UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

Dipartimento di Scienze Fisiche, Informatiche e Matematiche

Corso di Laurea in Informatica

Analisi sull'Utilizzo di Twitter per la Promozione di Mostre d'Arte

Babal Preet Kaur

Tesi di Laurea

Relatore:
Prof. Riccardo Martoglia

Anno Accademico 2016/2017

Parole chiavi

Hashtag

Twitter

Promozione

API Twitter

Tweepy

Indice

Introduzione **1**
Capitolo 1 ..... **4**
1 Obiettivo 4
1.1 Problema iniziale 4
1.2 Metodi adottati 6
1.3 Svolgimento 7
Capitolo 2 ..... **9**
2 Mezzi utilizzati 9
2.1 Twitter Streaming API 9
2.1.1 Twitter API keys 10
2.1.2 API function 11
2.1.3 Status 12
2.2 Tweepy 16
Capitolo 3 ..... **19**
3 Raccolta e manipolazione dei dati 19
3.1 Stabilire un collegamento 19
3.2 Streaming dati 22
3.3 Dataset finale 22
Capitoli 4 ..... **25**
4 Analisi dei dati 25
4.1 Contenuto del messaggio 25
4.1.1 Spam 29
4.1.2 Sentiment 30
4.2 Andamento temporale 32
4.3 Influenza 34
4.4 Popolarità 36
Conclusione e sviluppi futuri **38**
Bibliografia **40**

Indice figure

Figura 2.1 Formato tweet visualizzato su Twitter..... 12
Figura 2.2 Tweet dopo la trasformazione nel formato JSON 13
Figura 2.3 Oggetto USER e i possibili campi 15
Figura 2.4 : Oggetto PLACE e i possibili campi 15
Figura 2.5 Oggetto COORDINATES e i possibili campi 16
Figura 2.6 Oggetto ENTITIES e i possibili campi..... 16
Figura 3.1 I moduli importati e le credenziali d'accesso 19
Figura 3.2 Definizione del listener e delle sue azioni 20
Figura 3.3 codice per autenticarsi, connettersi e raccogliere i tweet filtrati. 21
Figura 3.4 Codice per trasforare il dataset 23
Figura 3.5 Risultato finale del tweet 24
Figura 4.1 Diffusione del messaggio con differenti caratteristiche per la mostra di Monet 27
Figura 4.2 Diffusione del messaggio con differenti caratteristiche per la mostra di Picasso..... 278
Figura 4.3 Diffusione del messaggio con differenti caratteristiche per la mostra di Klimt 278
Figura 4.4 Grafico sulla distribuzione dello spam nel dataset 29
Figura 4.5 Grafico sulla distribuzione del sentiment nel dataset 30
Figura 4.6 Grafico sulla distribuzione del tipo di sentimento nel dataset..... 31
Figura 4.7 Grafico che mostra i retweet in funzione dell'intervallo orario 32
Figura 4.8 Grafico sul numero dei retweet per fasce orarie..... 33
Figura 4.9 Grafico sul numero dei follower per fasce orarie 34
Figura 4.10 Grafico dell'influenza rispetto al sentimento per ogni mostra 35
Figura 4.11 Grafico sull'numero dei follower raggiuti da un tweet nelle varie ore della giornata..... 36

Introduzione

La globalizzazione ha reso il mondo una comunità globale, che è più vogliosa di conoscere città e culture straniere. In ogni periodo dell'anno, si possono incontrare turisti di diverse nazionalità, in città grandi o piccole, desiderose di ampliare la propria conoscenza.

Una meta comune per tutti i turisti sono i musei, fonte principale di arricchimento culturale. Al giorno d'oggi si possono visitare nei musei quasi ogni genere di mostre. Esse possono interessare un gran numero di persone con interessi e provenienze differenti. Queste mostre temporanee hanno durate che vanno da qualche settimana a un paio di mesi e presentano per i musei un costo elevato. Per riuscire a sostenere gli eccessivi costi e raggiungere un pubblico vasto i tradizionali media non sono ottimali.

Negli ultimi anni con l'aumento dei social media, la promozione di eventi si è focalizzata sul migliorare l'uso di questi mezzi per la pubblicità. Uno dei metodi più usati è associare un hashtag specifico ad un evento. In questo modo si potrà parlare di un evento includendo quel tag nel testo. Questo utilizzo di specifici hashtag attira l'attenzione della gente, che per capire il motivo dell'utilizzo andrà a vedere a cosa è associato ed eventualmente si informerà sull'evento. Questo meccanismo viene anche utilizzato per la promozione di mostre. Consentendo ai musei di risparmiare sul costo della pubblicizzazione e raggiungere anche persone fuori dal territorio nazionale.

Non solo l'uso di hashtag su questi social media permette una promozione ottimale. Sono importanti anche altri fattori per massimizzare la promozione su queste piattaforme. La fascia oraria in cui viene pubblicato un tweet può diventare decisiva sulla sua diffusione. Se un tweet venisse pubblicato in un periodo quando il flusso di utenti è basso, solo pochi lo vedrebbero e lo condividerebbero, mentre se lo stesso tweet venisse pubblicato quado il flusso di utenti sul social è alto, quel tweet raggiugerebbe molti più visitatori e avrebbe una diffusione maggiore. Molti networking site limitano il numero di caratteri che gli utenti possono pubblicare. Per questo il contenuto del testo che viene pubblicato è molto importante. Il messaggio può contenere parole di apprezzamento o disgusto, è importante sapere come le persone reagiscono di fronte a certi tipi di messaggi. Questi tweet possono essere condivisi alimentando il sentimento, di approvazione o di dissenso, tra gli altri utenti.

In questa tesi verificheremo se l'utilizzo di Twitter è efficiente per la promozione di una mostra. Quanto è apprezzato l'utilizzo di hashtag nei tweet. Se l'orario di pubblicazione va ad incidere sulla sua diffusione. Inoltre quanto il contenuto del tweet sia importante nella sua diffusione.

La tesi è strutturata su due parti principali:

- Approccio al problema:
  - Capitolo 1 – *Obiettivo*, nel quale viene descritto l'approccio da noi utilizzato per la risoluzione del problema;
  - Capitolo 2 – *Mezzi utilizzati*, vengono descritti gli strumenti utilizzati per il raccoglimento dei dati.
- Risoluzione:
  - Capitolo 3 - *Raccolta e manipolazione dei dati*, come sono ottenuti i dati dalle API di Twitter e i filtri applicati;
  - Capitolo 4 - *Analisi dei dati*, spiegazione delle analisi fatte e i risultati ottenuti.

Per ultimo si trovano le conclusioni sul lavoro svolto e possibili sviluppi futuri.

Parte I

Approccio al problema

Capitolo 1

1 Obiettivo

In questo capitolo viene descritto l'approccio da noi utilizzato per la risoluzione del problema. Quali sono stati i motivi per cui si è scelto di affrontare questo problema in un certo modo rispetto ad altri.

1.1 Problema iniziale

I musei sono da sempre stati una delle mete preferite dai turisti., che li visitano in tutte le stagioni. Una volta i musei ospitavano mostre, che duravano per lunghi periodi di tempo o c'erano allestimenti, che rimanevano fissi nelle strutture, a tal punto da etichettare in quella determinata specializzazione. Oggi invece, pur rimanendo fedeli a una determinata tipologia, cercano di variare gli eventi offerti al proprio pubblico. Per attirare più gente dai gusti differenti, cercano di ospitare tante mostre che vanno da una durata di un paio di settimane a qualche mese. Questo implica che la stessa mostra potrà fare il giro di tanti paesi, e si potrà far conoscere da una mole di persone più elevata.

Poiché l'obiettivo è quello di rivolgersi a un vasto genere di visitatori di differenti origini, i tradizionali metodi di promozione, ovvero al radio, i giornali, i volantini ecc. non sono sufficienti. Si deve valutare la possibilità di trovare nuovi metodi che riescano a raggiungere persone di diverse etnie.

Negli ultimi anni gli organizzatori di eventi, si sono spostati su piattaforme virtuali per promuovere eventi, mediante l'uso di siti web e social network. A differenza dei media tradizionali che sono molto costosi e si limitano a un pubblico ristretto, i media virtuali non hanno costi elevati. Basta creare un sito o una pagina dell'evento su un qualsiasi sito di networking.

Con il boom dei social media le persone hanno preso l'abitudine di interagire l'una con l'altra virtualmente. Essi utilizzano vari metodi, dalla condivisione di un semplice messaggio testuale a contenuti multimediali come foto, video immagini digitali ecc... In tutto questo, un ruolo molto importante, è quello dell'hashtag. Questo è un tipo di etichetta (tag) utilizzato su alcuni servizi web e social network come aggregatore tematico, la cui funzione è di rendere più facile per gli utenti trovare messaggi su un tema o contenuto specifico [1]. Dare una o più parole chiavi per identificare un certo evento o movimento può essere determinante per la sua popolarità. Più una parola sarà attraente (catchy) e sintetizzante, più le persone partecipanti saranno spinte a utilizzarlo, diffondendolo tra altri, che a loro volta, estrani all'evento, agiranno nello stesso modo.

I promotori di mostre prendono in considerazione soprattutto questo aspetto dei social media e ne sfruttano l'effetto a loro vantaggioso, cioè utilizzando un hashtag per promuovere una determinata mostra presente nel loro museo. Il basso costo della promozione online riesce a compensare l'elevato costo dei continui lavori dovuti alle mostre temporanee. Inoltre il target di persone a cui il messaggio promozionale arriva è elevato.

Gli hashtag non sono solo l'unico fattore che incide sulla popolarità di un post. Ci sono diverse variabili, che fanno aumentare il numero di condivisione o il numero delle volte in cui è piaciuto un messaggio. Per esempio il contenuto può essere assai importante. Il discorso all'interno del post può voler condividere sentimenti positivi o sentimenti negativi. Sia un testo positivo, in quanto apprezza l'oggetto del messaggio, sia un testo negativo, che lo critica invece, possono essere visti e diffusi, tra i seguaci dell'utente originario, e rispecchiare positivamente o negativamente

quel oggetto. Altro fattore incidente sono i contenuti multimediali, che possono essere utilizzati per aumentare la diffusione del messaggio. D'altro canto, se la presenza di tali contenuti è eccessiva può anche ottenere soltanto l'effetto contrario

La vita media di un messaggio pubblicato su un social è di circa un giorno, ovviamente le prime ore saranno quelle decisive per la sua diffusione. Quindi sarà importante analizzare l'esistenza di determinati periodi di tempo durante la giornata in cui è conveniente pubblicare il proprio messaggio rispetto ad altri.

