Knowledge-Based Systems 80 (2015) 163–179

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

CARESOME: A system to enrich marketing customers acquisition and retention campaigns using social media information

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

[a] Department of Computer Science and A.I., University of Granada, Granada E-18071, Spain
[b] Department of Computer Science, University of Jaén, Jaén E-23071, Spain
[c] Iwate Prefectural University (IPU), Iwate, Japan

ARTICLE INFO

Article history:
Received 30 November 2014
Received in revised form 20 December 2014
Accepted 30 December 2014
Available online 9 January 2015

Keywords:
Intrinsic impact
Extrinsic impact
Social CRM
Customers retention
Customers acquisition
Localized social media
Ubiquitous insights

ABSTRACT

The enabling of geo-localization for Social Media content opens the door to a new set of applications based on the voice of the customer. For any company it is critical to understand both their own and their competitors' strengths and weaknesses in all locations where they offer a service. With this motivation we created a Customers Acquisition and REtention system based on SOcial MEdia (CARESOME). Our system extracts and separates all social media interactions in a given location by market player and communication purpose and quantifies the impact of each single interaction over a given time period. To model the impact of the social media interactions, CARESOME relies on a set of metrics based on both intrinsic and extrinsic components—including Entity Engagement Index, Differential Perception Factor, Tie-Strength and Number of Exposed users—. In addition to the definition of our impact quantification metrics, we provide a thorough discussion about the design decisions taken to build our system. To illustrate the behavior of our system, we show-case a real world scenario from the airline industry based on two major airports in Great Britain.

1. Introduction

Social Media (SM) started as a space where anybody with an account could interact with any other user, share content, express their own personal views, etc. without being subjected to any kind of censorship. As a side effect of this democratization of the Web, the relationship between a company and its customers and stakeholders went through an unprecedented transformation [1]. For the first time, customers could engage in a near real time manner with companies and brands [2]. The advent of SM radically changed the way customers engage with service providers or product vendors. Any customer could express in an unfiltered way his/her opinion about a brand, a service, a price increase, etc. and the result of it was publicly available in a near real time manner to other customers or customers-to-be. The *killer application* of SM in the consumer market has been the customer empowerment. The customer feedback, that used to be trapped in the traditional offline *word-of-mouth* modus operandi, is now available to each and every user willing to know more about the quality of service of any company in the world. SM made these communication barriers fall and changed the customer-company engagement rules for ever, as different types of business are using customer data for better comprehension on customers data [3].

Companies had been left with no choice but embracing the new customers' engagement channel and developing customers' acquisition and customers' retention strategies on top: leaning back and not doing anything was no longer an option [4]. Early adopters managed to build up a new form of competitive advantage relying on both own customers' binding and competitors' customers caption strategies. Spotting signs of own customers' satisfaction decay in a given location made companies trigger local customers' retention campaign, for example in [5], stating that purchasing behaviors can be significance altered by supplying consumers with seller-generated information. Likewise, localized satisfaction decay in competitors' customers base pushed companies to implement highly aggressive regional campaigns to take advantage of the weak spot by mitigating diffusion [5].

The need for geo-localized systems to monitor the customer satisfaction at a local scale and to assess the impact of customers' interactions with the brand over SM, emerged [6]. Early warning systems—the equivalent in other domains—have been increasingly adopted in the field of disaster prevention as the sensorial technique allowed for semi-automatic monitoring. There are countless applications for early detection of earthquakes [7–9], pandemics

* Corresponding authors. Tel.: +34 958 24 4258; fax: +34 958 24 3317.
E-mail addresses: juan.bernabe-moreno@webentity.de (J. Bernabé-Moreno), cporcel@ujaen.es (C. Porcel), viedma@decsai.ugr.es (E. Herrera-Viedma).

http://dx.doi.org/10.1016/j.knosys.2014.12.033
0950-7051/

[10,11], flood and other natural hazards [12]. In the financial domain fast alerting system have been employed for a wide range of purposes: for example, all variety of economic indicators have been used at a macro level to assess the vulnerability of emerging and existing markets [13–15] and to detect financial crisis in their early stages [16,17], but also at a much more micro level to detect for example critical transactions [18], etc.

When the access to the world wide web (WWW) escaped the desktop boundaries and became mobile and pervasive—mainly because of WiFi and the third generation mobile cellular system for networks based on the GSM standard, Universal Mobile Telecommunications System (UMTS)—, the SM platform providers leveraged the geo-location of the interactions as a *highly enriching* additional information source. There was such a demand for location enriched user interactions, that new platforms less focused on content but more focused on location, like Foursquare emerged and quickly started conquering the market. Well established content and community based platforms traditionally positioned as the medium where customers engaged with brands (such as Facebook or Twitter) immediately reacted enabling the localization of the user interactions to improve the user experience. The geo-localization of customers' interactions opened a new door for companies to better understand their own customers' base and develop strategies to take customers away from competitors in their own favor. Understanding the impact of each and every interaction over SM on one hand and triggering on time the appropriate reaction proved to be two essential factors for succeeding in a customers retention or customers acquisition strategy [19]. Even if companies are heavily investing in standing up SM care teams, the SM adoption makes the handling of each and every SM interaction far from scalable, which introduces the need for a way of quantifying their impact.

In this paper we present our Customer Acquisition and REtention system over SOcial MEdia, which leveraging Big Data technologies [20], implements a framework based on geo-localized Tweets, to measure the impact of the interactions created in a location on a brand or institution, or any kind of entity. Our system takes a holistic view over all factors that play a role in the impact perception within a SM context, such as authors engagement with the entity, followers exposure to the interactions, Tie-Strength between authors and followers, etc. As outcome, CARESOME produces a set of metrics for a location over a given time frame aggregated by communication purpose category to enable the response by different company departments (e.g.: customer care is likely to focus on the categories criticism and complaints, while marketing would rather be interested in measuring the impact of a new campaign based on positive feedback, etc). Additionally, the result of these metrics is packed into impact categories to enable faster decision making, as time to reaction is proven to be a critical success factor in every early-warning like system. In other words, CARESOME turns the information extracted from the different SM channels into actionable insights for companies to steer their customer acquisition and loyalization campaigns based on opinions of customers.

