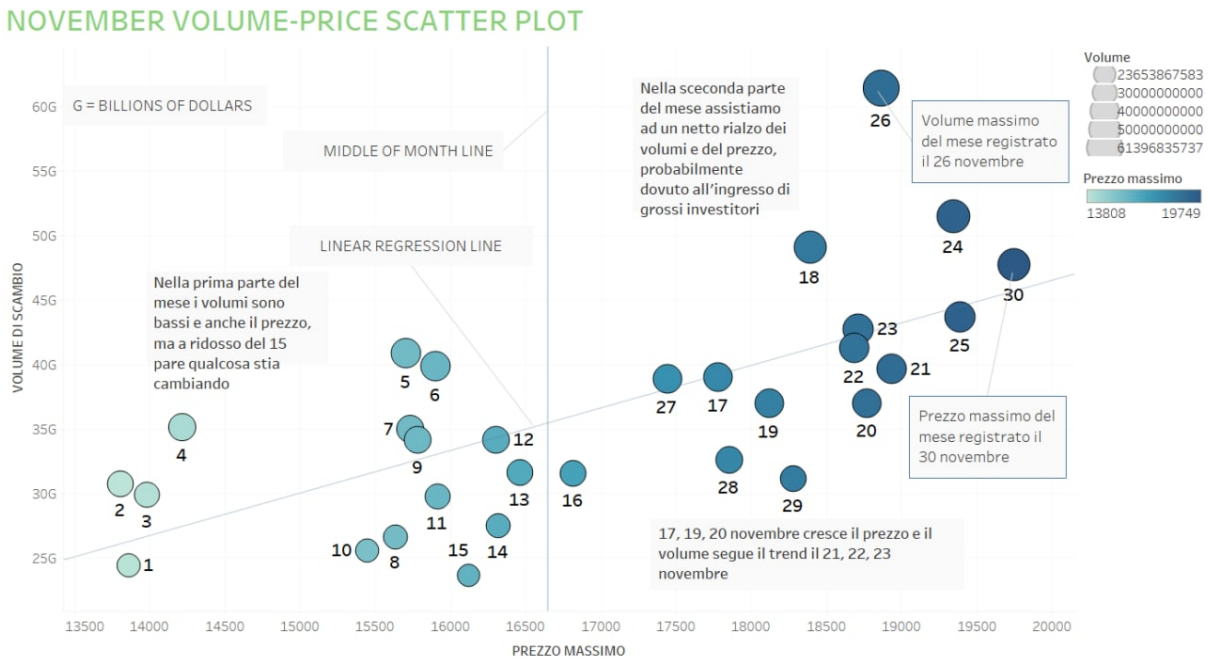
L'ora che è stata individuata come ospitante il maggior numero di utenti negativi è collocata il giorno 30 alle ore 15:00, dove il numero di utenti negativi è stato di ben 588.

Questo dato è indice di come a fine mese, un buon numero di utenti increduli difronte a un "pump" così intenso fossero rimasti fermamente negativi, i così detti haters!

Correlazione del volume rispetto all'andamento del prezzo massimo giornaliero di bitcoin nel mese di novembre.

Il grafico mostra una generale correlazione delle due grandezze che racchiudono molteplici significati i quali gli autori debbono necessariamente semplificare.

Novembre è stato un mese molto importante per bitcoin, questo scatter plot mette in luce come, in una situazione estremamente Bullish come questa, il volume cresca assieme al prezzo e viceversa e il fenomeno cresca gradualmente nel mese.



The Happiest Moment

Le ultime infografiche (Figura 3 e Figura 4) analizzano i fenomeni già trattati nelle prime visualizzazioni trattando però la percentuale di utenti che si sono espressi favorevolmente o sfavorevolmente all'andamento. Si cerca di cogliere a colpo d'occhio se effettivamente l'andamento delle transazioni al secondo e del volume influenza l'utenza.