In questa tesi si parlerà di mostre d'arte temporanee. Si è cercato di capire se l'utilizzo di un hashtag, per una mostra d'arte, è utile alla sua promozione online. Una semplice parola anticipata da un cancelletto (#) riesce a suscitare interesse tra le persone che lo vede. Per diffondere il messaggio tra più utenti è utile capire se ci sono fasce orarie con maggiore o minore affluenza. Si cerca inoltre di studiare se la pubblicazione di un messaggio in un certo orario favorisce la sua popolarità. Inoltre si fanno anche alcune analisi sui diversi tipi di contenuti, e in che modo questi vadano ad interferire sulla diffusione di un messaggio riguardante una mostra. Di conseguenza si valuta se utilizzare un certo tipo di messaggi vada ad aiutare la promozione della mostra stessa.

1.2 Metodi adottati

Negli ultimi anni l'aumento dei cosiddetti social media è stato notevole. Esistono vari tipi di networking site, che permettono di fare cose diverse. Alcuni danno la possibilità di condividere foto, altri articoli, altri piccoli pensieri. Si può dire che c'è un social (termine comunemente usato per descrivere un servizio di rete sociale) per ogni gusto. Molti di questi servizi mettono a disposizione dei propri utenti la cosiddetta API, Application Programming Interface, uno strumento che rende l'interazione tra programmi e web service molto facile. Molte web company offrono API ai developer per interagire con i loro servizi ed accedere dati in modo programmatico [2]. Nella maggior parte dei casi l'accesso a queste API è a

pagamento. Twitter, uno dei social network più popolari attraverso il quale le persone comunicano tramite brevi messaggi chiamati tweet [3], offre un servizio di API in parte gratuito. Si possono ricevere alcuni servizi completamente gratis, mentre per altri è richiesto un abbonamento.

È stato utilizzato Twitter per capire se l'uso dei social media, tramite l'uso di hashtag, pubblicazione di determinati tipi di contesto in determinati orari, è utile per creare buzz tra il pubblico per una determinata mostra. Ogni giorno su Twitter vengono pubblicati centinaia di migliaia di tweet, contenenti argomenti di ogni genere. L' obbiettivo della ricerca è stato quello di determinare quelli più significativi per la ricerca stessa e filtrarli dal resto. Per determinare i tweet più significativo sono stati utilizzati diversi mezzi. La API, messe a disposizione dal social network, permette di ricevere i tweet pubblicati, filtrando quelli richiesti. I tweet filtrati saranno poi a loro volta soggetti ad altri filtri, in modo tale da poter estrapolare informazioni riguardanti la popolarità della mostra.

L'approccio utilizzato è quello di raccogliere i tweet con determinato hashtag, usato dai promoter per pubblicizzare una mostra e procedere a delle statistiche relative alla mostra in questione.

1.3 Svolgimento

I dati raccolti tramite API Twitter per ogni tweet contengono diverse informazioni su esso e sull'utente che lo ha pubblicato. Nel caso di retweet si possono anche trovare informazioni riguardanti l'utente che ha pubblicato il tweet originale.

Sui dati ottenuti sono state fatte diverse considerazioni. Questi dati contengono informazioni sull'andamento del tweet, ovvero quanto un determinato tweet sia piaciuto ai suoi visitatori (quanti favourite ha avuto) o quanto questo tweet si sia diffuso tra gli utenti di Twitter (quanti retweet). Combinando questi dati in modi appropriati si può fare un'assunzione sulla popolarità e riuscire a dire se l'uso di hashtag è utile alla promozione di una mostra. L'orario di pubblicazione del tweet

originale e del retweet sarà utile per determinare la fascia orari esatta in cui la gente tende a retwittare più velocemente. Inoltre si può scoprire per quanto tempo in media i tweet raccolti sono stati popolari.

Applicando un filtro che sul testo del tweet si può scoprire che tipo di parole contiene, cioè se sono positive o negative. Ottenendo il sentimento del testo e i dati sulla sua diffusione si può capire se per una determinata mostra si è diffusa un'opinione favorevole o contraria.

Capitolo 2

2 Mezzi utilizzati

In questo capitolo vengono descritti gli strumenti utilizzati per il raccoglimento e la successiva analisi, dei dati relativi ai tweet. Si vedrà il modo in cui i tweet vengono raccolti e la loro forma finale che si andrà ad analizzare. In particolare i diversi passaggi, che bisogna seguire per far si che un tweet diventi da quello che si vede sullo schermo quando si visita il sito di Twitter, a quello che viene messo nel dataset che si andrà ad analizzare.

2.1 Twitter Streaming API

Alla API, che Twitter mette a disposizione dei propri utenti, non tutti gli account possono connettersi. Solamente gli utenti che si identificano come developer possono utilizzare tale servizio. L'accesso a questa API è consentito solo esclusivamente se si è in possesso delle cosiddette Twitter API keys: esse sono quattro e servono per autenticarsi come developer di Twitter, si vedrà nel capitolo 2.1.1 come ottenerle. Ci sono diversi passaggi che bisogna seguire per ottenere da Twitter i dati necessari per fare analisi utili. Nei capitoli seguenti andremo a vedere in dettaglio cosa offre l'interfaccia di Twitter ai suoi utenti e quali informazioni si possono ricevere da lui e il loro formato.

2.1.1 Twitter API keys

Twitter API keys sono le chiavi necessarie per ricevere dati da Twitter. Queste chiavi permettono ad un utente di identificarsi come developer ed accedere alla Twitter Streaming API (o API di Twitter), fondamentali per ricevere i tweet richiesti. Questi ultimi sono in uno specifico formato offerto da Twitter. Questo formato, in cui vengono ricevuti i tweet, si chiama JSON (JavaScript Object Notation), un semplice formato per lo scambio di dati. Per le persone è facile da leggere e scrivere, mentre per le macchine risulta facile da generare e analizzarne la sintassi [4]. Nel capitolo 2.1.2 verrà spiegato in modo dettagliato i dati ottenuti e il significato dei vari campi presenti.

Le quattro chiavi necessarie per la connessione sono: API key, API secret, Access token e Access token secret. Per ottenere le Twitter API keys, bisogna:

- avere un account Twitter o crearne uno nuovo;
- andare sul sito https://apps.twitter.com/ e accedere con le proprie credenziali Twitter;
- creare una nuova applicazione Twitter.

Ad ogni applicazione Twitter vengono assegnati le chiavi e i token necessari per effettuare uno streaming di dati. Queste chiavi sono personali, identificano i cosiddetti developer di Twitter. Quando si crea un'applicazione, a questa vengono assegnati automaticamente API key e API secret, che vengono chiamate API keys. Mentre, quelli che sono definiti come, i token d'accesso, non vengono assegnati automaticamente. Per ottenere Access token e Access token secret si deve associare i token all' applicazione. Solo ottenendo tutte e quattro le chiavi si hanno le credenziali per ottenere una connessione a API Twitter, dal quale si possono raccogliere in seguito i dati relativi ai tweet.

2.1.2 API function

Le API di Twitter permettono di fare moltissime operazioni sui tweet. Si possono ottenere dati in due modi: tramite un account gratuito o effettuando un abbonamento. Ovviamente con un abbonamento si hanno dei privilegi sulla raccolta dei dati. In seguito si possono vedere quali sono alcune delle operazioni che si possono fare, per l'uno o l'altro metodo [5]:

- Ricevere un campione di tweet in tempo reale, per fare brevi analisi.
- Ricercare tweet, fra quelli degli ultimi 7 giorni con un account normale o fino a tweet datati 2006 pagando un abbonamento. In questo modo si ricevono tweet che hanno già avuto una esistenza e si possono ricavare informazioni più dettagliate sul loro andamento.
- Filtrate tweet in tempo reale, metodo utilizzato per ricavare il dataset utilizzato in questa tesi. Questo metodo presenta alcuni parametri, tra i quali:
    - `Name` (parametro richiesto) contiene la descrizione del tweet
    - `Follow` (parametro opzionale): una lista di ID degli user dei quali si vogliono ricevere tweet. Questi possono essere tweet creati dallo user, i retweet dello user, i commenti ai tweet creati dello user, i retweet ai tweet creati dello user. Non conterrà tweet in cui viene menzionato o citato lo user.
    - `Track` (parametro opzionale): lista delle parole che dovranno essere contenute nel tweet. Basta che una sola parola della lista sia presente nel tweet, per ricevere il suo status.
    - `Language` (parametro opzionale): lista delle lingue in cui devono essere i tweet.

2.1.3 Status

Ogni tweet ricevuto è praticamente un blocco di oggetti presenti su Twitter. I dati relativi ai tweet ricevuti sono anche chiamati status. Ogni status è composto da una lista di attributi “root-level” e attributi “parent” i quali hanno attributi “child”. Esempi di attributi “root-level” sono `id`, `created_at` e `text`, invece esempi di attributi “child” possono essere `user`, `entities` e `place`.

Si può notare come un tweet tradizionale, che si vede normalmente su www.twitter.com, Figura 2.1, quando viene interpretato da JSON diventa un insieme di caratteri, Figura 2.2.

