Spatial Analysis of Social Media Response to Live Events

The Case of the Milano Fashion Week

Marco Brambilla, Stefano Ceri, Florian Daniel, Gianmarco Donetti
Politecnico di Milano, DEIB
Via Ponzio 34/5. I-20133 Milano
{firstname.lastname}@polimi.it

ABSTRACT

Social media response to catastrophic events, such as natural disasters or terrorist attacks, has received a lot of attention. However, social media are also extremely important in the context of planned events, such as fairs, exhibits, festivals, as they play an essential role in communicating them to fans, interest groups, and the general population. These kinds of events are geo-localized within a city or territory and are scheduled within a public calendar. We consider a specific scenario, the Milano Fashion Week (MFW), which is an important event in our city.

We focus our attention on the coverage of social content in *space*, measuring the propagation of the event in the territory. We build different clusters of fashion brands, we characterize several features of propagation in space and we correlate them to the *popularity* of involved actors. We show that the clusters along space and popularity dimensions are loosely correlated, and that domain experts are typically able to understand and identify only popularity aspects, while they are completely unaware of spatial dynamics of social media response to the events.

Keywords

Social media analysis; live events;

1. INTRODUCTION

Thanks to the wide adoption of smartphones, which enable continuous sharing of information with our social network connections, the online response to popular real world events is becoming increasingly significant, not only in terms of volumes of contents shared in the social network itself, but also in terms of velocity in the spreading of the news about events with respect to the time and to the geographical dimension. It has been noted that social signals are at times faster than media news with highly impacting events, such as terrorist attacks or natural disasters.

This work deals more specifically with the problem of social media response to a scheduled and popular real world event, the Milano Fashion Week occurred from the 24^{th} to the 29^{th} February of 2016, analysing the behaviour of users who re-acted (or pro-acted) in relationship with each specific fashion show during the week.

MFW, established in 1958, is part of the global "Big Four fashion weeks", the others being located in Paris, London and New York [4]; it is organized by *Camera Nazionale della Moda Italiana*, who manages and fully co-ordinates about 170 shows, presentations and events, thus facilitating the work of showrooms, buying-offices, press offices, and public relations firms. 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. MFW represents the most important meeting worldwide between market operators in the fashion industry.

We formulate our problem as the analysis of the response in time and space of popular events on social media platforms, by correlating each fashion brand to a specific class of social responses in the two dimensions. We search for patterns and indicators of such social media responses that enable us to understand how time and space play a role and if the specific fashion brand can be linked to such patterns and indicators.

Our goal is to describe and characterize the social media response to the events which appear in the official calendar, in terms of spatial features of dispersion/concentration of social media signals related to each specific fashion show. Based on our analyses, we build different clusters of stakeholders (fashion brands). We also consider online popularity of fashion brands, and we show that space and online popularity provide different angles on the reaction of the public to the event.

The paper is organized as follows: Section 2 presents the data collection and preparation approach. Section 3 presents the definition and evaluation of several features for measuring the dispersion in space of the social response. Section 4 describes the clustering of brands based upon those features and compares resulting classes with brand popularity, whose measure is also discussed in terms of social interaction. Section 5 contains related work, and Section 6 concludes.

2. DATA COLLECTION & PREPARATION

We initially extracted posts by invoking the social network APIs of Twitter and Instagram; for identifying the social reactions to MFW, we used a set of 21 hashtags and keywords provided by domain experts in the fashion sector, i.e., researchers of the *Fashion in Process* group (FIP) of Politec-

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3051698

nico di Milano[1]. We focused on 3 weeks, before, during and after the event. In this way, we collected 106K tweets (out of which only 6.5% geolocated) and 556K Instagram posts (out of which 28% geolocated); eventually, we opted for considering only Instagram posts, as they represent a much richer source for the particular domain of Fashion with respect to Twitter.

