# POLITECNICO DI MILANO

Corso di Laurea Magistrale in Computer Science and Engineering

Dipartimento di Elettronica, Informazione e Bioingegneria



# Analysis of social media response in time and space to live events during the "Milano Fashion Week"


Relatore: Prof. Marco Brambilla
Correlatore: Prof. Stefano Ceri

Tesi di laurea di:
Gianmarco Donetti
Matr. 836883



Anno Accademico 2015 - 2016

*A mia mamma e mio papà.*

# Abstract

The Information Age is strongly evolving due to the social media strength and diffusion, where people increasingly get in touch not only with friends, but also with celebrities, news media broadcaster and politicians, with different purposes, based on the type of connection the two actors have, and to the spread of the smartphones, which able ourselves to remain always in touch with our social connections mentioned before. In this always connected world, the response to popular real world events is becoming through the years much more significant and meaningful, in terms of volumes of contents shared in the social network itself, but also in terms of reaction velocity in the spreading content phase with respect to the time and to the geographical dimension. This work deals exactly with the problem of social media response to a specific popular real world event, the Milano Fashion Week occurred from the $24^{th}$ to the $29^{th}$ February of 2016, analysing the behaviour of such users that are acting in re-action (or in pro-action) to the specific events.

The study forks in different sub-analysis: the first chapter focuses on the study of different authors attributes, attempting to find some correlation between the popularity these users obtained in this specific case study, the influential strength they have already gained in terms of number of followers and the generated content volume. Defining the popularity score of a user as the summation of all the likes and comments (in the Instagram scenario) or all the likes and retweets (in the Twitter scenario) we have stored in our database, we present a comparison between the different measures mentioned before, showing that both for Instagram and Twitter the strongest correlation is between the total number of posts someone shares with his (or her) community and the number of likes, comments and retweets he (or she) received in response to all the posts, while this correlation totally disappear or even increase in a negative way if we compare again the generated posts volume versus the average number of likes and comments or retweets received per post. We also find a nice positive correlation between the influential strength, defined as the number of followers, and the popularity score, that is: as the number of follower increase, the feedback in terms of likes, retweets and comments will increase, too. At last, we find no correlation at all between the number of posts generated and the influential strength.

After this study on these specific authors attributes, we focus ourselves on

the social response to the different events belonging to the Milano Fashion Week schedule with respect to the time axis and the geographical axis, respectively. We have been able to build some clusters in these two different dimensions of study, finding some similar patterns in the different types of response. In the time scenario we face the objective of trying to predict the labelling of the clustering phase a-priori, with the only poor information related to the brands and the events.

The fourth chapter tries to make a comparison between the results obtained in the analysis of the previous two chapters. We show how the levels of accuracy between the different juxtapositions are not so strong, demonstrating how each subspace of analysis, that are the Time Response one, the Geo Response one and a new one, which could be referred as the Popularity Response one, plays a strong and relevant role in characterizing the brands involved in the event.

# Sommario

L'Era dell'Informazione sta fortemente evolvendosi grazie all'impatto e alla diffusione dei mezzi di comunicazione sociali (chiamati più comunemente in inglese social media), grazie ai quali le persone entrano in contatto sempre di più non solo con veri e propri amici, ma anche con celebrità, diffusori e fonti di notizie e persino politici, con scopi diversi, basati sul tipo di rapporto che intercorre tra le due figure considerate, e grazie alla diffusione degli smartphone, che ci rendono in grado di rimanere sempre in contatto con le relazioni sociali appena menzionate. In questo mondo ormai sempre connesso, la risposta ad eventi popolari dal vivo sta diventando, attraverso gli anni, sempre più significativa e interessante, sia in termini di volumi di contenuti generati nelle stesse reti sociali, che in termini di velocità di reazione nella fase di propagazione di contenuti, facendo riferimento alle dimensioni temporale e geografica. Questo lavoro affronta proprio il problema della risposta creatasi nei social media a un evento dal vivo molto popolare, nello specifico la Settimana della Moda di Milano tenutasi tra il 24 e il 29 di Febbraio 2016, analizzando il comportamento degli utenti che hanno agito in reazione (o pro-azione) all'evento specifico.

Questo studio si divide in differenti sotto-analisi: il primo capitolo si concentra sull'analisi di alcuni attributi propri degli autori, ricercando delle correlazioni tra la popolarità che questi utenti hanno riscosso nello specifico scenario preso in considerazione, la forza influenza che gli stessi utenti detengono in termini di numero di seguaci e il volume di contenuti da essi generato. Definendo il punteggio di popolarità di un utente come la somma di tutti i "mi piace" e i commenti (per quanto riguarda Instagram) o di tutti i "mi piace" e le ri-condivisioni (per quanto riguarda Twitter) che abbiamo immagazzinato nella nostra base di dati, presentiamo quindi una comparazione tra le misure menzionate prima, mostrando come, sia per Instagram che per Twitter, le grandezze con correlazione più evidente risultano quelle di numero totale di messaggi che un utente condivide con la sua comunità virtuale e il numero di apprezzamenti, commenti e condivisioni l'utente stesso riceve a risposta delle sue micro-pubblicazioni, mentre tale correlazione scompare completamente, se non addirittura cresce in maniera negativa, nell'accostare le misure di numero totale di messaggi condivisi e valore medio di popolarità riscossa per singolo messaggio. Abbiamo inoltre avuto dei risultati di discreta correlazione tra la forza influenza, definita come il numero di seguaci,

e il punteggio di popolarità, che sta a significare: all'incrementare del numero di seguaci, un utente riesce a riscuotere molta più popolarità. Infine, non abbiamo trovato correlazione alcuna tra il volume di contenuti condivisi e la forza influenza.

Una volta completato questo studio su specifici attributi degli autori dei contenuti e dei messaggi, ci siamo concentrati sulle risposte sociali ai diversi eventi propri del calendario della Settimana della Moda, rispetto all'asse temporale e all'asse geografico. Siamo stati in grado di costruire dei raggruppamenti dei vari marchi coinvolti all'evento in queste due differenti dimensioni di analisi, scoprendo dei modelli molto simili nei diversi tipi di reazione. Per quanto riguarda la risposta temporale ai diversi eventi, ci siamo inoltre posti l'obiettivo di costruire dei modelli di predizione del tipo di risposta, sfruttando solamente le scarse informazioni relative ai brand e, soprattutto, agli eventi stessi.

Nel quarto capitolo cerchiamo di confrontare i diversi risultati ottenuti finora, nei due capitoli precedenti. Dopo aver mostrato come i livelli di accuratezza di giustapposizione non risultano molto elevati, abbiamo sottolineato come ciascun universo di analisi (ossia quello di risposta temporale, quello di risposta geografica, e un nuovo universo di risposta di popolarità), interpreta un ruolo significativo nella caratterizzazione di tutti i brand coinvolti all'evento.

# Contents

# Chapter 1

# Introduction

## 1.1 Context and problem statement

While the Internet and the World Wide Web have always been used to facilitate social interaction and connections, the emergence and rapid propagation of Web 2.0 functionalities during the first decade of the new millennium enabled an evolutionary leap forward in the social component of web use. This and the falling costs for on-line data storage made it feasible for the first time to offer masses of Internet users access to an array of user-centric spaces that they could populate with user-generated content, along with a correspondingly diverse set of opportunities for linking these spaces together to form virtual social networks. In order to define "social media" for our current purposes, we synthesize and aggregate definitions presented in the literature [33, 7] and identify the following commonalities among current social media services:

1. Social media services are (currently) Web 2.0 Internet-based applications;
2. User-generated content is the lifeblood of social media;
3. Individuals and groups create user-specific profiles for a site or app designed and maintained by a social media service;
4. Social media services facilitate the development of social networks on-line by connecting a profile with those of other individuals and/or groups.

Tapping latent demands, social media services quickly emerged as both business and social phenomena. Facebook, launched in 2004, has now reached more than 1.7 billion active monthly users worldwide. Twitter, started in 2006, currently has 313 million monthly active users, with the 82% active users from mobile platforms. Instagram, born in 2010, today has more than 500 million active monthly users, with 95 million photos or videos shared per day. These two new social media differ from the others mostly because of the asymmetric relationships that can

be installed among users. This property gave the two social media a very close role to news media broadcaster, just like in a publisher-subscriber scenario.

Twitter[28] users follow others or are followed. The relationship of following and being followed requires no reciprocation. A user can follow any other user, and the user being followed need not follow back. Being a follower on Twitter means that the user receives all the messages (called tweets) from those the user follows. Common practice of responding to a tweet has evolved into well-defined mark-up culture: RT stands for retweet, '@' followed by a user identifier address the user, and '#' followed by a word represents a hashtag. This well-defined mark-up vocabulary combined with a strict limit of 140 characters per posting conveniences users with brevity in expression. The retweet mechanism empowers users to spread information of their choice beyond the reach of the original tweet's followers.

Instagram[22] is a popular photo or capturing and sharing mobile application. It offers its users a unique way to post pictures and videos using their smartphones, apply different manipulation tools in order to transform the appearance of an image, and share them instantly on multiple platforms (e.g., Twitter) in addition to the user's Instagram page. It also allows users to add captions, hashtags using the '#' symbol to describe the pictures and videos, and tag or mention other users by using the '@' symbol (which effectively creates a link from their posts to the referenced user's account) before posting them.

In addition to its media capturing and manipulation functions, Instagram also provides similar social connectivity as Twitter that allows a user to follow any number of other users. The users following other Instagram users are called "followers". Besides, users can set their privacy preferences such that their posted photos and videos are available only to the user's followers that requires approval from the user to be his/her follower. By default, their images and videos are public, which means they are visible to anyone using Instagram app or Instagram website. Users consume photos and videos mostly by viewing a core page showing a "stream" of the latest photos and videos from all their friends. They can also favourite or comment on these posts. Given these functions, we regard Instagram as a kind of social awareness stream[32] like other social media platforms such as Facebook and Twitter.

In this context we formulate our problem as the analysis of the impact of big popular events on social media platforms, trying to evaluate the social media response to real-world events, appreciate some facts and findings in the specific scenarios, and discover patterns and relevant indicator of social media response to big popular occurrences.

In this way, we select a specific vertical domain to mine, that is the Milano Fashion Week occurred from the $24^{th}$ to $29^{th}$ of February in 2016, as our case study. Milano Fashion Week, established in 1958, is part of the global "Big Four fashion weeks", the others being Paris Fashion Week, London Fashion Week and New York Fashion Week[8]. The schedule begins with New York, followed by Lon-

don, and then Milan, and ending with Paris. This is the most prestigious event organized (partially) by Camera Nazionale della Moda Italiana, with its two yearly. The woman collection fashion shows are the most awaited moment of the international fashion system. Camera Nazionale della Moda Italiana manages and fully co-ordinates all the events, so facilitating the work of showrooms, buying-offices, press offices, and public relations firms. Milano is the prestigious location that hosts more than 170 shows, presentations and events, promoting the maisons that have made famous *Made In Italy* in the world and supporting new talents that make of the fashion world a sector in continuous evolution. In this fascinating and full of creativity scenario, Camera Nazionale della Moda carries out essential functions like drawing up the calendar of the shows and presentations, managing the relations with the Institutions, the press office and creation of special events. Milano Fashion week, with its two annual editions of September-October (Spring/Summer Collection) and February-March (Autumn/Winter Collection) represents the most important meeting between prêt-á-porter and market operators and it awards Milan with a refined example of the perfect union between creativity and organization.

In this domain our goal is to describe and characterize the social media response to the events in the official calendar, in terms of reactions with respect to the time axis, at first, and then with respect to some spatial features of dispersion and concentration of the social signal related to specific fashion shows. After these studies, we inspect some correlation between the different and orthogonally characterizations of the brand-events couples considered.

## 1.2 The data

The whole work is based on a database of posts and medias shared on Twitter and Instagram, built thanks to a scraper and a crawler. Firstly, we launch our scraper and start collecting information from Twitter[39], with two different requirements for the query:

1. The first for the time window: we have decided to not take only into account the short period of the Milano Fashion Week event, but also to include some days before and after the official schedule. In this way, our days of analysis goes from the $17^{th}$ of February to the $7^{th}$ of March, 2016;

2. The second for the content of the messages, in terms of hashtags that have to match the text of the tweets. We create our set of hashtags with the help of domain experts, that provide us 21 different tags.

With these specifications and filters on the query to be submitted to the scraper, we built our first database. We instantly made a brief overview of some measures on the database, in order to understand better our data and maybe also to improve the collection built so far a little in terms of contents and volumes.

This first database contains 72846 different tweets. The percentage of geo-located posts is about 6.7%, for a total of 4882 located post. We can now have a look at the top authors and top hashtags per presence in our database, as shown in the Figures 1.1a, 1.2a attached at the end of this chapter. Analysing the hashtags presence, we can see that some of the top hashtags found were not included in the query we have launched in order to obtain the data. Moreover, there are some tags we have looked for that have no matchings at all in the entire data-set, as shown in Table 1.1. Then, we decide to have a look in the top-hashtags we analysed before and find which of them were in the query set and which not. In this way, we find some nice hashtags that were not present in the initial query set, nice in terms of semantic and presence in the micro-messages we have stored.

Indeed, we reformulate the query set of hashtags including some of the new ones found before and removing the useless ones with no matchings at all, and we also extend the time window in order to capture more movements on the social networks and to increase the number of geo-located posts in absolute value, useful for our future analysis.

Our final database for the Twitter scenario contains 106278 tweets, with a percentage of nearly 6.5% of geo-located tweets, that corresponds to 6921 different posts. The new time trend is shown in Figure 1.4a, where we can also have a first look at the presence of scheduled events from the official calendar of the Milano Fashion Week 2016. In the end, we can see the different top 15 from the point of view of the authors and their volumes of content shared and the hashtags presence.

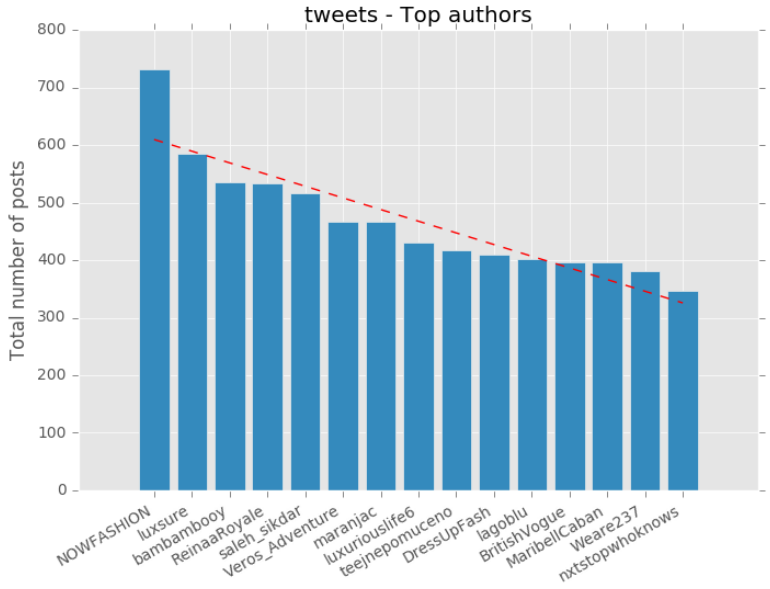
Now we have also all that we need in order to query the Instagram API[23]. Then, we set the same time window and the same hashtags in our query, and we obtain a pool of Instagram posts containing 556045 different media, with time trend shown in Figure 1.4b, with the 27.986% of them reporting geo-locations information, that is a total of 155613 posts. Immediately, we can notice the differences in the number of media collected from the two social media and, specially, the big difference in the percentages of geo-located posts. This can be explained identifying different behaviours from the Twitter users and the Instagram users. Indeed, in both the two social media, the geo-tagging of the post is not forced to the user from the application and we can think about the Instagram users as more inclined in sharing and showing off their activities, the place where they are taking the pictures and proving they were actually there. We report at the end of the chapter the histograms referring to the top authors per volume shared and the top hashtags per frequency in our database, as we did for Twitter.

## 1.3 Structure of the work

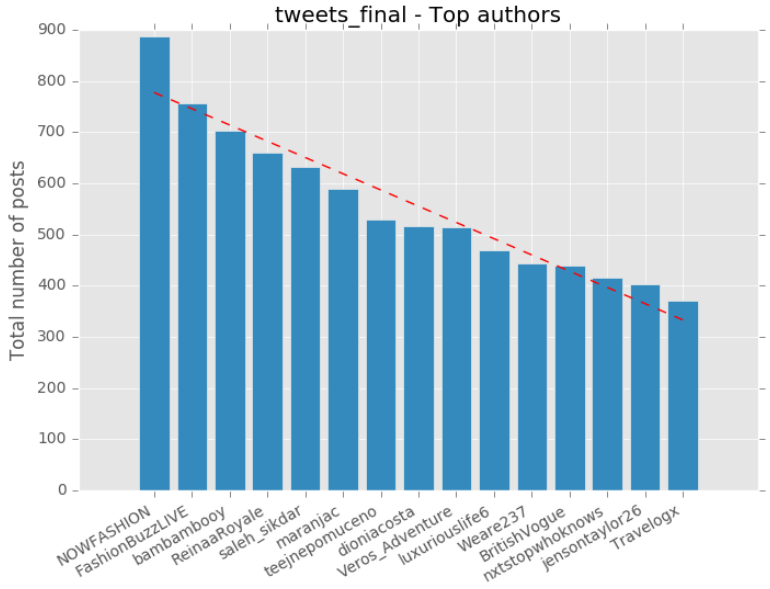
In this section, we introduce the main structure of the work.
In chapter 2 we report the state of the art related to the various methods adopted all over the work.
In chapter 3 we introduce our case study by analysing the correlation between some measures related to the users involved in our scenario.
In chapter 4 we present the results of the Time response analysis.
In chapter 5 we present the results of the Geo response analysis.
In chapter 6 we report the findings on the correlation between the different characterizations of the brand-events couples we have computed so far.
In chapter 7 we present our conclusions, with a short summary of the work done, some critical discussions on the results and possible application of future works.

Table 1.1: Occurrences of the first query hashtags set.

| **Hashtag** | **Presence** |
|---|---|
| #cameramoda | 92 |
| #MFWLIVE | not found |
| #mfw | 13196 |
| #mfw2016 | 569 |
| #mwfw | 59 |
| #MFWFW16 | 1 |
| #mwfw2016 | 4 |
| #mfw16 | 1351 |
| #AW16 | 10052 |
| #rtw | 4576 |
| #fw16 | 6558 |
| #milanfashionweek2016 | 165 |
| #milanfashionweek | 3645 |
| #milanowomensfashion | not found |
| #milanwomanfashionweek | not found |
| #milanwomanfashionweek2016 | not found |
| #settimanadellamodamilano | not found |
| #vfno2016milan | not found |
| #whitemilan | 10 |
| #whitemilano | 106 |
| #viatortona | 13 |

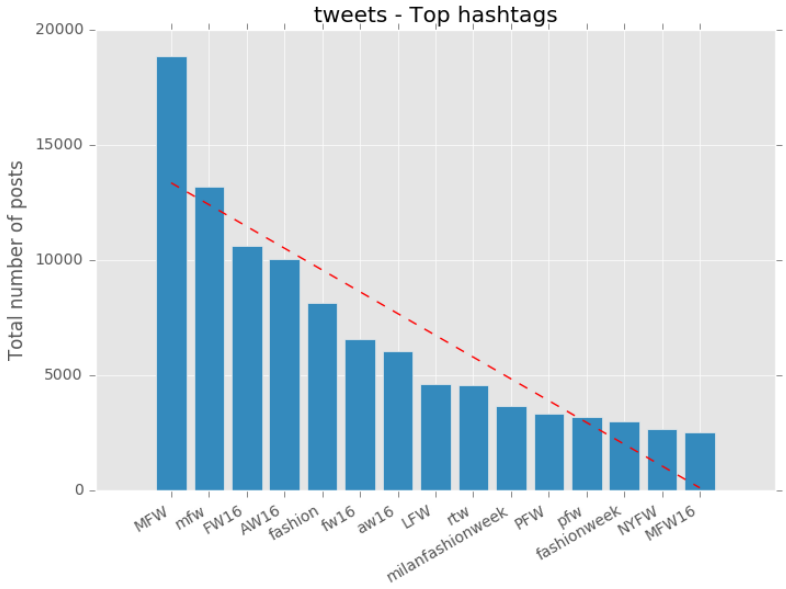



(a)

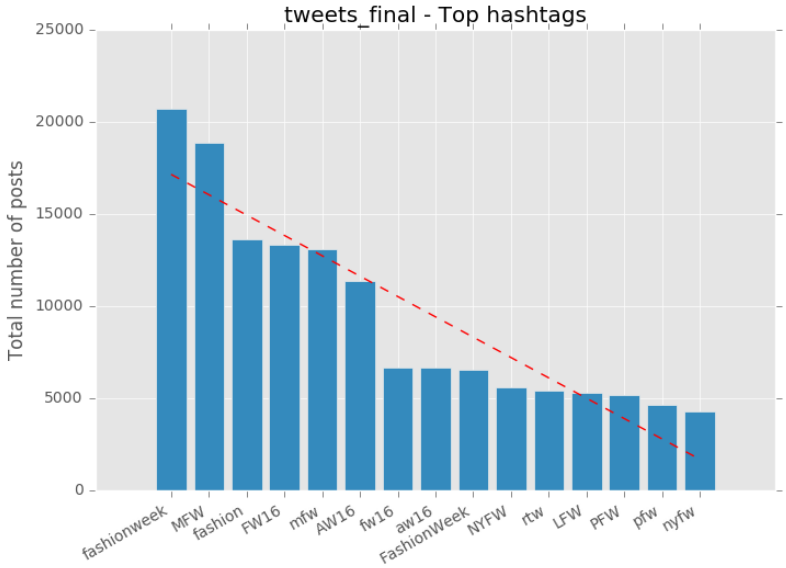



(b)

Figure 1.1: Top 15 authors on Twitter sorted per number of posts collected, with the first query for (a) and the second query for (b).