Figura 2.1 Formato tweet visualizzato su Twitter

```
{"created_at":"Fri Mar 23 14:08:51 +0000 2018","id":977185409129066496,"id_str":"977185409129066496","text":"RT @ViennaInfoB2B: The
@belvederemuseum shows \"Vik Muniz. Verso\", the first solo #exhibition of Vik Muniz in Vienna. The artist has
dedicat\u2026","source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e","truncated":
false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_re
ply_to_screen_name":null,"user":{"id":4496852834,"id_str":"4496852834","name":"\u30d1\u30ea\u30b8\u30e
3\u30f3\u30cc\u30c0\u30eb\u30af","screen_name":"brnejaen6","location":"\u65e5\u672c \u5bcc\u5c71\u770c","url":null,"description":"\u
30c4\u30a4\u30c3\u30bf\u30fc\u521d\u5fc3\u8005\u3067\u3059\u3002\u30e8\uff0d\u30ed\u30c3\u30d1\u306e\u4e16\u754c\u907a\u7523\u3084\u
82b8\u8853\u3001\u30b5\u30c3\u30ab\u30fc\u3001\u6d77\u5916&\u56fd\u5185\u65c5\u884c\u304c\u4e3b\u306a\u8da3\u5473\u3067\u3059\u3002\
u3088\u308d\u3057\u304f\u304a\u9858\u3044\u3057\u307e\u3059!","translator_type":"none","protected":false,"verified":false,"followers
_count":120,"friends_count":139,"listed_count":0,"favourites_count":4564,"statuses_count":861,"created_at":"Tue Dec 08 12:55:00
+0000 2015","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":"ja","contributors_enabled":false,"is_translator":false,"p
rofile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/
bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":fal
se,"profile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":
"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/
profile_images\/675673959178756097\/2Tu3bBaa_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/6756739
59178756097\/2Tu3bBaa_normal.jpg","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"
notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"Fri Mar 23
13:44:02 +0000 2018","id":977179165412732928,"id_str":"977179165412732928","text":"The @belvederemuseum shows \"Vik Muniz. Verso\",
the first solo #exhibition of Vik Muniz in Vienna. The artist has de\u2026 https:\/\/t.co\/ERMpZkfjlb","display_text_range":[
0,140],"source":"\u003ca href=\"http:\/\/www.falcon.io\" rel=\"nofollow\"\u003eFalcon Social Media Management \u003c\/a\u003e","trun
cated":true,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,
"in_reply_to_screen_name":null,"user":{"id":190237407,"id_str":"190237407","name":"ViennaTouristBoard","scree
n_name":"ViennaInfoB2B","location":"Vienna","url":"http:\/\/b2b.vienna.info","description":"Official Page of the Vienna Tourist
Board. The Market & Media Management team tweets for you.","translator_type":"none","protected":false,"verified":false,"followers_co
unt":11506,"friends_count":2786,"listed_count":321,"favourites_count":3312,"statuses_count":5928,"created_at":"Mon Sep 13 12:56:06
+0000 2010","utc_offset":3600,"time_zone":"Vienna","geo_enabled":true,"lang":"de","contributors_enabled":false,"is_translator":false
,"profile_background_color":"FFFFFF","profile_background_image_url":"http:\/\/pbs.twimg.com\/
profile_background_images\/453479460713402368\/3w3ZJWMm.jpeg","profile_background_image_url_https":"https:\/\/pbs.twimg.com\/profile
_background_images\/453479460713402368\/3w3ZJWMm.jpeg","profile_background_tile":false,"profile_link_color":"E02222","profile_sideba
r_border_color":"FFFFFF","profile_sidebar_fill_color":"FFFFFF","profile_text_color":"0D0202","profile_use_background
_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/921376522426085377\/
RsK7tG5y_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/921376522426085377\/
RsK7tG5y_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/190237407\/1520425124","default_profile":false,
"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place
":null,"contributors":null,"is_quote_status":false,"extended_tweet":{"full_text":"The @belvederemuseum shows \"Vik Muniz. Verso\",
the first solo #exhibition of Vik Muniz in Vienna. The artist has dedicated himself to investigating the reverse sides of famous
#paintings, among them works of #Gustav Klimt &amp; #EgonSchiele: https:\/\/t.co\/WHorsg0q3K \u00a9 H.Hurnaus https:\/\/t.co\/
EAongNwydQ","display_text_range":[0,277],"entities":{"hashtags":[{"text":"exhibition","indices":[62,73]},{"text":"paintings","indice
s":[176,186]},{"text":"Gustav","indices":[208,215]},{"text":"EgonSchiele","indices":[228,240]}],"urls":[{"url":"https:\/\/t.co\/
WHorsg0q3K","expanded_url":"http:\/\/fal.cn\/49ht","display_url":"fal.cn\/49ht","indices":[242,265]}],"user_mentions":[{"screen_name
":"belvederemuseum","name":"Belvedere Museum","id":28537619,"id_str":"28537619","indices":[4,20]}],"symbols":[],"media":[{"id":97717
9160912236544,"id_str":"977179160912236544","indices":[278,301],"media_url":"http:\/\/pbs.twimg.com\/media\/
DY-jRQAWkAAmxGt.jpg","media_url_https":"https:\/\/pbs.twimg.com\/media\/DY-jRQAWkAAmxGt.jpg","url":"https:\/\/t.co\/
EAongNwydQ","display_url":"pic.twitter.com\/EAongNwydQ","expanded_url":"https:\/\/twitter.com\/ViennaInfoB2B\/
status\/977179165412732928\/photo\/1","type":"photo","sizes":{"thumb":{"w":150,"h":150,"resize":"crop"},"small":{"w":600,"h":399,"re
size":"fit"},"large":{"w":600,"h":399,"resize":"fit"},"medium":{"w":600,"h":399,"resize":"fit"}}}]},"extended_entities":{"media":[{"
id":977179160912236544,"id_str":"977179160912236544","indices":[278,301],"media_url":"http:\/\/pbs.twimg.com\/media\/
DY-jRQAWkAAmxGt.jpg","media_url_https":"https:\/\/pbs.twimg.com\/media\/DY-jRQAWkAAmxGt.jpg","url":"https:\/\/t.co\/
EAongNwydQ","display_url":"pic.twitter.com\/EAongNwydQ","expanded_url":"https:\/\/twitter.com\/ViennaInfoB2B\/
status\/977179165412732928\/photo\/1","type":"photo","sizes":{"thumb":{"w":150,"h":150,"resize":"crop"},"small":{"w":600,"h":399,"re
size":"fit"},"large":{"w":600,"h":399,"resize":"fit"},"medium":{"w":600,"h":399,"resize":"fit"}}}]}},"quote_count":0,"reply_count":0
,"retweet_count":3,"favorite_count":3,"entities":{"hashtags":[{"text":"exhibition","indices":[62,73]}],"urls":[{"url":"https:\/\/
t.co\/ERMpZkfjlb","expanded_url":"https:\/\/twitter.com\/i\/web\/status\/977179165412732928","display_url":"twitter.com\/i\/web\/
status\/9\u2026","indices":[117,140]}],"user_mentions":[{"screen_name":"belvederemuseum","name":"Belvedere Museum","id":28537619,"id
_str":"28537619","indices":[4,20]}],"symbols":[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low
","lang":"en"},"is_quote_status":false,"quote_count":0,"reply_count":0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[
{"text":"exhibition","indices":[81,92]}],"urls":[],"user_mentions":[{"screen_name":"ViennaInfoB2B","name":"ViennaTouristBoard","id":
190237407,"id_str":"190237407","indices":[3,17]},{"screen_name":"belvederemuseum","name":"Belvedere Museum","id":28537619,"id_str":"
28537619","indices":[23,39]}],"symbols":[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"en","timestamp_ms":"152
1814131420"}
```


Figura 2.2 Tweet dopo la trasformazione nel formato JSON

In seguito sono spiegate in dettaglio cosa vogliono dire alcuni dei campi che si possono osservare nella Figura 2.2:

`created_at`	string	contiene l'orario di quando è stato pubblicato il tweet, settato automaticamente nel fuso orario UTC
`id`	int (64)	rappresenta l'identificatore univoco del tweet
`id_string`	string	rappresentazione in formato string del ID del tweet
`text`	string	testo del tweet

`user`	*	dati dell'utente che ha postato il tweet
`coordinates`	**	può essere nullo, rappresenta la posizione geografica da dove è stato spedito il tweet
`place`	**	può essere nullo, rappresenta il posto associato a quel tweet, non è necessario che il tweet sia stato postato in quel posto
`retweet_count`	int	numero di volte il tweet originale è stato retwittato
`favorite_count`	int	può essere nullo, indica quante volte ai Twitter user è piaciuto quel tweet
`entities`	***	entità contenute nel testo tweet
`favorited`	boolean	può essere nullo, il tweet è piaciuto a un utente autenticato
`retweeted`	boolean	il tweet è stato retwittato almeno una volta da un utente autenticato
`lang`	string	può essere nullo, indica la lingua in cui è stato pubblicato il tweet, viene rilevata automaticamente da Twitter valutando il testo del tweet

```
{
  "user": {
    "id": 2244994945,
    "id_str": "2244994945",
    "name": "TwitterDev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https://dev.twitter.com/",
    "description": "Your official source for Twitter news",
    "verified": true,
    "followers_count": 477684,
    "friends_count": 1524,
    "listed_count": 1184,
    "favourites_count": 2151,
    "statuses_count": 3121,
    "created_at": "Sat Dec 14 04:35:55 +0000 2013",
    "utc_offset": -25200,
    "time_zone": "Pacific Time (US & Canada)",
    "geo_enabled": true,
    "lang": "en",
    "profile_image_url_https": "https://pbs.twimg.com/"
  }
}
```


Figura 2.3 Oggetto USER e i possibili campi

* `user` contiene i metadati pubblici relativi all'account Twitter e descrive i tweet. Gli user possono essere chiunque e qualsiasi cosa. Sono appunto le entità che possono: twittare, retwittare, citare tweet, seguire altri user, creare liste, avere una timeline e essere menzionati.

In caso di retweet o citazioni di tweet, lo user rappresentato è quello top-level, ovvero l'utente che ha retwittato o citato il tweet. Sarà poi JSON ad includere dentro la lista degli attributi di user un secondo user che rappresenterà l'utente che ha twittato il post originale.


```
"place":
{
    "attributes":{},
     "bounding_box":
    {
        "coordinates":
        [[
                [-77.119759,38.791645],
                [-76.909393,38.791645],
                [-76.909393,38.995548],
                [-77.119759,38.995548]
        ]],
        "type":"Polygon"
    },
     "country":"United States",
     "country_code":"US",
     "full_name":"Washington, DC",
     "id":"01fbe706f872cb32",
     "name":"Washington",
     "place_type":"city",
     "url": "http://api.twitter.com/1/geo/id/0172cb32.json"
}
```


Figura 2.4 : Oggetto PLACE e i possibili campi

** `place` presente sempre in presenza di geo-tag, indica il nome del posto e il paese.

```
"coordinates":
{
    "coordinates":
    [
        -75.14310264,
        40.05701649
    ],
    "type":"Point"
}
```

Figura 2.5 Oggetto COORDINATES e i possibili campi

**coordinates presente solo se a un tweet viene assegnato un posto preciso, ed ha un valore che indica la longitudine e uno che indica la latitudine

```
"entities":
{
    "hashtags":[],
    "urls":[],
    "user_mentions":[],
    "media":[],
    "symbols":[]
    "polls":[]
}
```

Figura 2.6 Oggetto ENTITIES e i possibili campi

*** entities mostra il metadata e le informazioni sui contenuti presenti nel tweet. Esso è formato da una serie di attributi i quali, possono essere a loro volta degli array.