We started our work presenting all the background our research is built upon (Section 2). In Section 3, we introduced the impact quantifying framework and define all relevant metrics. Section 4 explains the CARESOME system architecture and Section 5 presents a real-world scenario and provides a discussion about the system performance and our design decisions. Section 6 closures this work pointing out future research lines and summarizing our conclusions.

2. Background

Nowadays, almost every company relies on SM as a communication channel to push company messages and offers, but also increasingly to obtain unfiltered feedback from both existing and prospective customers. Many studies have focused on different aspects of the SM adoption: Kaplan et al. [21] highlighted the need for the integration of SM with traditional media to reach customers more efficiently, while defending the advantages of SM to engage with customers in a time-close and high-efficient manner. Mangold et al. [22] built upon the idea of considering SM as integral part of the promotion mix, emphasizing the benefit of a less controlled environment to better understand customers.

Several papers focused on researching the role of SM in business and corporations. Jansen et al. in [23] analyzed the corporate image impact of all interactions related to a brand created over the Twitter channel. In [24], Li et al. explained the positive impact of the user engagement over the Twitter company channels on the corporate reputation. In [25] Java et al. demonstrated how similar intentions foster connectivity between users and community building around brands and institutions. Plenty of studies shed light on how companies shall deal with SM related issues like trust and distrust within online communities [26] and protection of user's information [27].

SM rapidly moved from being *yet another channel* in the communication strategy of a company to be labeled as a *game changer* to engaging with customers: Hennig et al. [1] explained how microblogging was shaking traditional business models by increasing the role of product quality, especially reducing the time window where product new adopters did not have any feedback on the product. Culnan et al. [28] pointed out the need for brands to create communities to exploit the full potential of the virtual customer environments. In [29] the link between SM engagement and profitability of online companies was analyzed by Chan and his co-authors. In [30], Rapp and his co-authors analyzed the role of SM from the seller, retailer and consumer perspective, demonstrating the value of the SM interactions for better conversion rate.

The effect of the Worth-of-Mouth (WoM) marketing has been extensively researched together in the SM context. Chevalier and his co-authors analyzed in [31] the effect of book reviews. Villanueva et al. [32] researched the differences in terms of loyalty and equity of customers being acquired through marketing-induced activities vs. WoM gained customers, pointing out performance differences. Bolton established back in 1998 [33] a modeled based on the link between customers retention and customers satisfaction and Rishika et al. [34] empirically proved the effect of increased SM engagement on the customer visit frequency and customer value.

As proved in [35,36], the spreading of bad news takes place really fast over the SM channel, which corroborates their value for the promptly detection of customers' complaints, service outages, etc. Countless papers built upon the fast news spreading aspect of SM: in [37], Sakaki et al. define an algorithm based on particle filtering for geo-location and spread for earthquakes early detection based on tweets. Also based on tweets, Culotta et al. suggest in [38] a method to detect epidemic expansion on early stages. In [39], Middleton and his co-authors present a near real time system to map crisis based on several geo-localization techniques of SM information. In the same research line, Yin et al. in [40] present a system that implements text mining and natural language processing (NLP) techniques to extract situation awareness information from Twitter to support crisis coordination and emergency response. In [41] Colbaugh and Glass employed a stochastic model for dynamic of the interactions based on the underlying network structure to generate useful predictions about the spread of information. The US Homeland department pioneered the usage of SM to collect real time information about incidents, quantify their extent, monitor their evolution and channel the proper response—programme SMART-C (SM Alerts and Response to Threats to Citizens)—[42].

Predicting (i.e., customers) behaviors in SM for management decision making is still challenging tasks [43–45]. The analysis of SM content and engagement to predict upcoming events has been also intensively researched. In [46] the Bothos, Apostolou and Mentzas explain how agents constantly analyzing social media content according to the Belief-Desire-Intentions paradigm can extract enough sentiments and assessments to enable informed decision making in the markets they operate.

Our Impact metrics, as we are going to explain later in this paper, relies on how influential a particular SM's user is. Modeling influence in SM channels has been subject of intense research over the last few years. Kwak [47] defined 3 metrics aimed at quantifying the *social influence*: the so called*propagation influence*, based on the Google Search PageRank algorithm [48], *followers influence*—more followers implies more influence—, and *re-tweet influence*—more re-tweets means more influence—. Ye and Wu [49] relied on the same set of metrics but changing the propagation influence by a much simpler to compute *reply influence*—the more replies one user receives, the more influential the user is—. Cha [50] also identified 3 influence drivers: the size of the user's audience or social network—*indegree influence*—, the generated content with pass-along value—*retweet influence*—, and the engagement in others' conversation—*mention influence*—. Romero et al. [51] develop a mechanism to quantify how the exposure to other users is making them adopt a new behavior. Yang et al. in [52] add a new dimension to the influence computing, namely the response immediacy in their influence modeling for an online health care community.

There have been several studies showing how the Tie-Strength between two SM users plays an important role in the perception of SM interactions. Marsden and his co-authors in [53] back in 1984 laid the foundations for measuring the Tie-Strength after Mark Granovetter introduced the concept in 1973 in his paper "The Strength of Weak Ties" [54]. In [55], a model to predict tie strength by exploiting social media interaction parameters is discussed. The work done by Haythornthwaite [56] confirmed that more strongly tied pairs communicate more frequently, maintain more and different kinds of relations and use more media to communicate. Grabowicz et al. in [57] analyzed the relationship between SM links and real-world tie-strength and Pan et al. in [58] attempted to quantify the role of tie-strength plays in scientific collaboration networks. Shin et al. presented a method to quantify the degree of user sociability in SM relying on the tie-strength [59].