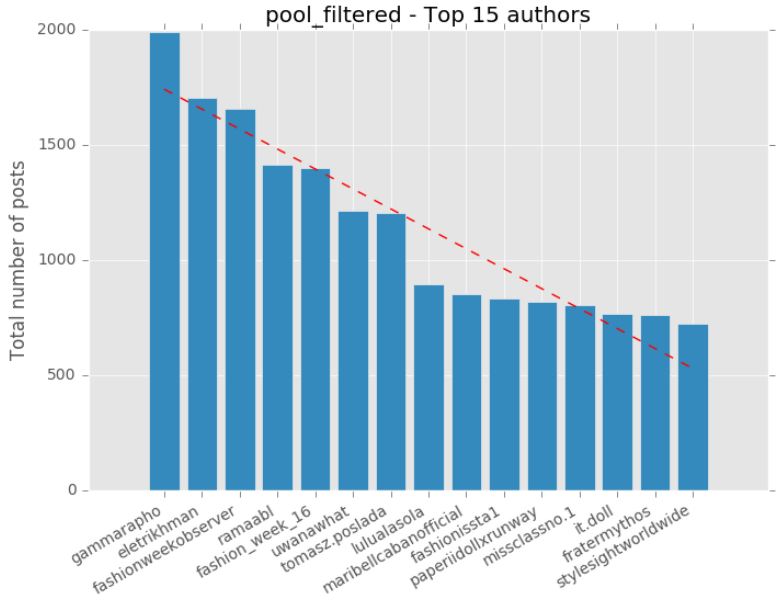



(a)

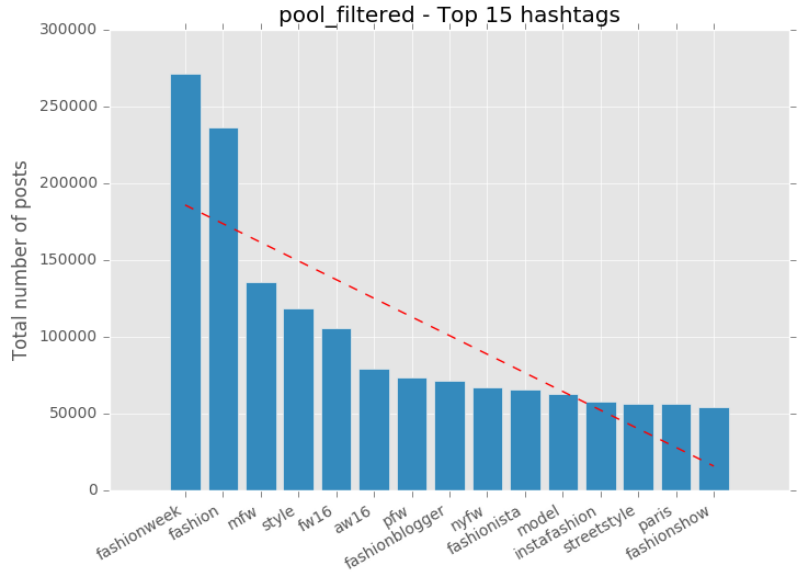



(b)

Figure 1.2: Top 15 hashtags on Twitter sorted per number of occurrences collected, with the first query for (a) and the second query for (b).

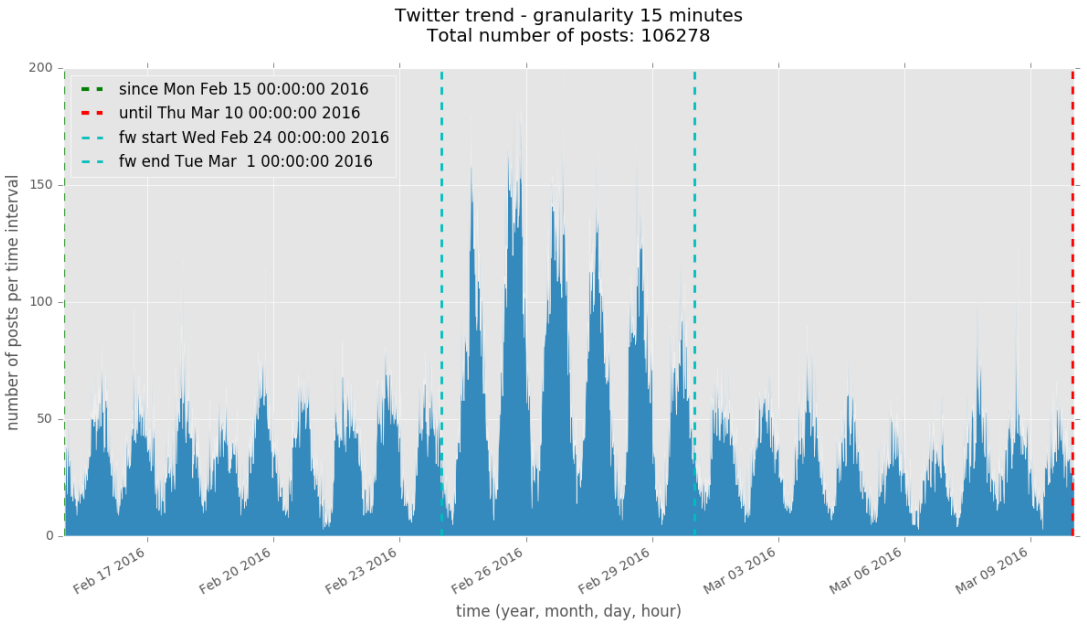



(a)

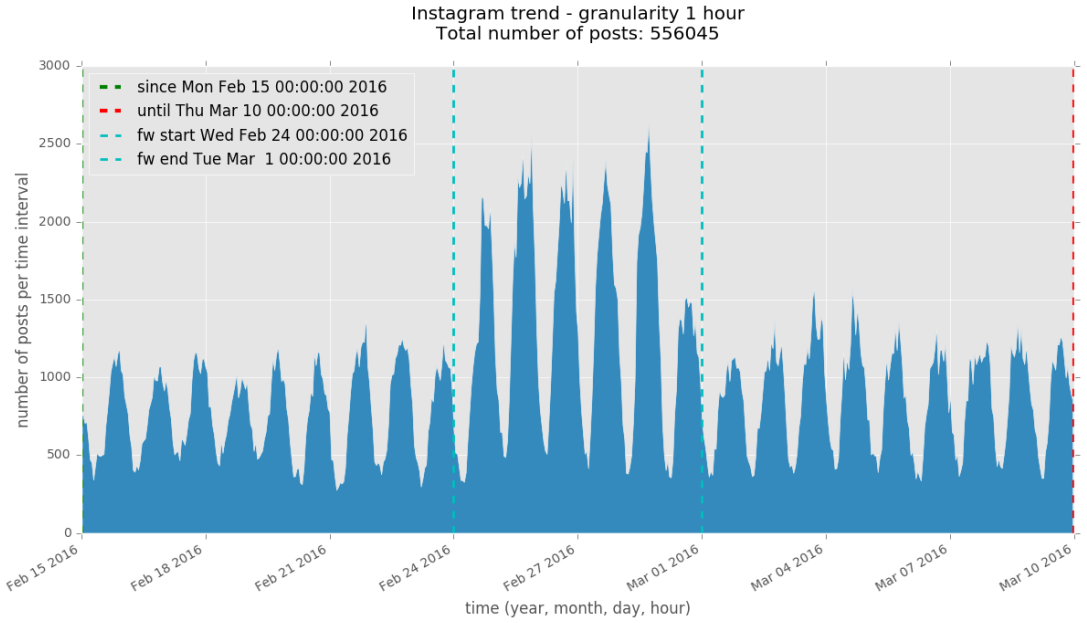



(b)

Figure 1.3: Top 15 authors sorted per number of posts collected together with top 15 hashtags sorted per number of occurrences collected, on Instagram.




(a) Twitter trend




(b) Instagram trend

Figure 1.4: Time trend and volumes both for Twitter (a) and Instagram (b). The granularities are 15 minutes and 1 hour and we can notice the lines delimiting the duration of the Milano Fashion Week and the lines related to our window of analysis.

# Chapter 2

# Related work

The goal of this chapter is to present the most relevant works related to the main objective we are going to face. The chapter is divided into three main sections, in order to group together similar problems.

At first we focus on the fashion world, after that we switch to the analysis of the social network users, and in particular their attributes, and then we conclude with some works about event response in the social media platforms.

## 2.1 Fashion

The work of Manikonda et al. [30] presents a qualitative analysis on the influence of social media platforms on different behaviors of fashion brand marketing, considering the top20 fashion brands and investigating how they use Twitter and Instagram by observing their native profiles. They analyze their styles and strategies of advertisement. The authors employ both linguistic and computer vision techniques while comparing and contrasting strategic idiosyncrasies. They also analyze brand audience retention and social engagement hence providing suggestions in adapting advertising and marketing strategies over Twitter and Instagram.

The study of Kim and Ko [26] set out to identify attributes of social media marketing (SMM) activities and examine the relationships among those perceived activities, value equity, relationship equity, brand equity, customer equity, and purchase intention through a structural equation model. Five constructs of perceived SSM activities of luxury fashion brands are entertainment, interaction, trendiness, customization, and word of mouth. Their effects on value equity, relationship equity, and brand equity are significantly positive. For the relationship between customer equity drivers and customer equity, brand equity has significant negative effect on customer equity while value equity and relationship equity show no significant effect. As for purchase intention, value equity and relationship equity had significant positive effects, while relationship equity had

no significant influence. Finally, the relationship between purchase intention and customer equity has significance.

The findings of de Vries et al. [15] show that different drivers influence the number of likes and the number of comments. Namely, vivid and interactive brand post characteristics enhance the number of likes. Moreover, the share of positive comments on a brand post is positively related to the number of likes. The number of comments can be enhanced by the interactive brand post characteristic, a question. The shares of both positive and negative comments are positively related to the number of comments.

An analysis from Chrisler et al. [14] was conducted of 977 tweets sent immediately before and during the 2011 Victoria's Secret Fashion Show that reference the show. Although the majority were idiosyncratic remarks, many tweets contain evidence of upward social omparisons to the fashion models. The authors say that there were tweets about body image, eating disorders, weight, desires for food or alcohol, and thoughts about self-harm. The results support social comparison theory, and suggest that vulnerable viewers could experience negative affect, or even engage in harmful behaviors, during or after viewing the show or others like it.

An article from Entwistle and Rocamora1 [17], based on two studies of the fashion industry, examines one of its key institutions, London Fashion Week (LFW). The authors argue that this event is a materialization of the field of fashion. They examine how LFW renders visible the boundaries, relational positions, capital and habitus at play in the field, reproducing critical divisions within it. As well as making visible the field, LFW is a ceremony of consecration within it that contributes to its reproduction. The central aim of this article is to develop an empirically grounded sense of field, reconciling this macro-structural concept with embodied and situated reality.

Finally, Okada et al. [34] develop a motion capture system using two cameras that is capable of estimating a constrained set of human postures in real time. They first obtain a 3D shape model of a person to be tracked and create a posture dictionary consisting of many posture examples. The posture is estimated by hierarchically matching silhouettes generated by projecting the 3D shape model deformed to have the dictionary poses onto the image plane with the observed silhouette in the current image. Based on this method, the authors have also developed a virtual fashion show system that renders a computer graphics-model moving synchronously to a real fashion model, but wearing different clothes.

## 2.2 Analysis of authors attributes

Cha et al. [11], using a large amount of data collected from Twitter, present an in-depth comparison of three measures of influence: indegree (the number of followers of a user, directly indicating the size of the audience for that user), retweets

(which they measure through the number of retweets containing one's name, indicating the ability of that user to generate content with pass-along value), and mentions (which they measure through the number of mentions containing one's name, indicating the ability of that user to engage others in a conversation). Based on these measures, they investigate the dynamics of user influence across topics and time, making several interesting observations. First, popular users who have high indegree are not necessarily influential in terms of spawning retweets or mentions. Second, most influential users can hold significant influence over a variety of topics. Third, influence is not gained spontaneously or accidentally, but through concerted effort such as limiting tweets to a single topic.

In Bakshy et al. [4]'s paper the authors investigate the attributes and relative influence of 1.6M Twitter users by tracking 74 million diffusion events that took place on the Twitter follower graph over a two month interval in 2009. They find that the largest cascades tend to be generated by users who have been influential in the past and who have a large number of followers. In spite of these intuitive results, however, they also find that predictions of which particular user or URL will generate large cascades are relatively unreliable. They conclude, therefore, that word-of-mouth diffusion can only be harnessed reliably by targeting large numbers of potential influencers, thereby capturing average effects. Finally, they consider a family of hypothetical marketing strategies, defined by the relative cost of identifying versus compensating potential "influencers". The results show that although under some circumstances, the most influential users are also the most cost-effective, under a wide range of plausible assumptions the most cost-effective performance can be realized using "ordinary influencers" - individuals who exert average or even less-than-average influence.

Finally, Kwaw et al. [28] have the goal of studying the topological characteristics of Twitter and its power as a new medium of information sharing. They have crawled Twitter and obtained 41.7M user profiles, 147B social relations, 4262 trending topics, and 106M tweets. In its follower-following topology analysis they have found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks. In order to identify influentials on Twitter, they have ranked users by the number of followers and by PageRank and found two rankings to be similar. Ranking by retweets differs from the previous two rankings, indicating a gap in influence inferred from the number of followers and that from the popularity of one's tweets. They have analyzed the tweets of top trending topics and reported on their temporal behavior and user participation. They have classified the trending topics based on the active period and the tweets and show that the majority (over 85%) of topics are headline news or persistent news in nature.

## 2.3 Social media event response

In this section we report some works that take into account the social websites as medium of diffusion of popular events.

Guan et al. [18] select 21 hot events, which were widely discussed on Sina Weibo[1] in 2011, and empirically analyze their posting and reposting characteristics. In comparison to Twitter's hot topics, the authors find that the reposting ratio of event-related weibos on Sina Weibo is much higher. Other relevant findings include that males are more actively involved in these hot events than females. Then for each event, they divide related weibos into original ones and reposting ones, and analyze their characteristics for different factors, namely picture, URL, gender and verification status. It is found that, for each event, the proportion of verified users in original weibos is significantly higher than that in reposting ones. In most of the 21 events, picture containing original weibos are more likely to be reposted, while those with URLs are less likely. Another finding is that original weibos posted by verified users are more likely to be reposted, while the gender factor has little effect on reposting likelihood. It is also shown that the distribution of reposting times fits a power law and the reposting depths are exponentially distributed.

Becker et al. have presented two relevant works in this field. In the first one [5], by automatically identifying events and their associated user-contributed social media documents, the authors show how they can enable event browsing and search in a search engine. They exploit the rich context associated with social media content, including user-provided annotations (e.g., title, tags) and automatically generated information (e.g., content creation time). Using this rich context, which includes both textual and non-textual features, they can define appropriate document similarity metrics to enable online clustering of media to events. As a key contribution of this paper, the authors explore a variety of techniques for learning multi-feature similarity metrics for social media documents in a principled manner. They evaluate their own techniques on large-scale, real-world data-sets of event images from Flickr[2], an Internet image community website.

In the second work [6], Becker et al. underline how user-contributed messages on social media sites such as Twitter have emerged as powerful, real-time means of information sharing on the Web. The authors explore approaches for analyzing the stream of Twitter messages to distinguish between messages about real-world events and non-event messages. The adopted approach relies on a rich family of aggregate statistics of topically similar message clusters, including temporal, social, topical, and Twitter-centric features.

The focus of the paper from Chen and Roy [13] is to detect events from pho-

[1] http://www.weibo.com/
[2] https://www.flickr.com/

tos on Flickr. The problem is challenging considering: (1) Flickr data is noisy, because there are photos unrelated to real-world events; (2) it is not easy to capture the content of photos. This paper presents the authors' effort in detecting events from Flickr photos by exploiting the tags supplied by users to annotate photos. In particular, the temporal and locational distributions of tag usage are analyzed in the first place. Then, they identify tags related with events, and further distinguish between tags of aperiodic events and those of periodic events. Afterwards, event-related tags are clustered such that each cluster, representing an event, consists of tags with similar temporal and locational distribution patterns as well as with similar associated photos. Finally, for each tag cluster, photos corresponding to the represented event are extracted.

The problem of event summarization using tweets is well faced by Chakrabarti and Punera [12], where they argue that for some highly structured and recurring events, such as sports, it is better to use sophisticated techniques to summarize the relevant tweets. The authors formalize the problem of summarizing event-tweets and give a solution based on learning the underlying hidden state representation of the event via Hidden Markov Models.

Calabrese et al. [9], adding the information given by cell-phone traces, deal with the analysis of crowd mobility during special events. They analyze nearly 1 million cell-phone traces and associate their destinations with social events. They show that the origins of people attending an event are strongly correlated to the type of event, with implications in city management, since the knowledge of additive flows can be a critical information on which to take decisions about events management and congestion mitigation.

Finally, Arcaini et al. [2] propose a procedure consisting of a first collection phase of social network messages, a subsequent user query selection, and finally a clustering phase, for performing a geographic and temporal exploration of a collection of items, in order to reveal and map their latent spatio-temporal structure. Specifically, both several geo-temporal distance measures and a density-based geo-temporal clustering algorithm are proposed. The approach can be applied to social messages containing an explicit geographic and temporal location. The algorithm usage is exemplified to identify geographic regions where many geotagged Twitter messages about an event of interest have been created, possibly in the same time period in the case of non-periodic events, or at regular timestamps in the case of periodic events. This allows discovering the spatio-temporal periodic and non-periodic characteristics of events occurring in specific geographic areas, and thus increasing the awareness of decision makers who are in charge of territorial planning.

# Chapter 3

# Analysis of the correlation between different authors attributes

## 3.1 Introduction

Once we have built the data set of social media posts, which are the tweets from Twitter and the photos with their description from Instagram, we can focus our analysis on the authors of these messages. We call authors all the Twitter or Instagram users that shared a specific media (stored in the data set) with their social network. Each author is characterized by some attributes and features: some of these attributes are simply extracted from the social media profile informations, some other are computed from these data. A nice thing to do now is a correlation analysis between some of these measures, trying to find something interesting and meaningful, with respect to the domain of study.

Firstly, let's specify these measures, recalling that each one of them is author-specific:

1. *Generated post volume*, the total number of post collected in our database, generated by the author;

2. *Popularity score*, a new measure defined as the sum of all the likes and all the comments in the Instagram scenario, retweets in the Twitter scenario, received by the author in the posts stored in our database;

3. *Average popularity score*, the same as before but considering the average number of likes and comments/retweets received in all the photo shared by the author and stored in the database;

4. *Strength score*, a measure referring to the power of the author to influence and to reach other people, computed as follows:

$$strength_i = \sqrt{follower_i} + 100 \times \mathbb{1}(verified_i) \tag{3.1}$$

with:

$$\mathbb{1}(verified_i) = \begin{cases} 1 & \text{if } verified_i \text{ is True} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

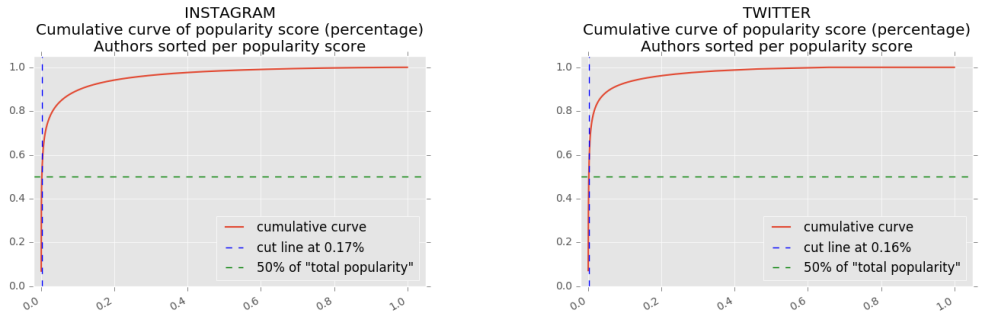
The attributes $verified$ is a field present only on Twitter that specifies whether the identity of the user has been verified by Twitter administrators; it is set to True only for "vip" authors and other relevant identities. We have to notice that in the Instagram scenario we do not have the $verified$ attributes, then the strength score is taking into account only the square root of the follower count.

Now we can study the cumulative curves of these measures (generated volume, popularity score and strength). This type of curve is obtained sorting the authors by the considered attribute and putting on y-axis the cumulative value of the attribute for each user, from the top authors to the low ones. Looking at the next Figures (from 3.1 to 3.3) we can notice how a very small amount of users contribute to the 50% of the considered attribute.

Looking at the popularity curves, we have a similar behavior in the two social networks: the cut line for Instagram is at 0.17% and for Twitter is at 0.16%, because of the big differences in this attribute for the users on both the two platforms. In fact, we have a lot of users (8372 of 23767, around the 35%) who never received likes or retweets on Twitter, and on Instagram about the 33% of users has the popularity score lower than 40, while the top 10 authors for this measure has a cumulative popularity score of 20556083, with an average of 2055608 popularity score per user, that is really far from the score of 40 we were taking into account before.

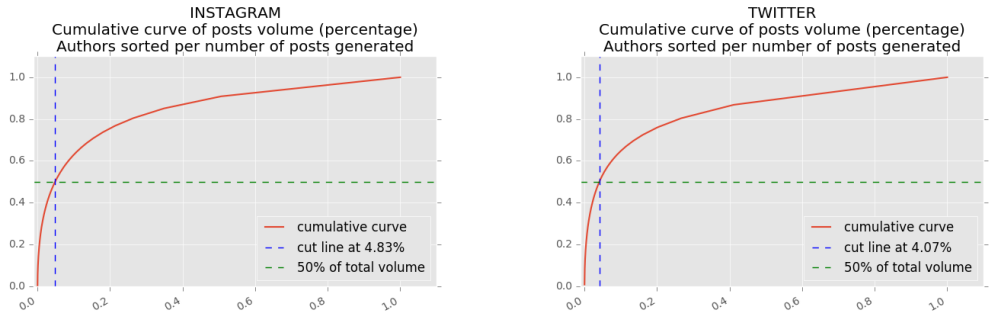
Again, we have similar results for the generated posts volume, but with much less power in the hands of the top authors, with the cut lines around 4-5%.

Finally, the elbow is smoothing on the strength score cumulative curves, even if with some differences. Indeed, Twitter seems to be more an oligarchic reign than Instagram, with cut lines at 7.82% and 14.58%, respectively.



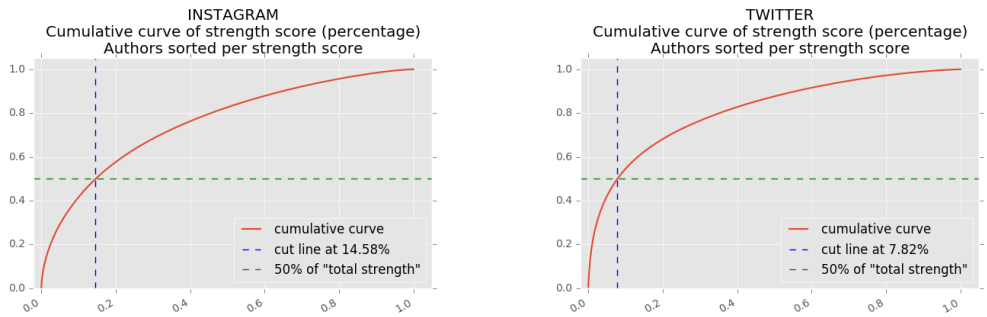

(a) (b)

Figure 3.1: Cumulative curve of popularity score on Instagram (a) and Twitter (b). The curve shapes are very similar, showing how a little bunch of users received most of all the likes and comments or retweets.