I più comuni sono:

- hashtags - lista dei hashtag contenuti nel testo del tweet
- urls – lista degli url presenti nel testo del tweet
- user_mention – citazione dei user nel tweet

2.2 Tweepy

Per procedere con la raccolta dei dati bisogna prima di tutto stabilire una connessione con Twitter Streaming API. È stata utilizzata una libreria Python per stabilire una connessione con le API. La libreria Python che si andrà ad utilizzare per questo scopo si chiama Tweepy (non è presente di default quindi bisogna installarla). Tweepy è una libreria open-source, ospitata su GitHub, e consente a Python di comunicare con la piattaforma Twitter e utilizzare la sua API [6]. L'accesso a Twitter viene effettuato tramite OAuth, un protocollo aperto, che permette l'autorizzazione di API di sicurezza con un metodo standard e semplice sia per applicazioni portatili, sia per applicazioni per pc fisso e web [7].

Alcune caratteristiche del metodo OAuth sul ricevimento dei dati da Twitter:

- Sicuro: non è necessario che gli utenti condividano le credenziali dell'account con applicazioni di terze parti.
- Standard: l'implementazione OAuth di Twitter è compatibile con moltissime librerie.

La libreria Tweepy permette di accedere al API Twitter. È possibile ottenere ogni oggetto e utilizzare ogni metodo che le API forniscono. Da noi Tweepy è stata utilizzata per monitorare i tweet in tempo reale e catturarne uno quando si verificava un determinato evento.

`StreamListener` è un oggetto di Tweepy, molto importante, che permette in tempo reale di leggere i tweet pubblicati e raccogliere quelli utili alla ricerca effettuara. `StreamListener` è un oggetto importante che ha tanti metodi. Uno dei metodi utilizzato da noi è `filter()`. Questo metodo permette di specificare i parametri visti nel capitolo 2.1.2, in modo da ricevere solamente i tweet che rispettano i parametri specificati.

Parte II

Risoluzione

Capitolo 3

3 Raccolta e manipolazione dei dati

Nei capitoli precedenti si è visto quali strumenti sono necessari per la realizzazione dell'obiettivo posto nel primo capitolo. In questo capitolo si andrò a vedere come sono stati utilizzati tali strumenti per ottenere e analizzare i dati.

3.1 Stabilire un collegamento

Come già accennato nel capitolo 2.2, Tweepy è una libreria Python, che tramite l'uso di OAuth permette l'autenticazione e una connessione con l'interfaccia di Twitter. Si vedrà in dettaglio come viene stabilita la connessione con l'API di Twitter.

```
#Import the necessary methods from tweepy library
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

#Variables that contains the user credentials to access Twitter API
consumer_key = "hMDOBrdmagxhZtauK7Tticilj"
consumer_secret = "FVPJGLoHeNY94Kj09KEj9PMdIHyAQhxsCoadnLKHR7xjlAqcSp"
access_token = "844847914170007552-u9bA2i3pYKc1QRXzwXnH0rup51L65FV"
access_token_secret = "StDOtp9BWN0VjOgx5yy9Eg5pOUkPxJhvPYT5pLlqnSSqu"
```

Figura 3.1 I moduli importati e le credenziali d'accesso

Nella Figura 3.1 (più sopra)più sopra come prima cosa si vedono i moduli che sono stati importi da Tweepy, per accedere e ricevere dati da API Twitter, questi moduli sono:

- `StreamListener`, modulo utilizzato per la recezione e la stampa i dati;
- `OAuthHandler`, adoperato per gestire il protocollo OAuth;
- `Stream`, impiegato per connetterci a Twitter streaming API.

In seguito si possono notare le variabili che contengono le credenziali per accedere a Twitter API. Queste credenziali, molto importanti, si possono ottenere dal sito https://apps.twitter.com/, seguendo le istruzioni nel capitolo 2.1.1. Le credenziali si dividono in due gruppi: API keys e OAuth access token. Le API keys sono API key e API secret, invece OAuth access token sono Access token e Access token secret. Queste 4 chiavi sono personali e sono associate a una sola applicazione Twitter (che viene creata accedendo al sito specificato prima). Insieme formano le credenziali per accedere all'interfaccia di Twitter.

```
file = open('today.txt', 'a')

#This is a basic listener that just save received tweets into a file.
class StdOutListener(StreamListener):

    def on_data(self, data):
        file.write(data)
        return True

    def on_error(self, status):
        print status
        return True

    def on_timeout(self):
        return True  # Don't kill the stream
```

Figura 3.2 Definizione del listener e delle sue azioni

Nella Figura 3.2 si può osservare la definizione della classe `Listener` che eredita `StreamListener`, un'istanza di tweepy.Stream. Un listener basico normalmente stampa i tweet ricevuti. La classe `Listener` stabilisce una sessione di streaming e indirizza i messaggi, tramite il metodo `on_data`, verso un file. Il modulo `on_error` individua gli errori che si possono presentare e ritorna True per evitare

che l'ascolto dei tweet si fermi dopo il verificarsi dell'errore. Il modulo `on_timeout` non fa terminare lo stream dopo la scadenza del tempo. Grazie a questo si possono ascoltare e raccogliere tweet per molti giorni consecutivi.

```
if __name__ == '__main__':

    #This handles Twitter authetification and the connection to Twitter Streaming API
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)

    #This line filter Twitter Streams to capture data by the keywords
    stream.filter(track=['#monet','#monetmilano', '#egonschiele', '#richardserra',
        '#kandinskij', '#monettopicasso', '#albertina', '#picasso', '#portraitscezanne',
        '#hokusai', '#theirmortalremains', '#ernstbeyeler', '#torneesquius', '#michelangelosebastiano',
        '#chrisofili', '#puntadelladogana', '#toulouselautrec'])
```

Figura 3.3 codice per autenticarsi, connettersi e raccogliere i tweet filtrati.

Il pezzo di codice presente nella Figura 3.3 permette l'autenticazione e la connessione allo streaming di API di Twitter. Il primo passo per procedere con l'autenticazione è passare al gestore OAuth le API keys tramite la seguente funzione: `OAuthHandler (consumer_key, consumer_secret)`. Per riuscire ad effettuare l'autenticazione completamente è fondamentale impostare i token OAuth al gestore, utilizzando `set_access_token(access_token, access_token_secret)`.

Per effettuare l'autenticazione è importante stabilire la connessione. Per fare ciò bisogna prima di tutto creare uno stream object. `Stream()`, uno stream object, è il metodo che abilita la connessione allo streaming di Twitter. Si passano allo `Stream()` come parametri il gestore OAuth contenente le credenziali e il `Listener`, definito prima.

Dopo aver stabilito la connessione il `Listener` è in grado di ricevere tutti i tweet che vengono pubblicati in tempo reale. Nel caso di studio non sono necessari tutti i tweet pubblicati, ma solo quelli contenenti le keyword: *#monet, #monetmilano, #egonschiele, #richardserra, #kandinskij, #monettopicasso, #albertina, #picasso,*

#portraitscezanne, #hokusai, #theirmortalremains, #ernstbeyeler, #torneesquius, #michelangelosebastiano, #chrisofili, #puntadelladogana, #toulouselautrec.

È stato inserito un filtro, che andrà a catturare solamente i tweet inerenti alla ricerca. Il metodo `filter(track= [lista hashtag])` crea il filtro desiderato sui tweet pubblicati in tempo reale, autorizzando il `Listener` a ricevere solamente quei tweet.

3.2 Streaming dati

Avviando lo script, descritto nel capitolo precedente, si è in grado di ricevere tutti i tweet che vengono pubblicati in tempo reale. È importante notare che in questo modo non si ottiene nessun tweet pubblicato in precedenza, ma bensì i possibili retweet che vengono pubblicati mentre si ascoltano i dati in stream. Si possono comunque ottenere dati che danno informazioni sul tweet originale e il suo andamento. Con i quali si possono tuttavia fare diverse analisi, ad esempi scoprire l'influenza che ha avuto il tweet.

3.3 Dataset finale

Si può notare, in Figura 2.2, che i dati raccolti per un tweet contengono informazioni aggiuntive, oltre al semplice testo. Molte di queste informazioni non riguardano direttamente il tweet, possono invece rivelarci dettagli sull'utente, oppure sono campi, che non forniscono informazioni o ancora nella maggior parte dei casi contengono informazioni nulle, per questo motivo non si possono fare statistiche complete poiché non sono presenti nella totalità dei status.

Per avere un dataset pulito e snello con solo informazioni che possono essere utili a capire l'andamento di un Tweet è stato deciso di tenere solamente alcuni campi dello status. Sono state fatte alcune considerazioni per favorire la scelta di un campo rispetto ad un altro.

Prima di tutto è stato considerato che tipo di statistiche si intende fare e poi quali dati fra quelli dello status sono utili per fare le analisi. Per passare dallo status completo ad arrivare ad ottenere solo determinati campi è stato utilizzato il seguente codice, presente in Figura 3.4.

```
# Import the necessary package to process data in JSON format
try:
    import json
except ImportError:
    import simplejson as json

# We use the file saved from last step as example
tweets_filename = 'today.txt'
tweets_file = open(tweets_filename, "r")
write_tweets = open('dataset.txt', "a")

for line in tweets_file:

        # Read in one line of the file, convert it into a json object
        tweet = json.loads(line.strip())

        if 'text' in tweet:  # only messages contains 'text' field is a tweet
            write_tweets.write("\nTextTW : %s\n"% {tweet['text']})
                                # This is the tweet's id
            write_tweets.write("Tweetid : %s\n" %{tweet['id']})
            write_tweets.write("Tweetid-str  : %s\n" % {tweet['id_str']})
            write_tweets.write("Username  : %s\n" % {tweet['user']['name']})
            write_tweets.write("Followers-count : %s\n" % {tweet['user']['followers_count']})
            write_tweets.write("Following-count : %s\n" % {tweet['user']['friends_count']})
            write_tweets.write("Favourite-count : %s\n" % {tweet['user']['favourites_count']})
            write_tweets.write("Created-at : %s\n" % {tweet['created_at']})
            write_tweets.write("Lang : %s\n" % {tweet['lang']})
            write_tweets.write("Favorited-count : %s\n" % {tweet['favorite_count']})
            write_tweets.write("Retweet-count : %s\n" % {tweet['retweet_count']})
        if 'retweeted_status' in tweet:
            write_tweets.write("Username-author : %s\n" % {tweet['retweeted_status']['user']['name']})
            write_tweets.write("Created-at-author : %s\n" % {tweet['retweeted_status']['created_at']})
            write_tweets.write("Screen-name-author : %s\n" % {tweet['retweeted_status']['user']['screen_name']})
            write_tweets.write("Followers-count-author : %s\n" % {tweet['retweeted_status']['user']['followers_count']})
            write_tweets.write("Following-count-author : %s\n"% {tweet['retweeted_status']['user']['friends_count']})
            write_tweets.write("Favourite-count-author : %s\n" % {tweet['user']['favourites_count']})
```