3. Framework definition

The ultimate aim of our framework is providing a means to quantify the impact in an efficient way, so that our metrics can be consumed near real time for decision making. The Impact of a SM interaction with a brand can be modeled from two perspectives: intrinsically—reflecting the impact perception on the SM interaction author—and extrinsically—which captures how the SM interaction impacted its author's SM network—.

The Impact computed over all users located in a place provides a really sensible Key Performance Indicator (KPI) to take decisions upon. In our approach, the Impact is provided in different categories, which perfectly maps with the way big corporations are usually structured in departments. For example, the complaint management department is interested in monitoring the impact over time of the complaints coming from a place over the SM channel, whereas marketing rather focuses on the monitoring of suggestions, criticism and engagement with running campaigns. CARESOME provides also the flexibility of defining own categories in the event that the standard ones are not suitable.

3.1. Preliminary definitions

Before starting with the definition of our framework, 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 Δt

$$U \equiv \{u_i\}\ (i = 1, \ldots, n),\ InLocation(u_i, L, \Delta t) \tag{1}$$

Definition 2. The Social Network for a given user u_i is defined as:

$$SN\ (u_i) \equiv \{u_j\}(j = 1, \ldots, n),\ \forall\ u_j \in SN(u_i),\ Follows(u_i, u_j) \tag{2}$$

$Follows(u_i, u_j)$ is a relation 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), there are others where it is not necessarily the case, like Twitter, where $Follows(u_i, u_j) \not\Leftrightarrow Follows(u_j, u_i)$.

The fact that a user u_j is part of the SN of another user u_i does not necessarily mean that u_j has to be located in the same location L of user $u_i : u_j \in SN(u_i), u_i \in U \not\Rightarrow u_j \in U$, as $u_j \in SN(u_i), u_i \in U \not\Rightarrow InLocation(u_j, L, \Delta t)$

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

$$SN(U) \equiv \cup_{i=1}^{|U|} SN(u_i) \tag{3}$$

Definition 4. A Social Media Interaction *it* represents the atomic piece of content authored by the user u_i during the time Δt in a Social Media Platform (e.g.: a tweet, a re-tweet).

The function $Author(u_i, it_i, \Delta t)$ returns *True* if u_i created the interaction it_i in the time period Δt, and *False* otherwise.

The time interval Δt 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*.

Definition 5. We define all user interactions (*Interactions*) for a given user u_i over a time interval Δt, as:

$$Interactions\ (u_i, \Delta t) \equiv \{it_i\}(i = 1, \ldots, n),$$
$$\forall it_i \in Interactions\ (u_i, \Delta t),\ Author(u_i, it_i, \Delta t) \tag{4}$$

Definition 6. The set of User Foreign Interactions ($ForeignInteractions(u_i, \Delta t)$) represents all Interactions with a direct mention to the user u_i but not authored by him/her:

$$ForeignInteractions\ (u_i, \Delta t) \equiv \{it_j\}$$
$$\forall it_j \in ForeignInteractions(u_i, \Delta t), \neg Author(u_i, it_j, \Delta t),$$
$$DirectMentioned(it_j, u_i) \tag{5}$$

$DirectMentioned(it_i, u_i)$ is a function retrieving *True* if the user u_i is explicitly mentioned in it_i. In Twitter, the User Foreign Interactions include re-tweets, mentioned and replies.

$$ForeignInteractions(u_i, \Delta t) \cap Interactions(u_i, \Delta t) = \emptyset$$

Let's illustrate it with one example; for a user with the user name *@user1*, a tweet created by *@user2* saying "*Happy birthday @user1*" represents a foreign interaction for *@user1*. If @user3 retweets it, it counts as well as a foreign transaction for *@user1*, who is mentioned in the tweet text, but also as a foreign interaction for *@user2*, as his/her tweet has been re-tweeted.

Definition 7. The set of Direct Mention Interactions is as a subset of *Interactions* $(u_i, \Delta t)$ defined as follows:

$$DirectMentionInteractions\ (u_i, \Delta t) \equiv \{it\},$$
$$\forall it_i \in Interactions\ (u_i, \Delta t), \exists u_j \mid it_i \in ForeignInteractions\ (u_j, \Delta t) \quad (6)$$

Intuitively, DirectMentionInteractions represents all the interactions created by the user u_i where any other user is explicitly mentioned. Obviously, $DirectMentionInteractions\ (u_i, \Delta t) \subseteq Interactions\ (u_i, \Delta t)$.

Definition 8. A *Social Media Entity E* is the representation of the set of all terms used by Social Media Users to interact with a real world entity such as a brand, a corporation, an institution, a club, etc. It includes for example social media account name(s), product names, company abbreviations or company slogans.

Definition 9. We define the set of *Interactions* for a given user u_i with the entity E over a time interval Δt as:

$$Interactions\ (u_i, E, \Delta t) \equiv \{it\},$$
$$\forall it_i \in Interactions\ (u_i, E, \Delta t),\ Author(u_i, it_i, \Delta t)\ \wedge related(it_i, E) \quad (7)$$

where $related(it_i, E)$ is a NLP membership function retrieving *True* if the interation it_i is connected to the entity E—intuitively, one or more words from the Entity defining set are mentioned in it_i—and *False* otherwise.

3.2. User-entity engagement

Based on the before mentioned definitions, we introduce the concept of "engaged", defined as a logical function:

$$Engaged\ (u_i, E, \Delta t) \equiv True, \exists it_i,$$
$$it_i \in Interactions\ (u_i, E, \Delta t), u_i \in U \cup SN(U) \quad (8)$$

where u_i is the user, E is the representation of the Entity, Δt is the time span specified consisting of two components (*t_startdate* and *t_enddate*), it_i represents a social media interaction and $Interactions\ (u_i, E, \Delta t)$ represents the interactions of the user u_i related to the Entity E in the time interval Δt, as we explained before. At user level, it's also possible to define a metric to quantify the level of engagement of the user with the Entity, the so called *Entity Engagement Index (EEI)*:

$$EEI(u_i, E, \Delta t) = \frac{|Interactions\ (u_i, E, \Delta t)|}{|\cup_{k=1}^{|E|} Interactions\ (u_i, E_k, \Delta t)|} \quad (9)$$

where u_i represents a given SM user, E is the representation of the Entity, $Interactions\ (u_i, E, \Delta t)$ is as defined before and $|\cup_{k=1}^{|E|} Interactions\ (u_i, E_k, \Delta t)|$ is the cardinal for the union set of all interactions with all possible entities created by the user u_i during the time span Δt.