(a) (b)

Figure 3.2: Cumulative curve of generated posts volume on Instagram (a) and Twitter (b). Also in this case, the curve shapes are very similar, with cut lines at 4.83% for Instagram and 4.07% for Twitter.




(a) (b)

Figure 3.3: Cumulative curve of strength score on Instagram (a) and Twitter (b). Now the two shapes are a bit dissimilar, with a more oligarchic reign in Twitter, and a smoother elbow on Instagram.

## 3.2 Method

In order to study the correlation between the different measures related to each author, we adopt two rank correlation coefficients from statistics. Indeed, the knowledge has to be extracted from the rank of the measures, studying the similarities among them. These two coefficients are:

1. *Spearman rank correlation coefficient* ($\rho$)[42]: a nonparametric measure of rank correlation (statistical dependence between the ranking of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. Suppose there are $n$ pairs of observations from continuous distributions. Rank the observations in the two samples separately from smallest to largest. Equal observations are assigned the mean rank for their positions. Let $u_i$ be the rank of the $i^{th}$ observation in the first sample and let $v_i$ be the rank of the $i^{th}$ observation in the second sample. Spearman's rank correlation coefficient $\rho$ is computed as follows:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} u_i - v_i}{n(n^2 - 1)} \tag{3.3}$$

2. *Kendall rank correlation coefficient* ($\tau$)[42, 27]: a statistic used to measure the ordinal association between two measured quantities. A tau test is a non-parametric hypothesis test for statistical dependence based on the tau coefficient. It is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities. The definition of Kendall's $\tau$ that is used is:

$$\tau = \frac{(P - Q)}{\sqrt{(P + Q + T)(P + Q + U)}} \tag{3.4}$$

   where P is the number of concordant pairs, Q the number of discordant pairs, T the number of ties only in the first ranking, and U the number of ties only in the second ranking. If a tie occurs for the same pair in both the first and the second ranking, it is not added to either T or U.

Intuitively, both the Spearman's $\rho$ and the Kendall's $\tau$ will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: $1^{st}$, $2^{nd}$, $3^{rd}$, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of -1) rank between the two variables.

## 3.3 Findings

In this chapter we report all the results coming from the analysis that we have done, and we prefer to fork the discussion in two different paths, one for Twitter

and one for Instagram, due to significant differences in the studied numbers, coming from the two distinct social media. In the end, we will rejoin our path, making some considerations about the comparison.

### 3.3.1 Instagram

Three different subsets from the database are analyzed, in order to capture more meaningful results.

The first is the one corresponding to the entire set of posts stored in the database, with no filters at all. In this way, we have 556045 media coming from 102715 different authors. Maybe, considering all the authors together we are influencing the analysis because of all those authors that generated a very small amount of media. In Figure 3.4 we can see the number of authors as a function of the popularity score obtained, considering only the authors of 2 or 1 post (from now on, we will call these users Small Authors). We can notice that the most of them are not so popular, with a mode of 1101 authors at 21 popularity score. Therefore, we decided to create a second subset (the first real subset, since the previous one was considering the entire set), removing from the analysis the Small Authors that are under the mode of the distribution curve in Figure 3.4, i.e. considering only those authors with a generated posts volume greater than 2 or a generated posts volume less or equal than 2 but with a popularity score at least of 21. In this way, we obtain 87108 Nice Authors, thus we are reducing the set of authors of 15607 users. Analyzing the results obtained in the correlation coefficients and making a comparison between the two different sets, we were not so satisfied because:

1. The correlations are not so strong;

2. The Nice Authors are not so 'nice' in terms of results, not showing many differences from the whole unfiltered set.

Then, another set was created, always in order to reduce the noise coming from Small Authors, with a new filter on the popularity score of each single post: cutting out all the posts with number of comments plus number of likes below 20, and then considering only the authors with at least 2 posts generated or less or equal than 2 but with a popularity score at least of 50, we are now considering a new set of Cool Authors, counting 57603 users.

In the next pages, we report all the scatter plots referring to the different correlation analysis and a table summarizing the value of the coefficients in the different cases. The plots are in a logarithmic scale in order to better understand and visualize the different behaviors.

From the first column we can notice a strong positive correlation between the number of posts generated and the number of likes and comments received overall. This reinforce the trivial law that says: the more you publish, the more

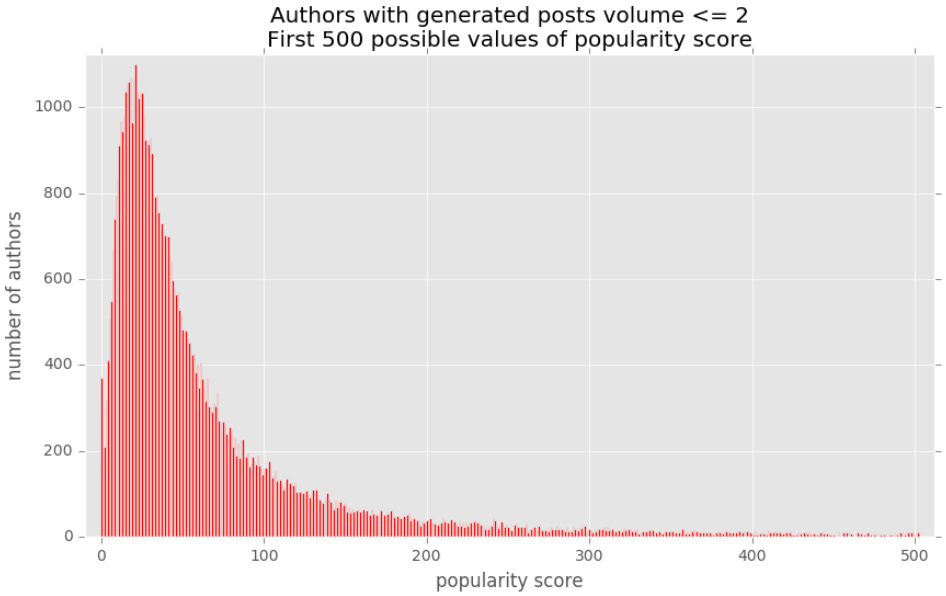



Figure 3.4: Considering the authors with generated posts volume less or equal than 2, this histogram represents the distribution of the numbers of authors with respect to the popularity score. The higher pick, i.e., the mode of the distribution, is at 21 popularity score, counting 1101 users.

you receive likes. This kind of law is really strong here, for the Instagram scenario, stronger than the Twitter scenario that we will see later.

The second column reports the correlation coefficients for the two measure of number of posts shared and average popularity score. In this case, we can notice the nearly absence of correlation in the first two set, while in the last set we are going in the direction of small negative correlation. Then we realize that with the increasing in the number of contents shared on the Instagram social media, we are also reducing the average number of likes and comments per post: summing up also the results of the previous column, we can conceive about:

- The presence of big authors, that publish a lot of media and, overall, get a lot of feedback (in terms of likes and comments) from the social network they have, but maybe in some post they receive a relative poor feedback;
- The presence of small authors, with a bunch of posts in our database, but with a lot of likes and comments collected in these few media.

The third column shows us that, for the database we have built and then the media we have collected, the strength score, which represents best the influential power of the user, is not correlated at all with the number of posts generated in this specific domain. Hence, we have authors with a lot of influential power that posts only a few media, and on the other hand we have some authors with a few followers but with a significant number of media shared on their small network.

The fourth and last column shows us another relevant aspect: in all the three different sets, the number of likes and comments related to the posts generated is positively correlated with the strength score. Then, trivially, users with more followers are more able to get more likes and comments, and they actually get them.

The pictures in the next pages will help us in the understating of the reasonings explained so far. We first attach the pictures related to the unfiltered set of all the authors collected in our database, then we append the figures related to the first filter, the Nice Authors, and finally we post the figures related to the second filter, the Cool Authors.

Table 3.1: Spearman rank correlation coefficient for different couples of attributes in the Instagram scenario.

| **Authors set** | **Generated posts volume VS Popularity score** | **Generated posts volume VS Average Popularity score** | **Generated posts volume VS Strength score** | **Popularity score VS Strength score** |
|---|---|---|---|---|
| *All together* | 0.6746 | 0.0901 | 0.1570 | 0.5698 |
| *Nice Authors* | 0.6384 | -0.1181 | 0.0528 | 0.5096 |
| *Cool Authors* | 0.6423 | -0.2791 | -0.0230 | 0.4509 |

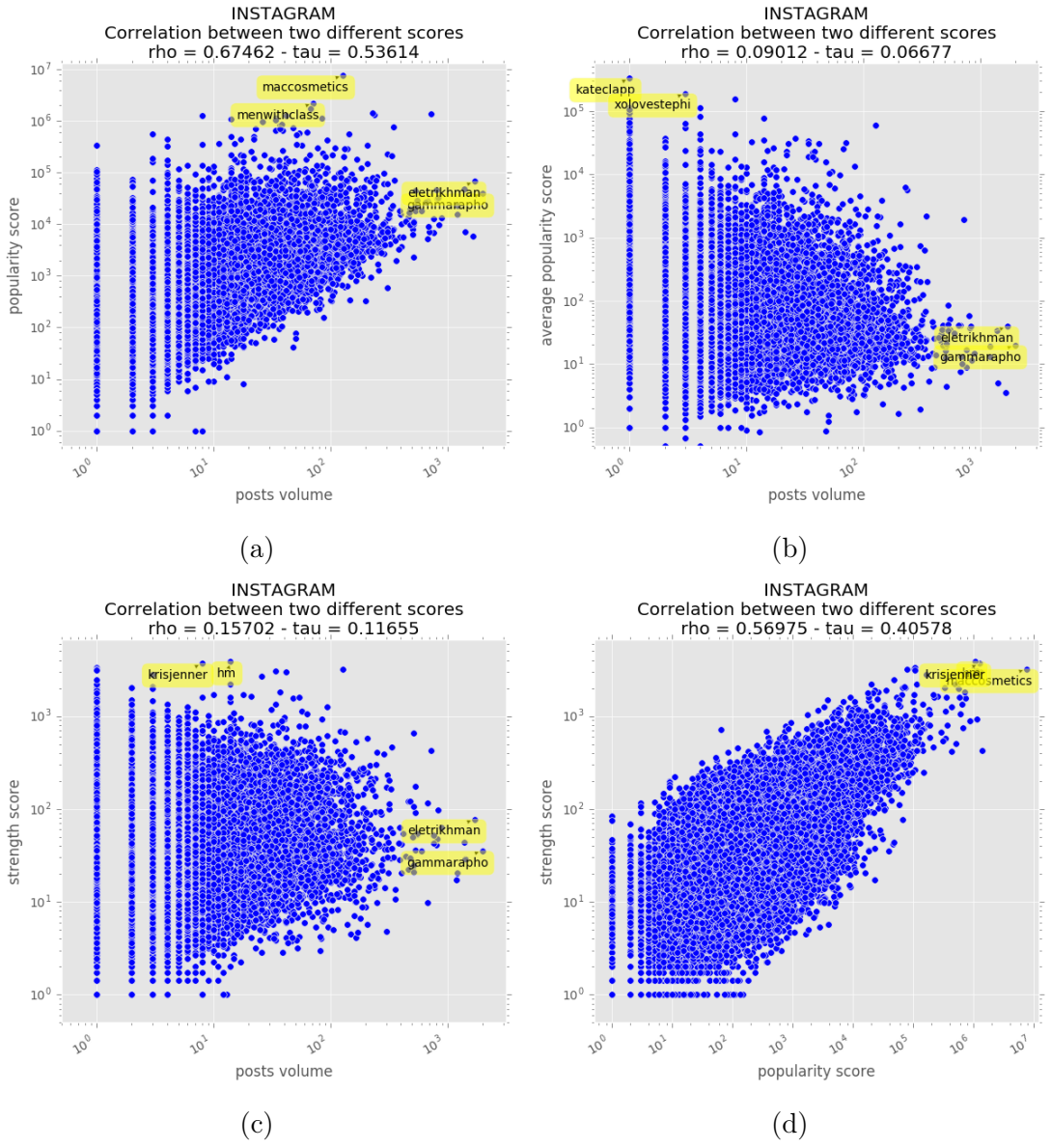



Figure 3.5: Considering all the authors, logarithmic scale scatter plots referring to the rank correlation between the different authors attributes. The figures axes report the two analyzed measures.

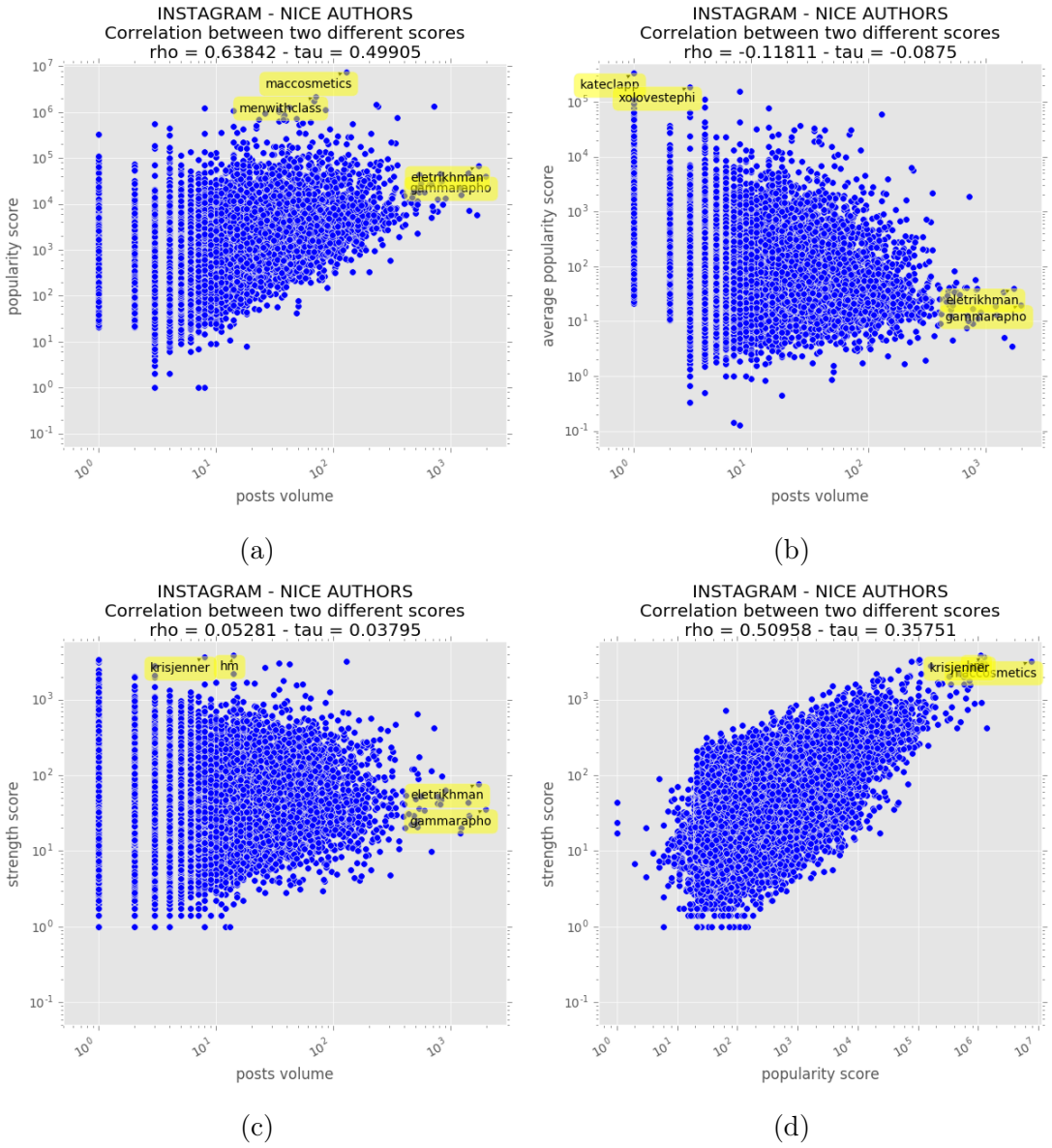



Figure 3.6: Considering the Nice Authors, logarithmic scale scatter plots referring to the rank correlation between the different authors attributes. The figures axes report the two analyzed measures.

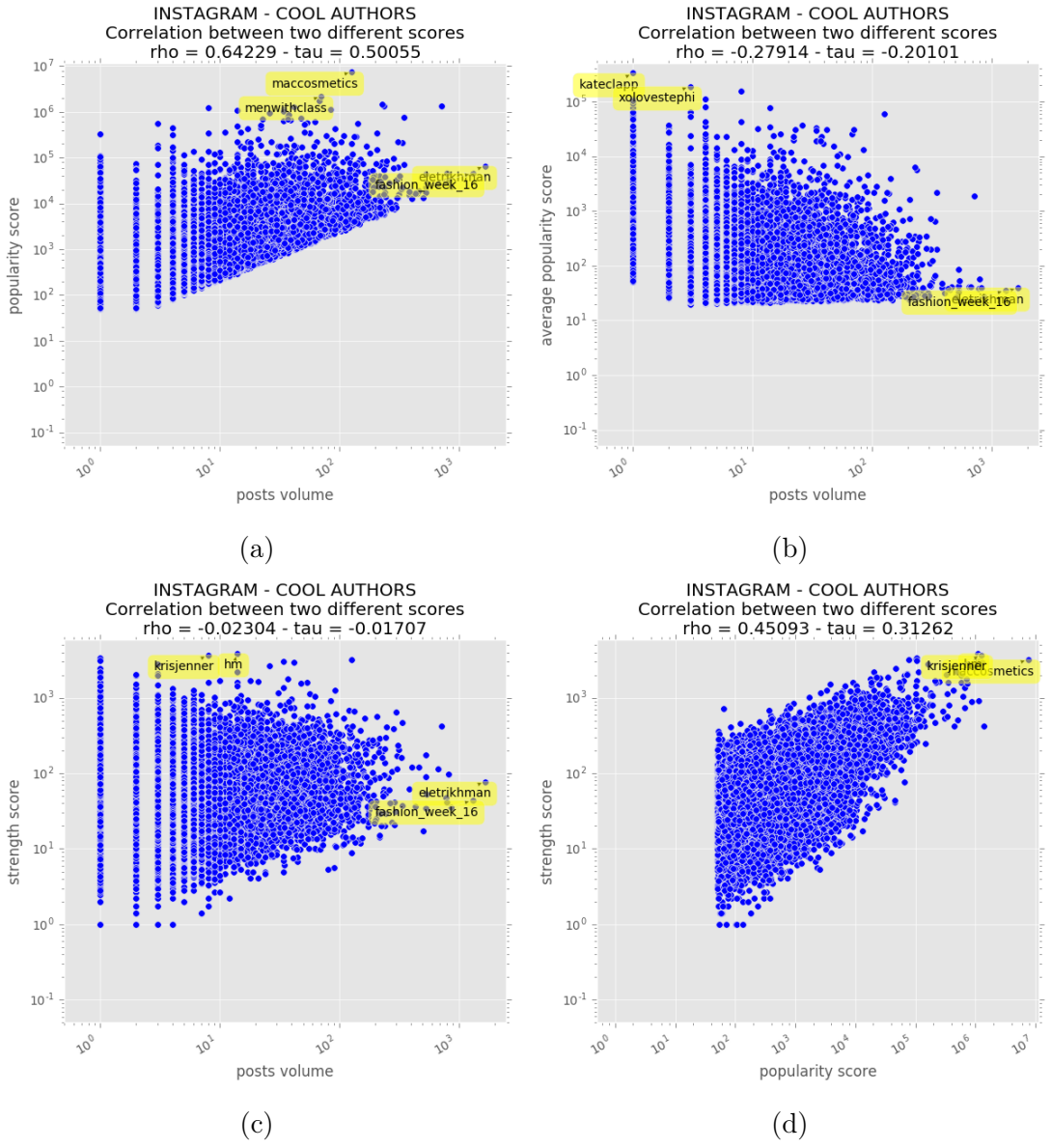



Figure 3.7: Considering the Cool Authors, logarithmic scale scatter plots referring to the rank correlation between the different authors attributes. The figures axes report the two analyzed measures.

### 3.3.2 Twitter

A very similar work has been done for Twitter, where the numbers are really different.

Also in this case, the first set includes all the 106286 tweets stored, from 23767 different authors. In Figure 3.8 we can see the number of authors as a function of the popularity score obtained, considering only the authors of 2 or 1 post (from now on, we will call these users Small Authors). We can notice that the most of them are not popular at all, with a mode of 7548 authors at 0 popularity score.

Following the same reasoning adopted for Instagram, that is much more strong here because of the presence of a lot of users taking no retweets and likes, we decided to create a second subset removing from the analysis the Small Authors with popularity score less or equal than 4, i.e. considering only those authors with a generated posts volume greater than 2 or a generated posts volume less or equal than 2 but with a popularity score at least of 5. In this way we obtain 8435 Nice Authors, thus we are reducing the set of authors of 15332 users.

Like in the Instagram scenario, analyzing the results obtained in the correlation coefficients and making a comparison between the two different sets, we were not so satisfied because:

1. The correlations are not so strong;
2. The Nice Authors are not so 'nice' in terms of results, not showing many differences from the whole unfiltered set.

Then, another set was created, with a new filter on the popularity score of each single post: cutting out all the posts with number of comments plus number of retweets below 2, and then considering only the authors with at least 2 posts generated or less or equal than 2 but with a popularity score at least of 5, we are now considering a new set of Cool Authors, counting 4681 users.

In the next pages, we report all the scatter plots referring to the different correlation analysis and a table summarizing the value of the coefficients in the different cases. The plots are in a logarithmic scale in order to better understand and visualize the different behaviors.

As before, we proceed with a column by column study on the obtained results. From the first one we can notice a strong positive correlation between the number of tweets generated and the number of likes and comments received overall. Like in the Instagram scenario, this reinforce the trivial law saying the more you publish, the more you receive likes. In this scenario, however the numbers are not so strong like the Instagram's ones, as we have anticipated before. Furthermore, we have a kind of anomaly in the Nice Authors case, due to the weakness of the applied filter and the sparsity of the points close to the ones we have pulled out of the analysis thanks to the filter.