Figura 3.4 Codice per trasforare il dataset

Poiché i dati sono nel formato JSON, bisogna importare la libreria dedicata ed effettuare la lettura. Si inizia leggendo dal primo file, il quale contiene i dati raw ottenuti da API Twitter, e scrivendo nel secondo file solamente i dati utili alle analisi. Questo passaggio serve per ridurre di dimensioni il dataset tenendo solo i dati strettamente necessari alle analisi.

```
TextTW : RT @ViennaInfoB2B: The @belvederemuseum shows "Vik Muniz. Verso",
the first solo #exhibition of Vik Muniz in Vienna. The artist has dedicat\u2026
Tweetid : 977185409129066496
Tweetid-str  : 977185409129066496
Username  : \u30d1\u30ea\u30b8\u30e3\u30f3\u30cc\u30c0\u30eb\u30af
Followers-count : 120
Following-count : 139
Favourite-count : 4564
Created-at : Fri Mar 23 14:08:51 +0000 2018
Lang : en
Favorited-count : 18
Retweet-count : 7
Username-author : ViennaTouristBoard
Created-at-author : Fri Mar 23 13:44:02 +0000 2018
Screen-name-author : ViennaInfoB2B
Followers-count-author : 11506
Following-count-author : 2786
Favourite-count-author : 4564
```

Figura 3.5 Risultato finale del tweet

Come si può notare nella Figura 3.5 il risultato è diminuito in termini di dimensioni e contiene informazioni strettamente necessarie alle statistiche che si andrà a fare. Confrontando la Figura 2.2 con la Figura 3.5 si può notare la differenza, sia in termini di dimensione, che in termini di comprensione visiva. Certamente i dati raccolti da Twitter, come si può notare in Figura 2.2, sono comprensibili solo dalla macchina, invece dopo la trasformazione i dati sono comprensibili, anche senza l'intervento del calcolatore.

Capitoli 4

4 Analisi dei dati

Ora che è stato ottenuto un dataset ridotto grazie alla manipolazione dei dati raw si può cominciare ad analizzare e verificare quanto vada ad incidere un tweet sulla diffusione di una certa mostra e quali fattori lo caratterizzano.

4.1 Contenuto del messaggio

I messaggi pubblicati possono avere vari tipi di contenuto. Il testo è il tipo di contenuto tradizionale, che può esprimere un giudizio sull'argomento trattato. Questo giudizio è stato chiamato *sentiment*, che a sua volta può essere positivo o negativo. Le critiche fatte su un argomento possono avere un importante rilevanza sulla sua diffusione. Oltre al testo, il messaggio può contenere immagini, audio, video, collegamenti ad altri siti, tutti questi argomenti sono espressi nello status sotto forma di *link*. La presenza di questi link può avere anch'essa un impatto rilevante sulla popolarità del tweet. Bisogna fare molta attenzione ad includere link, citare utenti ed includere tag riferiti ad un certo argomento, perché un eccessivo utilizzo di questi contenuti potrebbe non essere apprezzato dagli utenti che visualizzano il messaggio. Sono stati chiamati *spam* i messaggi che fanno un abbondante utilizzo di questi materiali. Nei paragrafi successivi si andrà ad

osservare il modo in cui questi tre gruppi (sentiment, link e spam) possano essere presenti nei vari messaggi pubblicati e come vadano ad espandersi tra i follower.

Tabella 4.1

	tweets		*followers*			*tweets*		*followers*		
Mostra	Yes	No	Yes	No		Yes	No	Yes	No	Mostra
	29,24%	70,76%	11,97%	88,03%	Sentiment presence	5,26%	94,74%	5,58%	94,42%	
	45,21%	54,79%	65,29%	34,71%	Spam	31,58%	68,42%	4,77%	95,23%	#orlando
#monet	94,84%	5,16%	84,19%	15,81%	Link presence	84,21%	15,79%	90,25%	9,75%	furioso
	27,92%	72,08%	11,38%	88,62%	Sentiment presence	66,67%	33,33%	40,90%	59,10%	
	34,80%	65,20%	74,28%	25,72%	Spam	33,33%	66,67%	56,45%	43,55%	#seuratsignac
#picasso	87,19%	12,81%	89,76%	10,24%	Link presence	100,00%	0,00%	97,34%	2,66%	vangogh
	23,60%	76,40%	16,30%	83,70%	Sentiment presence	22,22%	77,78%	0,08%	99,92%	
	17,23%	82,77%	18,84%	81,16%	Spam	11,11%	88,89%	99,82%	0,18%	
#klimt	95,88%	4,12%	88,34%	11,66%	Link presence	100,00%	0,00%	99,99%	0,01%	#expomagritte
	27,14%	72,86%	32,52%	67,48%	Sentiment presence	55,17%	44,83%	97,41%	2,59%	
	38,57%	61,43%	30,68%	69,32%	Spam	24,14%	75,86%	0,23%	99,77%	
#chagall	95,71%	4,29%	86,79%	13,21%	Link presence	79,31%	20,69%	93,95%	6,05%	#beyondcaravaggio
	16,51%	83,49%	40,57%	59,43%	Sentiment presence	11,11%	88,89%	17,85%	82,15%	
	40,37%	59,63%	45,44%	54,56%	Spam	33,33%	66,67%	26,08%	73,92%	
#cezanne	100,00%	0,00%	80,92%	19,08%	Link presence	100,00%	0,00%	100,00%	0,00%	#mexicofrida
	21,54%	78,46%	7,74%	92,26%	Sentiment presence	23,96%	76,04%	74,16%	25,84%	
	26,77%	73,23%	7,96%	92,04%	Spam	32,29%	67,71%	83,62%	16,38%	
#goya	86,46%	13,54%	95,49%	4,51%	Link presence	76,04%	23,96%	93,83%	6,17%	#escher
	26,42%	73,58%	84,88%	15,12%	Sentiment presence	8,33%	91,67%	2,19%	97,81%	
	32,08%	67,92%	90,44%	9,56%	Spam	41,67%	58,33%	45,64%	54,36%	
#miro	74,53%	25,47%	94,27%	5,73%	Link presence	83,33%	16,67%	97,47%	2,53%	#eschermilano
	27,37%	72,63%	19,75%	80,25%	Sentiment presence	19,55%	80,45%	14,51%	85,49%	
	38,95%	61,05%	62,59%	37,41%	Spam	51,02%	48,98%	57,17%	42,83%	
#klee	85,26%	14,74%	61,59%	38,41%	Link presence	97,45%	2,55%	82,20%	17,80%	#vangogh
	9,94%	90,06%	26,47%	73,53%	Sentiment presence	12,50%	87,50%	19,17%	80,83%	
	32,75%	67,25%	21,06%	78,94%	Spam	9,38%	90,63%	15,83%	84,17%	
#vermeer	97,66%	2,34%	92,87%	7,13%	Link presence	81,25%	18,75%	97,77%	2,23%	#davidbowieis
	19,83%	80,17%	15,53%	84,47%	Sentiment presence					
	22,31%	77,69%	27,62%	72,38%	Spam					
#pollock	85,12%	14,88%	71,52%	28,48%	Link presence					

La tabella 4.1 mostra la presenza dei diversi contenuti nei tweet riguardanti le mostre. Inoltre indica come questi messaggi vengano recepiti dagli utenti, comunicando cosa viene apprezzata dagli utenti che seguono una determinata mostra. In seguito si osserva in dettaglio l'andamento dei contenuti riguardanti tre mostre.

Per capire se le caratteristiche del tweet hanno effetti sulla sua diffusione si è andati a vedere in dettaglio i tipi di contenuti presenti nel messaggio. Di seguito

viene proposta una lista, con tre mostre, in cui vengono descritti gli effetti delle varie caratteristiche sulla diffusione dei tweet tra i follower:

- Per *#monet* abbiamo notato che 88% dei follower totali sono stati raggiunti con il 70% dei tweet, che non contengono un sentimento; il 65% dei follower sono stati raggiunti dal 45% dei tweet contenente spam e l'84% dei seguaci sono venuti in contatto con il 94% dei tweet che al loro interno avevano un link.
- Per *#picasso* abbiamo notato che 88% dei follower totali sono stati raggiunti con il 72% dei tweet, che non contengono un sentimento; il 74% dei follower sono stati raggiunti dal 34% dei tweet contenente spam e l'89% dei seguaci sono venuti in contatto con il 87% dei tweet che al loro interno avevano un link.
- Per *#klimt* abbiamo notato che 83% dei follower totali sono stati raggiunti con il 76% dei tweet, che non contengono un sentimento; il 81% dei follower sono stati raggiunti dal 87% dei tweet non contenente spam e l'88% dei seguaci sono venuti in contatto con il 95% dei tweet che al loro interno avevano un link.