The Entity Engagement Index can also be expressed as a share of the interactions related to one entity over all interactions:

$$EEI(u_i, E, \Delta t) = \frac{|Interactions\ (u_i, E, \Delta t)|}{|Interactions\ (u_i, \Delta t)|} \quad (10)$$

3.3. Social media communication intent

Behind each and every posts or tweet or, in general, piece of content authored by a user in a Social Media platform there is an underlying communicative purpose: praise a piece of information or a company or an action, express some criticism, make a direct complaint, request information, provide an answer, etc. In the same way we introduced before the concept of *Social Media Entity*, we now provide the definition for *Communication Purpose Category*

Definition 10. A *Communication Purpose Category P* is the representation of the set of all terms in all varieties of forms used by Social Media Users to express a particular communicative intention (such as praise, criticism, information inquiry, complaints, etc).

Even if the boundaries might not be crisp, we can assign each interaction to a *leading Purpose Category* within the set of purpose categories considered PC:

$$\forall it_i \in Interactions\ (u_i, E, \Delta t), \exists p_k, Purpose(it_i) = p_k, p_k \in PC \quad (11)$$

where it_i represents a SM interaction, $Interactions(u_i, E, \Delta t)$ is the set of all interactions created by u_i over Δt, p_k is a the leading Communication Purpose, *PC* is the set of all Communication Purpose Categories.

$Interactions(u_i, P, \Delta t)$ represents the set of all interactions authored by a user u_i over the period of time Δt whose leading Purpose Category is *P*.

3.4. Differential perception factor, exposure and tie strength

Based on the concepts introduced in the previous Sections 3.2 and 3.3, we can define the building blocks for the metrics to quantify the impact created by the users located in a given area over time, and thereby enable the early reaction and steering of marketing retention and acquisition campaigns.

We introduce the so called *Differential Perception Factor* modeled as *Purpose Share* (see the previous Def. 10), which allows for latterly defining a correction factor to remove the SM behavioral bias:

$$DPF(u_i, E, P, \Delta t) = \frac{|Interactions\ (u_i, E, \Delta t) \cap Interactions\ (u_i, P, \Delta t)|}{|Interactions\ (u_i, P, \Delta t)|} \quad (12)$$

To make it more intuitive, let's bring up one example: let's assume that a given user in a location started posting complaints over Twitter about the bad services provided by his/her mobile operator. If the same user was very active posting complaints about many other companies such as the local transportation service, the internet provider, the employer, certain celebrities, etc., the Purpose Share for *Complaints* would be rather low. On the other hand, if the same user hardly ever complaints about anything, a single interaction pointing out his/her discontent with the mobile operator would be perceived as something rather serious and more significant.

The impact measure of a social media interaction originated in a particular area shall consider the number of users that are exposed to this content, no matter if they are in the same area or some where else.

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

$$Exposed\ (u_i, u_j, E, \Delta t) = \begin{cases} True, & u_j \in SN(u_i), \exists it_k, it_k \in Interactions(u_i, E, \Delta t), \\ & P(read(u_j, it_k, \Delta t)) \geqslant Threshold \\ False, & otherwise \end{cases} \quad (13)$$

where $P(read(u_j, it_k, \Delta t))$ is the probability that the user u_j reads the content posted in the interaction it_j in the designated time Δt. The $Threshold \in [0, 1]$ is defined to narrow down the selection.

The reason why we introduce the concept of *Exposed User* is to address the fact that not all the SM content created by the social

network of a particular user *is consumed by* the user. The subset of users exposed to the topic can then be defined as:

$$ExposedUsers\ (u_i, E, \Delta t) \equiv \{u\},\ \forall u_j,\ Exposed(u_i, u_j, E, \Delta t) = True,\ u_i \in U \quad (14)$$

An additional yet quite relevant aspect we incorporate to the Impact definition is the relationship between the author of the social media interaction and the user in his/her SM network. Depending on this relationship, the level of perceived relevance might vary. For example, if a given user u_i is a good friend of $u_j, u_j \in SN(u_i)$, the relevance, the u_j perceives u_j's posts to have is higher than it would be if there was practically no link between these users apart from the fact that u_j is part of the SM network of u_i. Thus, we define Tie-Strength between two social media users as:

$$TieStrength\ (u_i, u_j, \Delta t) = \frac{\#(ForeignInteractions(u_j, \Delta t) \cap DirectMentionInteractions(u_i, \Delta t))}{\#(DirectMentionInteractions(u_i, \Delta t))} \quad (15)$$

$u_i \in U, u_j \in SN(u_i)$ Basically, tie strength from user u_i on user u_j is the ratio between the interactions created by u_i where u_j has been particularly mentioned and all interactions created by u_i mentioning somebody. $TieStrength(u_i, u_j, \Delta t)$ is not necessarily $TieStrength(u_j, u_i, \Delta t)$. This metric is supported by the definitions Definitions 6 and 7.

Instead of taking the subset $DirectMentionInteractions(u_i, \Delta t)$ in the previous definition, it would be also possible taking the entire set $Interactions(u_i, \Delta t)$, but the tie strength would return rather lower numbers, as many users just broadcast messages to their entire network without explicitly mentioning anybody in particular. The Fig. 1 shows a fictive time line over 4 days for the users X, Y and Z and 3 other users A, B and C in $SN(X)$. The values for the metrics required to compute the Tie Strength for this example can be found in Fig. 2 with the entire set of combinations for the SM users A,B,C,X,Y and Z.

