Expert Systems with Applications 42 (2015) 3381–3395

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# A new model to quantify the impact of a topic in a location over time with Social Media

J. Bernabé-Moreno [a], A. Tejeda-Lorente [a], C. Porcel [b,*], E. Herrera-Viedma [a]

[a] University of Granada, Department of Computer Science and Artificial Intelligence, Granada, Spain
[b] University of Jaén, Department of Computer Science, Jaén, Spain

ARTICLE INFO

*Article history:*
Available online 20 December 2014

*Keywords:*
Social Media impact
Topic Engagement
Topic Exposure
Social Media sensor
Geo-located Social Media
Social network analysis
RFM
Information extraction

ABSTRACT

Social Media can be used as a thermometer to measure how society perceives different news and topics. With the advent of mobile devices, users can interact with Social Media platforms anytime/anywhere, increasing the proportion of geo-located Social Media interactions and opening new doors to localized insights. This article suggests a new method built upon the industry standard Recency, Frequency and Monetary model to quantify the impact of a topic on a defined geographical location during a given period of time. We model each component with a set of metrics analyzing how users in the location actively engage with the topic and how they are exposed to the interactions in their Social Media network related to the topic. Our method implements a full fledged information extraction system consuming geo-localized Social Media interactions and generating on a regular basis the impact quantification metrics. To validate our approach, we analyze its performance in two real-world cases using geo-located tweets.



## 1. Introduction

The usage of Social Media (SM) is an ever-growing phenomenon (White, 2013). Media consumers are increasingly shifting from classic (printed) media to digital platforms. As a result the communication stops being one-way with clearly defined *author*/*reader* roles. With the advent of the web 2.0, the definition of *author* started to blur. The blogosphere empowered readers to make their own contributions to the content published by a given author, which radically increased the information richness, adding further perspectives and points of view. Simultaneously, media started to be democratized, as anybody could start a blog and the visibility of the blog in the search engines was determined a priori by the number of people that considered the blog to be relevant outside the realm of the paid search (Page, Brin, Motwani, & Winograd, 1998).

The SM platforms based on the concept of *micro-blogging* took it to the next level, as everybody could be an author and a reader anytime. The *push-first*, *comment-later* paradigm so popular in the blogosphere started to look old-fashioned. Rather, anybody was empowered to initiate a communication, enrich an existing thread, jump from a thread to another one, ignore, criticize, share richer content like pictures, videos, etc. The ease of publishing, sharing and consuming content boosted the adoption of these Social Media platforms as the place to talk anytime about everything with everybody. The best example is Twitter, which has become a communication platform for almost all the digital world (Kwak, Lee, Park, & Moon, 2010). By March 2012 the platform counted 140 million active users creating an average of 340 million tweets a day (Bennett, 2012). The night of November 7th, during 8:11 and 9:11 pm when the world wanted to share the results of the US elections, an average of 9965 Tweets per Second (TPS)[1] resulted in the creation of more than 35 million tweets within one hour.

With the advent of wireless internet technologies based on WiFi hot-spots and mobile communication networks, the Social Media content creation became more pervasive. The access to the digital media was no longer exclusive to desktops; the rise of the smartphones and mobile data packages enabled the always-on era and opened the door to a new set of insights based on the location where the user interacted with the social network. As the proportion of geo-located SM interactions increased, the *geo-fencing* or delimitation of the location boundaries where the SM dialog took

* Corresponding author.
*E-mail addresses:* juan.bernabe-moreno@webentity.de (J. Bernabé-Moreno), atejeda@decsai.ugr.es (A. Tejeda-Lorente), cporcel@ujaen.es (C. Porcel), viedma@decsai.ugr.es (E. Herrera-Viedma).

[1] https://blog.twitter.com/2012/bolstering-our-infrastructure

http://dx.doi.org/10.1016/j.eswa.2014.11.067
0957-4174/

place became more accurate. These new capabilities led to more meaningful and representative analytic results, to the point that the SM activity could be taken as a good indicator of what is happening anytime anywhere.

The influence and impact in the SM channel has been matter of research almost from the advent of the modern SM networks and platforms. Yet, the research community mainly focused on understanding and modeling the impact of a particular user or a particular group of users on their own and foreign social networks. Our intent here is to prove that the impact of a given topic can be measured, quantified and monitored over time. Obviously, this topic centric geo-located impact measuring would open a new window of possibilities in different domains, such as understanding the performance of marketing campaigns on a given area, or understanding the affinity of local communities to certain marketing offers. Likewise, the generated insights can be used in the area of recommender systems and applied in different scopes, especially in e-commerce and digital media (Porcel, Tejeda-Lorente, Martnez, & Herrera-Viedma, 2012; Tejeda-Lorente, Porcel, Peis, Sanz, & Herrera-Viedma, 2014). Unlike the metrics typically used to assess the SM influence of a particular user, which mainly rely on well-defined entities and parameters present standard-wise in social network platforms (like *User*, *Friend*, *Follower*, etc.), there's no entity to represent a *topic*. Thus, modeling techniques need to be applied, which introduces a new level of complexity entering in the realm of semantic web (Berners-Lee, Hendler, & Lassila, 2001) and Natural Language Programming (NLP) (Manaris, 1998). Although our aim was not solving any NLP problem, we implemented a system to extract the required information from the Social Media networks and to apply the quantification methodology for a topic which relies on a whole set of NLP components.

The purpose of this paper is to define a new method to quantify the impact of a topic during a period of time on a given place based on how the users located in this place are exposed to the topic over their social network and how they actively interact or engage with the topic themselves. In other words, we want to turn the Social Media platform Twitter into a topic impact thermometer. Our method relies on the well-established industry-standard Recency, Frequency and Monetary schema (RFM) (Bult & Wansbeek, 1995). RFM models have been employed in the industry for almost 30 years to identify and segment the customer base in countless companies across industries based on following questions: *How recently? How often? How much value?*. In our case we rely on the same RFM components to make the value modeling for topic impact dependent on the time and on the number of interacting users. Each component consists of a set of metrics based on the number of users interacting with the topic in a location, the Engagement of these users with the topic – computed by the share of the content they produce related to the topic –, and their Exposure to the content their network creates related to the topic, with the option of creating an aggregate index as well.

This paper is structured as follows: firstly the background information where we briefly review the related work is presented. Then, we introduce our method together with metrics to quantify the impact of a topic on the Social Media channel. After that, we present a system that implements our metrics and then we show some practical examples of topic impact quantification. Finally, we share our conclusions and point out future work on this topic.

## 2. Background and related work

In this section we provide all the background information and related work to base our research, starting with the review of impact modeling and topic diffusion, introducing the RFM model and finally discussing the approaches to topic modeling and information extraction in Social Media.

### 2.1. Topic diffusion and Social Media impact

The diffusion of news or topics in the social networks has been subject of intense research especially in the last years (Cavusoglu, Hu, Li, & Ma, 2010; Centola, 2010; Stieglitz & Dang-Xuan, 2013). Although the methodology we propose in this article is not intended to explain the dynamics of the topic propagation in the social networks, rather to provide a measure for the impact, there are common elements used in both researching lines to understand the contribution of a given user based on how active she/he is, the handling of the variation over the time of the topic-related activity and the semantic definition of the topic. Guille and Hacid (2012) defined three dimensions playing a role in the propagation of a topic: social, semantics and temporal to model the probability of dispersion. The social dimension is defined taking into account the users' activity index, the ratio of directed tweets to the user, the mentioning rate and whether the user being mentioned is directly related to the mentioner. On the other hand, the semantics is based on the presence of a keyword in the message being propagated. The temporal dimension is provided as a computation of the user activity in 6 partitions of the day, but probably leaving the door open to finer time granularity.

Rajyalakshmi, Bagchi, Das, and Tripathy (2012) demonstrated the role of the strong links in the virality of the topics by modeling the diffusion with a stochastic approach, identifying as driving parameters the users activity time and the fading out effect – represented as a weight decay for a topic as time passed by. In their work, two cases are clearly separated: users creating instances of a global topic or users copying it from their network – local social network effect vs. the overall trending effect. Romero, Meeder, and Kleinberg (2011b) established a mechanism relying on Exposure curves to quantify the impact Exposure to other users in making them adopt a new behavior (e.g.: turning them from passive to active contributors or to start using a hash-tag, etc.). In addition, there have been several approaches to model the influence of a particular user in his/her own and in the global Social Media network. Ye and Wu (Ye & Wu (2010)) defined 3 different metrics to quantify the social influence: followers influence – the higher the number of followers, the higher the influence –, reply influence – the more replies one user receives, the more influential the user is –, and re-tweet influence – the more re-tweets, the more influent. Kwak (Kwak et al., 2010) suggested also 3 metrics but substituted the reply influence by one inspired by the Google Search PageRank algorithm (Page et al., 1998) to allow the propagation of influence. Depending on the metric applied the ranking of the top users varied. Romero, Galuba, Asur, and Huberman (2011a) demonstrated that influent users are those whose contributions are not just consumed but also forwarded and therefore overcome the so called passivity and more interestingly, that the popularity of an user and its influence do not quite often correlate. Cha, Haddadi, Benevenuto, and Gummadi (2010) differentiated 3 kinds of influence for a Social Media user: due to the size of the user's audience or social network indegree influence –, due to the generated content with pass-along value retweet influence, which is also aligned with the passivity activity work presented by Romero et al. (2011b) and due to the Engagement in others' conversation mention influence – and all of them are present as component for either Exposure or Engagement when applicable in our approach. The use of geo-localized SM interactions to provide information about local communities is a field of incipient research. In Scellato, Noulas, Lambiotte, and Mascolo (2011) the authors provide an extensive description of the social spatial properties of location based social networks. In Backstrom, Sun, and Marlow (2010), the authors rely

on the spatial proximity in combination with the social proximity to make geographical predictions. Another remarkable example (Cheng, Caverlee, Lee, & Sui, 2011) shows how to use location sharing services to explore and trace footprints.