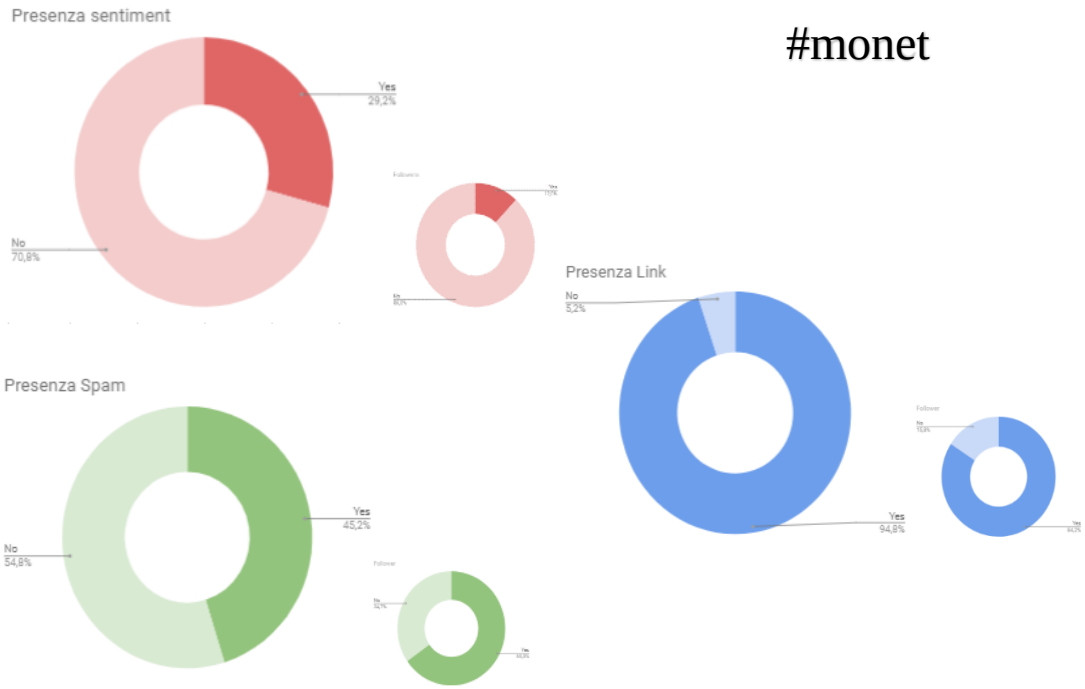

Figura 4.1 Diffusione del messaggio con differenti caratteristiche per la mostra di Monet

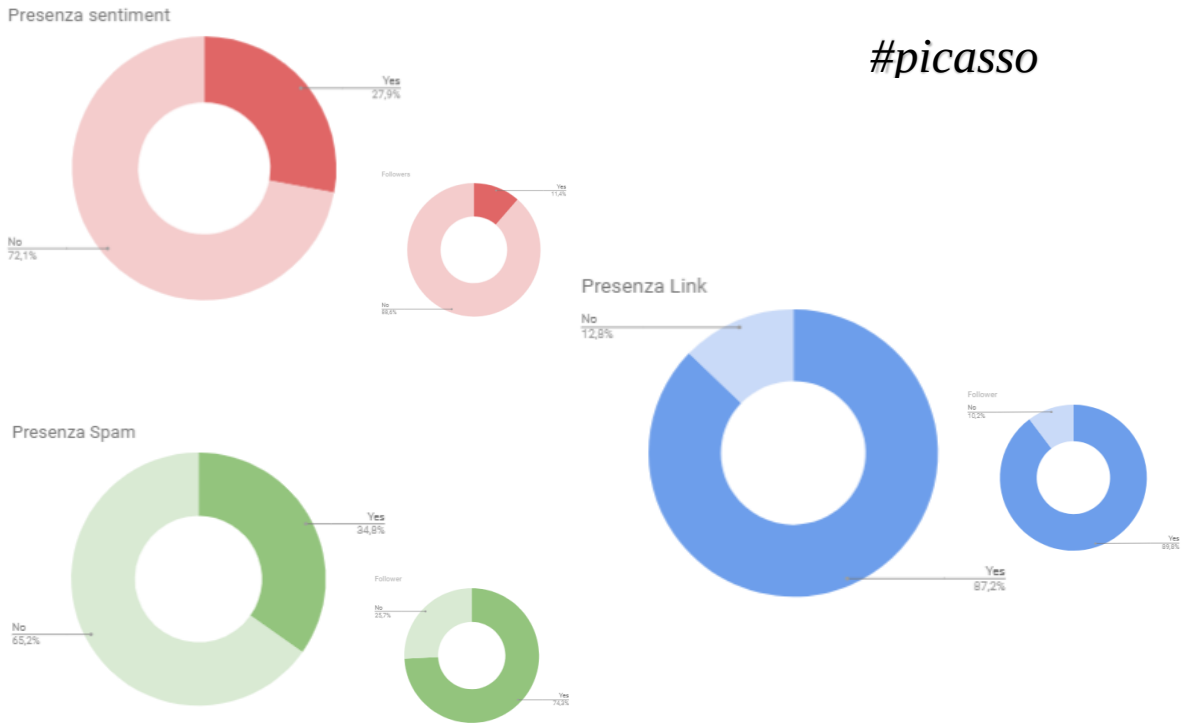

Figura 4.2 Diffusione del messaggio con differenti caratteristiche per la mostra di Picasso

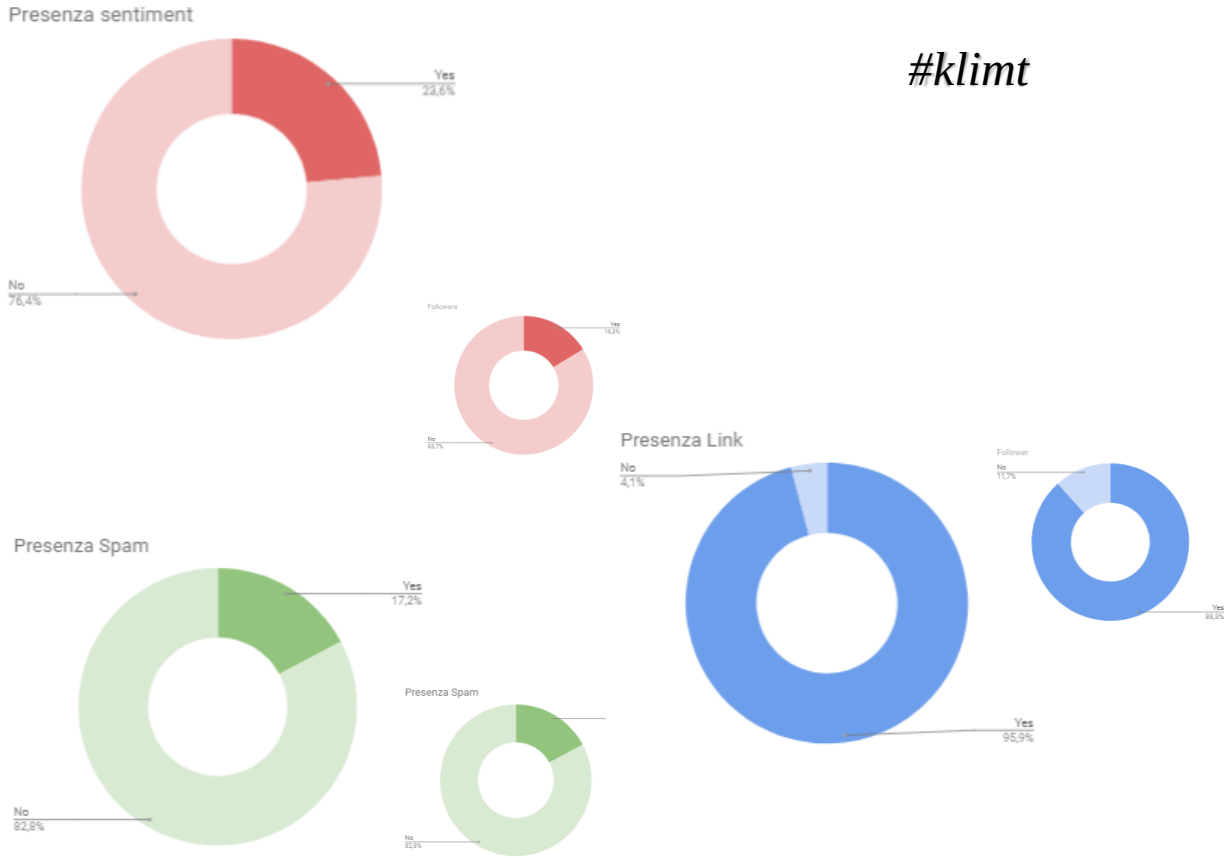

Figura 4.3 Diffusione del messaggio con differenti caratteristiche per la mostra di Klimt

4.1.1 Spam

Sui social media ci sono molte persone che fanno un uso eccessivo degli hashtag solo per far notare il proprio tweet, che può o no essere inerente all'argomento a cui si riferisce l'hashtag. Nel insieme di tweet raccolti, sicuramente ne sono stati catturati molti, che non interessano alla analisi dei sentiment, poiché non contengono abbastanza parole per esprimere un giudizio. Quindi si è deciso di creare dei filtri per cercare di individuare questi casi speciali. In questo modo si nota che nei tweet non utili, chiamati spam, si è riscontrato un utilizzo eccessivo di hashtag, mention, e link, e sono privi di un vero e proprio testo di senso compiuto. È stata quindi scelta una soglia di hashtag, mention e link massimi sopra i quali un tweet può essere considerato spam, il filtro di soglia elimina i tweet con:

- più di 4 hashtag;
- più di 3 mention;
- più di 2 link;
- più di 5 hashtag, mention e link sommati insieme.

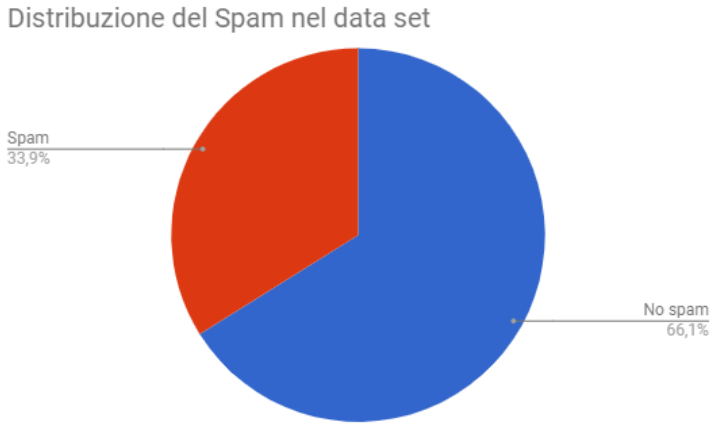

Figura 4.2 Grafico sulla distribuzione dello spam nel dataset

Si può notare nel grafico in Figura 4.1, che più del 30% dei dati raccolti contiene spam. Si è deciso di eliminare questi dati da alcune analisi future, ad esempio per il sentiment (trattato nel capitolo successivo), poiché a causa dell'alto numero di contenuti come hashtag, mention e link, non c'è abbastanza spazio, in un tweet di

soli 120 caratteri, per esprimere qualcosa di personale riguardo la mostra. In fine si può dire che anche se questi dati non contengono un testo caratterizzante, sono comunque utili per far conoscere l'hashtag e connettere con più utenti.

4.1.2 Sentiment

Il contenuto del tweet racchiude parole, che a loro volta possono esprimere un sentimento. Per una mostra è importante sapere se il testo del tweet è stato positivo o negativo, in questo modo si può sapere se una mostra è stata gradita, molti più tweet positivi, o non è stata gradita, più tweet negativi. È stato osservato il comportamento tutti i tweet, guardando come si suddivide, nel caso in cui sia presente, il sentiment nel dataset, in modo da capire quando quest'ultimo vada ad incidere sui nostri dati.