3.5. Intrinsic and extrinsic impact metrics

Our suggestion for modeling the impact created by an particular user in a place over his/her SM channels relies on 2 components: the first one focuses on just the user's behavioral aspects and posted SM content—intrinsic component—whereas the second one takes into account the interaction with the SM network of the user—extrinsic component—.

Based on the *Differential Perception Factor* and the *Entity Engagement Index*, we define the intrinsic component:

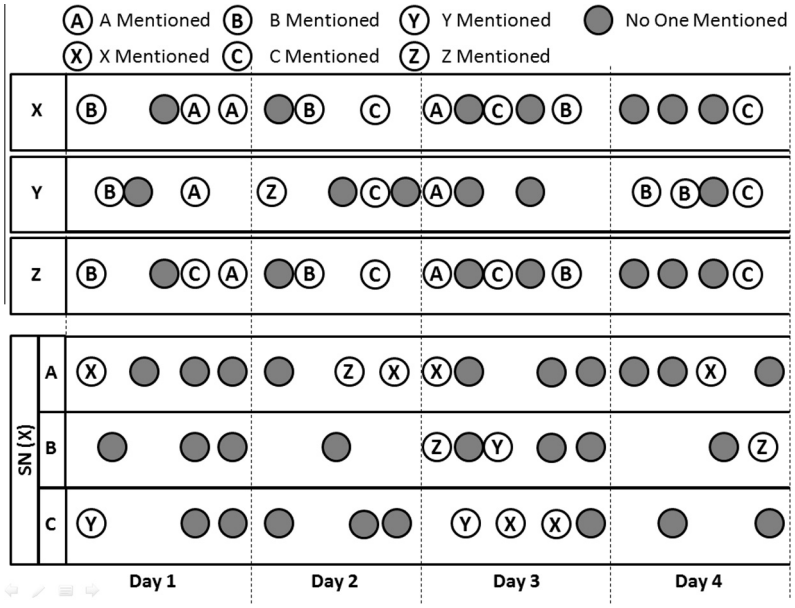

Fig. 1. Social media tie strength computing sample.

	Interactions			TieStrength						
	Total	Direct Mentions	Foreign		X	Y	Z	A	B	C
X	16	9	6	X	-	0,00	0,00	0,33	0,33	0,33
Y	14	8	2	Y	0,00	-	0,13	0,25	0,38	0,25
Z	16	9	3	Z	0,00	0,00	-	0,22	0,33	0,44
A	15	5	7	A	0,80	0,00	0,20	-	0,00	0,00
B	11	3	8	B	0,00	0,33	0,67	0,00	-	0,00
C	12	4	9	C	0,50	0,50	0,00	0,00	0,00	-

Fig. 2. Tie strength metrics based on the example in Fig. 1.

$$Intrinsic\ Impact(u_i, E, P, \Delta t) = \Im(EEI(u_i, E, \Delta t), DPF(u_i, E, P, \Delta t)) \quad (16)$$

The extrinsic component requires the joint computation of the *Exposed Users* set and the *Tie Strength*:

$$Extrinsic\ Impact(u_i, E, P, \Delta t) = \Im(ExposedUsers(u_i, E, \Delta t)), TieStrength(u_i, SN(u_i)) \quad (17)$$

Which can be implemented as an addition of the Tie Strength with u_i of all users in the exposed to the interactions created by u_i in the time period under analysis:

$$Extrinsic\ Impact(u_i, E, P, \Delta t) = \sum_{j=0}^{\#ExposedUsers(u_i, E, \Delta t)} TieStrength(u_i, u_j, \Delta t) \quad (18)$$

The resulting impact is then a combination of both intrinsic and extrinsic components:

$$Impact(u_i, E, P, \Delta t) = \Im(IntrinsicImpact(u_i, E, P, \Delta t), ExtrinsicImpact(u_i, E, P, \Delta t)) \quad (19)$$

The $\Im$ function is usually a simple product but can also be implemented in a more sophisticated way giving for instance different weights to the components.

3.6. Making the results actionable

The underlying complexity to the metrics computing might compromise the overall performance of the system, delivering highly accurate results but not quick enough to take decisions upon. Thus, we provide ways of obtaining actionable results in shorter time when the use case forces the trade-off between accuracy and time-to-results to be decided in favor of the later. Higher precision implies higher latency, which might be appropriate for batch analysis, but not meet the requirements for an early-warning fast-reaction system. The simplifications introduced in this section are designed in a way that the metrics' values inflate—increase of false positive situations—, which from the business perspective is more acceptable than the other way around (rather alerting on something that maybe is not that important than not alerting about an important situation at all).

We have identified the complexity drivers and suggested alternative ways of computing the previously defined metrics (see Fig. 3) when the time to results is more critical. Our approach to approximation for the before presented metrics is described below:

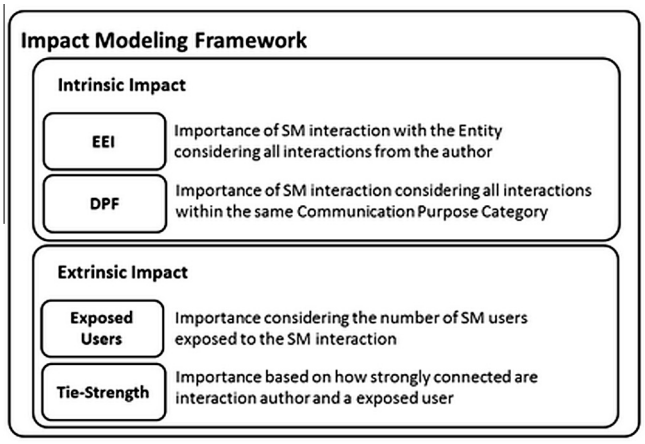

Fig. 3. Overview of the meaning of the metrics defined in our framework.