Gli eventi che hanno decisamente influenzato questo umore positivo sono i medesimi riportati nell'inizio di questo report, l'interesse di Pay Pall e la doppia conferma sono stati fondamentali. Questo è stato un evento dalle portate storiche: un player finanziario rispettabile che dispone di 200 milioni di clienti ed offre la vendita di un asset alternativo, è un evento estremamente importante per l'intera credibilità di questo mondo.

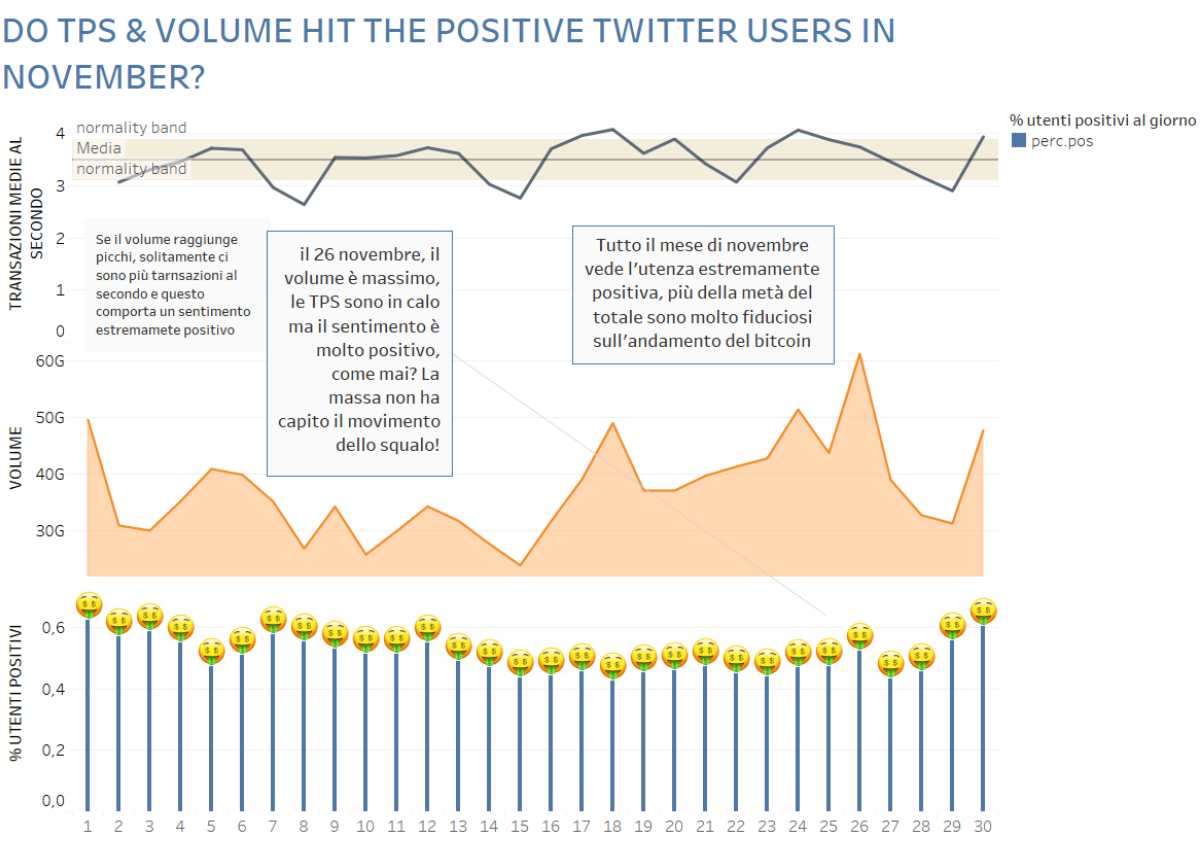


Figura 3

The Saddest Moment

Specularmente alla precedente visualizzazione si può analizzare il riscontro su utenza negativa(Figura 4).