### 2.2. RFM background

The Recency, Frequency and Monetary (RFM) models were developed as a logical step in the evolution of marketing segmentation techniques. When the shotgun approaches (marketing everything to everybody) proved inefficient in terms of returns, the marketing campaigns started separating customers in segments based on socio-demographics attributes. Taking as segmentation criteria the customers' purchasing behavior proved more sensible indicator to response to campaigns than the former socio-demographic segmentation (Hughes, 2005). Especially the Recency last time that a purchase was committed –, Frequency how many purchases have been committed – and Monetary value of the purchases committed – are usually employed to create a triple of scores per customer. The RFM models segment the customers relying on these scores, so that each segment is targeted in a particular much more tailored way. RFM approaches present also known limitations, like the risk of over-soliciting high-ranked customers, but this is rather a limitation related to the way of applying the findings of the model, not to the model itself, and therefore, it has no effect due to the way we want to apply it. Kumar (2008) pinpointed 3 limitations of RFM-based approaches to model customer behavior: it does not reveal any information about customers' loyalty which again is not an issue in our case, as the loyalty of an user to a topic is out of the scope of the metrics defined –, does not predict the next buy our model does not need to predict the next time the user is going to engage or be exposed to a topic – or the expected profitability over the time as our metrics work backwards, the predictive capabilities are not relevant. Another metric traditionally related with RFM is the so called Customer Lifetime Value (Fader, Hardie, & Lee, 2005; Khajvand, Zolfaghar, Ashoori, & Alizadeh, 2011; Sohrabi & Khanlari, 2007) or the predicted value a customer is going to generate in her entire life time (Farris, Bendle, Pfeifer, & Reibstein, 2010). We have not focused on Customer Lifetime Value like metrics as part of this study.

### 2.3. Social Media usage for topic extraction and trend detection

The Social Media platforms where users continuously post relevant messages referred to an ever changing huge variety of topics are the perfect playground for researchers to develop automatic topic uncovering algorithms. Blei, Ng, and Jordan (2003) started a new research line with their Latent Dirichlet Allocation, a multinomial probabilistic soft clustering of words based on co-occurrence. Many other researches have taken it as reference for topic extraction adding some improvements like in AlSumait, Barbará, Gentle, and Domeniconi (2009), where a method was suggested to prevent the generation of junk topics, or in Chuang, Ramage, Manning, and Heer (2012) it was pointed out the need for supervision by domain experts of the generated topics set, etc. Jones (1972) set the basis for the topic extraction based on the well-known Inverse Document Frequency (Salton & McGill, 1986). Latent semantic indexing (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) has also been vastly employed Automatic topics and trend detection have been also subject of research. Using Twitter particularly we found countless works oriented to extract topics and trends: for example, in Cataldi, Di Caro, and Schifanella (2010) it was suggested an approach based on topic aging theory to extract the emerging topics only considering the authority of the user based on Page Rank; in Naaman, Becker, and Gravano (2011) the authors proposed a taxonomy of trends but specific to a given area relying on the extended geo-location capabilities of the Social Media conversation and suggested a trend featurization based on the associated messages to explain the different trends in local communities. In Sakaki, Okazaki, and Matsuo (2010) the concept of *social sensor* was coined in order to detect events almost real time analyzing tweets. Mathioudakis and Koudas (2010) created a system called TwitterMonitor to detect and anticipate trends over the Twitter Stream.

## 3. Defining a new method for quantifying Social Media impact

In this section we present a new method and its metrics for quantifying how a topic impact the people of a place based on the RFM paradigm.

The motivation behind adopting the RFM approach in SM is twofolded. On one hand, it calls out and addresses separately each and every topic impact driver to latterly produce a compound metric or score: topics that are being discussed recently have a higher impact than topics that are no longer in the focus of attention – their relevance experiments a decay –; the repetition or how frequently a topic has been interacted with is also an indicator of impact, as well as the quality of the interaction with the topic – the value or Monetary of the topic. On the other hand, a broadly adopted, well-established approach that has already been in use for longer than a decade helps making the insights we create more actionable and directly usable for the industry. Unlike the traditional RFM approach, where Recency is typically computed as how many *time units* back in the time (typically days) from the present moment (today/now), our approach is designed to be more generic: the time span is defined beforehand and the Recency is always calculated taking as reference the extreme of the time span closer to the present. Additionally, and leveraging the emerging geo-localized nature of the Social Media interactions, our aim is creating a topic impact score for a specifying region, which unlocks a new set of possibilities – like impact comparison in two different cities or countries, etc. Even if our metrics are design to overcome the existence of users with different levels of SM activity, there are several factors that might introduce certain bias, like the access to the internet of a particular region, the SM affinity – typically older population is less affine, etc. – and other minor factors. Even for certain areas where the SM users constitute a less representative portion of the population, our model provides a result, but with higher volatility and less reliability, and should be interpreted as such.

### 3.1. Preliminary definitions

Before starting with the definition of our methodology, a set of concepts to support our metrics needs to be established:

**Definition 1.** The set $U$ represents the set of Social Media users from which we have evidence they have been in the location $L$ ($InLocation(u_i, L, \Delta t)$) we are monitoring during the time period under analysis $\Delta t$

$$U \equiv \{u\},\ \forall u_i \in U,\ InLocation(u_i, L, \Delta t) \tag{1}$$

**Definition 2.** The social network for a given user $u_i$ is defined as:

$$SN\ (u_i) \equiv \{u\},\ \forall u_j \in SN(u_i),\ Follows(u_i, u_j) \tag{2}$$

$Follows(u_i, u_j)$ is a function representing a SM connection between the users $u_i$ *and* $u_j$, so that $u_i$ is exposed to the SM content generated by $u_j$. $Follows(u_i, u_j)$ is not always commutative; although in several SM platforms it is the case (e.g.: Facebook or Linked.in).

**Definition 3.** The set $SN(U)$ represents the set of all the users being followed by the users in $U$:

$$SN(U) \equiv \{u\},\ \forall u_i \in SN(U),\ \exists u_j \in U \,|\, u_i \in SN(u_j) \tag{3}$$

**Definition 4.** We define all user interactions (*Interactions*) for a given user $u_i$ over a time interval $\Delta t$, as:

$$Interactions(u_i, \Delta t) \equiv \{it\},\ \forall it_i \in Interactions(u_i, \Delta t),\ Author(u_i, it_i, \Delta t) \tag{4}$$

A Social Media interaction represents the atomic piece of content generated by the user $u_i$ during the time $\Delta t$ in a Social Media platform (e.g.: a tweet, a re-tweet). Thus, $Author(u_i, it_i, \Delta t)$ is a function that retrieves *True* if $u_i$ created the interaction $it_i$ in the time period $\Delta t$, and *False* otherwise. The time interval $t$ might be measured in weeks, days or hours, depending on the use case and consists of two extremes: *t_startdate* and end date *t_enddate*.

We call *activeInteractions* to those made by any user $u_i \in U$ in the location and *passiveInteractions* the ones made by any user $u_j \in SN(U)$ the users in the location are exposed to.

**Definition 5.** We define the set of *Interactions* for a given user $u_i$ with the topic $T$ over a time interval $\Delta t$ as:

$$Interactions(u_i, T, \Delta t) \equiv \{it\},\ \forall it_i \in Interactions(u_i, \Delta t),\ (u_i, it_i, \Delta t) \wedge related(it_i, T) \tag{5}$$

where $related(it_i, T)$ is a NLP membership function retrieving *True* if the iteration $it_i$ is connected to the topic $T$ – intuitively, one or more words from the semantic field for the topic $T$ are mentioned in $it_i$ – and *False* otherwise.

**Definition 6.** We define *Contributor* to a topic $T$ as the user $u_i$ who created at least one interaction $it_i$ with the topic $T$ over a time interval $\Delta t$

$$Contributor(u_i, T, \Delta t) \equiv True,\ \exists it_i,\ it_i \in Interactions(u_i, T, \Delta t),\ u_i \in U \cup SN(U) \tag{6}$$

**Definition 7.** A user $u_i$ is *exposed* to a topic $T$ over the time span $\Delta t$, when there is at least one $u_j$ in its social network $u_j \in SN(u_i)$ who contributed to the topic.

$Exposed(u_i, T, \Delta t)$ is a logical function defined as:

$$Exposed(u_i, T, \Delta t) = \begin{cases} True, & \exists u_j,\ u_j \in SN(u_i),\ Contributor(u_j, T, \Delta t)\ is\ True \\ False, & otherwise \end{cases} \tag{7}$$

where $SN(u_i)$ represents all social network users connected to $u_i$

### 3.2. Recency in Social Media

In the traditional usage of the RFM model, *Recency* has always been used as indicator for the last time a customer or prospect interacted with the brand, purchased a product, added a product to the online shopping cart, etc. *Recency* is traditionally used to assign a score to customers depending on how long ago the last interaction took place. In our case, we will provide the Recency metric as an indicator to express how *up-to-date* the topic is. Thus, a topic that is *hot* today is going to have a much higher Recency score than a topic the community stopped talking about weeks or months ago. As we are modeling a topic and not a particular user, we are going to suggest an aggregated approach. Depending on the topic in question, the interactions in the Social Media channel might be quite sparse, which introduces the need to work with thresholds. As pre-defined absolute thresholds might diminish the suitability for generic scenarios, we are going to set thresholds as a minimum of the topic share in a time unit (e.g.: a day), defined as follows:

$$Share(T, \Delta t) = \frac{\sum_{i=1}^{\#U}(\#Interactions(u_i, T, \Delta t))}{\sum_{i=1}^{\#U}(\#Interactions(u_i, \Delta t))} \tag{8}$$

which is the sum of SM interactions created over the time period $\Delta t$ by the users geo-located in the place under analysis related to the topic $T$ vs. the total number of SM interactions (those that are not related with the topic $T$ as well).

For example, if we set the threshold to 0.1 per day, the *Recency* would start counting up when the amount of posts related to the topic a day goes over 0.1 share of all interactions in the given day. This threshold can be adjusted to our convenience depending on the max number of different topics to be considered per day, the volatility tolerance, the topic in question, etc. The *Share* can be also defined in terms of Users engaged with a particular topic vs. all the active users in the particular day (which removes the bias introduced by overly active users, overly passive users, etc.).

As explained before and unlike the traditional RFM model, our methodology is designed to work for any specific time frame and location in a more generic way. Hence, *Recency* does not have to necessarily be always measured from the present (today) back in the past. Rather, our definition requires the two extremes of a time interval $\Delta t$, so that both can be in the past. *Recency* is measured taking as reference the second extreme of the interval (*t_enddate*) over a time span to *t_startdate*. Although the most common time unit is the *day*, it is possible to adjust our approach to work in *weeks* or even in finer granularity units like *hours*, which is only advisable when the volume of interactions/time unit is sufficient to avoid volatility situations.

Based on the *share* concept to define *Thresholds* and the Definition 6, we define *Recency* as follows:

$$Recency(T, \Delta t) = \begin{cases} \frac{t_i - t_{startdate}}{t_{enddate} - t_{startdate}}, & if\ \exists t_i,\ t_i \in [t_{startdate}, t_{enddate}] \\ 0, & otherwise \end{cases} \tag{9}$$

where $t_i$ is the first time unit closer to $t_{enddate}$, so that $Threshold \leq \sum u_i,\ Contributor(u_i, T, t_i) = True,\ u_i \in U \cup SN(U)$.

### 3.3. Frequency in Social Media

In our method, *Frequency* is designed to measure how often interactions with the topic are registered during a defined time period. The more interactions with the topic, the higher the Frequency and the higher the overall impact of the topic on the users located in the place under analysis.