3.6.1. Intrinsic impact approximation

Removing the SM behavioral bias is *DPF's* main job, but computing it requires fairly complex time-consuming NLP operations to assign each and every SM interaction made by the SM user to the appropriate communication purpose. The *DPF* can be simplified as follows at the expend of keeping the potential SM behavioral bias:

$$DPF(u_i, E, P, \Delta t) \approx 1 \quad (20)$$

Alternatively, the system could store and return a counter for each communication purpose category for the user, under the assumption that the SM behavioral bias does not strongly change over time. It would substantially simplify the computation of the *DPF* metric (see Fig. 5).

The *EEI* requires the scanning of the latest interactions created by the user u_i and the flagging of those that are related to the Entity *E*. This step can be spared by approximating the EEI by a value specific to the user, to the location or just generic for all users. As the engagement with an Entity *E* can strongly vary driven by events of all kinds, the use of a pre-computed *EEI* value for a given user u_i based on historic data might lead to slightly less accurate results.

3.6.2. Extrinsic impact approximation

Unlike the components of the Intrinsic Impact, Exposed Users and Tie-Strength require a joint simplification, as both refer to the user authoring the SM interaction u_i and the user being exposed to it $u_j, u_j \in SN(u_i)$. Computing the set of exposed users requires extracting all interactions of the entire $SN(u_i)$ during the period Δt for further computing of the exposure window for each user in $SN(u_i)$. The Tie-Strength calculation is performed by analyzing all direct mentioned interactions and the foreign interactions for the involved pair of users, which again requires the pulling and scanning of all interactions during the period under analysis.

Both Tie-Strength and Exposure are adjusting factors of the total number of followers the user u_i has in his/her Social Network. To avoid the single computation at user level of Exposure window and Tie-Strength, we can work with predefined value distributions, dividing the Tie-Strength values range into intervals and multiplying the total number of followers of u_i by the proportion our distribution function assigns to each interval. These distributions can be based on frequency of occurrence by value interval.

$$TieStrength(u_i, \Delta t) \approx \#SN(u_i) * \sum_{k=0}^{\#D} k * d(k),\ d\ defined\ by\ D \quad (21)$$

where d is the chosen distribution consisting of $\#D$ intervals, $d(k)$ is the value associated with the interval k and $\#SN(u_i)$ is the cardinal of the followers of u_i. For example in Fig. 6, the distribution D is defined in 4 intervals $k \in 5, 10, 22, 63$ with following weights $d(5) = 1, d(10) = 0.75, d(22) = 0.5$ and $d(63) = 0.25$. In the example, $\#SN(u_i) = 135$, which results in a TieStrength of $135 * 0,2143 = 28,9305$. The TieStrength value is always bigger or than 1; if for particular privacy settings our crawler cannot retrieve the number of followers or just because a particular user does not have any follower, we assume that every user at least follows him/herself.

The same procedure can be applied for approximating the cardinal of the ExposedUsers set:

$$\#ExposedUsers(u_i, \Delta t) \approx \#SN(u_i) * \sum_{k=0}^{\#D'} k * d'(k),\ d'\ defined\ by\ D' \quad (22)$$

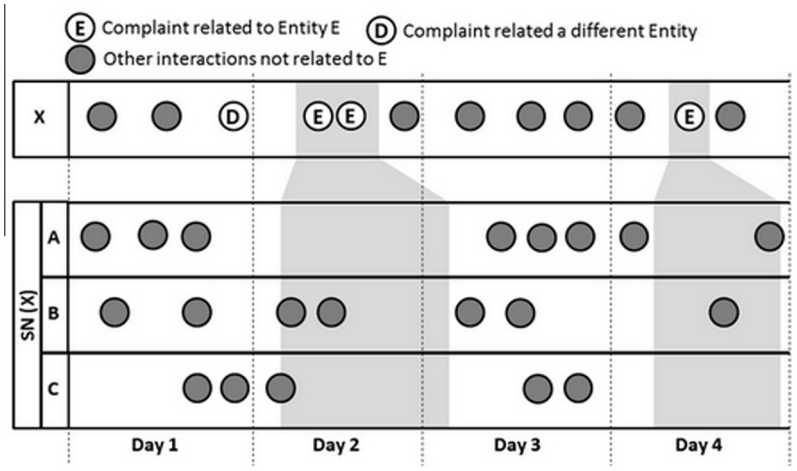

Fig. 4. Fictive social media interactions set to illustrate the impact computation.

			A	B	C
DPF (X, E, P, Δt)	0,75	*Tie Strength (X,*	0,10	0,50	0,20
	day 1	**day 2**	**day 3**	**day 4**	**Aggregated /Averaged**
ExposedUsers(X, E,	Ø	{B,C}	Ø	{A,B}	{A,B,C}
# ExposedUsers(X, E,	0	2	0	2	3
EEI(X, E,	0	0,67	0	0,33	0,25
IntrinsicImpact (X, E, P, Δt)					
Avg	0	0,35	0,00	0,30	0,1625
Sum	0	0,70	0,00	0,60	0,325
ExtrinsicImpact (X, E, P,Δt)	0	0,50	0,00	0,25	0,1875
Impact(X,E,P,Δt)					
Avg	0	0,18	0,00	0,08	0,175
Sum	0	0,35	0,00	0,15	0,125

Fig. 5. Impact metrics based on the example in Fig. 4.

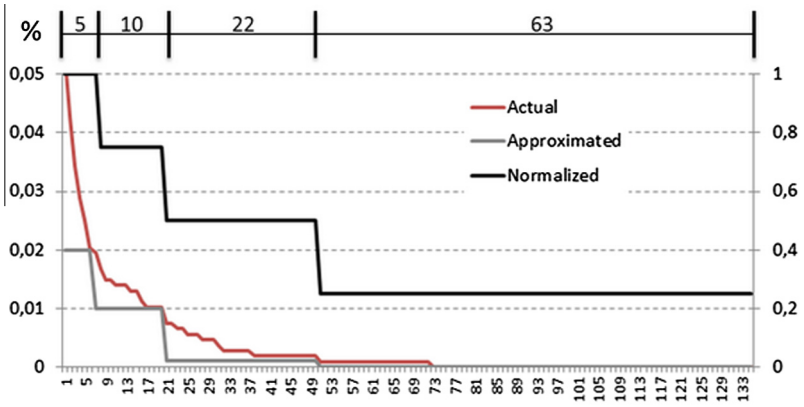

Fig. 6. Example of tie-strength approximation by a weighted distribution.

where d' is the chosen distribution consisting of $\#D'$ intervals, $d'(k)$ is the value associated with the interval k and $\#SN(u_i)$ is the cardinal of the followers of u_i.