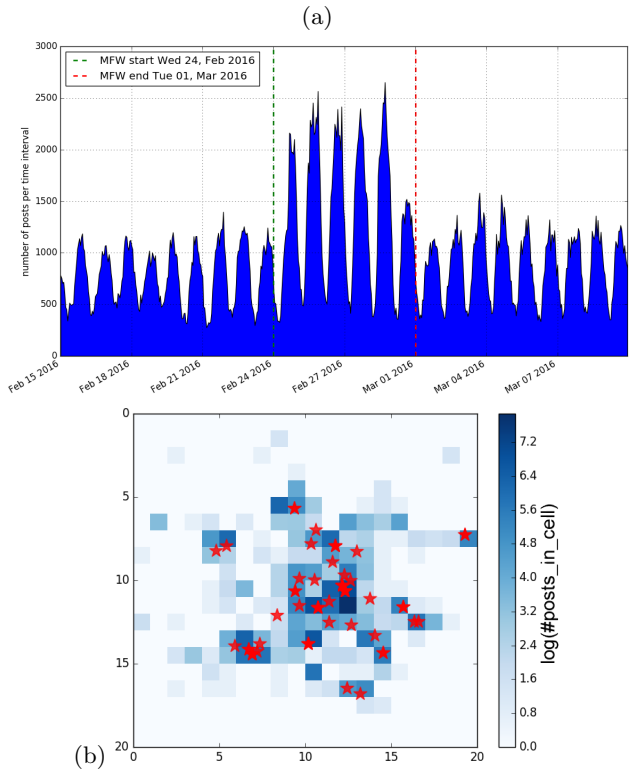

Figure 1: Temporal overview for the three analyzed weeks of Instagram posts (a) and map representing the geographical distribution of events (represented by red stars) and post density (b).

We performed an initial analysis of the content, for associating each post with the corresponding event. In this specific scenario, the task was simple because each event was directly associated with a fashion brand, mentioned in the posts; the characterization of the brands was again provided by the FIP experts. For instance, for identifying the posts related to the Gucci catwalk, which was held on February 24th at 2:30pm in Milano, Via Valtellina 17, we collected the posts containing the hashtags and keywords *#Gucci* and *Gucci*, filtering the posts through suitable regular expressions. This allowed us to collect 7718 Instagram posts related to that specific event. Figure 1 shows the temporal and geographical distribution of the posts.

3. SPATIAL RESPONSE ANALYSIS

We focused on geographical dispersion of social media response. We have two different spatial signals: (1) the calendar events; and (2) the volume of social media posts on the Web with geographical information attached, i.e., latitude and longitude. Given these two signals, several features can

[1]http://www.fashioninprocess.com/

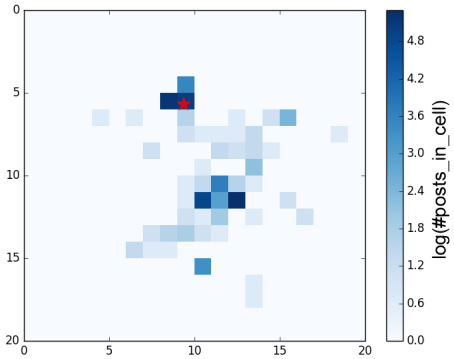

Figure 2: Social signal related to the events of Gucci. We report the position of the event (red star) and the heatmap of the post density over the cells.

be computed in order to describe the spatial dispersion of posts following an event. We focused on *fashion shows*, a specific type of event during MFW. As before, we run our analysis brand by brand.

To obtain a more compact and aggregate description of the data, we didn't consider the geographic coordinates as continuous values. Instead, we built a grid of cells above the area of Milano city and assigned each post to the appropriate cell. The grid has a square shape, with sides of $10km$, divided into 20 rows and 20 columns, for a total of 400 cells of $500m \times 500m$ (which is enough to discriminate among the most important spots in downtown Milano). For each brand, we associated to each cell of the grid the number of posts geo-localized within the grid; also the calendar events are assigned to their own specific cell. In order to evaluate how dispersion is changing over time, we considered for each brand four different time windows, with a temporal duration respectively set to 3, 6, and 24 hours since the beginning of the fashion show, plus the whole analyzed period. In Figure 2 we can see the heatmap related to *Gucci*, with a time window including all the days of analysis.