Based on the type of interaction, we distinguish *Frequency of Exposure* or *Passive Frequency* and *Frequency of Engagement* or *Active Frequency*.

The *Frequency of Exposure* for a topic in a given period of time can be expressed as the number of users exposed to the topic per time unit. The subset of users exposed to the topic can then be defined as:

$$ExposedUsers(T, \Delta t) \equiv \{u\},\ \forall u_i,\ Exposed(u_i, T, \Delta t) = True,\ u_i \in U \tag{10}$$

Thus, the *Frequency of Exposure* can be defined as the number of users exposed over the time period:

$$Frequency\_Exposure(T, \Delta t) = \frac{\#ExposedUsers(T, \Delta t))}{length(\Delta t)} \tag{11}$$

Additionally, we define the *Frequency of contribution* as the total number of users with active interactions with the topic in the specific time frame based on the set of all contributing users to the topic:

**Table 1**
Exposure and Engagement categories.

| | |
|---|---|
| Exposure categories | |
| Disconnected Exposure | A Social Media user is exposed to the interactions of his/her social network (or further users directly linked to her). As the Social Media world does not stop but users regularly disconnect, there is so much content created within a particular social network that does not even get ever read by the user if online. With the proper set of web analytics in place, one could determine whether the user actually clicked on a piece of content generated in her social network or even model the probability of having read the piece of content based on session start time and session duration |
| Connected Exposure | When there are interactions related to the topic within the user's particular Social Media network and the user himself/herself was online within the time window around the time of interaction with the topic. Active or online can be understood as connected and/or interacting with the Social Media platform (creating content, comments, etc.) |
| Explicitly mentioned | User mentioned in a post related to the topic created by another. Unlike the previous categories of Exposure based on the broadcasting of a message in the user's social network, this kind refers to peer-to-peer delivery of a message in the social channel: from a particular user to another particular user yet keeping it accessible to the entire social network of both. Even if we cannot talk of an activity directly triggered by the user with the topic, the fact that a different user within his/her social network posted a piece of content about the topic and mentioned him/her on it, is going to increase the possibility of reading the post: (example from Twitter: user_2 posted: "*sorry mate, your team did not have any chance against #manu @user_1*". The user_1 gets a notification which most likely makes her reading the post from user_2 talking about a football match) |
| Engagement categories | |
| Active response | The user actually answered or commented a post created by another user within his/her Social Media network about the topic. (e.g.: based on Twitter: user_2 posted: "*sorry mate, your team did not have any chance against #manu @user_1*" user_1 replied "*@user_2 Chelsea FC for ever! #cfc*") |
| Active forwarding | The user just confirms that he/she feels identified with a piece of content generated by another user within her Social Media network about the topic we are analyzing (depending on the Social Media platform as a "I like" or as a *Retweet*). |
| Actively initiated | The user starts talking about a topic within his/her social network. She is the initiator and the one who brought up the topic into her social network. We see this one as the highest level of Engagement |

$$Frequency\_Contribution(T,\Delta t) = \frac{\#ContributingUsers(T,\Delta t)}{length(\Delta t)}, \quad \forall u_i \in ContributingUsers(T,\Delta t), \; Contributor(u_i,T,\Delta t) = True \quad (12)$$

Putting both metrics together, we get to the envisioned *Frequency* metric:

$$Frequency\,(T,\Delta t) = \frac{1}{2}(Frequency\_Exposure\,(T,\Delta t) + Frequency\_Contribution\,(T,\Delta t)) \quad (13)$$

In order to normalize these metrics, we make both relative to the quantification of the total Exposure and the total contribution. For that we just put in relation the previously obtained Frequency metric to the total number of users that could have been contributing or exposed:

$$Frequency\_Penetration\,(T,\Delta t) = \frac{1}{N} Frequency\,(T,\Delta t) \quad (14)$$

### 3.4. The Monetary component – Value in Social Media

Unlike *Recency* or *Frequency*, the *Monetary* or *value* component requires certain modeling decisions about which factors need to be considered to which extent. When we talk about *value* for a SM topic, one could think of reach (Bogart, 1967). Measuring reach only, might leave certain aspects of the impact modeling unaddressed, like the *quality* of the audience where the topic is active, the level of Engagement with the topic, the latent Exposure among others.

Intuitively, *value* shall measure the number of SM users impacted by the topic – be it passively or actively –, quantify the intensity of this impact and put it in relation to the set of total users that could have been impacted. In the subsequent sections we are going to present our approach to model the different facets of a topic value to latterly consolidate everything into a single combined metric.

#### 3.4.1. Modeling the Exposure/Engagement of a particular user with a topic

In our methodology we consider the level of Exposure to the topic (concept inspired by the group contagion theory proposed in Barsade (2002)) and the level of user Engagement with the SM content related to the particular topic as the main components for modeling value. *Exposure* builds upon the Definition 7 and includes all scenarios where a given user could potentially read SM content related to a topic. As soon as the user adopts an active role towards the topic (creates or forwards content related with the topic), the user becomes a *contributor* (see Definition 6) and we speak of *Engagement*. The Table 1 presents the different categories we are going to use to attribute different intensity levels for both concepts.

To enable finer granularity when we measure the intensity within the *Engagement* categories, we additionally use the type of content. For example, one user taking a picture and posting it with a message about the weather shows more Engagement that just a text. The same applies for links: a user sharing a link about a topic suggests that the user read about the topic already somewhere else and shared the reference to this content indicating a higher level of Engagement as well.

To model the degree of Engagement and Exposure related to a topic based on the categories defined in Table 1 and including the different content types, we apply a weighting schema. To apply the weights we partition the set of all active interactions of a given user $u_i$ with the topic $T$ in the time interval $\Delta t$ by the type of content $C$ on one hand and by Engagement Category $En$ on the other hand: The types of content we considered $\{Text\ only, Contains\ links, Contains\ video\ or\ picture\} \in C$ constitute complete partitions of $Interactions(u_i,T,\Delta t)$, so that $\forall it_i \in Interactions\ (u_i,T,\Delta t),\ \exists! c_j,\ ContentType(it_k) = c_j,\ c_j \in C$. And so do the Engagement categories $\{Active\ Response, Forwarding, Active\ Initiating\} \in En$, so that $\forall it_i \in Interactions(u_i,T,\Delta t), \exists! e_j,\ EngagementCategory(it_k) = e_j,\ e_j \in En$

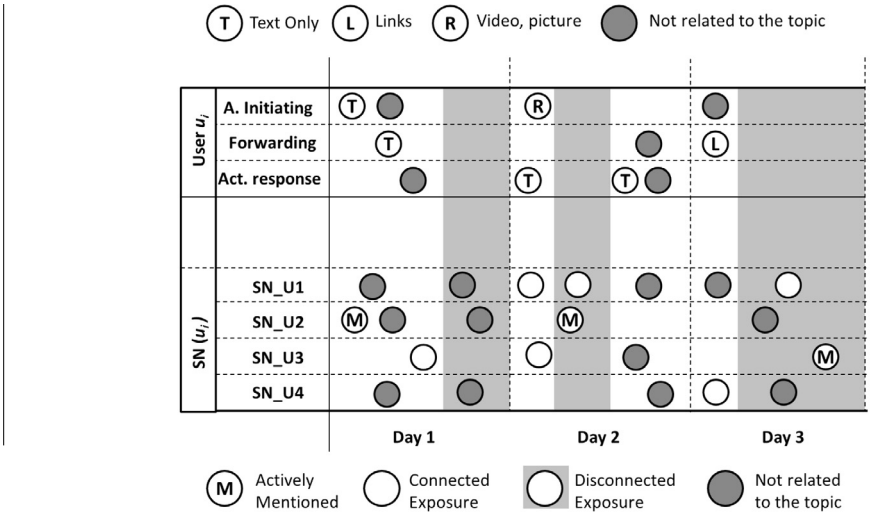
Let's express the different partitions of the Interactions set for a given user based on types of content as

$Interactions(u_i,T,\Delta t|c_k), c_k \in C$ and the partitions based on Engagement categories as $Interactions(u_i,T,\Delta t|e_k),\ e_k \in En$.

Based on both partitions, we define *Engagement* by weighting the different *Engagement category–content type* pairs, as follows:

$$Engagement(u_i,T,\Delta t) = \sum_{j=1}^{\#(En)} \sum_{k=1}^{\#(C)} w(c_k,e_j) \,\#\, ((Interactions(u_i,T,\Delta t|c_k) \cap Interactions(u_i,T,\Delta t|e_j)) \quad (15)$$

Similarly, the set of all Interactions created by the social network of a given user $u_i$, $SN(u_i)$ related to the topic $T$ in a period of time $\Delta t$ can be partitioned based on Exposure categories for the user $u_i$ $\{DisconnectedExposure, ConnectedExposure, ActivelyMentioned\} \in Ex$, so that $\forall it_k \in Interactions(SN(u_i),T,$




**Engagement**

| | | Actively Initiating | Forwarding | Active Response |
|---|---|---|---|---|
| Weights | Rich Content | 1 | 0,8 | 0,6 |
| | Links | 0,8 | 0,64 | 0,48 |
| | Text Only | 0,7 | 0,56 | 0,42 |
| # Topic Interactions | Rich Content | 1 | | |
| | Links | | 1 | |
| | Text Only | 1 | 1 | 2 |

*Engagement* 3,74
*Total Interactions* 11
*Engagement Index* 0,34

**Exposure**

| | | |
|---|---|---|
| Weights | Disconnected | 0,2 |
| | Connected | 0,5 |
| | Antively Mentioned | 1 |
| # Topic Interactions | Disconnected | 2 |
| | Connected | 4 |
| | Antively Mentioned | 3 |

*Exposure* 5,4
*Total Interactions* 21
*Exposure Index* 0,257

**Fig. 1.** Example of timeline with all interactions of user $u_i$ and his/her social network $SN(u_i)$.

$\Delta t), \exists! e_j,\ ExposureCategory(it_k) = e_j,\ e_j \in Ex$. $Ex$ is the set of all partitions based on Exposure or Exposure categories. Based on these partition, we can define *Exposure* as follows:

$$Exposure(u_i, T, \Delta t) = \sum_{k=1}^{\#(SN(u_i))} \sum_{j=1}^{\#(Ex)} w(e_j) \ \#\ (Interactions(u_k, T, \Delta t) \cap Interactions(u_k, T, \Delta t | e_j)) \quad (16)$$

### 3.4.2. Active and Passive Impact based on Engagement and Exposure

The metrics previously defined to quantify *Engagement* and *Exposure* work on absolute terms. The definition of Impact upon them puts both metrics into relation to the total number of interactions created by the user or generated within the SN of the user. To address that, we define *Active Impact* or *Engagement Index* and *Passive Impact* or *Exposure Index* as follows:

$$ActiveImpact(u_i, T, \Delta t) = \frac{Engagement(u_i, T, \Delta t)}{\#Interactions(u_i, \Delta t)} \quad (17)$$

$$PassiveImpact\ (u_i, T, \Delta t) = \frac{Exposure(u_i, T, \Delta t)}{\#Interactions(SN(u_i), \Delta t)} \quad (18)$$

The resulting value component of the impact of a topic $T$ on a given user $u_i$ over a period of time t is a combination of *Passive Impact* and *Active Impact*:

$$ImpactValue(u_i, T, \Delta t) = \frac{1}{k+l}(k * ActiveImpact(u_i, T, \Delta t) + l * PassiveImpact(u_i, T, \Delta t)) \quad (19)$$

where $k$ represents the weighting for the Active Impact and $l$ represents the weighting for the Passive Impact ($l$ and $k$ are positive numbers). Depending on how the use case gives priority to the Engagement over Exposure, the values for the weights $l$ and $k$ are defined.