This approach still reflects the differential link strength distribution within a given user's social network, but does not differentiate users overly tied to their networks from users almost not mentioning anybody in their tweets (just broadcasting information without ever meaning anybody). It could be achieved by introducing a weighting factor in the distribution $d(k) * w(k)$, however, computing this weighting factor requires analyzing case by case the Tie-Strength of the users u_i with each user $u_j, u_j \in SN(u_i)$. In this case, the complexity involved is similar to the one required for computing the original Tie-Strength metric. This can be achieved by account ranking or other type of graph SM ranking.

Another alternative for approaching *#ExposedUsers* is just multiplying the $\#SN(u_i)$ by a factor $\#ExposedUsers(u_i, \Delta t) \approx SN(u_i) * K,\ K \in [0,1]$, which would be consistent with the binary character we expressed in the definition of Exposed Users (subsection 3.4). This option is not the best choice for approximating the Tie-Strength, as it would no longer reflect how differentially strong the links between different users are.

3.6.3. Working with levels

The complexity of decision making for business stakeholders based on a priori resulting large numbers quantifying the impact might make the CARESOME output difficult to consume. Thus, we suggest the mapping of the impact metric values to categories, as many as different action making scenarios are defined in the use case or are meaningful for the business. Each category or level is defined by a *min* and a *max* value for the metrics. Typical category schemes are the traffic light inspired RAG (Red, Amber, Green), some Likert-inspired [60] (e.g.: *Very Strong, Strong, Medium-light, Light, No Impact*).

The suggested *Levels* might no provide similar results if defined globally. Sometimes, setting up the defining Levels max–min pairs specifically for a location soften the differences between geographical areas.

4. CARESOME system architecture

CARESOME is designed to pull the SM content generated for a set of predefined locations over time and measure the impact of all SM interaction on a defined Entity (company, institution, brand, etc). These metrics can then be used to understand when a customer retention campaign is required in a particular location, when the entity's image is damaged or weakened in the location and competitors can execute promotional actions with higher conversion chances, etc. Additionally, CARESOME offers (after proper configuration), the possibility of monitoring similar impact metrics on immediate competitors, which allows for identifying weak points on locations that can be targeted more aggressively by acquisition campaigns.

Fig. 7 shows the modules of the system as well as their input data sources and the data storage used for the information exchange between them (*Tweets Harvester, Tweets Classifier, User Data Collector, Metrics Generator*). In the following subsections we are going to describe each and every step from the data gathering to the metrics presentation stage, explaining which modules are involved and providing details about the implementation (see Fig. 8).

4.1. Tweets harvester

Relying on the Twitter Search API,[1] the component *Tweets Collector* periodically extracts all SM interactions generated in a SM location and stores them into a database for further processing. The location is typically defined as a pair of geo-location coordinates—latitude and longitude—and a radius. A pre-filtering by language can also be applied to the harvester to just pick tweets in a given language. The Geo-Gazetteer and Geo-Coder components help allocation SM interaction with missing GPS coordinates to the right areas.

4.2. Tweets classifier

The role of the classifier consists of flagging the previously gathered tweets that are related to the Entity we are analyzing on one hand, and assigning a Communication Purpose Category to these tweets on the other hand. The *Entity Flagger* component

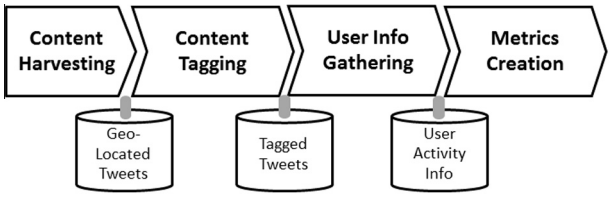

Fig. 7. System structure.

is configured by the so called *Entity Definition File*. All potential terms pointing out the relationship between the tweet content and the Entity shall be part of this file. These terms can be names of SM accounts—like the official brand account, the entity news account, the entity accounts specific to a country, etc—. For example, taking the airline *British Airways* from UK, we would have the Official British Airways Global account @BritishAirways, the account for the North American Customers Care @BristishAirways, the accounts related to official and unofficial news and press releases related to the airline @BA_Headlines and @BritishAirNews, accounts for "haters" like @We_hate_BA, etc. All relevant hash-tags shall also be included, for example the ones used by the company for running marketing campaigns (*#UnGroundedThinking*, etc.), the ones referencing the entity itself (*#BAirlines, #BritishAirlines, etc.*), the ones defined by customers to spread their lack of satisfaction (*#BAsucks*), etc. Also the name of the services offered by the company and/or name of the products—in our example, flight numbers like BA0177, etc—. Depending on the scope of the analysis, sub-brands might be also part of this file (such as @flybmi for the British Midland International airline). Additionally, typical n-grams with for example the slogan of the company or of a particular campaign, etc are included (e.g.: "Learning to fly").

As in SM due to the brevity but also due to the typing speed, the spelling mistakes are quite frequent—getting even worse with the adoption of small screen devices and the sometimes unwanted effect of the automatic spelling corrector—, our flagging component implements a tolerance threshold given by a string similarity function [61] to accept spelling mistakes (e.g.: *birtish airways* or *british airways* with a similarity over 0.7 would not be rejected if the threshold was set to 0.6).