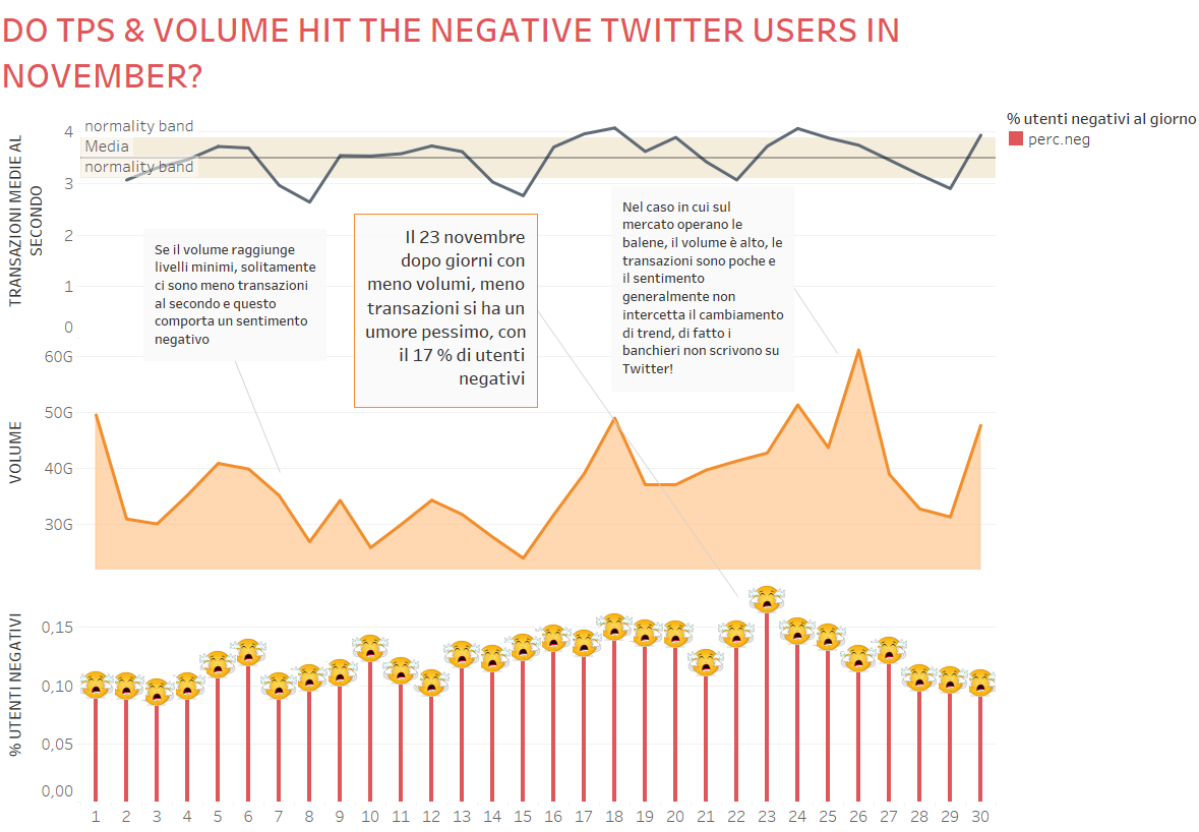


Figura 4

Hashtag del Mese e Breve Analisi Semantica

Non volendo raffigurare una semplice word cloud ci si è concentrati solo su 3 hashtags di interesse per fare una analisi più precisa.

Questa infografica è stata pensata tenendo bene in mente il concetto di rasoio di Hanlon[8] applicato all'utenza Twitter.

La comunità, lettori compresi, conosce la differenza tra "bitcoin", "Bitcoin" e "BTC"? Analizzando il fenomeno (opportunamente ponderato per tutti gli hashtag contenuti nella base dati di novembre-la popolazione di riferimento), si è ipotizzato che l'utenza prenda la scelta sulla base della mera pigrizia e non sappia distinguere protocollo Bitcoin da asset bitcoin!