This is the definition we suggest for the *Monetary* or *Value* component in our RFM methodology.

Fig. 1 explains how both Engagement and Exposure Indexes are obtained for a fictive user's time-line over 3 days applying the formulas (17) and (18). If we weighted both components in (19) with 0.5 each, the total impact of the Topic on the user $u_i$ would be 0.298.

The formula (19), which defines the Impact at user level, can be extended to all the users $U$ in the location:

$$ImpactValue(U, T, \Delta t) = \frac{1}{\#U} \sum_{i=1}^{\#U} ImpactValue(u_i, T, \Delta t) \quad (20)$$

Building on top of Romero et al. (2011b), our impact modeling assigns different Engagement levels depending on whether the user initiates the topic within her social network or just engages with a topic currently discussed in her network. Unlike Kwak et al. (2010), our approach just considers the inbound diffusion component, what we called Exposure but omitting the outbound diffusion – the geo-located users are not necessarily exposed to the activity of their followers according to the way the online Social Media platforms are designed on one hand and the Engagement of the followers of geo-located users does not contribute to the overall Engagement in the location under analysis on the other hand. The Exposure and Engagement components we suggest

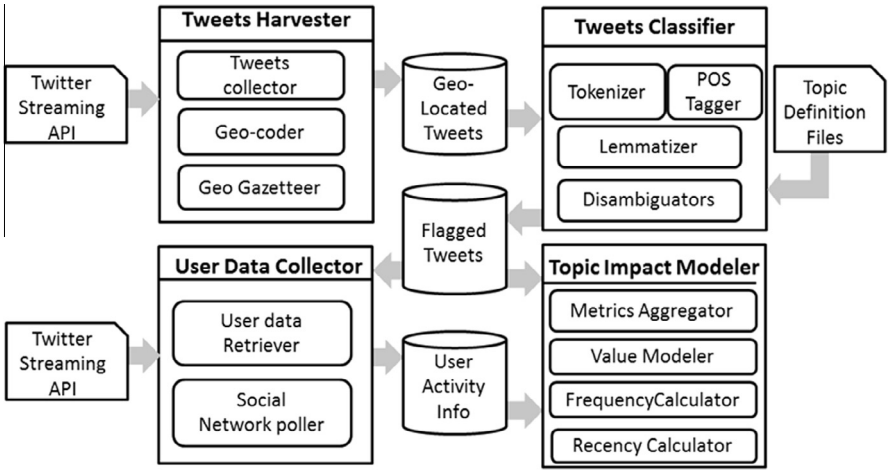



**Fig. 2.** System modules overview.

relies on the activity/passivity concept introduced in Romero et al. (2011a), evolving into the set of metrics we defined above. As we mentioned before, we've chosen Exposure and Engagement to model our value metric because both represent a quantification of the topic intensity on the users located on a particular location.

### 3.5. Quantifying the topic impact using the Social Media RFM model

Our Social Media RFM model provides a single value quantifying the impact of a topic in a SM channel within a period of $\Delta t$ as a combination of the metrics discussed above:

$$Impact(U,T,\Delta t) = F(Recency(U,T,\Delta t),\ Frequency(U,T,\Delta t),\ Value(U,T,\Delta t)) \quad (21)$$

The function $F$ can be any combination of the three components, varying from a simple average to a weighted average to more complex scenarios. The definition of $F$ should be made dependent on the use cases (i.e.: the Recency component might be ignored or weighted very low for recurring topics like "weather" under the assumption that every body post about the weather regularly, whereas other scenarios like a particular event like "New York Marathon"). As aforementioned, having a single index, allows for combining the impact of different topics, in different locations over different periods of time. For example, you could compare the topic "Pope election in Rome during the conclave weeks" with "Royal wedding of Prince William the week of the 29th of April 2011 in London".

## 4. System architecture

Even if the metrics presented so far can be applied to each and every Social Media platform – provided there are means of getting access to the required information –, the system we implemented focuses on Twitter only. We've chosen Twitter over other existing networks because of the ease of information extraction (no constraints in terms of the need for being connected to users to retrieve them over the APO), because of the variety of the topics discussed unlike other purpose specific SM networks – like Linked.in –, because it's broadly adopted, and because the geolocation capabilities are extensively developed.

The system polls the geo-located tweets from the publicly available Twitter Search API,[2] flags those tweets that are related to the topic, extracts the information about the users involved in these tweets, including their social network and finally applies the set of metrics for impact modeling. The system consists of 4 different modules in charge of different labors all along the process.

Each module consists of a set of components with a clearly defined function. In the following sections we are going to describe how the different modules work and what the role of the components being involved is.

### 4.1. Tweets harvester

This module performs poll-requests from the Twitter Search API to store the tweets into a local data base for further processing. The tweets are selectively picked for a given area which is configured in the harvester, namely the one we want to perform the topic impact analysis over time. Additionally, the Twitter API supports the filtering by language (e.g.: only tweets in English), but even it would make the later NLP much easier, it might disregard the interactions of all users related with the topic in the target area for being in a foreign language. We opted for a work-around that does not filter out the tweets by language upfront, yet does not introduce the need for applying NLP techniques in all identified languages, as we are going to explain in the next section (see Fig. 2).

The harvester also provides the capability of assigning the interactions to already standard geographical output systems (like postal sectors, output areas, census units, etc. depending on the country).

A gazetteer supports both correction of inaccurate information and spelling mistakes in addresses, etc. To implement these functionalities our system relies on existing geo-coding API's provided by the major web mapping providers[3] – which usually are free of use up to a limit of request per day. The geo-mapper component takes as input the shape files (polygon lines) describing the output geographical units of a system of choice (e.g.: postal sectors) and applying the standard point-in-polygon algorithm (Sutherland, Sproull, & Schumacker, 1974) establishes the mapping of the interaction or tweet to a standard geographical unit. The outcome of the harvester is a collection of full-fledge[4] tweets with a time stamp, a pair of geographical coordinates and potentially the link to a particular standard geography unit.

[2] Available at https://dev.twitter.com/docs/api/1/get/search

[3] Google Geocoding API: https://developers.google.com/maps/documentation/geocoding/. The location API provided by Bing Maps REST Services: http://msdn.microsoft.com/en-us/library/ff701715.aspx

[4] Full-fledge because all the meta information coming from the Twitter API has not been discarded (see https://dev.twitter.com/docs/platform-objects/tweets)

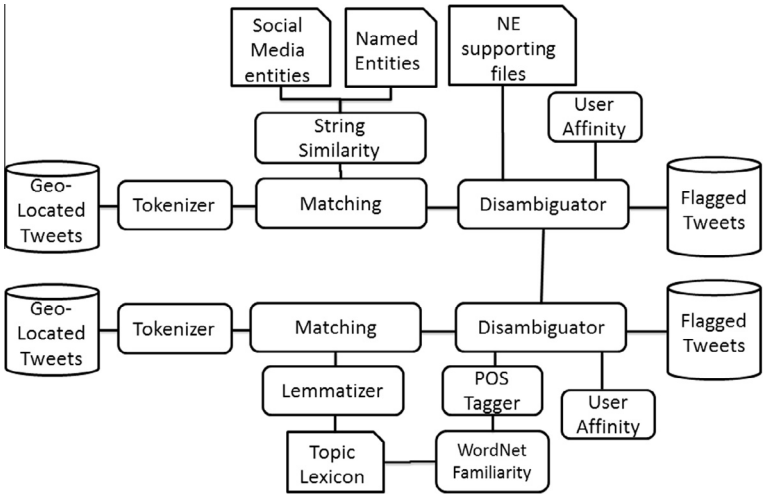



**Fig. 3.** Double-pipe tweets classifier.

### 4.2. Tweets classifier

The mission of this module is basically separating all the harvested tweets that are related to the topic from the others.

Running the system to produce the impact metrics for a given topic requires the non-trivial task of gathering and structuring the set of keywords that qualifies the topic. We suggest following sources:

- Social Media Entities related to the topic: Set of official accounts, nicknames, hashed tags, etc. users mention in their interactions with the topic (e.g.: for the topic "tennis", we would have RafaelNadal for Rafa Nadal, DjokerNole for Novak Djokovich, etc.). For completeness it should include both official accounts and those that are not official but with high levels of activity.
- Topic Named Entities: set of named entities related to the topic (e.g.: "Rafael Nadal", "Noval Djokovic", etc.).
- Topic Lexicon File: containing the set of non-named entities related to the topic (e.g.: in the *tennis* domain: "ace","match ball", "set", "advantage", etc.).

The orchestration of the steps required to perform the Tweets classification is designed to minimize the number of *false positives* as soon as possible in the process and increase the overall performance of the classifier. Thus, the number of tweets remaining from one step to the subsequent one is intended to shrink. Likewise, the steps involving higher complexity are pushed towards the end, whereas the simple and efficient ones are done at the beginning.

The way the classifier works in 2 phases, is explained in Fig. 3 both pipes share the same tokenizer: each geo-located tweet is tokenized applying a sentence tokenizer first and a word tokenizer later (based on O'Connor, Krieger, & Ahn (2010)) both adapting the Punkt Tokenizer (Kiss & Strunk, 2006) to deal with Social Media texts. The modified tokenizer provides the stop words removal as well. The Social Media and Named Entities pipe intents to match each and every reference term listed in the Social Media and Named Entities files applying a string similarity algorithm (Yang, Yuan, Zhao, Chun, & Peng, 2003), which delivers a similarity score. The matching module in our system then implements thresholds – which differs depending on the source – to support the fact that the Social Media content is often full of spelling errors, which is likely to happen even more frequently when it comes to named entities of foreign people (e.g.: staying in tennis, Nalbandian is often spelled as Nabandian even by renowned tennis Twitter accounts) (Clark, Roberts, & Araki, 2010). Even if our unigram based approach might look simple compared with more sophisticated approaches like hierarchical Dirichlet bigram language models (MacKay & Peto, 1995) or based on semantic *gists* (Griffiths, Steyvers, & Tenenbaum, 2007), the suggested approach for the disambiguation allows for keeping the topic model simple and therefore less processing intensive.