The next column reports the correlation coefficients for the two measure of number of tweets shared and average popularity score. In this case, we can notice

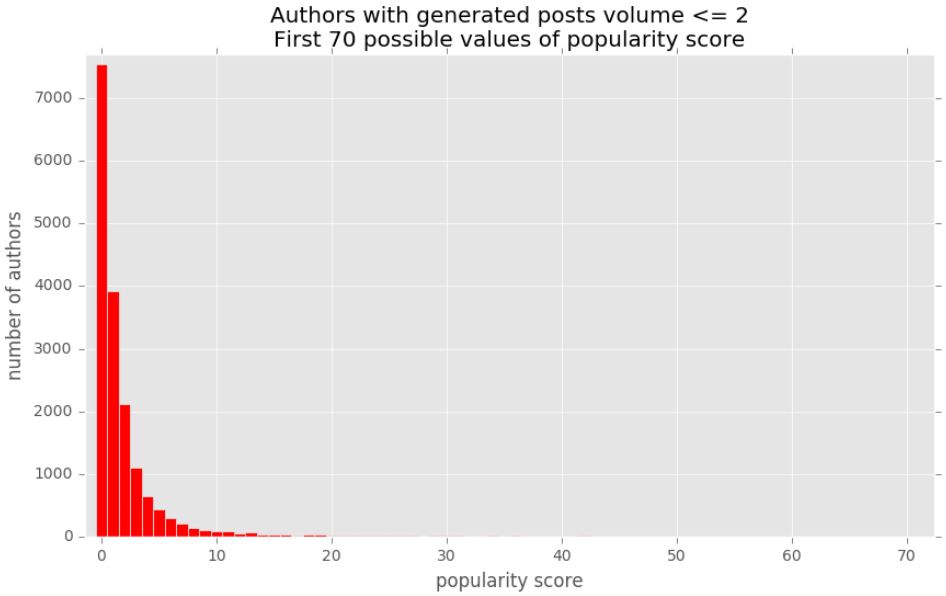



Figure 3.8: Considering the authors with generated posts volume less or equal than 2, this histogram represents the distribution of the numbers of authors with respect to the popularity score. The higher pick, i.e., the mode of the distribution, is at 0 popularity score, counting 15332 users.

the nearly absence of correlation in the first unfiltered set of authors, while in the second case and in the third case we are going in the direction of discreet negative correlation. Then, we realize that, also for the Titter scenario, with the increasing in the number of contents shared, we are also reducing the average number of likes and retweets per single post: summing up also the results of the previous column as done before for Instagram, we can conceive about:

- The presence of big authors, that publish a lot of media and, overall, get a lot of feedback (in terms of likes and comments) from the social network they have, but maybe in some post they receive a relative poor feedback;
- The presence of small authors, with a bunch of posts in our database, but with a lot of likes and comments collected in these few media.

The third column shows us that, for the database we have built and then the media we have collected, the strength score, which represents best the influential power of the user, is not correlated at all with the number of tweets generated in this specific domain. Hence, we have authors with a lot of influential power that posts only a few media, maybe because these users were not domain-specific, and on the other hand we have some authors with a few followers but with a significant number of media shared on their small network, almost surely because of their interest in this specific domain.

In the end, the fourth and last column shows us another relevant aspect: in all the three different sets, the number of likes and retweets received for the tweets generated is positively correlated with the strength score, that is the influential power of the user. Then, trivially like in the Instagram scenario, users with more followers are more able to get more likes and comments, and they actually get them.

The pictures in the next pages will help us in the understating of the reasonings explained so far. We first attach the pictures related to the unfiltered set of all the authors collected in our database, then we append the figures related to the first filter, the Nice Authors, and finally we post the figures related to the second filter, the Cool Authors.

Table 3.2: Spearman rank correlation coefficient for different couples of attributes in the Twitter scenario.

| **Authors set** | **Generated posts volume VS Popularity score** | **Generated posts volume VS Average Popularity score** | **Generated posts volume VS Strength score** | **Popularity score VS Strength score** |
|---|---|---|---|---|
| *All together* | 0.4650 | 0.1140 | 0.1104 | 0.3104 |
| *Nice Authors* | 0.0969 | -0.4655 | -0.1169 | 0.4593 |
| *Cool Authors* | 0.5543 | -0.3446 | 0.0771 | 0.4172 |

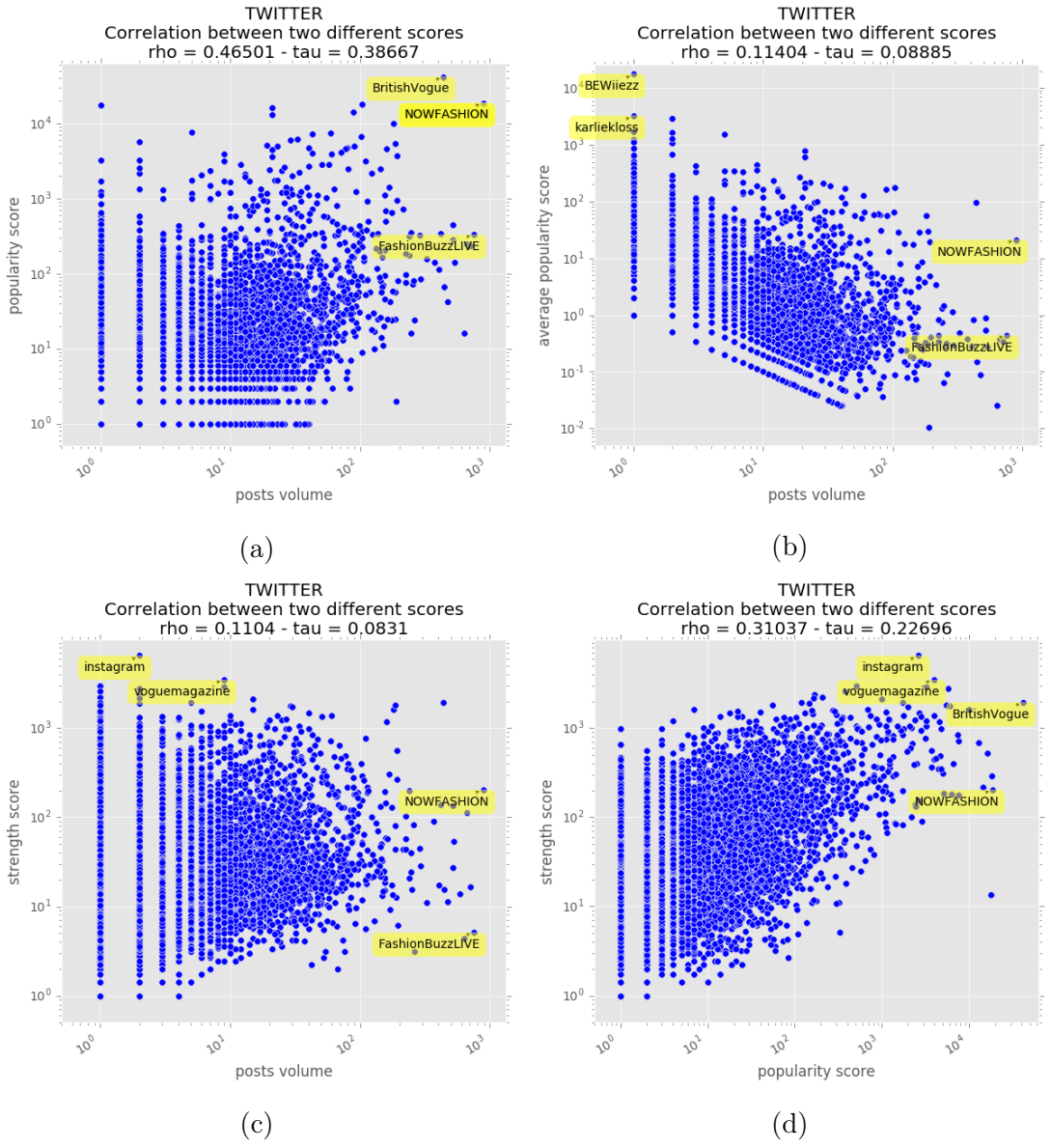



Figure 3.9: Considering all the authors, logarithmic scale scatter plots referring to the rank correlation between the different authors attributes. The figures axes report the two analyzed measures.

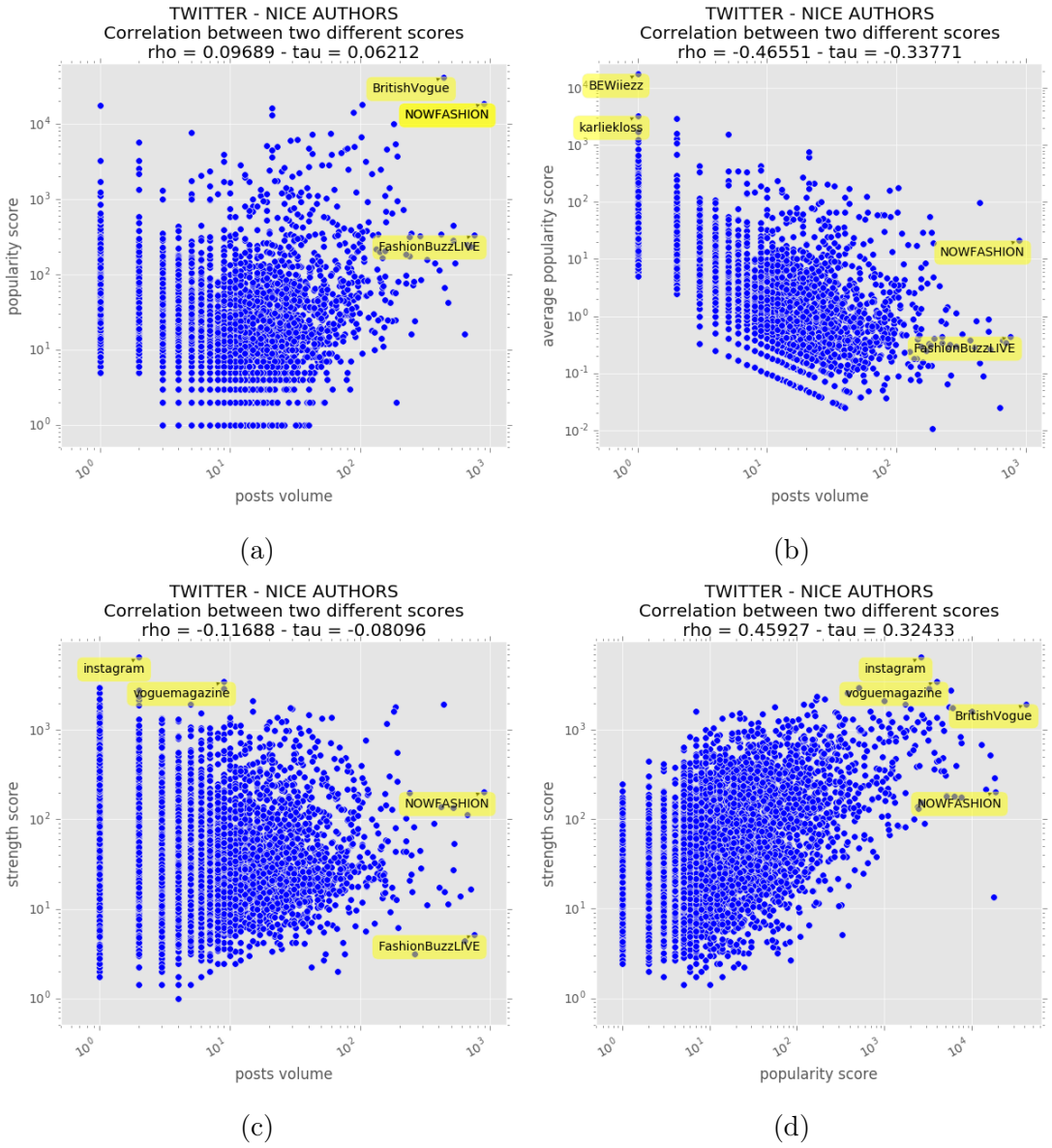



Figure 3.10: Considering the Nice Authors, logarithmic scale scatter plots referring to the rank correlation between the different authors attributes. The figures axes report the two analyzed measures.

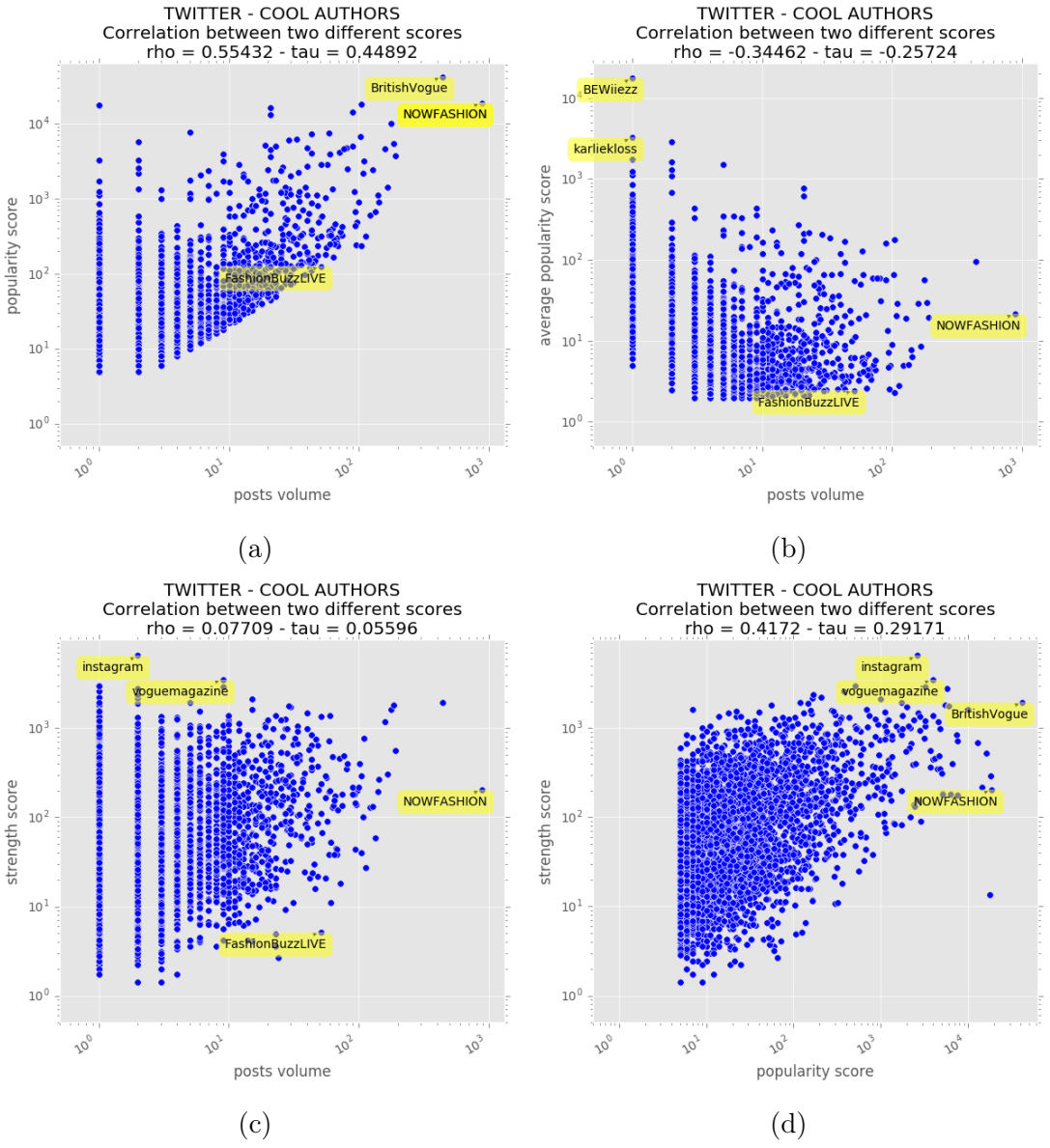



Figure 3.11: Considering the Cool Authors, logarithmic scale scatter plots referring to the rank correlation between the different authors attributes. The figures axes report the two analyzed measures.

### 3.3.3 Comparison between the two social media

In this last section, we provide a brief comparison of the results discussed in the previous two sections, in particular reading in parallel the two Tables 3.1 and 3.2. Like in the two distinct studies, we proceed with the following column by column analysis, merging the results when possible.

The first two columns report us a common behavior for the two scenarios, from which we can extrapolate the trivial law of the more you publish, the more you receive likes. This law is a bit more respected in the Instagram world, but also in the Twitter one the correlation of the two distinct measures is positively high.

Also the second columns show similar correlation coefficients: for both the social media, the number of posts and the average popularity feedback, in terms of likes and comments or retweets, are weakly and negatively correlated.

The strength score, which represents the influential power of the user, is not correlated at all with the number of posts generated in this specific domain, as shown by the third columns, and also in this case we have concordance in results from Twitter and Instagram.

In the end, the fourth and last columns show that the number of likes and comments or retweets related to the posts generated is positively correlated with the strength score, that is the number of followers of the user. Then, users with more followers have more possibility to get more likes and feedback, and they actually get them, both in Twitter and in Instagram.

# Chapter 4

# Time response analysis

## 4.1 Introduction

In this chapter we want to focus on the different types of response that the social media have shown in terms of generated posts volume with respect to the scheduled events in the Fashion Week calendar. Describing the problem with more details, we can say that, in this type of analysis, we have two different temporal signals, or time series, one for the calendar events and the other one for the volumes of social media posts on the web, and we try to analyse a sort of causality between the two curves. This kind of work has been done brand by brand, referring only to one brand at a time and then selecting only the Milano Fashion Week events and the social media posts related to that specific brand

The next study is about the results of the statistics tests performed before, in order to find the causality between the two signals. We try to perform a simple clustering on these results with the purpose of providing similarities in the response between different events.

The final problem we face is to consider the cluster labelling computed before and take into account the supervised machine learning problem of classifying each brand-event tuple (with all their relevant attributes) into the cluster labelling performed with the unsupervised learning technique of clustering adopted before.

## 4.2 Method

The main hard works in this chapter are the causality tests between two different time series, the clustering between the results obtained in the previous section and the ideation of a prediction model that could forecast the type of response for each event.

In order to perform the causality tests, the first thing we need is a couple of time series, that from now on we will call also "signals". Each test is performed on a single brand, and so we need to select from the calendar and the dataset

of social posts only the events and the posts related to a specific brand. Firstly, we have to define the calendar signal: for each event in the schedule we have its start time and its duration, then we define the calendar time series as:

$$calendarSignal(\Delta t) = \sum_{event \in Calendar} \mathbb{1}_{event}(\Delta t) \tag{4.1}$$

where:

- $\Delta$t is the time interval of analysis;
- $Calendar$ is the set of all the events in the schedule;
- $\mathbb{1}_{event}(\Delta t)$ is the indicator function for the specific event in the time window $\Delta$t, i.e. the presence of that event in that time window.

If this work is easy on the calendar signal, indeed it is enough to select only the events organized by a specific brand, it is a bit more difficult on the social signal, defined as the number of posts in the specific time window. In this case we used significant brand-specific regular expressions: if the matching with the post text is positive, then the post is assumed to be related to that specific brand.

### 4.2.1 Predictive causality test

Now that we have the technique to obtain the time series of each brand, we can perform the predictive causality tests. We want to find some causality relationship between the events calendar signal and the social media signal and the adopted method is a series of Granger Causality tests between the two different time series.

The null hypothesis for the test is that the calendar signal (that will be denoted as $x_2$) does not Granger cause the social media signal (that will be denoted as $x_1$). Granger causality means that past values of $x_2$ have a statistically significant effect on the current value of $x_1$, taking past values of $x_1$ into account as regressors. We reject the null hypothesis that $x_2$ does not Granger cause $x_1$ if the p-values are below a desired size of the test. In other words, a time series $x_2$ is said to Granger-cause $x_1$ if it can be shown, usually through a series of t-tests and F-tests on lagged values of $x_2$, that those $x_2$ values provide statistically significant information about future values of $x_1$.

We focus on the F-test performed by the function "grangercausalitytests"[1], recalling that an F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. The test statistic in an F-test is the ratio of two scaled sums of squares reflecting different sources of variability. These sums of squares are constructed so that the statistic tends to be greater when the null hypothesis is not true.

[1] http://statsmodels.sourceforge.net/0.6.0/generated/statsmodels.tsa.stattools

### 4.2.2 Clustering on the tests results

Once we have collected all the results from the F-tests, we try to perform a clustering among them. The adopted technique in this section is a standard k-means clustering. The k-means algorithm clusters data by trying to separate samples in $k$ groups of equal variance, minimizing a criterion known as the within-cluster sum-of-squares or inertia. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields.

The k-means algorithm divides a set of $N$ samples $X$ into $K$ disjoint clusters $C$, each described by the mean $\mu_j$ of the samples in the cluster. The means are commonly called the cluster "centroids"; note that they are not, in general, points from $X$, although they live in the same space. The k-means algorithm aims to choose centroids that minimise the within-cluster sum of squared criterion:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (\| x_i - \mu_j \|^2). \tag{4.2}$$

Inertia, or the within-cluster sum of squares criterion, can be recognized as a measure of how internally coherent clusters are. It suffers from various drawbacks:

- Inertia makes the assumption that clusters are convex and isotropic, which is not always the case. It responds poorly to elongated clusters, or manifolds with irregular shapes;

- Inertia is not a normalized metric: we just know that lower values are better and zero is optimal. But in very high-dimensional spaces, Euclidean distances tend to become inflated (this is an instance of the so-called "curse of dimensionality").

In basic terms, the algorithm has three steps:

1. Initialization step: the first step chooses the initial centroids, with the most basic method being to choose $k$ samples from the dataset $X$.

   After initialization, k-means consists of looping between the two other steps:

2. Expectation step: the first phase in the repeat loop assigns each sample to its nearest centroid, i.e. where it is expected to be assigned;

3. Maximization step: the second phase in the loop creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids are computed and the algorithm repeats these loop until this value is less than a threshold. In other words, it repeats until the centroids do not move significantly.

### 4.2.3 Classification problem

Now that each event is labelled with the cluster which it belongs, we can perform the supervised problem of predicting the labelling itself, adopting a validation strategy. We are trying to build a prediction model for the cluster labelling defined from the Granger test result, given the simple information related to brands and events. We adopt different techniques in order to do this, therefor we briefly introduce all of them.

#### Naive Bayes classifier for multivariate Bernoulli models

The first method we approach is a Naive Bayes classifier, specific for the case of multivariate Bernoulli models. It implements the Naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a Bernoulli Naive Bayes instance may binarize its input (depending on the binarize parameter).

The decision rule for Bernoulli Naive Bayes is based on:

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i). \tag{4.3}$$

#### Logistic Regression

The second model we approach comes from the linear ones, that is regularized Logistic Regression, which is very specific for probability estimation problem. In this model, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. The used implementation can fit a multiclass (with one-vs-rest paradigm) logistic regression with optional regularization, implied by the parameter $C$.

As an optimization problem, binary class norm-2 penalized logistic regression minimizes the following cost function:

$$\min_{w,C} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \log(\exp(-y_i(\mathbf{X}_i^T w + c)) + 1). \tag{4.4}$$

#### Cross-Validated Logistic Regression

We also try to improve the Logistic Regression performances through Cross-Validation. The function is given 10 values in a logarithmic scale between 0.0001 and 10000 for the parameter C, the best hyperparameter is selected by the cross-validator Stratified-KFold. For a multiclass problem, the hyperparameters for

each class are computed using the best scores got by doing a one-vs-rest in parallel across all folds and classes. Hence this is not the true multinomial loss.