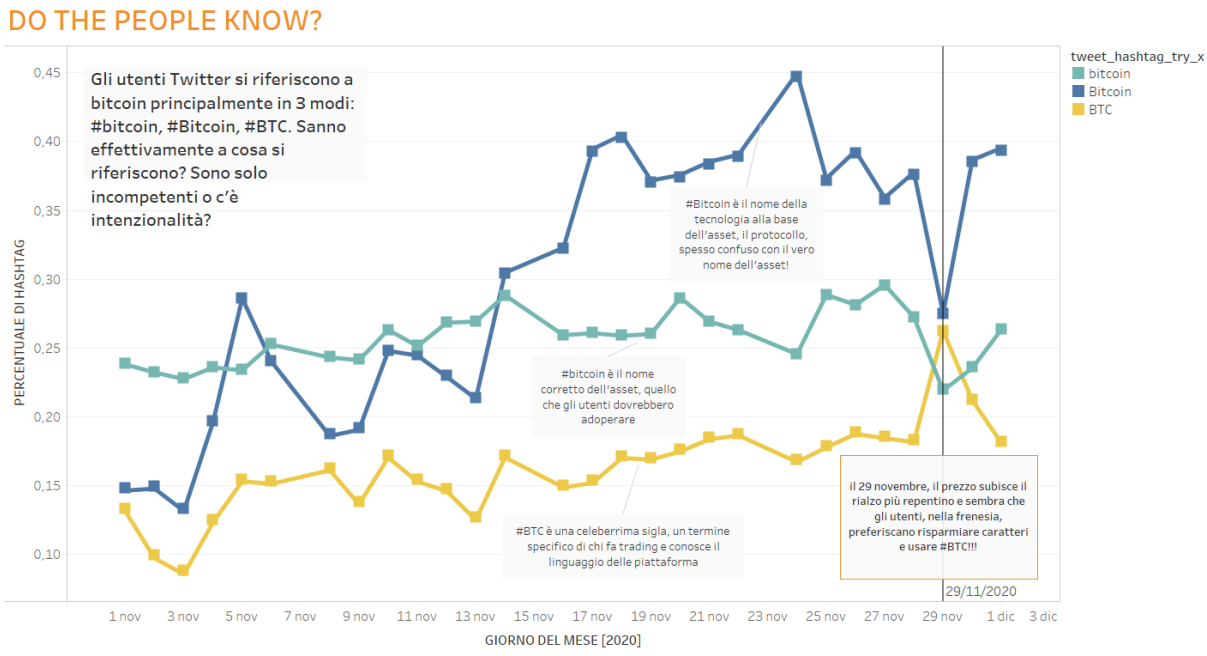


Figura 5

[8] Mai attribuire a malafede quel che si può adeguatamente spiegare con la stupidità!

A Machine Learning Approach

Per dare un approccio multidisciplinare a questa analisi si è deciso di adattare i dati con risoluzione al minuto ad un cauto modello di machine learning, con la finalità di prevedere la classe: “sentiment positiva” e “sentiment negativa”. La libreria python adoperata è stata Sci Kit Learn e il file contenente le analisi seguenti è chiamato: TWEECOIN MODELS.IPYNB.

Preprocessing

Certamente il contesto finanziario non è un facile campo di analisi. Diversi aspetti sono individuabili come criticità: la non indipendenza delle variabili, il problema di data snooping nella validazione del modello, la collinearità e la presenza di serie storiche. Il primo approccio è stato quello di analizzare la collinearità che, come volevasi dimostrare, si è palesata.