For all the messages tagged positively as belonging to the topic and added the set of matching terms or keywords, a disambiguation process takes place. The disambiguation relies on a semi-automated supporting list of homonyms for the named entities:

1. When the term to disambiguate has not been the only one part of the tags for a tweet and one or more of the other tags was univocally related to the topic, the context was sufficient for the disambiguation (e.g.: based on topic "football" *'RT bbc5live: 12:45: Our 1st #BPL commentary of day – Newcastle v Liverpool...a point will take Brendan Rodgers side top #NUFC #LFC'*, tagged with 'lfc,nufc,lfc,liverpool,rodgers' – *Rodgers* is disambiguated by the presence of other tags related to the topic).
2. When just a single term to disambiguate is part of the tagging set for a tweet and this term is a Named Entity, we applied a technique based on the expansion of the named entities by related terms like surname, name, alias, etc. inspired by (Fujita & Fujino, 2013). E.g.: *'@Theleaguemag a young steve bruce !!'*, *Bruce* disambiguated by 'Steve Bruce'.
3. When the disambiguation is not possible as none of the points mentioned above can be applied, the term is marked for a new run of the disambiguation once the pipeline for the topic lexicon is done and also matching lexicon terms might have been added to the tweet.
4. If no disambiguation is possible based on the tweet itself, we try to leverage the affinity of the tweets from the same author with the topic. If the affinity is high, the chance of the tweet to be related to topic is higher, even if there's some room left to interpretation and we might have to accept certain tolerance. The affinity of the user to the topic is calculated as a ratio of the positive tweets over the total number of tweets that have been harvested for the user.

The topic lexicon pipe works likewise in 2 phases, the matching phase and the disambiguation phase. As input for the disambiguation, the associated polysemy coefficient is calculated using the so called WordNet familiarity (Miller, 1995). Each term is basically given the number of synsets, which helps us understanding when

a disambiguation is required. To disambiguate lexicon terms, we apply following approaches:

1. Part of Speech based rules: after the POS tagging, for which we rely on the tailored tagger for Social Media (e.g.: *'was reading about arthritis drugs and apparently sometimes your hair falls out what if this happens to me'*, 'readingfc,reading' (*reading* – VBG – Verb, gerund/present participle – disambiguated to non-related to the topic) Reading . *'Reading FC is gonna have a hard year'* – NNP – Proper singular noun – disambiguated as related to the topic 'football').
2. Presence of multiple terms of the terms reference set (both Entities and other lexicon terms).
3. Author's affinity to the topic (as explained before).
4. Supervised disambiguation for the most frequently identified stand-alone terms.

Even if our approach to disambiguation is not proven to work in 100% of the cases, the number of false positives is to certain extent balanced by the terms that have not been considered in the modeling topic sources. Gathering all the terms that could identify a topic is hard task whose complexity increases depending on how dynamic topics are.

### 4.3. User data collector

This component is in charge of polling the social network of all the authors of the tweets identified by the first pass of the classifier (*socialNetworkUsersSet*). Additionally, the harvester comes again into picture to retrieve the tweets of each and every user in each author's social network to enable the Exposure analysis. Once all tweets of all users in the social network for the time-period being analyzed have been stored, the classifier acts again to flag those belonging to the topic (*socialNetworkTweetsSet*).

Another important task carried out by this component is the implementation of the Exposure window based on each user's interaction and each user's social network activity index.

### 4.4. Topic impact modeler

After gathering all the Social Media content and classifying it according to how related to the topic under analysis it is, this module applies the metrics computing the Exposure, the Engagement and Recency.

This module implements a pre-processing stage consisting of following steps:

- Contributors' flagging: setting the contributors' flag for all the harvested users, if they authored any of the tweets flagged as related to the topic (as defined in Definition 6).
- Exposed Users flagging: considering the *socialNetworkUsersSet* and taking as input the outcome of the previous step to determine who in any user's social network is a contributor, as defined in Definition 7.
- Content categorization: classification of the tweets flagged by the Tweets Classifier and including the *socialNetworkTweetsSet* based on Exposure and Engagement categories explained in the Section 3.4.2.

Applying Eq. (8), the *Recency Calculator* computes the share of the topic based on the ratio of flagged Tweets vs. all harvested Tweets, which is used as a reference only. The share is going to provide an indication of how reliable and how significant the metrics are. The share is also calculated in the particular time units specified (e.g.: day) backwards from the *t_enddate* until we reach a time unit whose share is greater than a threshold. The threshold selection in particular but also the need for the Recency component in general, depend very much on the specific use case. The Recency value is calculated according to Eq. (9).

**Table 2**
Topic characterization: football in UK.

| Topic characterization: football in UK | | | | | |
|---|---|---|---|---|---|
| Type | Entity group | #Terms | Type | Entity group | #Terms |
| *Social Media Entities* | Club Official Accounts | 49 | *Named Entities* | Players | 570 |
| | Players Official Accounts | 315 | | Teams | 44 |
| | Managers Official Accounts | 29 | | Managers | 29 |
| | Club Official Hash-tags | 52 | | Clubs | 49 |
| *Lexicon* | Football terminology | 72 | | | |

The *Frequency Calculator* then takes over to compute both the Frequency Exposure as defined in Eq. (11) relying on the pre-computing stage results for Exposed Users and the Frequency Contribution, like in the formula (12). The results are then combined into the overall Frequency calculation as defined in Eq. (13). The last step accomplished by this component is normalizing the result as described in the formula (14) to provide the Frequency Penetration value.

The *Value Modeler* is in charge of producing the value component of the framework. This component computes first the Topic Engagement or active impact as defined in (17) and the Passive Impact or Topic Exposure (18) later. Both computations require the previous content categorization we described as a pre-processing step above.

As a final step, the *Metrics Aggregator* pulls all the metrics together to generate a single value applying a concrete implementation with a particular set of weights per component as defined in the function (21), if a general score is desired.

## 5. Evaluating the Social Media RFM model

The purpose of this section is to evaluate how the suggested RFM model behaves in different scenarios to prove both sensibility and usage of the set of defined metrics.

Our proposal to the validation consists of two real world topics to ensure the coverage of all the variety of events a topic might manifest. For each one of the suggested topics we are going to present different scenarios carefully selected to thoroughly demonstrate the performance of our framework while highlighting the role of each metric in the different cases.

Our first topic revolves around two events that shook the hearts of multitudes within the space of one week: the deaths of the famous American actor *Paul Walker*[5] on November the 30th 2013 and the decease of the charismatic Peace Nobel Price winner, South Africa first black president and anti-apartheid icon *Nelson Mandela*,[6] just 6 days later. These one-off events are going to help us demonstrate the role of Recency and Frequency taking different time analysis windows (centered on the day of the death, the week after, the week before, etc.). Additionally, we are going to use provide an impact comparison of both deaths in the considered locations.

As second topic we chose *Football*, much wider in scope but highly suitable to prove the performance of our metrics due to the following reasons: recurrence – there are regular matches

[5] http://www.bbc.com/news/world-us-canada-25173331
[6] http://www.bbc.com/news/world-africa-25249520

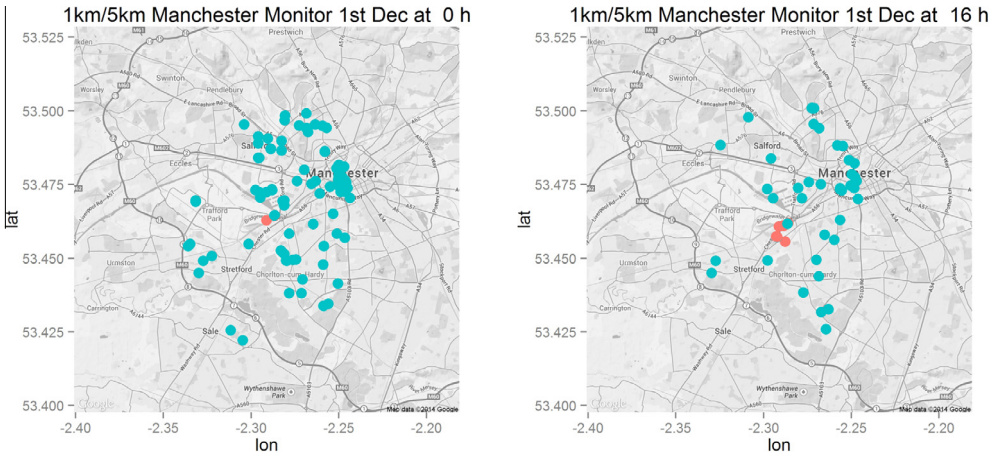



**Fig. 4.** Hourly distribution of tweets gathered on Dec. 1st by the 1 and 5 km harvesters.

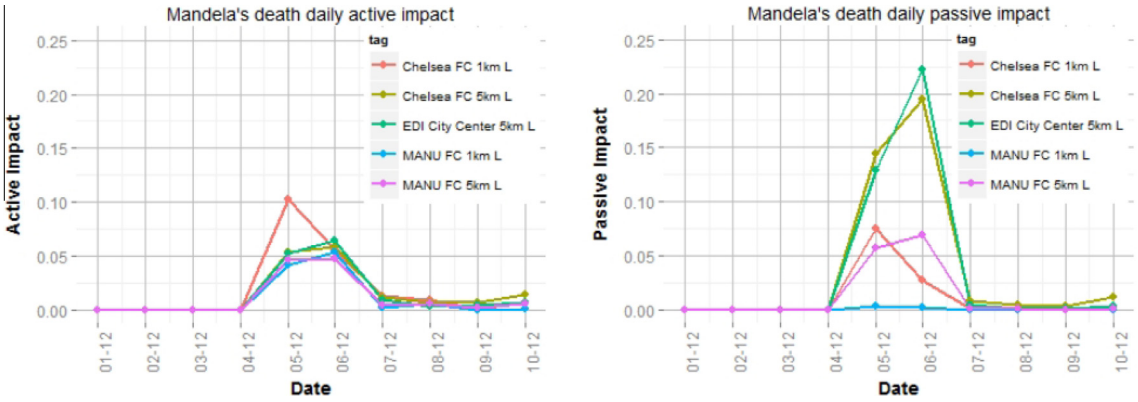



**Fig. 5.** Mandela's death active (a) and passive (b) impact.

coming every week –, variety of scenarios – team playing at home, as visitor, national championship, Champions League, etc. –, fine granularity in time – which allows for Engagement and Exposure calculations for hourly intervals –, popularity – with a lot of Social Media content generated about the topic and therefore, less volatile – and easy to model – with a rather large volume of SM and Named Entities and comparatively small lexicon, which natively reduces the need for disambiguation and therefore, the number of false positives. When the complexity of the topic increases, for example due to the presence of more ambiguous terms, the number of terms required to ensure a proper coverage, the need for remodeling for rapid changing topics, etc., the quality of the topic model might be affected and consequently the quality of the metrics we suggest in this paper. Depending on the complexity drivers, more advance modeling techniques could be applied to tackle particular problems, like disambiguation, etc. but they escape the scope of this paper.