### Support Vector Machine

Support vector machines are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

- Effectiveness in high dimensional spaces;
- Still effective in cases where number of dimensions is greater than the number of samples;
- Use of a subset of training points in the decision function (called support vectors), so it is also memory efficient;
- Versatility: different Kernel functions can be specified for the decision function.

We adopt a One-versus-Rest strategy, then the problem can be formulated at each iteration as a 2-class problem. Given training vectors $x_i \in \mathbb{R}^p$, with $i = 1, ..., n$, in two classes, and a vector $y \in \{1, -1\}^n$, the Support Vector Classifier solves the following primal problem:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \\ &\text{subject to:} \qquad y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ &\qquad\qquad\quad \xi_i \geq 0, \ i = 1, ..., n. \end{aligned} \tag{4.5}$$

Its dual is:

$$\begin{aligned} &\min_{a} \frac{1}{2} a^T \mathbf{Q} a - e^T a \\ &\text{subject to:} \qquad y^T a = 0, \\ &\qquad\qquad\quad 0 \leq a_i \leq C, \ i = 1, ..., n. \end{aligned} \tag{4.6}$$

where $e$ is the vector of all ones, $C > 0$ is the upper bound, $\mathbf{Q}$ is an $n \times n$ positive semidefinite matrix, $\mathbf{Q}_{ij} = y_i y_j K\langle x_i, x_j \rangle$, where $K\langle x_i, x_j \rangle = \phi(x_i)^T \phi(x_i)$ is the kernel function. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function $\phi$.

The decision function is:

$$sign(\sum_{i=1}^{n} y_i a_i K\langle x_i, x \rangle + b) \tag{4.7}$$

### Decision Tree

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Some advantages of decision trees are:

- Simplicity in the understanding and in the interpretation. Trees can be visualised;
- Ability to handle both numerical and categorical data;
- Use of a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic;
- Good performance even if its assumptions are somewhat violated by the true model from which the data were generated.

The disadvantages of decision trees include:

- Decision-tree learners can create over-complex trees that do not generalise the data well. This is called overfitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem;
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

### Random Forest

This is a perturb-and-combine technique specifically designed for trees. This means a diverse set of classifiers is created by introducing randomness in the classifier construction. The prediction of the ensemble is given as the averaged prediction of the individual classifiers.

In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.

#### Baselines

We build also some simple baseline models, in order to make a smart comparison and to better understand the performances of the cleverer models, described previously. In particular, we try two different dummy classifiers:

1. Most frequent strategy: it always predicts the most frequent label in the training set;
2. Stratified strategy: it generates predictions by respecting the training set's class distribution;
3. Random strategy: it generates predictions uniformly at random.

## 4.3 Findings

In this chapter, the analysis is done only for the Instagram scenario, due to the bigger volume of posts obtained and to the complexity of making a comparison between the two different social media, also because of the complexity of the work itself.

### 4.3.1 Predictive causality test

We want to find some causality relationship between the events calendar signal and the social media signal. In order to be more clear, Figure 4.1 shows the social response to the event of *Versace* of $26^{th}$ February at 20:00.

From now on, we state the following: we have no preferences of any kind in choosing some brands with respect to others; we will use different examples that clearly show the messages we want to report, picking different brands "more or less randomly".

From this first example we can notice a strong reaction in the social media relatively close to the scheduled events: indeed, we have a peak of about 180 Instagram posts in the time window starting when the event is just done, and then the number of posts per time window decreases rapidly.

In Figure 4.2 the same graph is reported but this time for *Prada*: we can notice how, for this specific example, the response is much more fast, and the peak on Instagram almost overlaps with the presence of the event in the schedule.

Keeping in mind these two graphs, we can have a look at the results of the Granger Causality tests performed for these two brands, expecting two slightly different outcomes. We report in Figure 4.3 the F-tests results with tables for some detailed numbers. We recall that Granger Causality is studied by performing some statistic tests (in this case we analyse F-tests) over two time-series, with different lags between them. In this study, the lag unit is the same as the granularity adopted before, that is 15 minutes. The outcomes of the statistical tests

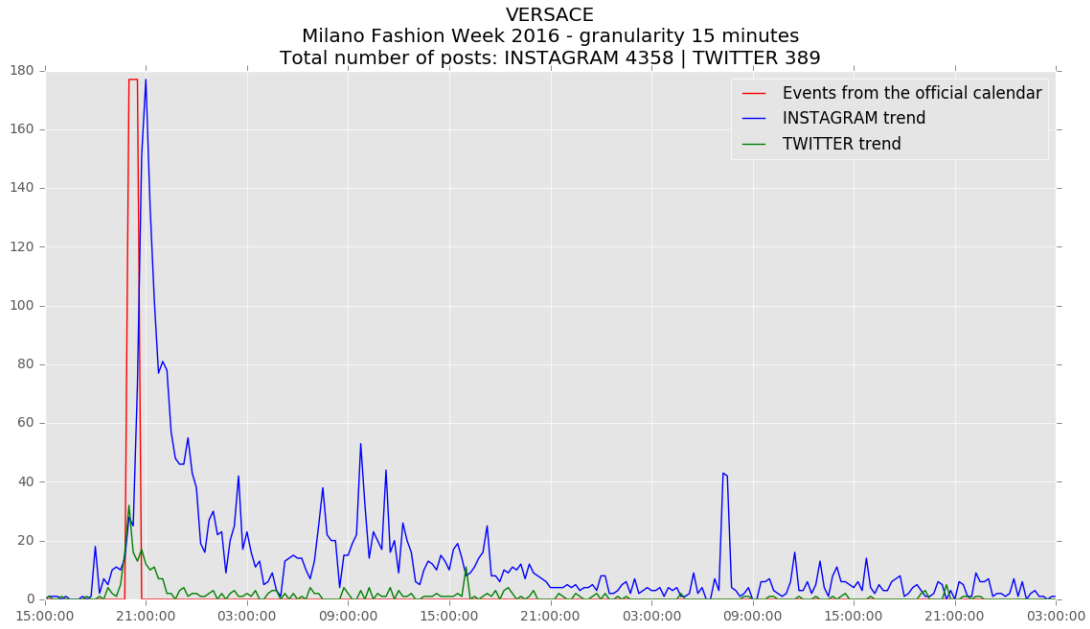



Figure 4.1: Social media response to Versace's event of $26^{th}$ February at 20:00. The blue line is for Instagram, the green one is for Twitter. The granularity is 15 minutes and the lines report the number of posts in the specific time window.

confirm us our expectation: the peak of causality correlation is higher before for *Prada*, and both the two different examples show statistical relevance in the low p-values and in the high F-tests.

We perform tests like these for all the brands that have one or more events in the Milano Fashion Week 2016 calendar and for which we have built ad-hoc regular expressions and collect the results. The next problem to face is to recognize similar behaviour in the social media response to the scheduled events, but this is the aim of the next subchapter.

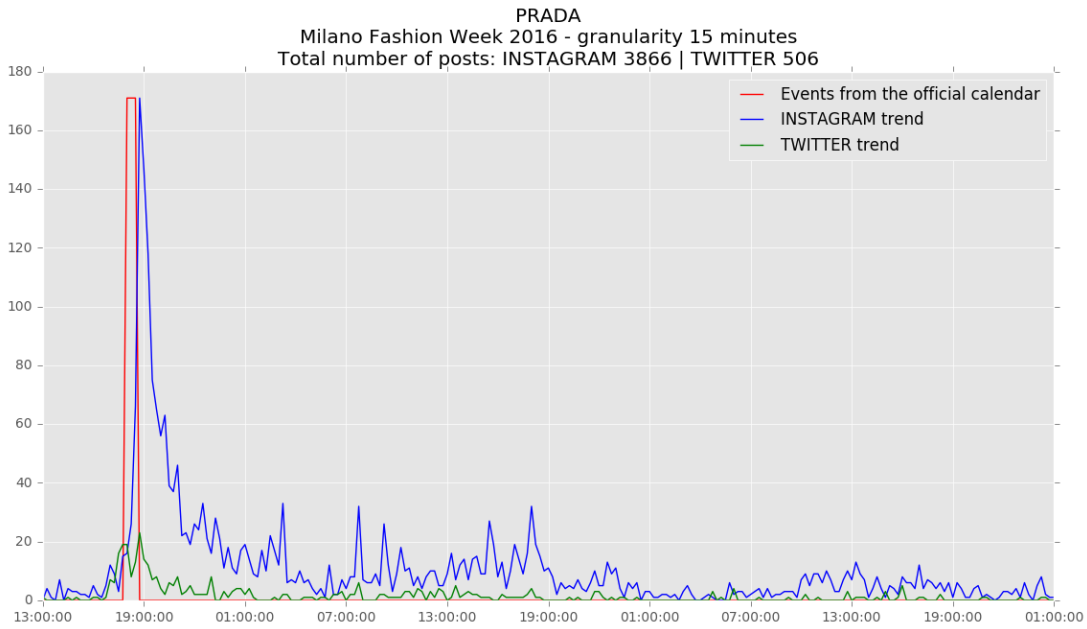
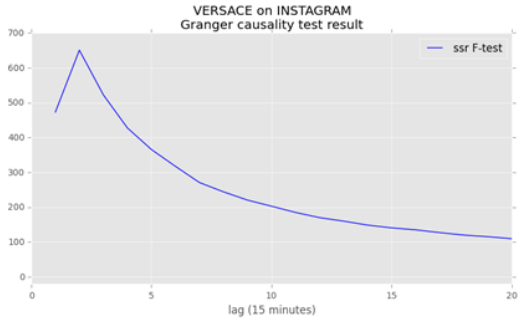
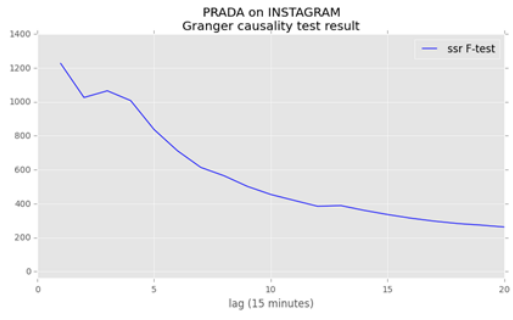



Figure 4.2: Social media response to Prada's event of $25^{th}$ February at 18:00. The blue line is for Instagram, the green one is for Twitter. The granularity is 15 minutes and the lines report the number of posts in the specific time window.




| lag | F_test | p_value |
|---|---|---|
| 1.0 | 472.379488577 | 2.02885449031e-95 |
| 2.0 | 650.514769059 | 1.36188585319e-224 |
| 3.0 | 522.013136052 | 1.38620376463e-258 |
| 4.0 | 427.150516939 | 2.8898581261e-275 |
| 5.0 | 365.228458774 | 2.44105101485e-288 |
| 6.0 | 316.901013939 | 4.70190948827e-296 |
| 7.0 | 270.327782435 | 9.67917256418e-294 |

(a) Versace Granger Causality test result




| lag | F_test | p_value |
|---|---|---|
| 1.0 | 1226.86550841 | 8.3642987685e-216 |
| 2.0 | 1025.8426601 | 4.00978737508e-319 |
| 3.0 | 1065.31101153 | 0.0 |
| 4.0 | 1007.47170703 | 0.0 |
| 5.0 | 837.219462677 | 0.0 |
| 6.0 | 712.190902275 | 0.0 |
| 7.0 | 613.958317737 | 0.0 |

(b) Prada Granger Causality test result

Figure 4.3: Granger Causality F-test results for Versace (a) and Prada (b), If Prada shows a more instantaneous reaction, Versace is a bit more relaxed. All the results seem to have a high statistical confidence, thanks to the low p-values and the high F-tests.

### 4.3.2 Clustering on the tests results

Once we have performed the Granger Causality tests for all the brand-events couples, we are able to recognize similar behaviors in the outcome curves presented before. Figure 4.4 shows all these curves, normalizing each peak at 1. This means that we are looking at the shape of these curves, and so we are not interested in the statistical relevance anymore but only in the type of response related to each brand event in the calendar. It is also shown an average response behavior from which we can denote a big variety in the response due to the span of the standard error bars.

In this moment, we can perform a clustering algorithm in order to group together similar curves and disjoin those curves that are different in shape. The adopted technique is k-means clustering in L-dimensional space, where $L$ is the number of lag analyzed in the previous tests (20 in our case), that is the number of points we have collected for each curve.

At this point we have to think about the question of how many clusters should we take into account. In order to answer this question, a preliminary analysis on the inertia has been taken: recalling that inertia (or within-cluster sum-of-squares) is a measure of how internally coherent clusters are, Figure 4.5 shows the slope of this criterion with respect to increasing the number of clusters. Trivially, the curve is monotonically decreasing, with the maximum value obtained when we set a global cluster, with every one inside, and the minimum is 0 when the number of clusters coincide with the number of elements to cluster.

From the plot in Figure 4.5 and from the pictures in Figure 4.6 we come out with the decision of picking k equal to 4. Indeed, k equal to 3 does not report the delayed response of some elements, while k equal to 5 add another cluster that is not relevant at all, that could be integrated in the clusters neat to it. Also the exact values of the within-cluster sum-of-squares is helpful in this step: in fact, 3 clusters get 23.82, 4 clusters get 17.92, 5 clusters get 14.87. The inertia

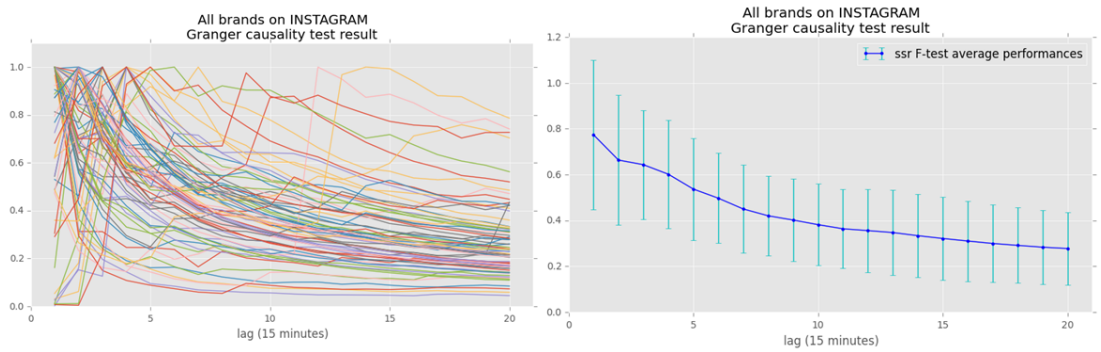



Figure 4.4: All Granger Causality test results in a normalized plot and the average among them. We can notice some different behaviours in the shapes, supported also from the height of the standard error bars in the average graph.

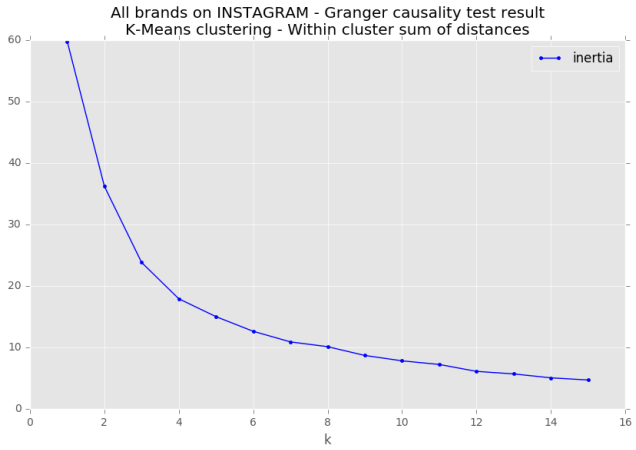



Figure 4.5: Curve representing the behaviour of the inertia with respect to the number clusters k. A good choice for this parameter seems to be in the interval from 3 to 5, where the slope of the curve is relaxing and getting more flat.

Table 4.1: Information on the inertia after clustering the results with different values for k.

| **Number of clusters ($k$)** | **Within-cluster sum-of-squares** | **Inertia gain in adding one more cluster** |
|---|---|---|
| *3* | 23.82 | 12.39 |
| *4* | 17.92 | 5.90 |
| *5* | 14.87 | 3.05 |

is decreasing very slow after k equal to 4, then our final choice for k, that is the number of clusters, is 4.

In the end, we have 4 different groups of (more or less) similar response, that can be described in this way, referring to the labelling colour:

- *Yellow*, with high immediate response;
- *Red*, with lagged response peak at 15 minutes;
- *Green*, with lagged response peak at 45-60 minutes;
- *Blue*, with an initial significant response but with another lagged response peak at 3 hours and 15 minutes.

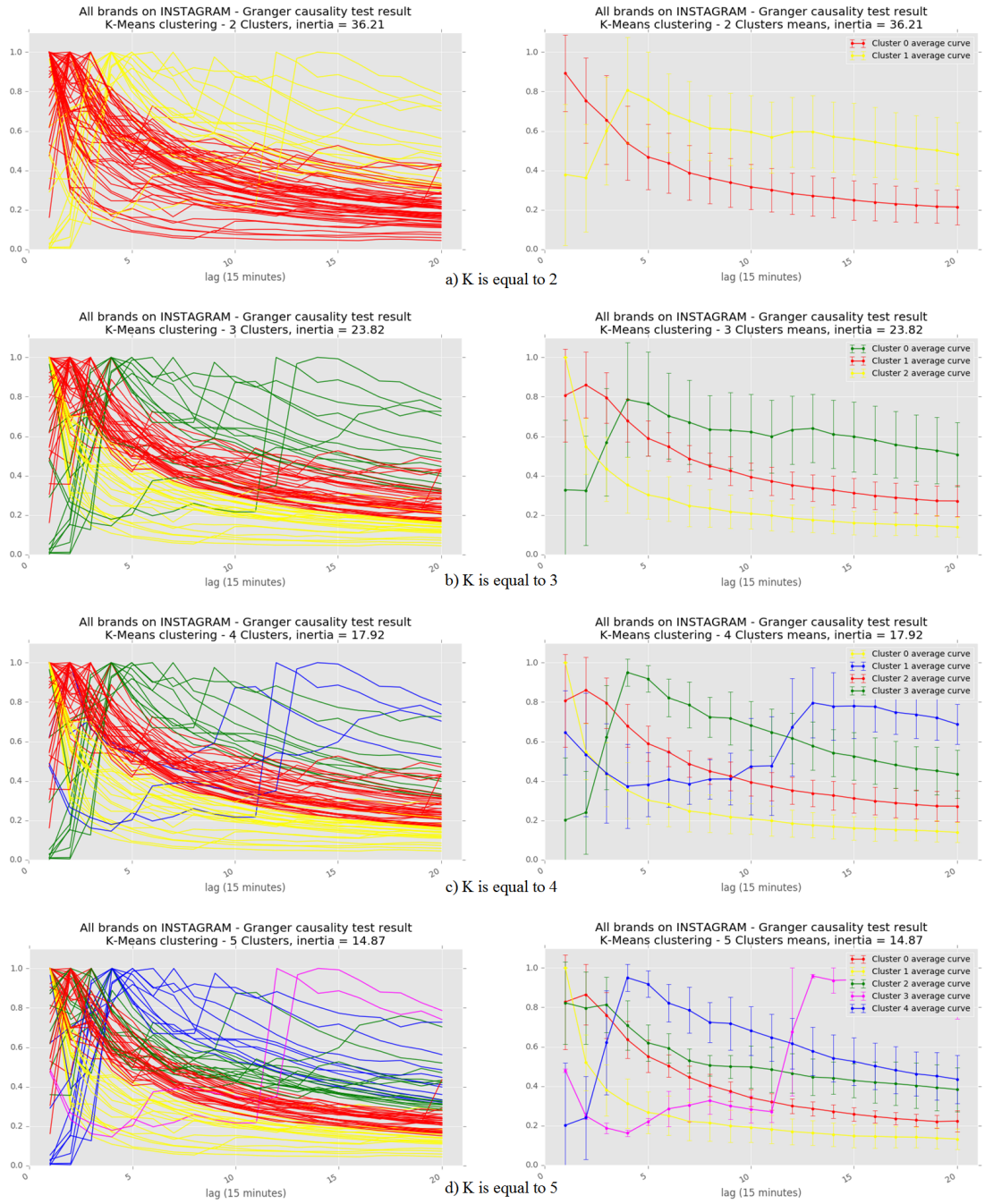



Figure 4.6: K-means clustering results for different values of k. In the left column we report graph showing all the curves and their cluster labelling, using different colours. In the right column we report the centroid of each cluster with the standard error bars. The best option seems to be 4, since all the relevant clusters are present. Indeed, in k equal to 3 we lose the group of the delayed response, while in k equal to 5 we add an unnecessary cluster that lies near other two clusters already present before.

At this point, we can pick from our clusters the most representative elements in order to have some real examples. These most representative elements are defined as the closest one to the related cluster center, within the cluster. Figure 4.7 reports as representatives:

- *Costume National*, from the yellow group;
- *Trussardi*, from the red group;
- *Alberta Ferretti*, from the green group;
- *Emporio Armani*, from the blue group.

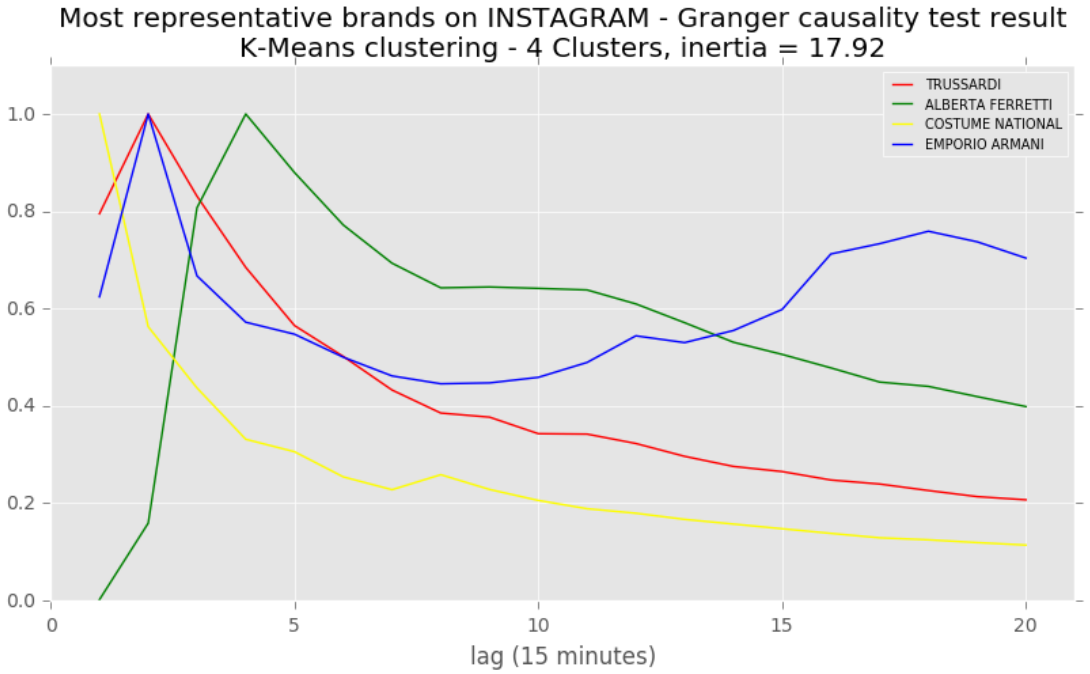



Figure 4.7: Most representative elements from each cluster.