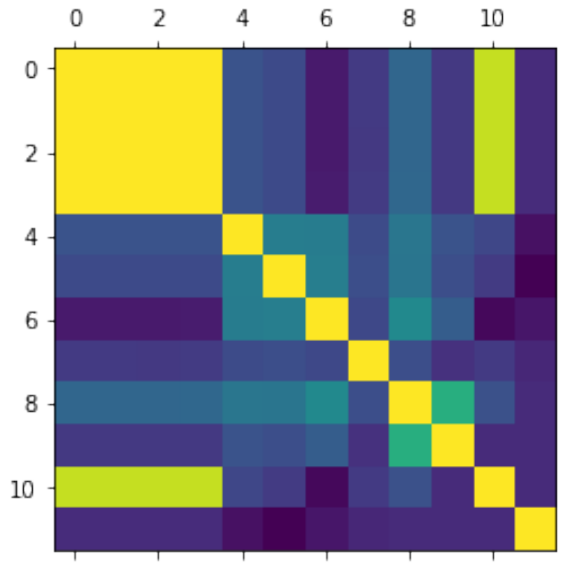


Figura 6

Ma questo fenomeno non deve certamente scoraggiare proprio perché tutto lo studio delle serie temporali tratta la naturale collinearità, che va sotto il nome di autocorrelazione. Quindi si è proceduto con la creazione di una finestra temporale (novembre) che rappresenti lo stato corrente del sistema dando piena osservabilità del fenomeno in una visione statica che “prescinde” dal fenomeno temporale; questo permette di ignorare cautamente la collinearità.

Le variabili che paiono in prima istanza in numero esiguo sono state aggiunte a nuovi attributi ottenuti tramite feature extraction. Questo è permesso dal fatto che la variabile tempo, è scomponibile nelle sue molteplici dimensioni che assumono un ruolo come variabili input. Si sono mantenute in analisi le seguenti features: 'open', 'close', 'low', 'high', 'neg', 'neu', 'pos', 'volume','tps', 'time_hour' ,'time_day', ‘time_min'.

E si presenta di seguito il risultato dell’applicazione dei seguenti modelli:

“LogisticRegression”,
"DecisionTreeClassifier",
"KNeighborsClassifier",
"LinearDiscriminantAnalysis",
"GaussianNB", "SVC" e un loro breve commento.

UnderSampling Table

La tecnica consiste nell'eseguire un campionamento casualmente dalla classe abbondate (“sentiment positiva”) di un numero pari alla cardinalità della classe rara in modo da livellare lo sbilanciamento e risolvere il class imbalance problem. Si riscontra un generale livello accettabile di accuracy per quanto ci siano casi sparsi di leggero overfitting e una casistica di lieve underfitting per il modello LinearDiscriminantAnalysis.

MODELLO	Accuracy on Training	Accuracy on Test
LogisticRegression	0,7559	0,738
DecisionTreeClassifier	1,0	0,8524
KNeighborsClassifier	0,9275	0,7557
LinearDiscriminantAnalysis	0,7376	0,7425
GaussianNB	0,6101	0,5454

MODELLO	Accuracy on Training	Accuracy on Test
SVC	0,7521	0,7196

OverSampling Table

È stato effettuato un tentativo agguntivo con il ricampionamento casuale dalla classe rara (“sentiment negativa”) con l'obiettivo di pareggiare la numerosità di osservazioni appartenenti alla classe complementare. Si riscontra un generale livello accettabile di accuracy per quanto ci siano casi di leggero underfitting. Il modello KNeighborsClassifier è il solo che presenta un risultato decisamente poco accettabile (estremo overfitting), ma ciò è ben spiegabile dalla forzatura introdotta: il modello ha la nomea di modello “pigro”, si concentra esclusivamente sul training per poi essere fallace nel test (in un conteso di ampia numerosità come quello indotto dall’oversampling).