### 5.1. The set up

We carried out the set up of our evaluation in three different steps: gathering of all the required information to create the topic definition files for the topics, configuration of the harvesters to start gathering geo-localized tweets in the locations of interest and definition of the impact aggregation function and weighting schemas.

For Mandela's and Paul Walker's deaths we just took the named entities of both personalities and the popular aliases people use to refer to them (e.g.: *Madiba* for Nelson Mandela).

The football topic definition file, due to its breath, has not been that straight away. We gathered the named entities from the official site of the Premiere League,[7] the official Twitter accounts from all the players[8] and teams.[9] The lexicon file has been manually created compiling several sources to extract the football specific terms (78 unique terms).

Table 2 shows the summary of the sources we've employed to characterize the topic we are currently analyzing: 393 players', managers' and clubs' official accounts, 52 hash-tags representing football clubs in the group *Social Media Entities*, 692 names of players, managers, clubs (*Named Entities*) and 72 additional terms related to football (*penalty, offsider, etc.*).

We set up 5 harvesting engines: 2 of them to monitor the activity on two well-known football stadiums: Stamford Bridge (Chelsea FC) and Old Trafford (Manchester United FC), 2 additional ones centered on both stadiums but with a much larger radius (5km) covering an important part of London and Manchester and a last engine also with a radius of 5 km covering the city of Edinburgh (a place a priori not so much related with the topic *football*).

Although our harvesters have been running for longer than 3 months, we are going to focus our analysis on the first two weeks of December 2013, where the vast majority of scenarios manifest. The harvesters gathered 1,088,627 tweets during these 2 weeks in the mentioned locations.

[7] http://www.premierleague.com/en-gb/clubs.html.
[8] http://www.transfermarkt.co.uk.
[9] http://footballersontwitter.com.

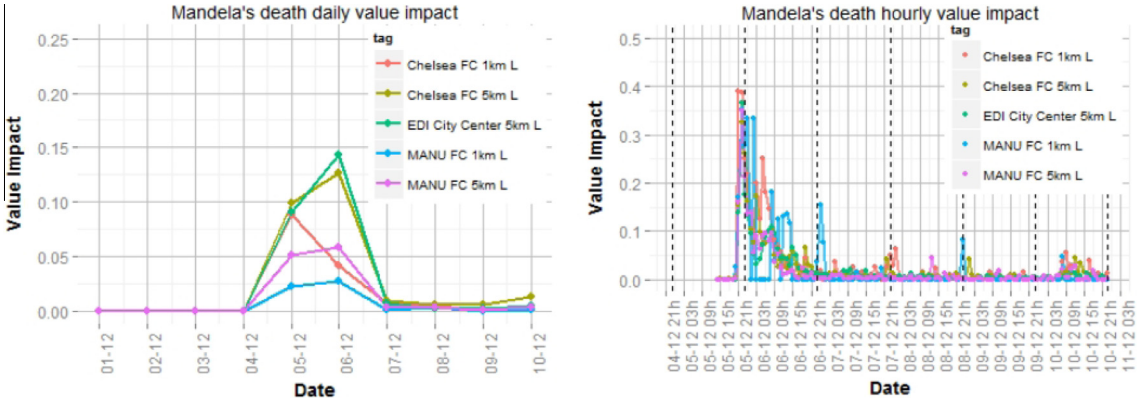



**Fig. 6.** Mandela's death daily (a) and hourly (b) impact value.

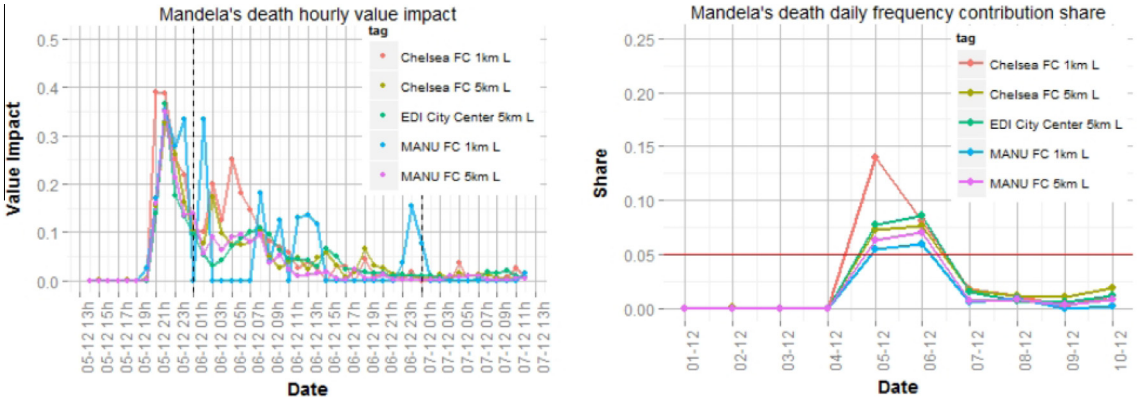



**Fig. 7.** Mandela's death hourly detailed impact value (a) and Frequency (b).

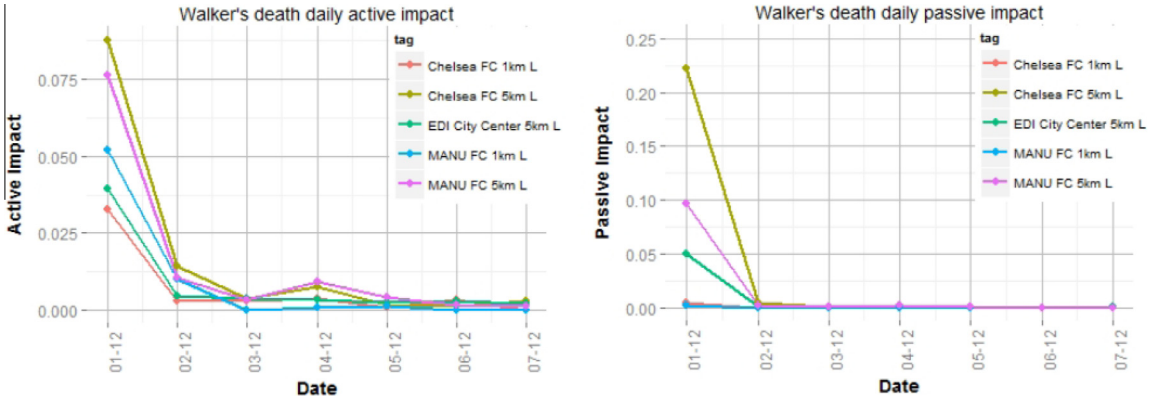



**Fig. 8.** Paul Walker's death active (a) and passive (b) impact.

As example of harvesting we show in Fig. 4 the geographical distribution of the tweets in Manchester for different hours for the 1st of December.

In order to make results comparable across all topics' analysis, we applied the same weighting schema for the Exposure Groups and Engagement Categories and Content Types. The definition of the weights is the one used in Fig. 1 to explain how the metrics are calculated.

The weightings to compute the Impact aggregating both Active and Passive components – see Eq. (19) – are going to be 0.5 for each one. Even if our suggested metric provides the flexibility of making a component prevail over the other one by increasing its weight, the examples we are presenting here do not require any special handling. For consistency reasons – especially to make the outcome comparable – we apply the same weighting schema to all examples.

The Aggregation function when we really require a single score (see Eq. (21)) is going to be defined using the Frequency Penetration as Frequency component and weighting the value component twice as much as the other two: 10 for Recency, 30 for Frequency and 60 for Value.

### 5.2. Topic 1: Impact modeling for Mandela's and Paul Walker's deaths

Paul Walker died in a car accident the 30th of November approximately at 23:30 GMT. Nelson Mandela died on the 5th of