### 4.3.3 Classification problem

In this final subchapter, we take into account the clustering performed before and considering the labelling outcome of the k-means algorithm as the prediction goal in a classification problem. The goal of this section is to build a prediction model for the cluster labeling defined from the Granger tests results, given the simple information related to brands and events, compare the results of different techniques and methodologies, taking into account some performance indicators.

#### Pre-processing phase

First of all, we must have a look at the data: Table 4.2 presents a portion of the grid file we have with all the information on the events and the schedule of the Milano Fashion Week 2016. Each datum in the classification problem is represented by a single row of this table, where the target is the attribute *label*. We have a total of 73 rows, that is 73 brand-events couples. Each brand is present only one time in the table, even if we have some brands with more than a single event in calendar. This is because the filtering on the posts data-set is performed using the text of the post, looking for matching with regular expression linked to the brand name. Then, if one brand is taking two or three different events from the official calendar, we are not able to find an accurate correlation with the specific event among the two or three we have from the schedule. More specifically, we have 68 brands taking 1 single event each, 4 brands taking 2 different events, and only 1 brand taking 3 different events.

For all the 68 couples of one-brand one-event there are no problems in defining the information related to the event, due to the unicity of the datum we are considering. Problems may arise when we have one-brand many-events relationships: we face these issues manually. For *Giorgio Armani*, *Emporio Armani* and *Jil Sander* has been easy because, looking at the calendar, we have, for each one of them, two different fashion shows simply one after the other, delayed 1 hour, in the very same place. Then, for these three cases, we collide the information of the two similar events into a single row, setting the number of events to 2, the starting date as the starting date of the first show, the ending date as the ending date of the second show. The case of *L72* has been a little more complicated, because it is taking two presentations in the same place, but in two different days. Then, we decide to lose the information on the start and the end date, maintaining the information on the type of show and location. The last case is the one related to the brand *Marni*. We have two fashion shows and a presentation, with two different locations and different times of exhibition. Then, we decide to lose all the information related to the events, making this datum really blurred and fuzzy, unfortunately.

Now, some pre-processing of the data was performed in order to convert categorical variable into indicator variables and to normalize the continuous valued

ones. The reasons to do these steps, especially the dummy translation, are the following:

1. Satisfy the model fittings, that often requires no categorical features;
2. Mine most knowledge as possible from these variables, such as the Date field. Indeed, different approaches are available in order to represent a datetime, such as timestamps, or spreading it out in different columns, one for each sub-field of a datetime (Year, Month, Day, Hour, ...). I prefer the explosion into indicator variables in order to better catch cyclical components in the response trend, that maybe are not so catchable with the absolute values described before.

### Fitting the models

The different models were fitted with different features in order to find the best combination of input variables for the final classification and probability estimation. It turned out that fitting the model trained with the features [$x^{start\_minute}$, $x^{start\_hour}$, $x^{start\_day}$, $x^{end\_minute}$, $x^{end\_hour}$, $x^{end\_day}$, $x^{class}$, $x^{live}$, $x^{type}$, $x^{invitation}$, $x^{open}$] was the best way.

We fit different models, as described in the previous chapter, in order to compare the results and pick the best one in the final classification step. The following Table 4.3 reports the main performance coefficients.

We also perform Leave One Out Cross Validation (LOO-CV) for all the techniques adopted, that is possible thanks to the small size of the data-set, in order to have better performance indicators for the different models.

The columns in the performances table refer to, from left to right:

1. the classifier adopted for the problem;
2. the time spent during the fitting phase, training the model;
3. the time spent for the final prediction, classifying the training samples;
4. the number of misclassified elements in the train itself;
5. the log loss on the train itself;
6. the number of misclassified elements in the test (recall, the test is one row at a time) adopting LOO-CV;
7. the average log loss on the test adopting LOO-CV;
8. the standard deviation of the log loss on the test adopting LOO-CV;
9. the average log loss on the train adopting LOO-CV;
10. the standard deviation of the log loss on the train LOO-CV.

Table 4.2: Some rows from the table containing all the data for the classification problem. Each row is referring to a brand-events couple.

| name | label | class | number_of_events | event_address | event_start | event_end | event_live | event_type | event_invitation | event_open |
|---|---|---|---|---|---|---|---|---|---|---|
| AU JOUR LE JOUR | 0 | emerged | 1 | piazza duomo, arengario | 28/02/2016 18:00 | 28/02/2016 18:30 | True | sfilate | True | False |
| DAMIR DOMA | 0 | emerged | 1 | via provata g. ventura, 14 | 28/02/2016 20:00 | 28/02/2016 20:30 | False | sfilate | True | False |
| GABRIELE COLANGELO | 0 | emerged | 1 | via brera, 28 | 27/02/2016 17:00 | 27/02/2016 17:30 | False | sfilate | True | False |
| LUCIO VANOTTI | 0 | emerged | 1 | piazza lina bo bardi, 1 | 29/02/2016 16:30 | 29/02/2016 17:00 | True | sfilate | True | False |
| MARCO DE VINCENZO | 0 | emerged | 1 | via turati, 34 | 26/02/2016 16:00 | 26/02/2016 16:30 | False | sfilate | True | False |
| MSGM | 0 | emerged | 1 | via compagnoni, 12 | 28/02/2016 15:00 | 28/02/2016 15:30 | False | sfilate | True | False |
| STELLA JEAN | 0 | emerged | 1 | piazza duomo, arengario | 28/02/2016 12:45 | 28/02/2016 13:15 | True | sfilate | True | False |
| VIVETTA | 0 | emerged | 1 | piazza duomo, arengario | 29/02/2016 14:30 | 29/02/2016 15:00 | True | sfilate | True | False |
| AIGNER | 0 | top | 1 | piazza lina bo bardi, 1 | 26/02/2016 19:00 | 26/02/2016 19:30 | True | sfilate | True | False |
| ALBERTO ZAMBELLI | 0 | top | 1 | piazza duomo, arengario | 28/02/2016 9:30 | 28/02/2016 10:00 | True | sfilate | True | False |
| ANTEPRIMA | 0 | top | 1 | via senato, 10 | 25/02/2016 14:00 | 25/02/2016 14:30 | False | sfilate | True | False |
| ANTONIO MARRAS | 0 | top | 1 | via compagnoni, 12 | 27/02/2016 11:30 | 27/02/2016 12:00 | True | sfilate | True | False |
| BLUGIRL | 0 | top | 1 | via san barnaba, 48 | 24/02/2016 10:00 | 24/02/2016 10:30 | False | sfilate | True | False |
| DIESEL BLACK GOLD | 0 | top | 1 | via valtellina, 7 | 26/02/2016 9:30 | 26/02/2016 10:00 | False | sfilate | True | False |
| DSQUARED2 | 0 | top | 1 | via san luca, 3 | 29/02/2016 9:30 | 29/02/2016 10:00 | True | sfilate | True | False |

Table 4.3: Performances of all the techniques adopted in the classification problem. From left to right we have: the classifier adopted, the time spent during the fitting phase, the time spent for the final prediction, the number of train errors, the log loss on the train, the number of errors adopting leave one out cross validation, the average log loss on the test (one row at a time) with standard error adopting leave one out cross validation, the average log loss on the train with standard error adopting leave one out cross validation.

| | *Time indicators* | | *Weak indicators* | | *Leave One Out Cross-Validation indicators* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **classifier** | **fit_time** | **prediction_time** | **train_errors** | **log_loss_train** | **loo_errors** | **loo_logloss_test_mean** | **loo_logloss_test_std** | **loo_logloss_train_mean** | **loo_logloss_train_std** |
| Bernoulli NaiveBayes | 0,0035 | 0,0015 | 22 | 1,39806 | 36 | 2,76323 | 4,17595 | 1,3901 | 0,0349 |
| Logistic Regression | 0,0225 | 0,0015 | 17 | 0,61651 | 33 | 0,76421 | 0,72045 | 0,6129 | 0,0123 |
| Cross-Validated Logistic Regression | 2,7521 | 0,0015 | 21 | 0,86406 | 32 | 0,56453 | 0,36066 | 0,8162 | 0,0666 |
| Support Vector Machine | 0,0190 | 0,0060 | 15 | 1,03000 | 38 | 0,49924 | 0,24064 | 0,9835 | 0,0444 |
| Decision Tree | 0,0035 | 0,0020 | 2 | 0,03798 | 37 | 17,03322 | 17,26817 | 0,0375 | 0,0044 |
| Random Forest | 1,1363 | 0,2067 | 2 | 0,25385 | 28 | 0,83452 | 0,79575 | 0,2507 | 0,0052 |
| Dummy MostFrequent | 0,0020 | 0,0020 | 36 | 17,03282 | 36 | 17,50636 | 17,26817 | 17,0328 | 0,2398 |
| Dummy Stratified | 0,0020 | 0,0015 | 48 | 20,81789 | 48 | 13,24806 | 16,79498 | 21,6853 | 1,8348 |
| Uniform Distribution | | 0,0005 | | 1,38629 | | | | | |
| Random Distribution | | 0,0020 | | 1,75860 | 60,47000 | | | 1,7586 | 0,0980 |

The different models adopted are the ones described in the previous chapter, where we talk about the method used for the classification problem. The number of errors for the random strategy is an average of the number of errors above 100 runs, in order to keep this measure more confident.

Now, we can analyse more in the details the performances of the different techniques. First thing to say, all these models are fast both in the training phase and in the prediction phase. Indeed, the data-set is really small. The only two techniques that are slower than the others are the Cross-Validated Logistic Regression and the Random Forest. The first because it has a cross-validating phase inside, and thus it has to choose the best value for the parameter $C$, among 10 possible candidate values. The second because a diverse set of classifiers is created by introducing randomness in the classifier construction, and the number of different estimators is set, by our choice, to 1000.

Then, we can look at the number of errors when using as test set the train self itself: this means that we are training and fitting our models on all the 73 samples we have, and then we are testing them on the very same samples. Unfortunately, this type of predictions leads to some overfitting, but with a small data-set like the one we have, we think that also a measure of this type has to be taken into account. For this performance indicator we have a predominance of the tree-based algorithms, that are Decision Trees and Random Forest. This performance accuracy is a clear example of overfitting: indeed, decision-tree learners can often create over-complex trees that do not generalise the data well, and adopting a strategy like the one we are considering, that is setting the test set coinciding with the train set, this is surely the case. However, each method is outperforming the baselines of Dummy Most Frequent and Dummy Stratified.

From the first log loss columns, where again test and train set are perfectly overlapping, we can notice the overfitting of the tree-based learners, together with nice performances in probability estimation from the other models, too.

Then all the measures referring to Leave One Out Cross-Validation comes. For these indicators (*loo_errors*, *loo_logloss_test_mean*, *loo_logloss_test_std*, *loo_logloss_train_mean*, *loo_logloss_train_std*) we have applied the technique of predicting each sample at a time, using all the remaining ones as training set. These could be more confident, accurate and truthful performance indicators, because the overfitting of the previous scenario, when colliding test and train set, is now avoided. Indeed, the number of errors are increasing: for this indicator, the best model is the Random Forest, that seems to beat the overfitting problem of the decision trees, from which it is based, with nice success. The 100 runs of random strategies give us an average of 60.47 errors over 73 (for every sample, it has to choose randomly among 4 possible target values), and every model we adopted is outperforming this baseline, even if the numbers of misclassified samples are not so low.

Considering the probability estimation indicators, that are the log loss, we immediately notice the overfitting of the Decision Tree, as we were expecting

from the previous analysis. Indeed, the average log loss on the test for this type of estimator is really close to the baselines of Dummy Most Frequent and Dummy Stratified. Together with the test log loss standard deviation, these measures are proving the high variance of the over-complex model built by the tree. The best model considering the log loss on the test samples (bot mean and standard deviation) is, without any doubt, the Support Vector Machine, with the lowest mean of the log loss itself and also the lowest standard deviation of the same measure. The remaining indicators of the log loss on the training sets underline once again the overfitting issue of the Decision Tree and discrete performance from all the models, overall, with no one overcoming the others.

Concluding, looking at the bigger picture, we notice that we have not a model that deserves a standing-ovation, giving more weight and importance to the Leave One Out Cross-Validation indicators. However, there are two models that can be preferred to the others: the first one is the Random Forest, which has a really low number of errors, only 2 over 73, when considering all the data-set as train and test set, and maintains the lowest number of errors (28 over 73) among the other techniques in the classification phase of the Cross-Validation. The second is the Support Vector Machine, which has the best performances in probability estimation, as we can see from the log loss indicators from the Leave One Out Cross-Validation indicators. Unfortunately, this technique ends up in classifying the different samples not so well, as reported from the 38 errors in the classification phase of the Cross-Validation.

# Chapter 5

# Geo response analysis

## 5.1 Introduction

In this chapter we want to focus on the geographical dispersion that the social media have shown in terms of generated posts volume with respect to the locations of scheduled events in the Fashion Week calendar. Describing the problem with more details, we can say that, in this type of analysis, we have two different spatial signals, the first one for the calendar events and the second one for the volumes of social media posts on the web with geographical information attached, like latitude and longitude, and we try to study some kind of measures of dispersion and concentration of these posts with respect to the location of the related event. With these two signals a lot of features can be computed in order to describe the event and the reaction on the web to it, in terms of spatial dispersion.

This kind of work has been done in the same way of the previous chapter, that is brand by brand, referring only to one brand at a time and then selecting only the Milano Fashion Week events and the social media posts related to that specific brand. In order to make a much clearer analysis and to set all the elements of study at the same level, we have considered not all the different types of events, but only the fashion shows, that remain the majority of events analysed so far, also in the previous chapter.

The next study is about looking at the attributes and features computed so far, in order to characterize the dispersion of the social media response, and performing some unsupervised learning on them, at first with the purpose of figuring out the principal components from the big set of all these features and only after this preliminary analysis trying to make a significant clustering between the brand-event couples with the aim of grouping together similar spatial behaviours.

## 5.2 Method

This chapter can be divided in two different parts: the first one concerning the characterization of the events in terms of features about dispersion, and the second one concerning the clustering problem among these new data.

Among this chapter, we are always talking about geo-referenced posts, that have a location in terms of latitude and longitude. We decide not to taking into account the continuous values coming from this type of measures, but to build a grid of cells above the area of Milano city, and assign each post to the appropriate cell. This in order to adopt a model that is more clear and easy to understand in the visualization step, but that is also able to satisfy the requirements of some coefficients and features we are going to compute. This grid has a square shape, with sides of $10km$, divided into 20 rows and 20 columns, for a total of 400 cells of $500m \times 500m$. In this way, each cell of the grid is containing a certain value, that is the number of posts shared from that specific cell. Trivially, also the calendar events are assigned to their own specific cell.

The values for the grid will be computed for each couple brand-event, with different time-scopes. Indeed, we decide to analyse four different time-windows, in order to capture the evolution of the dispersion of the social media movements over the time, specifically how and how much dispersion is changing. The widths of these different windows are:

- 3 hours, from the start time of the fashion show;
- 6 hours, from the start time of the fashion show;
- 24 hours, from the start time of the fashion show;
- the entire time window stored in the dataset, of 24 days.

With this type of model, all the following measures can be easily computed.

### 5.2.1 Building the set of features

Now that we have the model for the type of data we want to study, we can compute different measures that reflect the dispersion or the concentration of the social media movement with respect to the events locations. The adopted measures for this kind of study are:

1. *Gini coefficient*;
2. *Average distance* of the social media signals from the event location;
3. Number of *alive*, *active* and *strongly active* cells.

Each measure is described more accurately in the following sections.

### Gini coefficient

The Gini coefficient is a measure of statistical dispersion. It was developed by the Italian statistician and sociologist Corrado Gini and published in his 1912 paper "Variability and Mutability" and it was proposed as a measure of inequality of income or wealth of a nation's residents. The formula adopted in order to compute this coefficient is the following:

$$G = \frac{1}{n}\left(n + 1 - 2\left(\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}\right)\right) \tag{5.1}$$

where:

- $n$ is the number of considered cells;
- the population is assumed uniform on the values $y_i, i = 1, ..., n$, indexed in non-decreasing order, with $y_i \leq y_{i+1}$.

In our specific scenario, we can see the cells of the grid in place of the set of people and the number of posts in one cell in place of a person's income.

The Gini coefficient measures the inequality among values of a frequency distribution (for example, levels of income, or number of posts shared from a cell in the grid). A Gini coefficient of zero expresses perfect equality, where all values are the same (for example, where every cell has the same number of posts published). In the opposite way, a Gini coefficient of 1 (or 100%) expresses maximal inequality among values (e.g., when all the posts are related to a single cell, leaving all the remaining 399 cells with no social media signal). It has to be said that for larger groups, values close to or above 1 are very unlikely in practice.

We have decided to compute the Gini coefficient on two different models:

- The first one, considering the entire grid of cells;
- The second one, considering only those cells that results alive (the definition of this term comes shortly) for at least one couple brand-event, in the specific time-window of analysis.

We make these distinctions because of the unfair concentration of posts in a few cells and because of the presence of a lot of cells that are "dead" for every couple brand-event. In this way, we are considering with the second model an average of 40 cells instead of 400.

### Average distance

We define the average distance of the posts from the event as:

$$avgDist = \frac{1}{\sum_{r=1}^{R}\sum_{c=1}^{C}\mathbf{G}_{r,c}} \sum_{r=1}^{R}\sum_{c=1}^{C}\mathbf{G}_{r,c} \times dist(\langle r, c\rangle, \langle e_r, e_c\rangle). \tag{5.2}$$

In the formula written above (5.2), the *dist* function is the distance computed between cells, in a Manhattan way, with parameters like the tuple $\langle r, c \rangle$ indicating the row and the columns index, and the tuple $\langle e_r, e_c \rangle$ for the event cell row and column index. The $R \times C$ matrix $\mathbf{G}$ contains the number of posts in each cell, with $R$ and $C$ standing for the number of rows and columns in which the grid is divided.

High values of this feature mean high dispersion of the social media signal, far away from the cell where the event is taking place, while low values of average distance means high concentration near the event location.

*Gini coefficient* and *Average distance* seem to give the same information, but that is not true; indeed, Gini coefficient is not taking into account the location of the event, while Average distance is, then the first feature measures the concentration regardless of the event location, while the second one measures the dispersion of the social signal from the specific cell of the event.

#### Alive, Active and Strongly Active cells

We have defined three types of cell, in order to capture the evolution of the dispersion of the social signal with a different perspective:

1. *Alive cells*, with a percentage of posts shared in the considered time-window more than 1% of the total number of posts in the grid in the same time-window;

2. *Active cells*, with a percentage of posts shared in the considered time-window more than 10% of the total number of posts in the grid in the same time-window;

3. *Strongly Active cells*, with a percentage of posts shared in the considered time-window more than 20% of the total number of posts in the grid in the same time-window.

We compute the results of these type of measures for all the pre-defined frames and we also compute the differences between subsequent frames (e.g. 3h - 6h) in terms of number of on/off for each specific type of cell.

### 5.2.2 Clustering over the computed features

Now that we have built a various set of features for our brand-event couples, we want to perform a clustering of these couples based on these very kind of attributes. In order to face this problem, our first aim is to visualize these pieces of information in order to understand better the data we are talking about. But, as we are going to see, we are considering data not displayable in all their dimensions, then a Principal Components Analysis should be helpful.

### Principal Components Analysis

Principal Components Analysis (PCA) is an unsupervised learning problem used mainly for dimension reduction, compression and visualization. Indeed, the principal components of a set of data in $\mathbb{R}^p$ provide a sequence of best linear approximations to that data, of all ranks $q \leq p$. Denote the observations by $x_1, x_2, ..., x_N$, and consider the rank-$q$, linear model for representing them:

$$f(\lambda) = \mu + \mathbf{V}_q \lambda \tag{5.3}$$

where $\mu$ is a location vector in $\mathbb{R}^p$, $\mathbf{V}_q$ is a $p \times q$ matrix with $q$ orthogonal unit vectors as columns, and $\lambda$ is a $q$ vector of parameters. This is the best parametric representation of the affine hyperplane of rank-$q$. Fitting such a model to the data by least squares amounts to minimizing the reconstruction error:

$$\min_{\mu_i, \{\lambda_i\}, \mathbf{V}_q} \sum_{i=1}^{N} \parallel x_i - \mu - \mathbf{V}_q \lambda_i \parallel^2 . \tag{5.4}$$

We can partially optimize for $\mu$ and the $\lambda_i$ to obtain:

$$\begin{aligned} \hat{\mu} &= \hat{x}, \\ \hat{\lambda_i} &= \mathbf{V}_q^T (x_i - \bar{x}). \end{aligned} \tag{5.5}$$

This leaves us to find the orthogonal matrix $\mathbf{V}_q$:

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \parallel (x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x}) \parallel^2 . \tag{5.6}$$

We can assume that $\bar{x} = 0$, otherwise we simply replace the observations by their centered versions $\tilde{x}_i = x_i - \bar{x}$. The $p \times p$ matrix $\mathbf{H}_q = \mathbf{V}_q \mathbf{V}_q^T$ is a projection matrix, and maps each point $x_i$ onto its rank-$q$ reconstruction $\mathbf{H}_q x_i$, the orthogonal projection of $x_i$ onto the subspace spanned by the columns of $\mathbf{V}_q$. The solution can be expressed as follows. Stack the centered observations into the rows of an $N \times p$ matrix $\mathbf{X}$. We construct the singular value decomposition of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{5.7}$$

This is a standard decomposition, where $\mathbf{U}$ is an $N \times p$ orthogonal matrix whose columns $u_j$ are called the left singular vectors; $\mathbf{V}$ is a $p \times p$ orthogonal matrix with columns $v_j$ called the right singular vectors, and $\mathbf{D}$ is a $p \times p$ diagonal matrix, with diagonal elements $d_1 \geq d_2 \geq ... \geq d_p \geq 0$ known as the singular values. For each rank $q$, the solution $\mathbf{V}_q$ to (5.6) consists of the first $q$ columns of $\mathbf{V}$. The columns of $\mathbf{UD}$ are called the principal components of $\mathbf{X}$. The $N$ optimal $\hat{\lambda}_i$ in the second formula in (5.5) are given by the first $q$ principal components (the $N$ rows of the $N \times q$ matrix $\mathbf{U}_q \mathbf{D}_q$).