We next defined three states of cell, in order to capture the evolution of the dispersion of the social signal within them:

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

We computed the number of alive, active and strongly active cells for all brands; we also computed the differences between subsequent durations (e.g. 3h - 6h) by counting how many cells changed their state.

We then computed different measures that reflect the dispersion of the social media signal over time, using:

1. *Gini coefficient*;

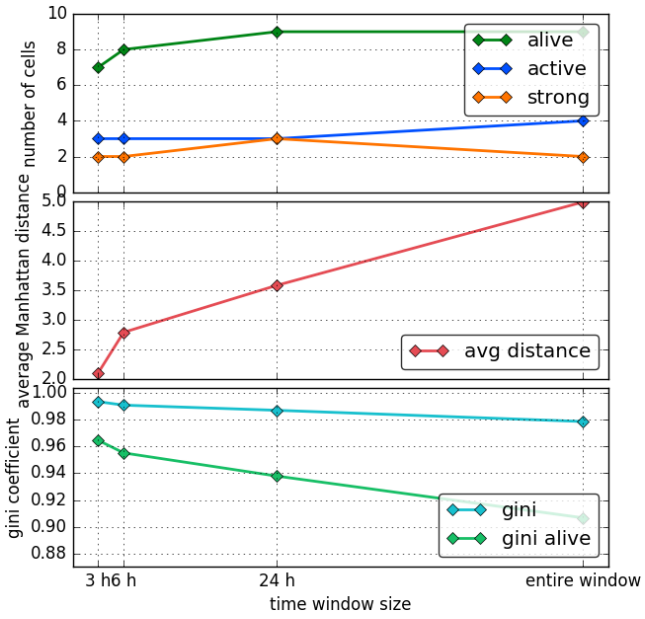

Figure 3: Spatial features of posts about Gucci in four different observation intervals: $3h{+}6h{+}24h{+}3w$**.**

2. *Average distance* of the social media signals from the event location;

3. Number of *alive*, *active* and *strongly active* cells.

Fig. 3 anticipates them for Gucci; each measure is described in the following sections.

3.1 Gini coefficient

The Gini coefficient is a measure of statistical dispersion[2]. computed by the following formula.

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

where:

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

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 the remaining 399 cells with no social signal). For large groups, values close to or above 1 are unlikely.

We computed the Gini coefficient on two different models:

- On the complete grid of cells;
- Only on those cells that were "alive" for at least one brand in the specific time-window of analysis.

We make these distinctions because of the extremely high concentration of posts in few cells and because of the presence of a lot of cells that are "dead" for every brand event, corresponding to suburbs. The second model has an average of 40 cells instead of 400.

[2]It was developed by the Italian statistician and sociologist Corrado Gini and published in his 1912 paper "Variability and Mutability", as a measure of inequality of income or wealth of a nation's residents; in our specific scenario, we associate cells of the grid to people and the number of posts in each cell to their income.

3.2 Average Manhattan Distance

We then defined the average distance of the posts from the event as:

$$avgDist = \frac{1}{\sum_{r=1}^{R}\sum_{c=1}^{C}\mathbf{G}_{r,c}}\sum_{r=1}^{R}\sum_{c=1}^{C}\mathbf{G}_{r,c}\times dist(\langle r,c\rangle, \langle e_r, e_c\rangle). \quad (2)$$

In the above formula, the *dist* function is the distance computed between cells in Manhattan way, with parameters like the tuple $\langle r, c\rangle$ indicating the row and the column 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 mean high dispersion of the social signal, far away from the cell where the event is taking place; low values mean high concentration near the event location. Although *Gini coefficient* and *Average distance* seem to give the same information, the former measures the concentration regardless of the location, while the latter measures the dispersion of the social signal from the specific cell of the event.