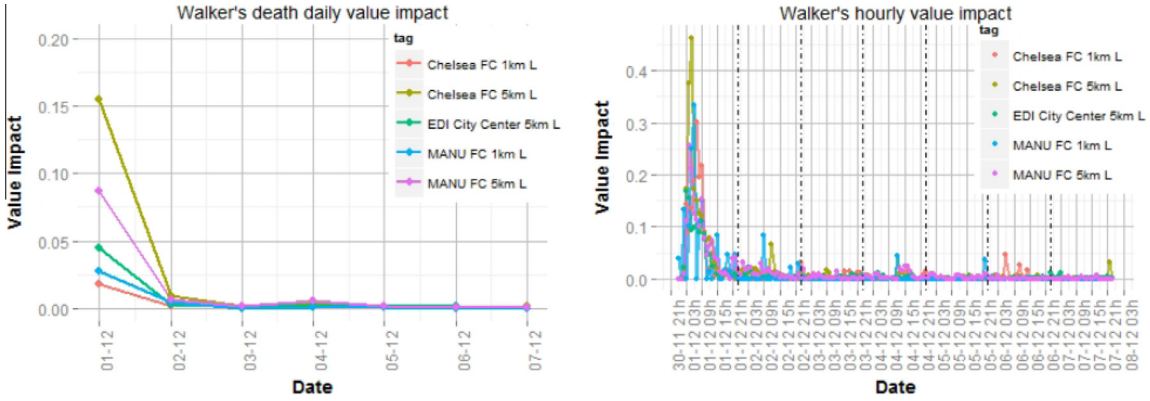



**Fig. 9.** Paul Walker's death daily (a) and hourly (b) impact value.

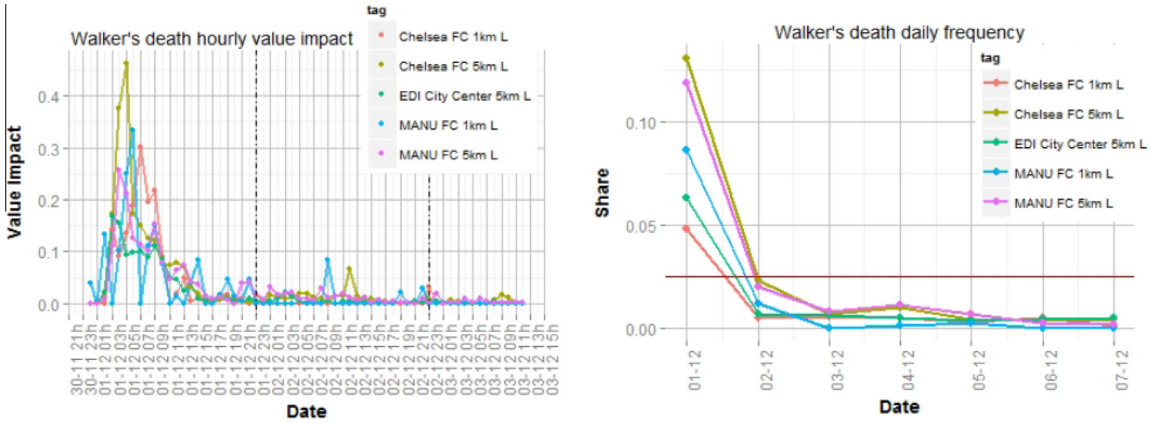



**Fig. 10.** Paul Walker's death hourly detailed impact value (a) and Frequency (b).

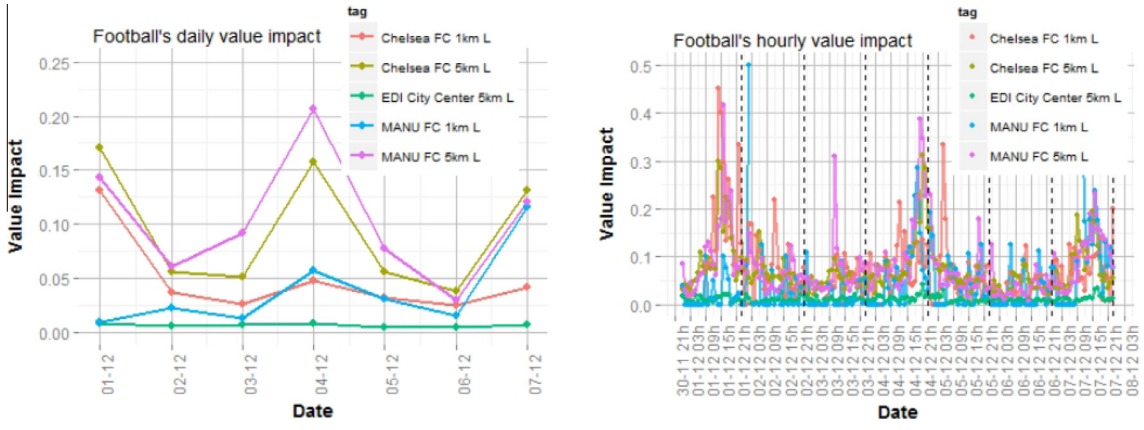



**Fig. 11.** Football daily (a) and hourly (b) impact value.

December short before 19:00 GMT after being hospitalized. The SM platforms echoed both events and our harvested gathered the users' reactions in all 5 locations.

As aforementioned, both events share the same pattern: one-off, high SM resonance, a very quick ramp-up phase to a peak and a rather short fade-out phase to practically disappear after a few days.

Fig. 5(a) shows the active impact (as defined in Eq. (17)) over all users identified per location. The small-radius harvester in Chelsea shows a higher score than the others in the first day but stabilizes one day after. The Passive Impact though, as displayed in Fig. 5(b) positions Edinburgh and the 5km Chelsea harvester, specially the day after the decease, much higher than the small-radius ones. The less Exposure of Manchester 1 km users is remarkable. Fig. 6(a) shows the daily aggregated view of the value, which is aligned with the results of the previous charts: Chelsea and Edinburgh more impacted by the death than Manchester.

Manchester is on the other hand, where Paul Walker's death caused a higher impact than in Edinburgh – see Figs. 8 and 9(a). The reason might be related to the different affinities of the Edinburghian and Mancunian local tweeting communities. Again, this is one example of the capabilities of our framework to understand localized preferences and people profiles. The long tail values in

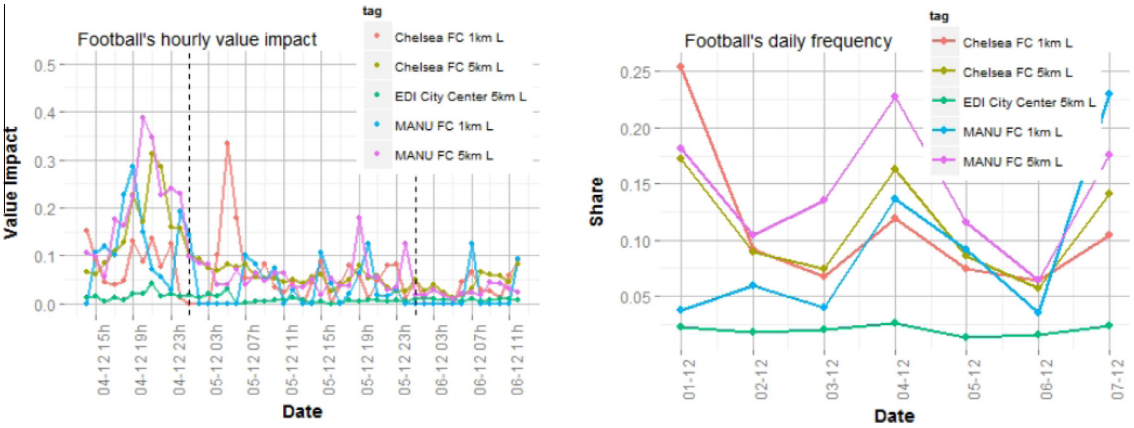



**Fig. 12.** Football hourly detailed impact value (a) and Frequency (b).

**Table 3**
Impact modeling of the Mandela's and Walker's deaths on 5 locations.

| Walker's vs. Mandela's death impact 1st–7th Dec. 2013 | | | | | |
|---|---|---|---|---|---|
| Topic | Harvester | Recency | Frequency | Value impact | Resulting impact |
| *Mandela's death* | *Chelsea* 1 km | 0.85 | 0.04 | 0.028 | 11.41 |
| | *Chelsea* 5 km | 0.85 | 0.02 | 0.039 | 11.44 |
| | *MANU* 1 km | 0.85 | 0.021 | 0.013 | 9.95 |
| | *MANU* 5 km | 0.85 | 0.021 | 0.021 | 10.40 |
| | *EDI* 5 km | 0.85 | 0.03 | 0.042 | 11.92 |
| *Walker's death* | *Chelsea* 1 km | 0.14 | 0.018 | 0.01 | 2.89 |
| | *Chelsea* 5 km | 0.14 | 0.033 | 0.035 | 4.54 |
| | *MANU* 1 km | 0.14 | 0.024 | 0.015 | 3.02 |
| | *MANU* 5 km | 0.14 | 0.034 | 0.025 | 3.97 |
| | *EDI* 5 km | 0.14 | 0.021 | 0.018 | 3.15 |

both cases show a very slight increase 4 days after the deceases, which might be triggered by TV media showing *TV specials* about these personalities and/or best movies in the case of Walker.

Figs. 7 and 10(a) provide the value overview by hour for both deaths. We observe a similar pattern: all monitors showing a peak which lasts for 4–5 hours to then go down. The hourly impact score right after the announcement of the tragic event topped 0.4, which is comparable with the highest impact values reached during football games in their local stadia (see in Fig. 12 the Manchester harvesters when the Manchester United–Everton match was kicked off at 07:45 pm in Old Trafford).

Figs. 7 and 10 present the daily Frequency for both events, each one with a suggested threshold for the Recency calculation. The Table 3 shows the value for the different components for one week. Even if the value component taken the death's day and the day after is similar for both Mandela and Walker, the fact that we are considering the entire week (1st to 7th Dec.) and Walker's death took place right at the beginning, make the Recency substantially differ in favor of Mandela's impact, which we see in the resulting score. The charts in Fig. 13 help understanding the hourly value distribution in both cases over the week per harvester.

In Fig. 14(a) we've taken the hourly impact value obtained for both topics the day after both personalities passed away. We show the difference between both values per hour and per location, to demonstrate how making the impact quantifiable enables the comparison. Thus, we can see that in the locations under analysis, Mandela's death had a greater impact during the first two to three hours, whereas Walker's death outperformed it in the next 4 hours especially in the greater Chelsea area. Both impact values get closer as the day proceeds.

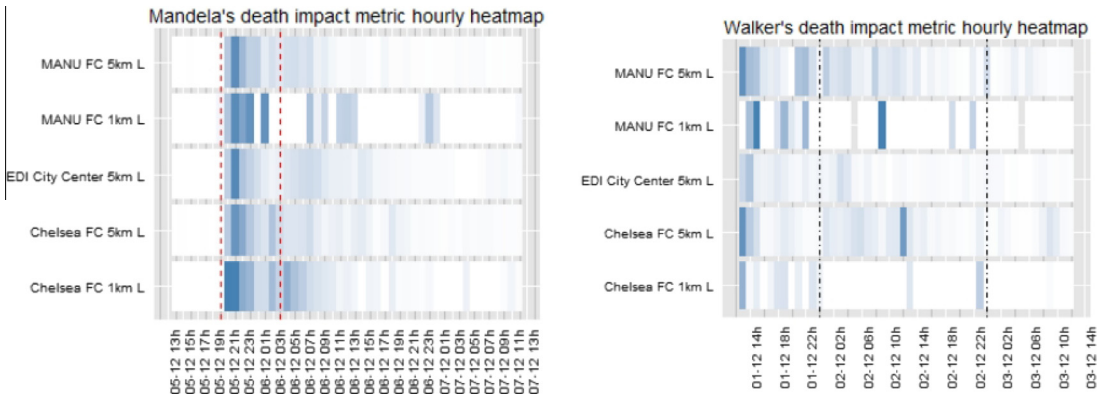



**Fig. 13.** Mandela's (a) and Paul Walker's (b) death hourly impact heatmap.

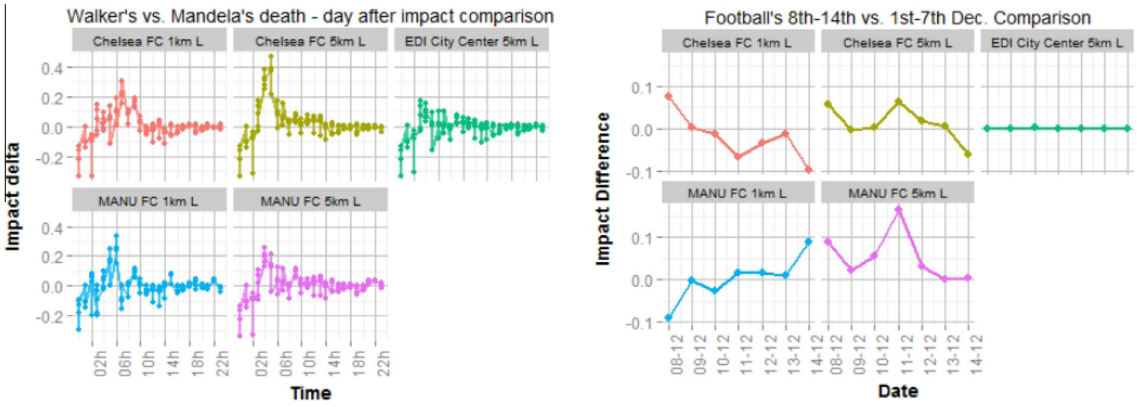



**Fig. 14.** Mandela–Walker day after death comparison (a) and Week over Week football comparison (b).

**Table 4**
Impact modeling of the topic football on 5 locations.

| Football impact 1st–7th Dec. 2013 | | | | | |
|---|---|---|---|---|---|
| Topic | Harvester | Recency | Frequency | Value impact | Resulting impact |
| *Football* | *Chelsea* 1 km | 1 | 0.387 | 0.051 | 24.721 |
| | *Chelsea* 5 km | 1 | 0.112 | 0.102 | 19.514 |
| | *MANU* 1 km | 1 | 0.087 | 0.042 | 15.207 |
| | *MANU* 5 km | 1 | 0.11 | 0.107 | 19.964 |
| | *EDI* 5 km | 0 | 0.01 | 0.01 | 0.9 |