PCA is used to decompose a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance, such that the linear combinations $\mathbf{X}v_1$ has the highest variance among all linear combinations of the features; $\mathbf{X}v_2$ has the highest variance among all linear combinations satisfying $v_2$ orthogonal to $v_1$, and so on.

#### K-Means clustering

After the analysis of the principal components, we perform a k-means clustering, as described in chapter 4.2.2 Clustering on the tests results on page 47, but in this case we are facing also the problem of selecting the features to utilize in the clustering phase. Indeed, we decide to take a threshold on the explained variance of the principal components to take into account in order to fit the k-means algorithm, but this does not modify the formulation of the problem and the algorithm steps of k-means.

## 5.3 Findings

In this chapter, the analysis is done only for the Instagram scenario, for the same reasons of the previous chapter, that are the bigger volume of posts obtained and the complexity of making a comparison between the two different social media. In this way, we can also make other further considerations, comparing the findings from this chapter with the results of the previous one.

The data-set is containing only the posts that are returning matchings with the brand-specific regular expressions and have geo-references saved. The specific numbers of posts come in the next analysis, since they are different for each brand. Furthermore, we are not taking into account the entire world atlas, but only a region of $10km \times 10km$ over the Milano city area, as shown in Figure 3-1. This square is divided into 400 equal cells, in a $20 \times 20$ grid. Each cell has sides of $500m$.

In this way, we are focusing on the Milano specific response, disregarding any far away regions. We have to say that the most of the social media signal we have stored is coming from this specific region, with small reactions from other parts of the world.

In order to show some examples, in the next Figures we can see the heatmaps related to *Gucci* and *Salvatore Ferragamo*, with time-window including all the days of analysis.

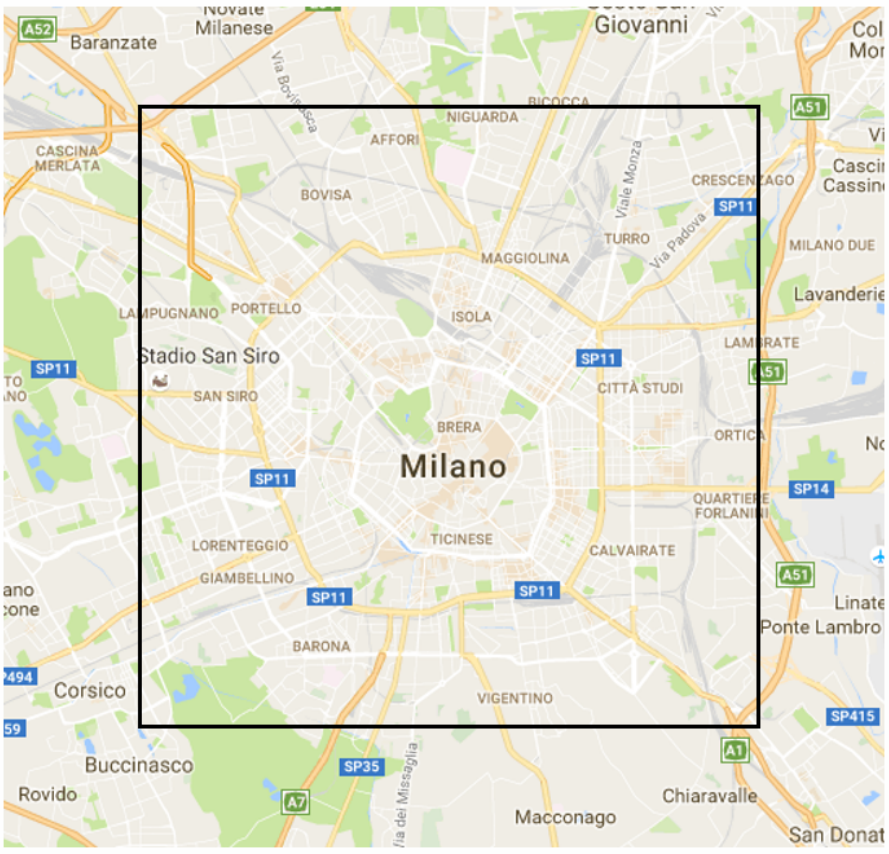
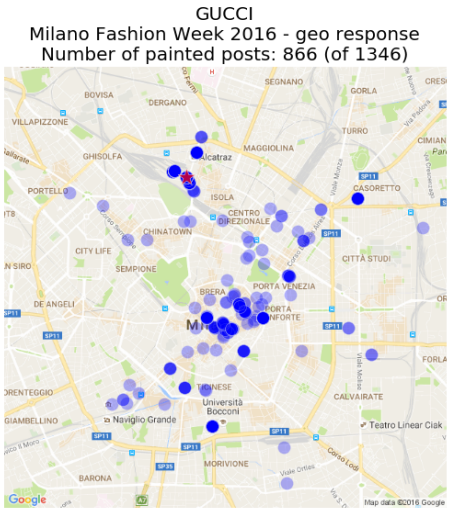
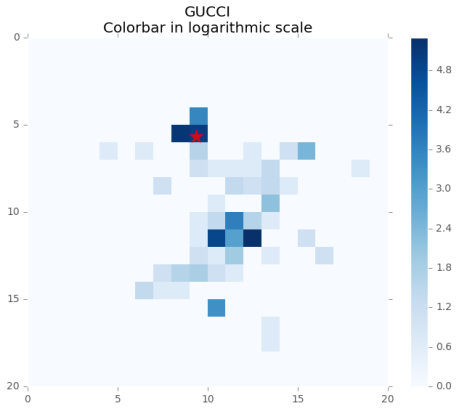



Figure 5.1: The black square represents the analysed region, that will be divided into a grid of 20 rows ad 20 columns. All the fashion shows analysed in this chapter are inside this area of $100km^2$

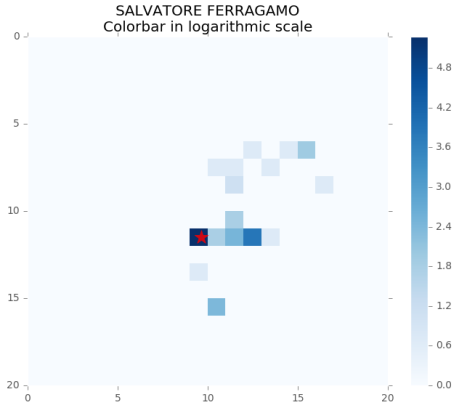



(a)




(b)


(c)




(d)

Figure 5.2: Social signal related to the events of Gucci and Salvatore Ferragamo. On the left side, we can see the real social movements (as blue points) over the 24 days of analysis and the events, located with a red star. The first number of posts is the number of blue dots, while the number between parenthesis is the number of posts all over the world. On the right side, we report the related heatmaps, where the colorbars are in logarithmic scale, in order not to give many relevance to the more active cells, and to be able to recognize all the alive cells.

### 5.3.1 Building the set of features

In the first steps of this analysis, we try to decorate the events from the Milano Fashion Week calendar with some knowledge coming out from the social media, in terms of geo-response, exploiting the heatmaps defined and presented above. We have defined some features that could explain the dispersion of the social signal and the evolution of such this measure, as described in the previous sections.

With the help of the plots in Figure 5.3, we can notice the behaviour of the geo-response to the events of *Gucci* and *Salvatore Ferragamo*. More or less, all the brands have similar behaviours in this kind of measures, as one can trivially expect. We can assert the following statements.

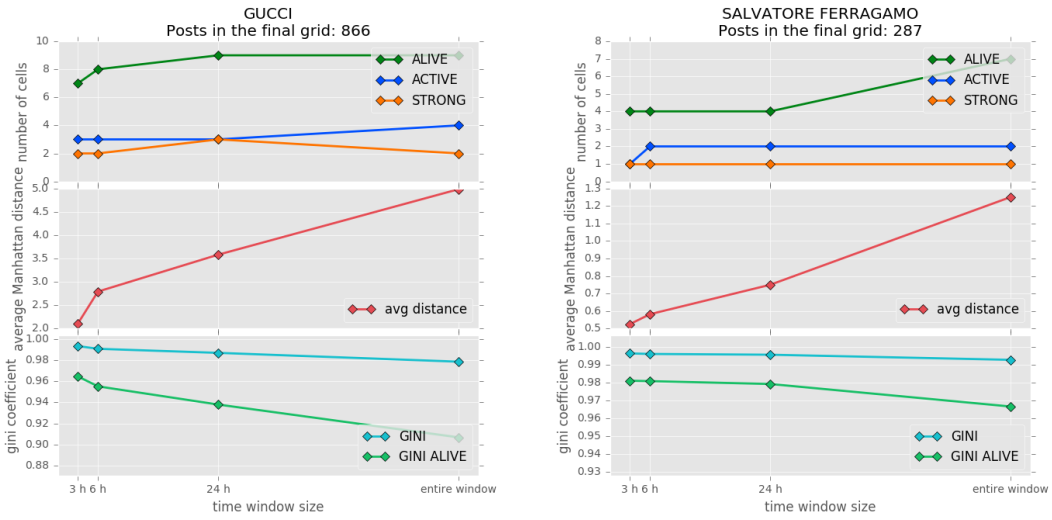
- As we increase the width of the time-window, the number of *alive cells* is increasing. On the other way, the number of *active* and *strongly active cells* is floating in the range from 1 to 3, with very few brands reaching 4 *active cells*.

- In the very first moments from the opening of the event the posts are shared near the very event location, but as we look at the bigger picture, including 24 hours or even the entire period of 24 days, the *average distance* is increasing, showing the growing dispersion of the social movement.

- The *Gini coefficient* proves how the concentration of the social signal remains always high, due also to the fact that the low percentage of users that allows Instagram to geo-tag their own photo is reducing the number of authors implied in this study, and so the few authors with high volumes of posts generated are biasing the results. However, looking at the *Gini alive coefficient*, that refers to the Gini coefficient computed only over the cells that result alive for at least one brand in the specific time-window, we can see a weak smoothing of the concentration strength with the increasing of the time-scope.

### 5.3.2 Clustering over the computed features

The next objective is to exploit this set of various features in order to cluster similar brands in terms of spatial response. Before any kind of analysis, the visualization step could help us in understanding the problem we are going to face. But we cannot visualize the data in all their dimensions, since we have got 42 different features for each couple brand-event, that correspond to a 42-dimensions hyperspace. Principal Component Analysis can help us in this very first step.




(a) (b)

Figure 5.3: Features from Gucci and Salvatore Ferragamo signals.

### Principal Components Analysis

In this section, we try to pull out from our model the best features in order to describe the data. Indeed, if we find these best attributes, we can visualize the data focusing only on these dimensions, that corresponds to the principal components of our not so simple data-set. The results of PCA show that the three features leading to the minimum reconstruction error are:

1. Number of *alive cells*, with time scope of 24 days;
2. *Average distance*, again with time scope of 24 days;
3. Number of *alive cells*, this time with time scope of 6 hours.

We can see the ranking of the top 10 attributes, sorted with respect to the explained variance, in Table 5.1. In this way, we can report the plot in the two principal components in Figure 5.4.

We can notice how there are some attributes among all the 42 that are really more relevant than the other ones: for example, choosing the first two principal components we are near to capture the 60% of the total variance, while if we want to reach the 90% of that, we should pick the first seven principal components.

Table 5.1: Ranking coming from PCA in terms of explained variance ratio.

| feature | Explained variance ratio |
|---|---|
| *alive* | 0,4577 |
| *avg_distance* | 0,1390 |
| *alive_6h* | 0,1132 |
| *alive_24h* | 0,0929 |
| *alive_3h* | 0,0509 |
| *active* | 0,0321 |
| *alive_off_from_6h* | 0,0239 |
| *alive_on_from_24h* | 0,0207 |
| *alive_on_from_3h* | 0,0156 |
| *active_6h* | 0,0117 |

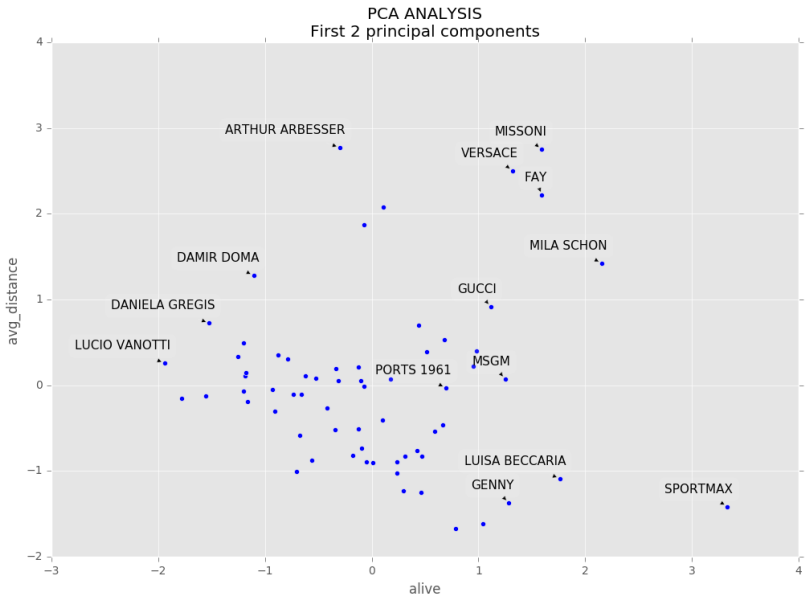



Figure 5.4: Visualization of the data in the two principal components. The PCA is compressing the data and then the values are transformed and not the real ones.

### K-Means clustering

Now that we have computed the features for our data and we have displayed them in a human-readable plot, we are going to form some clusters of brand-event couples from that.

At this point we have to decide the number of clusters that k-means should build, and in order to answer this question we run the algorithm with different k, from 1 to 15, and then we look at the inertia trend for all these values of k, as shown in Figure 5.5. In this moment, we have not unbalanced ourselves in the choice of the subspace of clustering yet, that is the number of components to use in the clustering phase . However, the inertia trends are more or less the same. Furthermore, for the choice of the number of clusters we have to recall our previous work in the time-response chapter, where the final value for k was 4. Also looking at the plot, this choice of 4 seems not a bad one, and then we set the value for k to 4.

For the clustering of the data, we have to make the further selection of the features to pick and to pass to the k-means algorithm in order to build the different groups. The choice we have taken is to pick all those features that help us in capturing at least the 85% of the explained variance, in such a way we could compress our data not losing a lot of information. In order to reach this threshold, we have to pick the 5 principal components, that are the top5 elements in Table 5.1. Then, we are running k-means algorithm exploiting only the features chosen. We are presenting the results in the two principal dimensions of number of *alive cells* and *average distance*, both with time scope of 24 days, and in other two non-principal components, that are the number of cells alive at 3 hours and the number of cells alive at 6 hours, in order to better understand the differences between the groups. In Figure 5.6 we report the clustering with 5 principal components, in the mentioned way.

In the end, we have 4 different groups of (more or less) similar spatial response, that can be described in this way, referring to the labelling colour:

- *Yellow*, with really few *alive cells* and low *average distance*;
- *Red*, with numbers of *alive cells* and *average distances* slightly more high than the *Yellow* cluster;
- *Green*, with the highest results for *alive cells* and *average distance*;
- *Blue*, a single element pop out from the *Green* cluster because of its high values in *alive cells*, both for 3 hours and 6 hours time scopes, but mostly for the time scope of the entire window.

We have to notice that in the previous brief summary of the different clusters we are taking into account only four relevant dimensions, but in the model

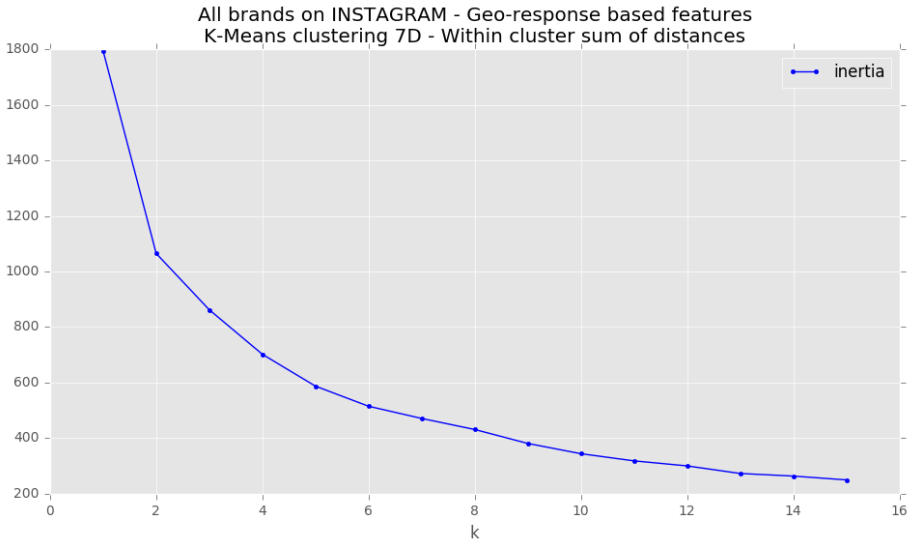
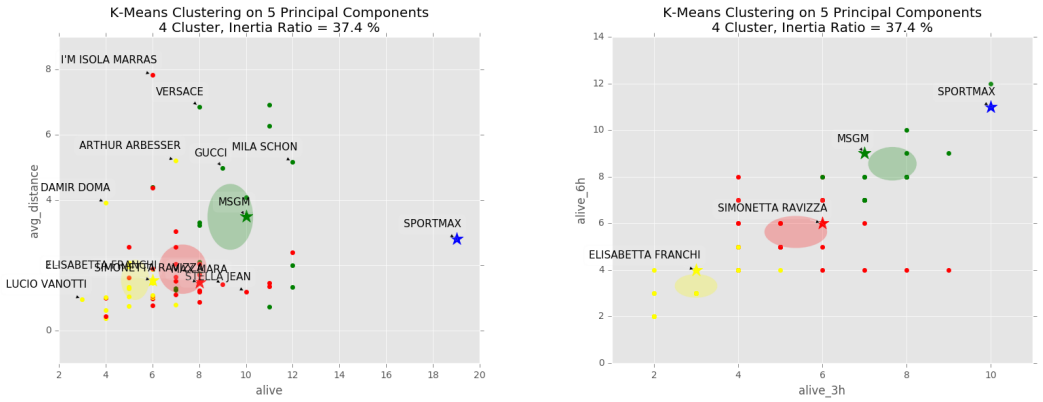



Figure 5.5: Curve representing the behaviour of the inertia with respect to the number clusters k. A good choice for this parameter seems to be in the interval from 3 to 6, where the slope of the curve is relaxing and getting more flat.



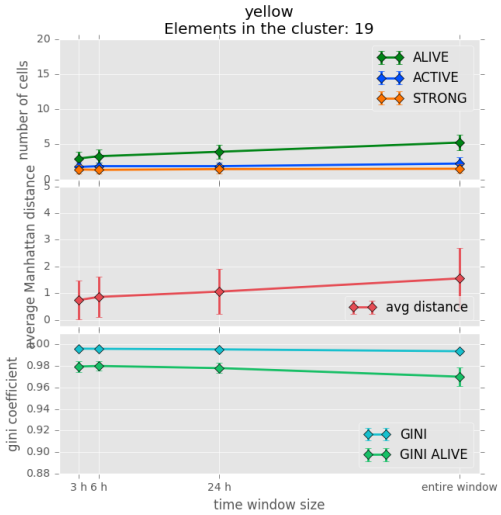

(a) (b)

Figure 5.6: K-means clustering results with the 5 principal components. In the left side we report graph in real-values on the two principal components, in the right side we have real-values of other two relevant components. The stars represent the most representative element within each cluster, in terms of distance from the cluster centroid.

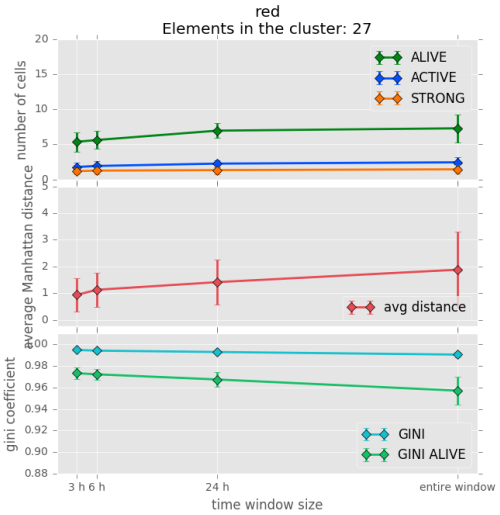
another relevant component is hidden but exploited, that is *alive_24h*, which characteristics are reflected in the clusters composition.

The most representative brands for each cluster are:

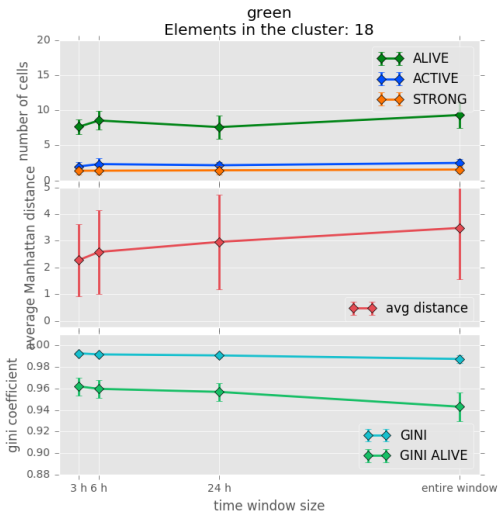
- *Elisabetta Franchi* for the Yellows, with 6 alive cells and average distance of 1.534;
- *Simonetta Ravizza* for the Reds, with 8 alive cells and average distance of 1.471;
- *MSGM* for the Greens, with 10 alive cells and average distance of 3.491;
- *Sportmax* for the Blues, with 19 alive cells and average distance of 2.81.



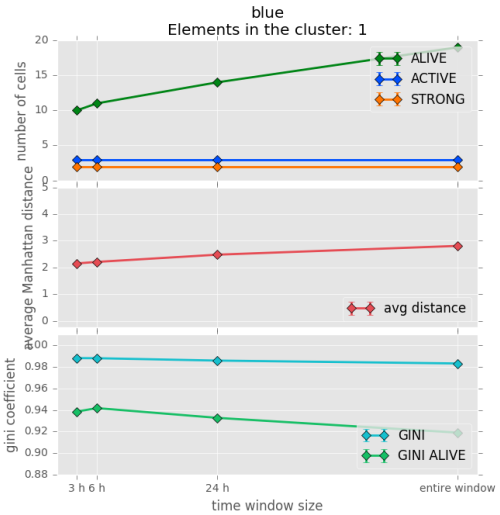

(a) Yellow cluster centroid




(b) Red cluster centroid




(c) Green cluster centroid




(d) Blue cluster centroid

Figure 5.7: All the features summaries for each cluster centroid.