3.3 Analysis of Cell States

In addition to measuring cells as *alive, active, and strongly active* at given times (3 to 6 to 24 hours), we also monitored their change of state, counting how many of them turn *on* or *off* by passing from 3 to 6 and from 6 to 24 hours.

3.4 General Observations about Features

By observing the collected features, we note that:

- As we increase the width of the time-window, the number of *alive cells* also increases. On the other hand, 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*.

- At the start of the event, posts are shared near the event location, but as we look at the bigger picture, including 24 hours or even the entire period of 3w, the *average distance* increases, showing the growing dispersion of the social signal. While the high activity around the location of the event is expected and does not require further investigation, our preliminary analyses give the impression that the subsequent birth of related active cells is correlated with locations in Milano that are highly visited in general; yet, this intuition needs further investigation to be confirmed.

- 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.

As an example, Figure 3 shows the geographical features of posts related to the Gucci event, at the different time intervals (we did not opt for uniform intervals, as these would have produced intervals with too few data for meaningful comparisons). One can notice that the dispersion and the number of activated cells usually increases over time.

4. CLUSTERING OF BRANDS BASED ON CELL STATES

4.1 Principal Components Analysis (PCA)

Given that we collected a large amount of analysis dimensions (42), we started with a Principal Components Analysis (PCA), an unsupervised learning method widely used for decomposing a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance.

The results of the application of PCA to our data are reported in Table 1. We can see the ranking of the top 10 attributes, sorted with respect to the explained variance ratio. Note that by choosing the first two principal components we are able to capture about 60% of the total variance, while if we choose the first seven principal components we capture about 90%.

Table 1: Ranking coming from PCA in terms of explained variance ratio.

feature	Explained variance ratio
alive cells	0,4577
average m. distance	0,1390
alive cells 6h	0,1132
alive cells 24h	0,0929
alive cells 3h	0,0509
active cells	0,0321
alive_off_from_6h	0,0239
alive_on_from_24h	0,0207
alive_on_from_3h	0,0156
active cells 6h	0,0117

4.2 K-Means Clustering

After the analysis of the principal components, we performed a k-means clustering. We decided to use the top 5 features from Table 1, as they capture at least 85% of the explained variance.

In order to decide the ideal number of clusters, we ran the k-means clustering algorithm with different k, from 1 to 15. Then we looked at the inertia trend for all these values of k; inertia (or within-cluster sum-of-squares) is a measure of internal coherence; the inertia curve is monotonically decreasing, with the maximum value corresponding to just one global cluster and the minimum value equal to 0 when the number of clusters coincide with the number of elements. We then picked k=4, as inertia decreases more slowly for $K \geq 4$.

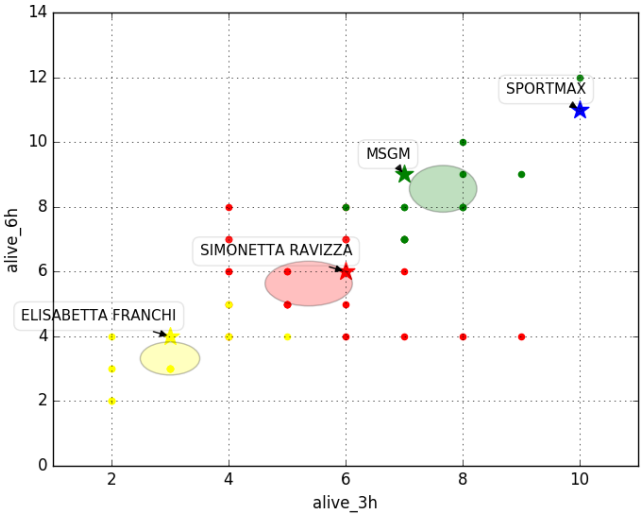

Figure 4: K-means clustering results with k=4 using the number of alive cells and the Manhatten distance as principle components; the stars represent the most representative element within each cluster, with minimal distance from the cluster centroid.