**Table 5**
Chelsea and Manchester United matches calendar.

| 1st–14th Dec. 2013 football fixtures | | | | |
|---|---|---|---|---|
| Date | Competition | Home | Result | Visitor |
| Sun 01.12.13 | Premier | **Chelsea** | 3 - 1 | Southampton FC |
| Sun 01.12.13 | Premier | Tottenham Hotspur | 2 - 2 | **Manchester United** |
| Wed 04.12.13 | Premier | Sunderland | 3 - 4 | **Chelsea** |
| Wed 04.12.13 | Premier | **Manchester United** | 0 - 1 | Everton |
| Sat 07.12.13 | Premier | Stoke City | 3 - 2 | **Chelsea** |
| Sat 07.12.13 | Premier | **Manchester United** | 0 - 1 | Newcastle United |
| Tue 10.12.13 | Champions | **Manchester United** | 1 - 0 | Shakhtar Donetsk |
| Wed 11.12.13 | Champions | **Chelsea** | 1 - 0 | Steaua Bucuresti |
| Sat 14.12.13 | Premier | **Chelsea** | 2 - 1 | Crystal Palace FC |

### *5.3. Topic 2: Impact modeling for the topic football*

The first two weeks of December have been very intense in terms of football matches for both Chelsea FC and Manchester United FC, as we can see in Table 5.

Fig. 11(a) shows very well the effect of a club playing in its home stadium in terms of impact. On Sunday the 1st, the active impact close to Old Trafford – harvester MANU 1 km – is very low compared with next Wednesday and Saturday. The same behavior is shown by the Chelsea 1 km harvester: high impact on the Sunday and pretty low on the next Wednesday and Saturday, which again comply with the fixtures given in the Table 5.

When we increase the radius and move away from the stadia, the differences between the impact when the local club plays at home or as visitor expectedly diminishes. It applies to both Active and Passive Impact and to the aggregation of both (Fig. 11) . It's remarkable the low impact of the topic football in the city of Edinburgh or just the low interest on the Premier league.

Fig. 14 (b) shows the week vs. week impact comparison per harvester. The impact of the home matches is obvious, but also the importance of Champions League in the 5 km harvesters. Edinburgh stays"unimpacted" by the topic football, as we said before.

In the Table 4, we present the impact results for the topic football in the defined week. We see the Chelsea 1 km harvester leading the table with a score of over 24 points. MANU FC showed a comparably lower impact – almost 10 points back. The 5 km MANU and Chelsea harvesters, not so sensible to when the local club plays in the local stadium, show remarkably similar results. Chelsea FC and MANU played 3 times each during the week, but Edinburgh stayed completely indifferent to that.

## 6. Conclusions

In this paper we present a new model built upon geo-localized Social Media interactions to quantify the impact of a topic on a particular location and to monitor how it changes over time. As a foundation for our model, weve chosen the well-known RFM paradigm and introduced the concepts of Exposure and Engagement of a particular user with a topic to model the Monetary or Value component. Concretely in the industry domain, our new Social Media RFM model could present a good performance in a variety of applications, ranging from event planning and marketing (campaign monitoring, topic affinity advertising, interest targeting) to

market research (media monitoring, geo-located panelists, news impact).

The introduction of the Exposure and Engagement metrics allows for modeling at user level both passive and active topic impact and allows for filtering and segmentation based on different user attributes as all the metrics are defined at user level. As show-cased with the implemented system and the football analysis, our metrics perform well even in hourly chunks; they are consistent over time (delivering similar results in similar situations in different periods) and easy to understand (as they reflect the nature of the social network but at individual level).

In a variety of scenarios or extreme cases, the Social Media RFM model is proven to be robust always delivering meaningful metrics as discussed and demonstrated with the examples we analyzed based on the system that has also been implemented as part of this paper. One of the strengths of the approach we suggest in this work is the fact that the topic impact comparison is supported in heterogeneous scenarios, for example with different topics over different time frames in different locations.

For the sake of simplicity, in our Social Media model we have considered the links between users as equally powerful in the Exposure calculation, which leaves the door open for improvement as different users might have differential influence power – popularity – on others which might result into a more realistic Exposure result. Another challenging aspect inherent to the source of the data itself in the suggested approach is the bias derived from using online-only social network users to assess the topic impact on a particular place. The fact that we focus on English speaking users only reinforces the bias, as a topic might well impact different cultures and nationalities with different intensity. These limitations certainly point to future research directions.

## Acknowledgments

This paper has been developed with the financing of projects TIN2010-17876, TIC5299, TIC-5991 and TIN2013-40658-P.

## References

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of lda generative models. In *Machine learning and knowledge discovery in databases* (pp. 67–82). Springer.

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on world wide web* (pp. 61–70). ACM.

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly, 47*(4), 644–675.

Bennett, S. (2012). Twitter now has more than 140 million active users sending 340 million tweets every day.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific American, 284*(5), 28–37.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Bogart, L. (1967). *Strategy in advertising*. New York: Harcourt, Brace & World.

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science, 14*(4), 378–394.

Cataldi, M., Di Caro, L., & Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining* (pp. 4). ACM.

Cavusoglu, H., Hu, N., Li, Y., & Ma, D. (2010). Information technology diffusion with influentials, imitators, and opponents. *Journal of Management Information Systems, 27*(2), 305–334.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194–1197.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Cheng, Z., Caverlee, J., Lee, K., & Sui, D.Z. (2011). Exploring millions of footprints in location sharing services. *ICWSM* (pp. 81–88).

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 443–452). ACM.

Clark, E., Roberts, T., & Araki, K. (2010). Towards a pre-processing system for casual english annotated with linguistic and cultural information. In *Proceedings of the Fifth IASTED International Conference, Vol. 711*, (pp. 044–84).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS, 41*(6), 391–407.

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 415–430.

Farris, P. W., Bendle, N. T., Pfeifer, P. E., & Reibstein, D. J. (2010). Marketing metrics: The definitive guide to measuring marketing performance. *Pearson Education*.

Fujita, S., & Fujino, A. (2013). Word sense disambiguation by combining labeled data expansion and semi-supervised learning method. *ACM Transactions on Asian Language Information Processing (TALIP), 12*(2), 7.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211.

Guille, A., & Hacid, H. (2012). A predictive model for the temporal dynamics of information diffusion in online social networks. In *Proceedings of the 21st international conference companion on world wide web* (pp. 1145–1152). ACM.

Hughes, A. (2005). *Strategic database marketing*. New York: McGraw-Hill.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study. *Procedia Computer Science, 3*(0), 57–63.

Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics, 32*(4), 485–525.

Kumar, V. (2008). Customer lifetime value the path to profitability. *Foundations and Trends(R) in Marketing, 2*(1), 1–96.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International conference on world wide web, WWW '10* (pp. 591–600). New York, NY, USA: ACM.

MacKay, D. J., & Peto, L. C. B. (1995). A hierarchical dirichlet language model. *Natural Language Engineering, 1*(3), 289–308.

Manaris, B. (1998). Natural language processing: A human–computer interaction perspective.

Mathioudakis, M., & Koudas, N. (2010). Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD international conference on management of data, SIGMOD '10* (pp. 1155–1158). New York, NY, USA: ACM.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM, 38*(11), 39–41.

Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology, 62*(5), 902–918.

O'Connor, B., Krieger, M., & Ahn, D. (2010). Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference* (pp. 161–172). Brisbane, Australia.

Porcel, C., Tejeda-Lorente, A., Martnez, M. A., & Herrera-Viedma, E. (2012). A hybrid recommender system for the selective dissemination of research resources in a technology transfer office. Information Science, *184* (1), 1–19.

Rajyalakshmi, S., Bagchi, A., Das, S., & Tripathy, R. M. (2012). Topic diffusion and emergence of virality in social networks. *CoRR*, abs/1202.2215:1212–2232.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011a). Influence and passivity in social media. In *Proceedings of the 20th International conference companion on world wide web, WWW '11* (pp. 113–114). New York, NY, USA: ACM.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011b). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on world wide web, WWW '11* (pp. 695–704). New York, NY, USA: ACM.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web, WWW '10* (pp. 851–860). New York, NY, USA: ACM.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York, NY, USA: McGraw-Hill, Inc..

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. *ICWSM, 11*, 329–336.

Sohrabi, B., & Khanlari, A. (2007). Customer lifetime value (CLV) measurement based on rfm model. *Iranian Accounting & Auditing Review, 14*(47), 7–20.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social mediasentiment of microblogs and sharing behavior. *Journal of Management Information Systems, 29*(4), 217–248.

Sutherland, I. E., Sproull, R. F., & Schumacker, R. A. (1974). A characterization of ten hidden-surface algorithms. *ACM Computing Surveys, 6*(1), 1–55.

Tejeda-Lorente, A., Porcel, C., Peis, E., Sanz, R., & Herrera-Viedma, E. (2014). A quality based recommender system to disseminate information in a university digital library. *Information Science, 261*, 52–69.

White, D. (2013). Social media growth 2006–2012.

Yang, Q. X., Yuan, S. S., Zhao, L., Chun, L., & Peng, S. (2003). Faster algorithm of string comparison. *Pattern Analysis & Applications, 6*(2), 122–133.

Ye, S., & Wu, S. F. (2010). Measuring message propagation and social influence on twitter.com. In *Proceedings of the second international conference on social informatics, SocInfo'10* (pp. 216–231). Berlin, Heidelberg: Springer-Verlag.