È stata utilizzata una libreria di Python per dare un valore, negativo o positivo, alle parole contenute nel messaggio pubblicato. AFINN è la libreria che permette di valutare le parole, per capire se un messaggio al suo interno contiene una critica positiva o negativa alla mostra a cui si sta riferendo.

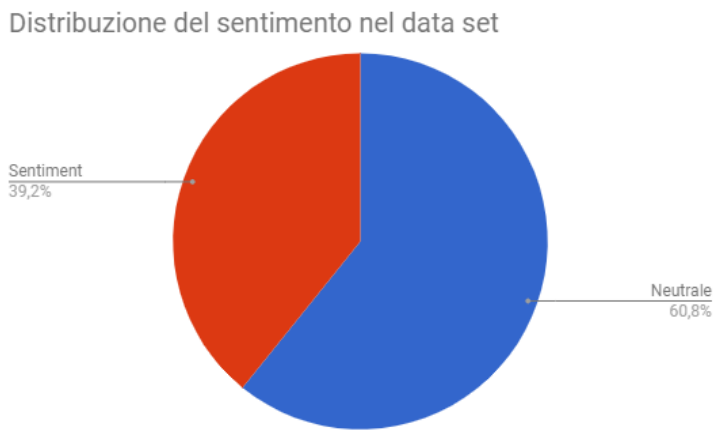

Figura 4.3 Grafico sulla distribuzione del sentiment nel dataset

Dalle analisi, Grafico 4.2, si può notare che la maggior parte, circa il 60%, dei messaggi postati non contengono nessun tipo di sentimento, mente il 40% dei

messaggi contiene un qualche tipi di sentimento, che può essere positivo o negativo. Già da questa prima analisi si può affermare che gli utenti che parlano di una mostra non esprimono forti sentimenti di favoreggiamento o antipatia verso di essa, ma bensì usano parole neutre per non esprimere un giudizio.

è stato analizzato in dettaglio che tipo di sentimento si può incrociare nel 40% dei tweet che ne esprimono uno. Grafico 4.3, mostra come si suddividono i tweet, tra positivo e negativo, nei dati che il filtro definito nel Capitolo 4.1.1 considera esprimere un sentimento. Si può osservare che solamente il 20% dei tweet presenta una percezione negativa e il restante 80%, la stragrande maggioranza, presenta un sentiment positivo.

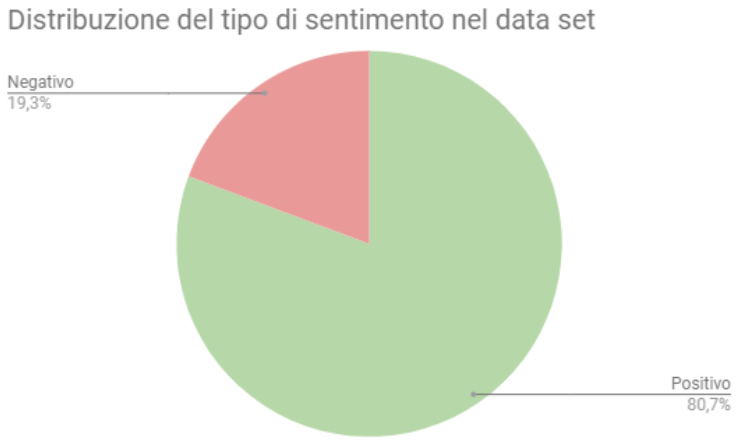

Figura 4.4 Grafico sulla distribuzione del tipo di sentimento nel dataset

Si può perciò affermare che, per le mostre d'arte, in generale i tweet vengono pubblicati senza esprimere un giudizio. Nei casi in cui questo non si verifichi, abbiamo, nella maggior parte dei casi, un senso di apprezzamento. Solo in pochissimi casi si possono trovare messaggi contenenti critiche negative.

4.2 Andamento temporale

Un fattore importante da conoscere è la vita media di un tweet, ovvero quanto resta visibile, tra il pubblico, prima di venire sepolto da altri tweet, più precisamente il margine di tempo dopo la sua pubblicazione in cui ha la massima popolarità. Per conoscere quanto "vive" in media un tweet su Twitter, è stato utilizzato il dataset totale. È stata calcolata la differenza d'orario tra la pubblicazione di un tweet è i suoi retweet. In questo modo si ottiene il grafico in Figura 4.7, che mostra dopo quante ore dalla pubblicazione sono stati pubblicati i retweet. Si può notare che le prime due ore sono quelle in cui ci sono più retweet, questo significa che nelle prime due ore dalla pubblicazione un tweet ha la maggior visibilità. Dopo questo arco di tempo il numero dei retweet scende drasticamente, portando, trascorse le 24 ore, la sua visibilità a zero.

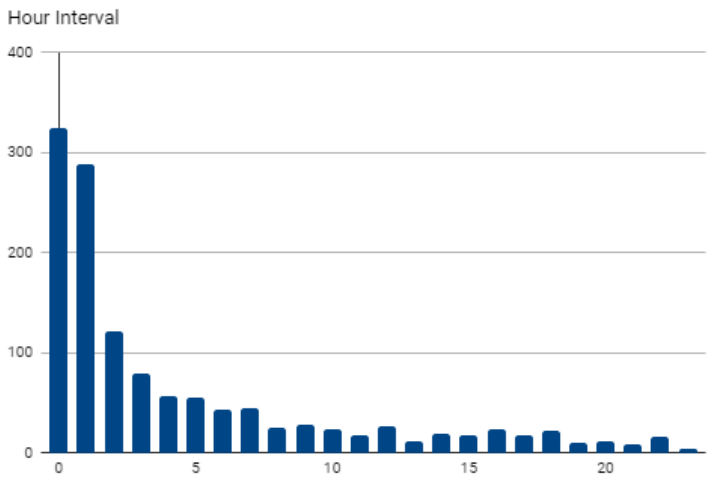

Figura 4.5 Grafico che mostra i retweet in funzione dell'intervallo orario

Poiché le prime ore, dopo la sua pubblicazione, sono fondamentali per diffondere un tweet, per i promotori delle mostre conoscere, se esiste, una certa fascia d'orario durante la giornata in cui l'affluenza dei utenti è la più alta, ovvero ci sono più pubblicazioni di tweet e di retweet, è fondamentale per utilizzare correttamente la piattaforma virtuale per diffondere i propri tweet. Avendo questa informazione si potranno pubblicare i tweet promozionali sugli account ufficiali dei musei in quelle fasce d'orario, che potranno raggiugere il maggior numero di utenti.

È stato quindi verificato se esistono dei orari favorevoli. Per fare ciò andiamo prima di tutto a dividere la giornata in 4 fasce orarie, ognuna con un intervallo di 6 ore:

- Notte, che va da mezzanotte alle cinque del mattino;
- Giorno, che va dalle sei alle undici del mattino;
- Pomeriggio, che va da mezzogiorno alle cinque del pomeriggio;
- Sera, che va dalle sei alle undici di sera.

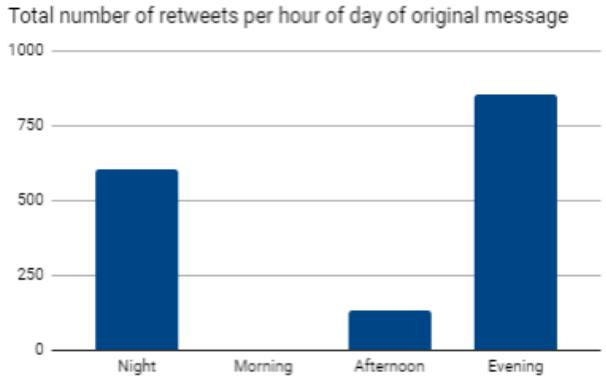

Figura 4.6 Grafico sul numero dei retweet per fasce orarie

Dal grafico in Figura 4.5 è evidente che ci sono fasce orarie in cui il numero dei retweet è elevato rispetto ad altre. Quindi si può affermare che ci sono fasce orarie nelle quali è più favorevole postare un messaggio per raggiungere un maggior numero di utenti. Più precisamente la sera e la notte sono i momenti in cui il numero dei retweet è elevato di conseguenza non sarebbe sbagliato dire che sono gli orari in cui l'affluenza dei utenti è maggiore rispetto ad altri orari, come il mattino.

Per confermare quanto già detto precedentemente è stata fatta una controprova. Ovvero è stato verificato il numero degli utenti che possono vedere un tweet pubblicato in quelle fasce orarie, tramite il numero dei followers. Avendo un altro

numero di retweet in certi orari si poteva aspettare che anche il numero degli utenti che questo messaggio può raggiungere, rispetti l'andamento già ottenuto.

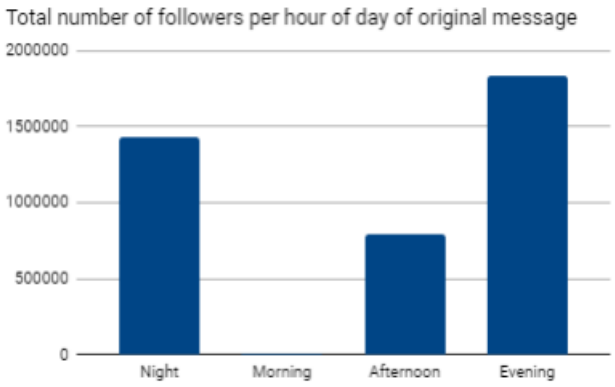

Figura 4.7 Grafico sul numero dei follower per fasce orarie

Esattamente come previsto, nel grafico in Figura 4.8, anche i followers totali aumentano come i retweet, grafico in Figura 4.9. Poiché con l'aumentare dei retweet aumentano anche i possibili utenti che potrebbero vedere questo messaggio, Più viene retwittato un messaggio più diventa popolare tra la gente. Concludendo possiamo dire che un tweet per diventare popolare deve raggiungere il maggior numero di utenti. Per questo deve essere pubblicato nelle fasce orarie in cui c'è un maggior numero dei seguaci dell'utente principale presente sul social network.

4.3 Influenza

Nel capitolo 4.1 è stata osservata la presenza del sentiment nel dataset. In questo capitolo si cercherà di capire se la presenza di sentiment, anche se in modo ridotto, può aiutare ad aumentare la sua influenza.