In Figure 4, we present the resulting 4 clusters, labeled with different colours and described as follows:

- *Yellow*, with really few *alive cells* and low *average distance*;
- *Red*, with higher number of *alive cells* and *average distance* than the *Yellow* cluster;
- *Green*, with highest number of *alive cells* and *average distance*;
- *Blue*, a single element with very high values in *alive cells* both for 3 hours and 6 hours time scopes - we conjecture an organized behavior of brand promoters.

The most representative brands for each cluster are:

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

4.3 Popularity Analysis

We next turned to a simpler observation, the *brand popularity*, in order to evaluate the popularity of a brand and see if it relates to the above clusters; we focused on 65 brands which were hosting fashion shows during MFW. In our analysis, We extracted from our Twitter and Instagram datasets a classic set of popularity features, related to each brand (mentioned in a post): the number of posts on Instagram, number of likes collected on Instagram, number of comments collected on Instagram, number of posts on Twitter, number

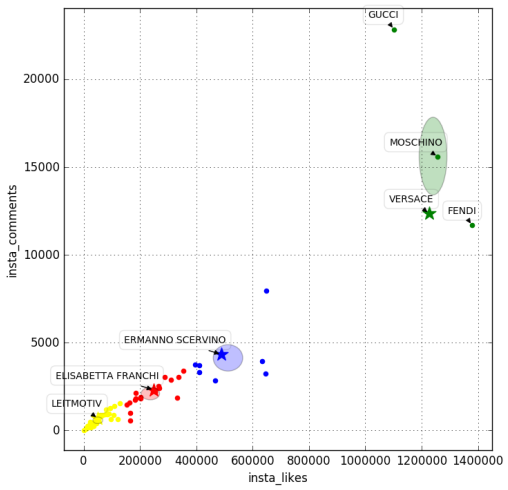

Figure 5: K-Means clustering result of our brands over the 2 principal components extracted from the social networks popularity analysis. The plot is in real values.

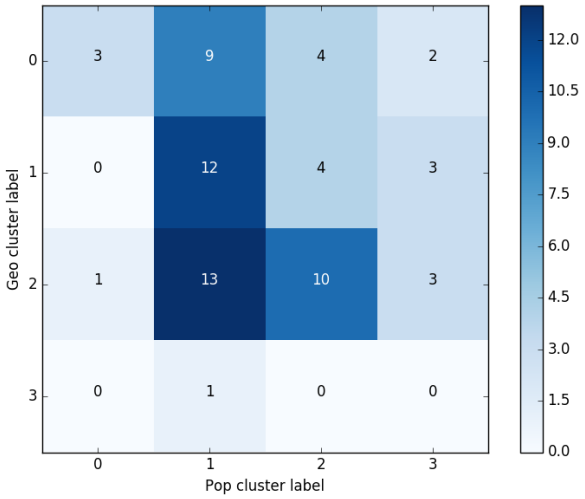

Figure 6: Matching matrix in comparing geo response versus popularity response.

of likes collected on Twitter, number of retweets collected on Twitter. We then performed a PCA to find the best features, and nonsurprisingly we noticed that likes on Instagram essentially dominate, as the 2 principal components are:

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

In the end, we run k-means over these attributes, asking again for 4 different clusters, in order to better compare our final results. Figure 5 shows the outcome of this clustering. 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. These results were confirmed by our experts in fashion design.

4.4 Cluster comparison

We then studied the correlation between the two clusterings. We tried to fix one clustering result in terms of brand-attached labels and renamed the other clustering labels with all the possible permutations in the set of adopted labels. For each re-labelling, we computed a measure of correlation between the two clustering results, assuming complete acknowledgement between same-name-labels, and then we took the best renaming permutation in terms of the specific measure adopted. We picked as statistic measure of validation the accuracy in juxtaposing one cluster to another. We recall that in a multiclass case the accuracy is measured as the sum of the true-matchings between the two clusterings.