# Chapter 6

# Cluster comparison

## 6.1 Introduction

The goal of this chapter is to find some correlation in all the clustering results we have obtained so far. Indeed, the final output of chapter 4 and chapter 5 are two different labellings of the brands we have in our scenario, coming out from different analyses. The problem that we are going to face in this section is to compare these different labellings, trying to catch similar patterns, whether they exist. In this way, we are proposing the model we have adopted in order to compare different clustering results, together with the new findings in terms of accuracy of juxtaposition.

## 6.2 Method

The problem we are going to face take some particular data as input for the analysis: these data are the output of the clustering phases of the previous chapters, then they are tuples of brand with the proper attached label. In our case, these labels are numbers, in the set $\{0, 1, 2, 3\}$, but we have to underline the fact that these numbers are meaningless, except for the sign of membership to a particular group. With this reflection, we can understand how same-name-labels (coming from two "orthogonal" clusterings) are completely incomparable. In other words, the relevant fact in clustering results is the knowledge of which elements are in the same group and which ones are not.

In this way, we are proposing a technique useful to compare different clustering results on the very same elements, that is independent of the fixed membership to the specific cluster. The basic idea is to fix one clustering result in terms of brand-attached labels and rename the other clustering labels with all the possible permutations in the set of adopted labels. For each re-labelling, we can compute a measure of correlation between the two clustering results, assuming complete acknowledgement between same-name-labels, and we can take the best renaming

permutation in terms of the specific measure adopted. We pick as statistic measure of validation the accuracy in juxtaposing one cluster to another one. We recall that we are in a multiclass case, then the accuracy will be the sum of the true-matchings between the two clustering results. The following lines of code should be helpful in understanding the algorithm.

```
1  def compare_clusters(c1, c2):
2      names_c1, names_c2 = [elem['name'] for elem in c1], [elem['name'] for elem
           in c2]
3      c1 = [elem for elem in c1 if elem['name'] in names_c2]
4      c2 = [elem for elem in c2 if elem['name'] in names_c1]
5      labels = set([elem['label'] for elem in c2])
6      permutations = itertools.permutations(labels)
7      result = []
8      for permutation in permutations:
9          re_label = {}
10         for l in range(len(labels)):
11             re_label[l] = permutation[l]
12         matchings = 0
13         for label in labels:
14             from_c1 = [elem['name'] for elem in c1 if elem['label'] == label]
15             from_c2 = [elem['name'] for elem in c2 if re_label[elem['label']] ==
                   label]
16             matchings += len(set(from_c1) & set(from_c2))
17         result.append({'permutation': permutation, 'matchings': matchings})
18     best = [r for r in result if r['matchings'] == max([r['matchings'] for r in
           result])][0]
19     re_label = {}
20     for l in range(len(labels)):
21         re_label[l] = best['permutation'][l]
22     c2 = [{'name': elem['name'], 'label': re_label[elem['label']]} for elem in
           c2]
23     matchings = []
24     for e1 in c1:
25         label = e1['label']
26         name = e1['name']
27         for e2 in c2:
28             if name == e2['name'] and label == e2['label']:
29                 matchings.append(name)
30     return (c1, c2, matchings, best['matchings']/len(c1), re_label)
```

Now that we have the best renaming we are also able to visualize the comparison with the help of a matching matrix. A matching matrix is the specific name for a confusion matrix, adopted in an unsupervised learning problem, and it is a specific table layout that allows visualization of the correlation in the two different results clustering. One clustering will be taken on the row side, while the other one will be taken on the column side. Each row will refer to the related predicted label of the first clustering, while each column will refer to the related predicted label of the second clustering. In this way, if all the elements are on the main diagonal, the two different clustering results will be totally correlated.

## 6.3 Findings

We have decided to compare the results coming out from three distinct analyses:

1. Time response clustering
2. Geo response clustering
3. Popularity response clustering

The first two of these analyses trivially refers to well described works at subsection 4.3.2 and subsection 5.3.2, while the third is a new one. We have simply extracted from our Twitter and Instagram databases a set of popularity features, related to each brand, that are the number of posts on Instagram, number of likes collected on Instagram, number of comments collected on Instagram, number of posts on Twitter, number of likes collected on Twitter, number of retweets collected on Twitter. Note that when we say "related to a specific brand", we are considering those posts with a matching on the brand-specific regular expression, as we have done in other studies before.

We now perform a PCA in order to find the best attributes that could describe our brands, considering only the 65 brands with fashion shows in schedule, finding that the 2 principal components are:

- Number of likes on Instagram (99.9% total var)
- Number of comments on Instagram (0.0025% total var)

In the end, we run k-means over these attributes, asking for 4 different clusters, in order to better compare our final results. Figure 6.1 shows the outcome of this analysis. The groups could be described as following: from the red cluster to the blue cluster we are going from the most unpopular brands to the most popular ones, in the two social media of Twitter and Instagram.

The incoming sections describe the comparison of these three different labellings, adopting the methods previously presented.

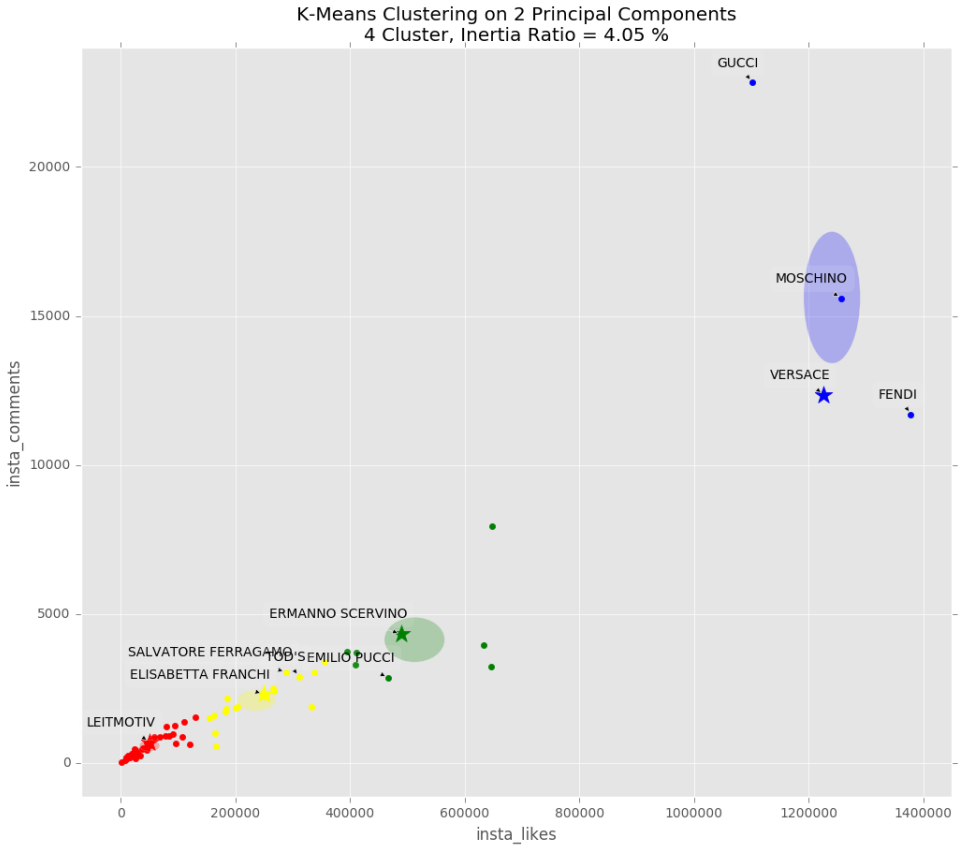



Figure 6.1: K-Means clustering result of our brands over the 2 principal components extracted from the social networks popularity analysis. The plot is in real values.

### 6.3.1 Time vs. Geo

Comparing the time response analysis clustering results with the geo response ones, we obtain the best juxtaposition with an accuracy of 44.62%, that produces the confusion matrix in Figure 6.2. In a few words, the best correlation is obtained juxtaposing:

- The red cluster from time, with lagged response peak at 15 minutes, with the green cluster from geo, with the highest results for average distance, the most dispersed one;
- The yellow cluster from time, with high immediate response, with the red cluster from geo, with average distances slightly more high than the yellow cluster;
- The green cluster from time, with lagged response peak at 45-60 minutes, with the yellow cluster from geo, with low average distance, the most concentrated one;
- The blue cluster from time, with an initial significant response but with the lagged response peak at 3 hours and 15 minutes, with the blue cluster from geo, the single element most dispersed.

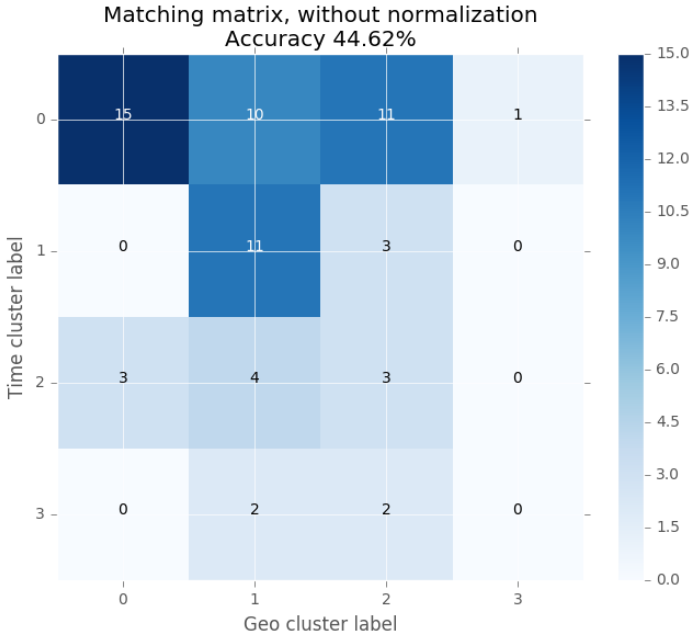



Figure 6.2: Matching matrix in comparing time response versus geo response.

### 6.3.2 Time vs. Popularity

Comparing the time response analysis clustering results with the popularity response ones, we obtain a nice juxtaposition with an accuracy of 41.54%, that produces the confusion matrix in Figure 6.3. In a few words, the best correlation is obtained juxtaposing:

- The red cluster from time, with lagged response peak at 15 minutes, with the red cluster from popularity, the most unpopular ones;
- The yellow cluster from time, with high immediate response, with the yellow cluster from popularity, the third ones for popularity;
- The green cluster from time, with lagged response peak at 45-60 minutes, with the blue cluster from popularity, the most popular ones;
- The blue cluster from time, with an initial significant response but with the lagged response peak at 3 hours and 15 minutes, with the green cluster from popularity, the ones just below the most popular.

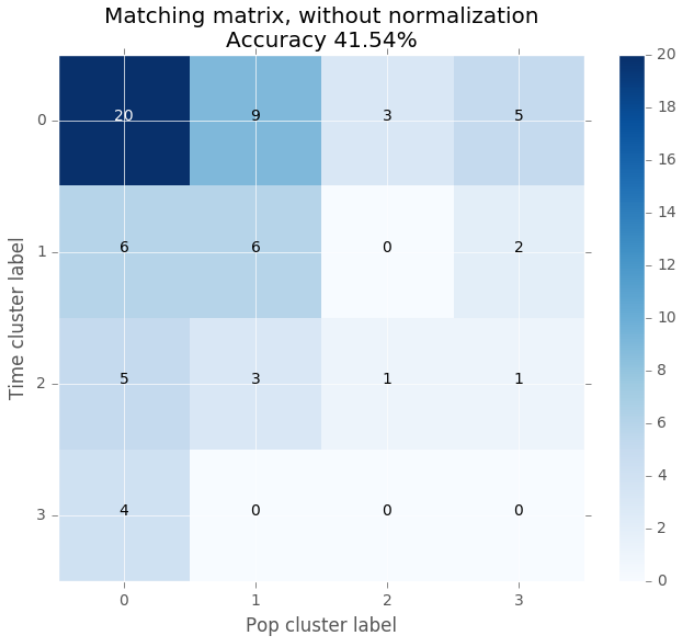



Figure 6.3: Matching matrix in comparing time response versus popularity response.

### 6.3.3 Geo vs. Popularity

Comparing the geo response analysis clustering results with the popularity response ones, we obtained a juxtaposition with an accuracy of 38.46%, that produces the confusion matrix in Figure 6.4. In a few words, the best correlation is obtained juxtaposing:

- The green cluster from geo, with the highest results for average distance, the most dispersed one, with the blue cluster from popularity, the most popular ones;
- The yellow cluster from geo, with low average distance, the most concentrated one, with the red cluster from popularity, the most unpopular ones;
- The red cluster from geo, with average distances slightly more high than the Yellow cluster, with the yellow cluster from popularity, the third ones for popularity;
- The blue cluster from geo, the single element most dispersed, with the green cluster from popularity, the ones just below the most popular.

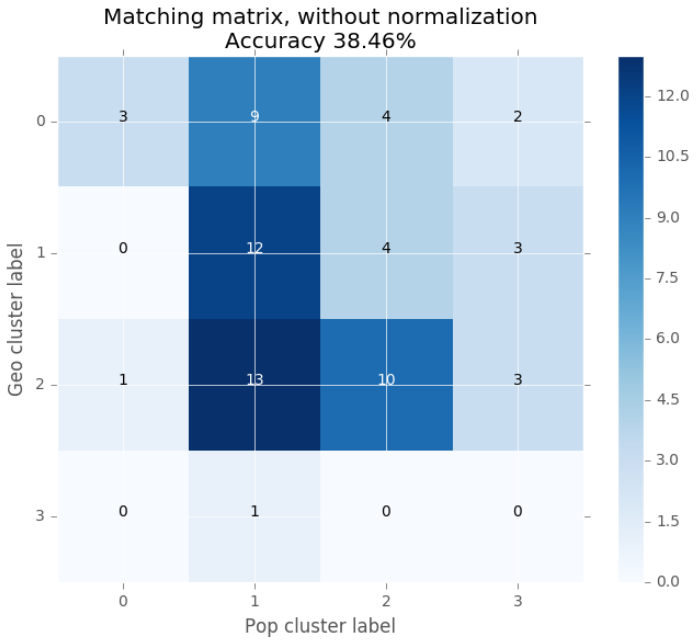



Figure 6.4: Matching matrix in comparing geo response versus popularity response.

### 6.3.4 Conclusion

Wrapping up all the previous comparisons, we have to say that, in some sense, we are not so satisfied of the results in terms of high accuracy. Indeed, we have not detected strong levels of correlation between the different groupings, and from the most trivial (but maybe most business-relevant) point of view we can notice how the categorization of the brands based on their returns in terms of popularity on the social media platforms is not precisely reflected in other categorizations based on the time or on the geo response.

However, the findings of this last part of the work state the fact that, in order to characterize and distinguish all the different brands taking part in such a popular-live event, considering the bigger picture could give more insights, bringing a big help in the exploration. As a matter of these results, a complete and well-defined brand categorization is possible only taking into account the three different analysis we have just proposed.

# Chapter 7

# Conclusions

In this project we have developed some detailed techniques and methods in order to study the social media response to live events.

In particular, we first have built a brief discussion on the correlation of some measures related to the authors of contents on the Twitter and Instagram platforms, showing how the volumes of messages and media shared on the social network and the influential power of the author are positively correlated with number of likes and comments or retweets received back. On the other hand, big amounts of posts shared are not leading to high average popularity values on the same posts, and, likewise, in a specific domain of interest, the influential power of the user is not correlated at all with the generated content volume.

Then, we have shown some statistical methods in order to characterize and, after that, categorize the brands and their own fashion shows with respect to the time response. In this study, we have also provided the results of this categorization. We have tried to make one step ahead, facing the supervised problem of classifying the type of time response of each brand-event tuple, given the scarce information on the events.

Again, we have indicated some techniques in order to characterize and, after that, categorize the brands and their own fashion shows with respect to the geo response. Like in the previous analysis, we have provided the results of this categorization.

After producing another grouping, this time based on a trivial popularity response on the social networks, we have juxtaposed the different results in the clustering of the brands, demonstrating how the different analysis underline different facts.

There are some points that could give more interesting results with future developments. At first, regarding the authors analysis it is possible to make one step forward in categorizing the users involved in this specific scenario. In this way, the actors offering the shows, that are the different brands, could look more carefully all the information related to the popularity and the influential strength of the more relevant authors and capture more relevant insights. In

the subsequent analyses, K-Means clustering has been widely adopted. This method has its own limitations, that could be overcome using a different and more complex technique like mixture models, that could help in grouping better the elements. Also the clustering works themselves could be proposed in other different ways: for example, in the time scenario could be nice to cluster the data over the real values of the statistic tests we have performed, and not on the pick-normalized ones, in order to capture another relevant information.

In the end, all the results in terms of volume and diffusion of contents on the social networks of Twitter and Instagram, could give the event organizer a better view of the social behaviour in response of such an event, that should be helpful for future happenings.

# Bibliography


[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. 1998.

[2] P. Arcaini, G. Bordogna, D. Ienco, and S. Sterlacchini. User-driven geotemporal density-based exploration of periodic and not periodic events reported in social networks. *Information Sciences*, pages 122–143, 2016.

[3] E. Bakshy, J.M. Hofman, W.A. Mason, and D.J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. 2011.

[4] E. Bakshy, J.M. Hofman, W.A. Mason, and D.J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. 2011.

[5] H. Becker, M. Naaman, and L. Gravano. Learning Similarity Metrics for Event Identification in Social Media. 2010.

[6] H. Becker, M. Naaman, and L. Gravano. Beyond Trending Topics: Real-World Event Identi

cation on Twitter. 2011.

[7] D.M. Boyd and N.B. Ellison. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, pages 210–230, 2008.

[8] Julie Bradford. *Fashion Journalism.* 2014.

[9] F. Calabrese, F.C. Pereira, G. Di Lorenzo, L. Liu, and C. Ratt. The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events. *Pervasive 2010*, pages 22–37, 2010.

[10] S. Casalegno, R. Inger, C. DeSilvey, and K.J. Gaston. Spatial Covariance between Aesthetic Value & Other Ecosystem Services. *PLoS ONE*, 2013.

[11] M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. *Association for the Advancement of Artificial Intelligence*, 2010.

[12] D. Chakrabarti and K. Punera. Event Summarization using Tweets. *Yahoo! Research*, 2011.

[13] L. Chen and A. Roy. Event Detection from Flickr Data through Wavelet-based Spatial Analysis. 2009.

[14] J. Chrisler, K. Fung, A. Lopez, and J. Gorman. Suffering by comparison: Twitter users' reactions to the Victoria's Secret Fashion Show. *Body Image*, pages 648–652, 2013.

[15] L. de Vries, S. Gensler, and P. Leeflang. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *Journal of Interactive Marketing*, 2012.

[16] C. Ding and X. He. K-means Clustering via Principal Component Analysis. *Proceedings of the 21 st International Conference on Machine Learning*, 2004.

[17] J. Entwistle and A. Rocamora1. The Field of Fashion Materialized: A Study of London Fashion Week. *Journal of Sociology*, pages 735–751, 2006.

[18] W. Guan, H. Gao, M. Yang, Y. Li, H. Ma, W. Qian, Z. Cao, and X. Yang. Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events. *Physica A*, pages 340–351, 2013.

[19] R. Hanna, A. Rohm, and V.L. Crittenden. We're all connected: The power of the social media ecosystem. *Kelley School of Business, Indiana University*, 2011.

[20] T. Hastie, R. Tibshirani, and J Friedman. *The Elements of Statistical Learning.* 2009.

[21] M. Helbich, M. Leitner, and N.D. Kapusta. Geospatial examination of lithium in drinking water and suicide mortality. *International Journal of Health Geographics*, pages 11–19, 2012.

[22] Y. Hu, L. Manikonda, and S. Kambhampati. What We Instagram: A First Analysis of Instagram Photo Content and User Types. *Association for the Advancement of Artificial Intelligence*, 2014.

[23] Instagram. Instagram API. https://www.instagram.com/developer/, May 2016.

[24] L. Jing, M.K. Ng, and J.Z. Huang. An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data. *IEEE Transactions on Knowledge and Data Engineering*, pages 1026–1041, 2007.

[25] A.M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Kelley School of Business, Indiana University*, 2009.

[26] A.J. Kim and E. Ko. Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business Research*, 2011.

[27] W.R Knight. A Computer Method for Calculating Kendall's Tau with Ungrouped Data. *Journal of the American Statistical Association*, pages 436–439, 1966.

[28] H. Kwaw, C. Lee, H. Park, and S Moon. What is Twitter, a Social Network or a News Media? *International World Wide Web Conference Committee*, 2010.

[29] V. Lorant, I. Thomas, D. Deliége, and R. Tonglet. Deprivation and mortality: the implications of spatial autocorrelation for health resources allocation. *Social Science & Medicine 53*, pages 711–719, 2001.

[30] L. Manikonda, R. Venkatesan, S. Kambhampati, and B. Li. Trending Chic: Analyzing the Influence of Social Media on Fashion Brands. *Department of Computer Science, Arizona State University*, 2016.

[31] Kakihara Masao. Grasping a Global View of Smartphone Diffusion: An Analysis from a Global Smartphone Study. *International Conference on Mobile Business*, 2014.

[32] M. Naaman, J. Boase, and C.H. Lai. Is it Really About Me? Message Content in Social Awareness Streams. *Rutgers University, School of Communication and Information*, 2010.

[33] J.A. Obar and S. Wildman. Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications policy, 39(9)*, pages 745–750, 2015.

[34] R. Okada, B. Stenger, T. Ike, and N. Kondo. Virtual Fashion Show Using Real-Time Markerless Motion Capture. *Corporate Research & Development Center, Toshiba Corporation*, 2006.

[35] Pagan. *VVIS*. PhD thesis, Politecnico di Milano, 2000.

[36] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks. Identifying causal relationships in case of non-stationary time series. 2014.

[37] L. Parsons, E. Haque, and H Liu. Subspace Clustering for High Dimensional Data: A Review. *Sigkdd Explorations*, pages 90–105, 2004.

[38] S. Sridharan, H. Tunstall, R. Lawder, and R. Mitchell. An exploratory spatial data analysis approach to understanding the relationship between deprivation and mortality in Scotland. *Social Science & Medicine 65*, pages 1942–1952, 2007.

[39] R. Volonterio. Twitter Scraper CLI. https://github.com/Volox/Scraper, April 2016.

[40] H.O. Zapata, M.A. Hudson, and P. Garcia. Identifying Causal Relationships Between Nonstationary Stochastic Processes: An Examination of Alternative Approaches in Small Samples. *Western Journal of Agricultural Economics*, pages 202–215, 1988.

[41] H. Zha, X. He, C. Ding, H. Simon, and M. Gu. Spectral Relaxation for K-means Clustering.

[42] D. Zwillinger and S. Kokoska. *CRC Standard Probability and Statistics Tables and Formulae*. 2000.

# List of Figures

# List of Tables