MODELLO	Accuracy on Training	Accuracy on Test
LogisticRegression	0,7376	0,8977
DecisionTreeClassifier	1,0	0,8945
KNeighborsClassifier	0,7555	0,1024
LinearDiscriminantAnalysis	0,7323	0,8968
GaussianNB	0,6033	0,8976
SVC	0,7374	0,8976

Questo modello può essere un grande aiuto per finalità di marketing (nel business del digital advertising legato alle crypto): prevedere l’umore dei clienti significa prevedere come, quanto e quando spenderanno le loro risorse. Gli autori si riservano di aver compiuto banalizzazioni estreme nel condurre questa analisi di Machine Learning, che possono aver determinato errori e confidano di poter ampliare la loro conoscenza in tema di serie storiche in un momento prossimo.

Conclusioni

Si è osservato come l'umore del popolo di twitter sia estremamente volubile, poco deterministico e molto scostante ma come sia possibile estrarre degli small patterns. Nell'abstract ci si è chiesto se è vero che nella stanza dei mercati entrano i sentimenti, la vicenda pare ora essere sdoganata: dalla stanza dei mercati sicuramente i sentimenti escono e copiosamente! Gli autori di questo paper si rendono conto di aver commesso errori in quanto implementato e riportato, una analisi di maggiore impatto avrebbe richiesto l'adozione di tecnologie real-time come Apache-Spark o Apache-Storm, avrebbe richiesto anche una maggiore conoscenza di modelli temporali, migliorie negli algoritmi di comprensione del linguaggio e una maggiore conoscenza del machine learning per il dominio specifico.

Confidando nell’aver fatto un lavoro impegnato e nel massimo dello sforzo, gli autori si riservano di poter un giorno migliorare il work flow e l’analisi.

Note per la configurazione

Prima di poter testare l'applicativo è necessario impostare alcuni parametri nel file Utilities.py:

1. *is_realtime*: True per eseguire la versione realtime; False altrimenti
2. *analysis_thread*[min=1; max=4] : Seleziona il numero di thread su cui dovrà essere effettuata l'analisi

Scraping Realtime:

3. *chrome_binary_location*: Percorso al file binary di Chrome
4. *chromedriver_path*: Percorso ai chromedriver

Scraping Non Realtime - Dati Storici:

5. *path_to_btc_csv*: Percorso alla tabella csv BTC
6. *path_to_tweet_csv*: Percorso alla tabella csv TWEET
7. *path_to_tps_csv*: Percorso alla tabella csv TPS

   Nota: i percorsi nel caso in cui le tabelle sono posizionate nella cartella Producer sono già impostati.

Istruzioni per la configurazione

Si riportano le istruzioni per la configurazione di pycharm.

1. Una volta aperto il progetto configurare l'interprete e scaricare le librerie
2. Bisogna ora aggiungere le configurazioni per l'esecuzione dei file ****Script.py*
3. Dovrebbero risultare quindi quattro configurazioni da poter eseguire:
   - DataProcessScript.py
   - DataStorageScript.py
   - ProducerHD.py (Producer Hist. Data)
   - RealTimeProducerScript.py
4. Una volta avviati i servizi di Apache e Mongo, l'avvio dell'applicativo dovrebbe avvenire come segue:
   1. DataStorageScript
   2. ProducerHD o RealTime
   3. DataProcessScript

Sitografia

I. https://twitter.com
II. https://coinmarketcap.com/currencies/bitcoin/
III. https://blockchair.com/bitcoin
IV. https://BlockChain.com

Ringraziamenti

Si ringraziano i seguenti docenti per il supporto ricevuto e i preziosi consigli in corso d'opera e in fase d'analisi: prof. Andrea Maurino, prof. Roberto Avogadro, prof. Andrea Seveso, prof Fabio Antonio Stella.

Si ringraziano Alexander Stedtfeld e Gabriel Stedtfeld, laureati in Economia che hanno fornito una preziosa consulenza per la criticità dell'integrazione delle fonti e un approccio pragmatico con le questioni concettuali.