Finora si sa anche dalle analisi fatte nei capitoli precedenti, che per ogni tweet si ha il numero di volte in cui è stato retwittato. Questa significa che ci sono stati degli utenti tra i seguaci della persona, che hanno pubblicato il tweet, che lo hanno gradito e a loro volta hanno deciso di condividerlo tra i propri seguaci, per dargli più

visibilità. Il valore dei retweet permetterà di sapere quanto è stato gradito un tweet su un dato artista o mostra.

In seguito si osserva se la presenza del sentiment va ad incidere in qualche modo sulla influenza che può avere una mostra d'arte. Per fare ciò si andrà a prendere per ogni mostra il suo sentiment totale (somma dei sentiment di tutti i suoi tweet) e lo si mette in rapporto al sentiment del dataset totale, questo valore che si ottiene è stato chiamato sentiment. Poi si andrà a trovare l'influenza, per il quale si mettono in rapporto i retweet di una mostra e la somma dei retweet di tutte le mostre. Si andrà a mettere per ogni mostra questi due valori in un grafico, avendo sull'asse delle x tutti i sentiment e su quelle delle y, tutte le influenze.

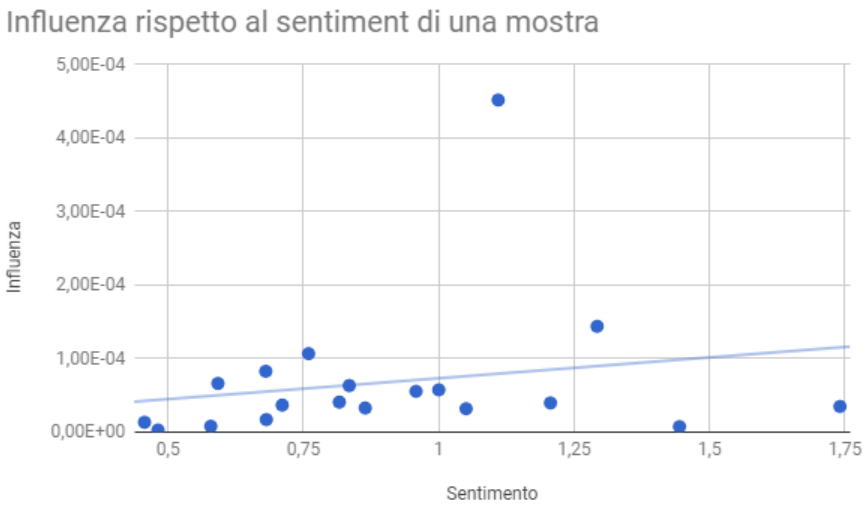

Figura 4.8 Grafico dell'influenza rispetto al sentimento per ogni mostra

Si può osservare nel grafico in Figura 4.10, come si comporta l'influenza rispetto al sentimento. Si può affermare che in generale se si ha un maggior numero di sentiment positivo, questo andrà ad incidere anche sulla influenza, che ha quella mostra sugli utenti, quindi con un sentiment maggiore anche l'influenza aumenterà, eccetto in casi speciali, come si può notare anche nel grafico. Anche in quel caso si può comunque confermare che avere un apprezzamento nel contenuto del messaggio aiuta ad aumentare l'influenza della mostra.

4.4 Popolarità

Fin ora si conosce che ogni utente ha un certo numero di follower, un tweet pubblicato da un utente sarà visibile a tutti i sui follower, i retweet di un tweet saranno visibile tra i followers dell'utente originale e di tutti gli utenti che lo hanno condiviso, aumentando notevolmente la rilevanza del messaggio. In questo capitolo si osserverà come aumenta il numero di followers raggiunti durante l'arco della giornata. Si spera che l'andamento rispetti ciò che è stato detto nel capitolo 4.2 per i grafici nelle Figure 4.8 e 4.9, ovvero un picco di followers negli orari con più affluenza e pochi utenti raggiunti negli orari in cui c'è poca folla sul networking site.

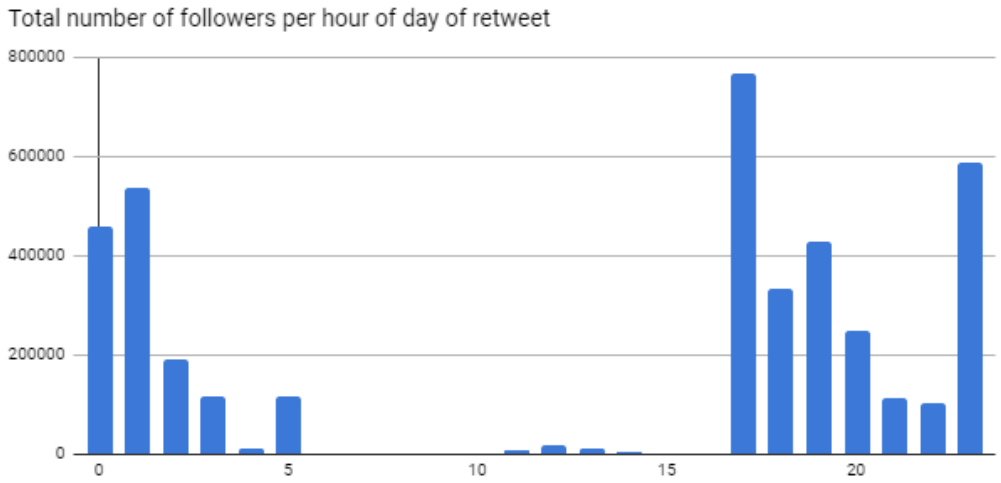

Figura 4.9 Grafico sull'numero dei follower raggiuti da un tweet nelle varie ore della giornata

Nel grafico 4.10, si può osservare come varia il numero dei seguaci nei vari orari della giornata. Come era stato già detto il grafico rispecchia largamente l'andamento riscosso nel capitolo 4.2. Si può notare che anche nelle fasce orarie, che sono state dichiarate avere il pico massimo di utenti, ci sono orari in cui non è propriamente conveniente pubblicare. Si può, dopo avere visto questi risultati dedurre che, è utile sapere quali sono gli intervalli durante la giornata in cui è

conveniente pubblicare un tweet, per aumentare le probabilità della sua diffusione. Pertanto si può affermare che è vantaggioso pubblicare più frequentemente tweet in quel intervallo, piuttosto che pubblicare tweet frequenti in tutta la giornata. Anche perché in quest'ultimo modo si può rischiare di ottenere il risultato opposto a quello desiderato.

Conclusione e sviluppi futuri

In questa tesi si è occupati dell'analisi sull'utilizzo di Twitter per la promozione di mostre d'arte temporanee all'interno dei musei. È stato osservato come l'utilizzo di hashtag per queste possa aiutare gli organizzatori a raggiungere un pubblico sempre più vasto, in modo tale da ottenere affluenza maggiore e quindi un probabile vantaggio e anche guadagno sulle difficoltà dovute alle spese di organizzazione elevate. Gli hashtag utilizzati sono stati di diversi tipi: alcuni erano specifici per l'evento altri invece erano generici, ovvero utilizzavano il nome di artisti.

È stato anche analizzato il contenuto del testo di ogni tweet, osservando come nella maggior parte dei casi i tweet pubblicati per una mostra non contenessero espressioni di critica. Nei casi in cui capitasse di trovare un sentiment nel tweet, questo per la grande maggioranza dei casi, sarà positivo. Anche se il sentimento non è sempre presente, nei casi in cui lo fosse, andrebbe a incidere sulla influenza, che può avere quella mostra, aumentandola. Un fattore molto importante, da non sottovalutare, è l'orario di pubblicazione. Se si vuole raggiungere il maggior numero di persone bisognerà pubblicare i tweet negli orari in cui l'affluenza su Twitter è più alta, poiché saranno attivi molti più seguaci, di chi posta il messaggio, che potranno vederlo e condividerlo, aiutando così ad alimentare la popolarità di una mostra.

Quindi si può dire che è possibile promuovere le mostre d'arte sui social network. Il modo migliore è l'utilizzo di un hashtag personalizzato per evitare confusioni. Per ottimizzare ulteriormente la promozione è fondamentale la pubblicazione frequente negli orari in cui ci sono più utenti attivi. Infine ha molta importanza anche invitare i visitatori a esprimersi un giudizio riguardo la loro esperienza.

In questa tesi non sono stati utilizzati ulteriori dati se non le API Twitter le quali forniscono i dati, con cui passare a molte analisi. Queste infatti andranno ad approfondire quelle che sono già state fatte oppure farne di nuove in futuro utilizzando differenti e più specifiche variabili, come ad esempio la distanza, che riescono a raggiungere i tweet, tramite l'uso di geo_tag, in modo da valutare quanti utenti ed eventuali turisti si potrebbero contattare tramite queste piattaforme. Ma anche l'oggetto di queste analisi può essere sostituito con uno, che potrebbe essere interessante allo stesso livello delle mostre, ad esempio riguardo l'organizzazione di eventi piccoli: così facendo i piccoli organizzatori sarebbero in grado di comprendere questi meccanismi della promozione dell'evento sul networking site ed eventualmente valutare su quali social è più conveniente pubblicizzare.

Bibliografia

[1] Wikipedia, «Hashtag,» Wikipedia, 01 10 2017. [Online]. Available: https://it.wikipedia.org/wiki/Hashtag. [Consultato il giorno 02 04 2018].

[2] A. Moujahid, «An Introduction to Text Mining using Twitter Streaming API and Python,» 21 07 2014. [Online]. Available: http://adilmoujahid.com/posts/2014/07/twitter-analytics/. [Consultato il giorno 02 04 2018].

[3] P. Gil, «What Is Twitter & How Does It Work?,» 05 02 2018. [Online]. Available: https://www.lifewire.com/what-exactly-is-twitter-2483331. [Consultato il giorno 02 04 2018].

[4] «Introduzione a JSON,» [Online]. Available: https://www.json.org/json-it.html. [Consultato il giorno 02 04 2018].

[5] Twitter, Twitter, Inc., 2018. [Online]. Available: https://developer.twitter.com/. [Consultato il giorno 02 04 2018].

[6] A. Novalić, «Introduction to tweepy, Twitter for Python,» 14 08 2013. [Online]. Available: https://www.pythoncentral.io/introduction-to-tweepy-twitter-for-python/. [Consultato il giorno 02 04 2018].

[7] Wikipedia, «OAuth,» Wikipedia, 16 09 2016. [Online]. Available: https://it.wikipedia.org/wiki/OAuth. [Consultato il giorno 02 04 2018].