We visualize the result of the comparison by means of a matching matrix, i.e. a confusion matrix that allows the visualization of the correlation in the two different results clusterings. One clustering will be taken on the row side, while the other one will be taken on the column side. Each row refers to the related predicted label of the first clustering, while each column refers to the related predicted label of the second clustering. In this way, if all the elements are on the main diagonal, the two different clusterings are totally correlated.

Comparing the geo-response analysis clustering results with the popularity response clustering, we obtained a juxtaposition with an accuracy of 38.46%, which produces the confusion matrix in Figure 6 (we number clusters incrementally from bottom-left to top-right in Figure 5). 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 higher 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.

As a conclusion, the different clusterings show only limited correlation. These numbers highlight that popularity alone is not a sufficient analysis dimension for understanding how brands and events interact during a large, city–wide event.

5. RELATED WORK

Social media event response. The work [11] selects 21 hot events, which were widely discussed on Sina Weibo, and empirically analyzes their posting and reposting characteristics. In the work [2], 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. The work [3] underlines 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 distinguish between messages about real-world events and non-event messages. In all these works no spacial analyses were run.

Spatial analyses of social media event response. The focus of [7] is to detect events from photos on Flickr by exploiting the tags supplied by users. In particular, the temporal and locational distributions of tag usage are analyzed. The problem of event summarization using tweets is well faced by [6], where the authors argue that for some highly structured and recurring events, such as sports, it is better to use sophisticated techniques to summarize the relevant tweets via Hidden Markov Models. The paper [5], adding the information given by cell-phone traces, deals with the analysis of crowd mobility during special events. They show that the origins of people attending an event are strongly correlated to the type of event. Finally, [1] proposes 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 paper aims at discovering the spatio-temporal periodic and non-periodic characteristics of events occurring in specific geographic areas.

Social Media Analysis for Fashion. The work [13] presents a qualitative analysis on the influence of social media platforms on different behaviors of fashion brand marketing. They analyze their styles and strategies of advertisement. The authors employ both linguistic and computer vision techniques. The study [12] sets out to identify attributes of social media marketing (SMM) activities and examines the relationships among those perceived activities, value equity, relationship equity, brand equity, customer equity, and purchase intention through a structural equation model. The findings of [9] show that different drivers influence the number of likes and the number of comments to fashion posts. Namely, vivid and interactive brand post characteristics enhance the number of likes. The analysis in [8] shows that many of the tweets during a 2011 Victoria's Secret Fashion Show were discussing the social status of the fashion models. The article [10] examines the London Fashion Week (LFW), arguing that this event effectively represents the field of fashion, as it shows the boundaries, relational positions, capital and habitus at play in the field. Finally, [14] develops 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.

6. CONCLUSIONS

We discussed how social content responds to live events in function of space, focusing on the Milano Fashion Week. We demonstrated that brands can be clustered into 4 classes of increasingly far-reaching responses, from most concentrated ones to most dispersed ones; we also showed that brand popularity alone is not sufficient for explaining dispersion. Our future work is to build a predictive model of the spacial dynamics of social content, and also attempt to correlate spreading with other features beyond brand popularity, e.g. by studying the profiles of each brand's social networks and specifically of Instagram. We also would like to understand better

7. ACKNOWLEDGMENTS

We wish to thank the FashionInProcess group of Politecnico (http://www.fashioninprocess.com/the-collective), and especially Paola Bertola, Chiara Colombi and Federica Vacca, who supported us in the definition of the domain-specific knowledge related to the event.

References

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

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

[3] H. Becker, M. Naaman, and L. Gravano. Beyond Trending Topics: Real-World Event Identication on Twitter. 2011.

[4] J. Bradford. *Fashion Journalism.* 2014.

[5] F. Calabrese, F. 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.

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

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

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.