The *Communication Purpose Flagger* works according to a similar input source (a definition file containing the terms for identifying a communication purpose category), but applies a more complex process. Each geo-located tweet is tokenized applying a sentence tokenizer first and a word tokenizer later (based on [62]) both adapting the Punkt Tokenizer [63] to deal with social media texts.The modified tokenizer provides the stop words removal as well, so that a lemmatizer takes over. The lemmatizer extracts the lemmas we then match against the input definition file. Both number and definition of categories depend on the particular business needs. For example, if there is a department specialized in handling complaints, a separate one running retention and acquisition campaigns, a third one in charge of improving the brand index, etc. makes sense understanding which communication purpose category maps to which action plan to be taken by which department and define them accordingly. In situations where a simple monitoring does the job, a sentiment-based separation of the tweets is sufficient.

Finding the defining terms for a particular communication purpose category is challenging because of the potential overlapping with another category and because of the underlying complexity in the Natural Language Processing. Analyzing previously generated user content related to specifically a given category in channels like an online Forum, a product review section, etc and performing n-gram extraction tasks [64,65] over a long history can help identifying the defining terms.

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

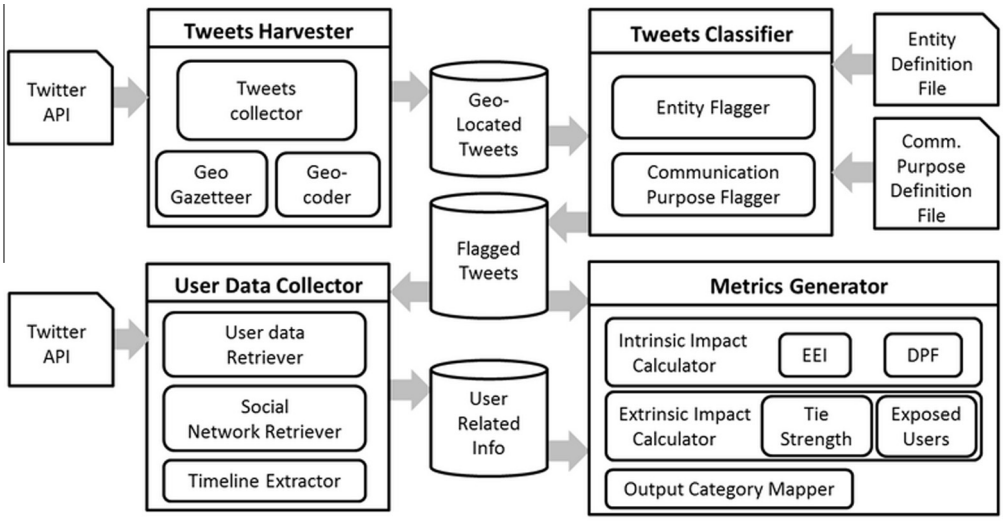

Fig. 8. CARESOME system modular architecture.

The *Classifier* module also implements a simple disambiguation mechanism relying on both Part of Speech tagging and the presence of more than one terms related to the Entity or Purpose Category. Additionally, for especial cases where the ambiguity impacts the name of the Entity *E*, a Naive-Bayesian classifier sufficiently trained helps separate senses, so that only tweets related to the Entity are flagged (e.g.: *Emirates* can be the name of the airline, but also the name of the Arsenal Stadium or even the country—UAE—).

4.3. User data collector

The purpose of this module is extracting all the information related to the authors of the flagged tweets to enable the computing of all relevant metrics. It's divided into different components, each one addressing a particular data gathering task. Each component can be also configured to apply just the data gathering required for the approximation described in the SubSection 3.6.3, instead of a more thorough yet slower data gathering required for the full-fledged metrics. The User Data Gathering Components are described below.

4.3.1. User data retriever

All relevant information about the user provided by the Twitter API, including number of followers, friends, retweets, etc are retrieved and persisted by this component. The system supports the filtering of certain accounts with a black-list mechanism, meaningful to exclude for example the company's employees accounts or the company's SM department accounts.

4.3.2. Network data retriever

To compute the Extrinsic Impact metrics, all kinds of information related to the Social Network of the author of any of the SM interactions related to the Entity of interest are required. This component extracts the entire SN framework for the identified SM authors in the previous module and persists them with a time-stamp. It allows for handling changes over time (new followers, followers leaving, etc.) which becomes especially critical when for performance reasons, the Tie-Strength is computed once per user and used going forward.

This component can also be configured to not retrieve anything if the both Tie-Strength and ExposedUsers are going to be approximated as explained in the subsection 3.6.3. The number of followers, which is the only input that is really required is made available by the previous component, as explained above.

4.3.3. Time-line extractor

Extracts the latest X tweets created by the Author, as well as the X latest tweets created by each user in the SN of the author to later enable the computing of Tie-Strength, as well as the exposure window.

4.4. Metrics generator

Once the set of required data has been gathered, processed and stored, the impact metrics are computed and transformed to be consumable in decision making scenarios. This is the purpose of this module, which relies on following three components.

4.4.1. Intrinsic impact calculator

The Entity Engagement Index and the Differential Perception Factor for the authors of the SM interactions related to the Entity under analysis flagged by the Tweets Classifier are created by this component. For all authors of the flagged tweets, the *EEI* is computed applying the formula (10). Similarly, the *DPF* is obtained as per the formula 12 for later combination of the results, as defined by the Intrinsic Impact Eq. (16).

If the system is set up to apply the approximations defined in Section 3.6, the DPF does not need to be computed (as 1 or other number close to 1 is always taken). Likewise, the EEI is taken as a configured value, which can be a generic one valid for all users or specific to a location which has been previously entered in the system and held in a look-up table (e.g.: the same for cities with a population between 50 K and 100 K). Alternatively, this module can implement a look-up table where the EEI is kept at user level from previous runs. The system can be set up in a hybrid mode, so that no approximation is done for users not present in the look-up table, but the values existing in the table are used as a kind of caching mechanism.

4.4.2. Extrinsic metric calculator

For the SM users who authored the tweets flagged as related to the Entity, both Tie-Strength and Number of Exposed Users are computed in this module. Applying the formula 15 over a set of current interactions created by the user u_i, the Tie-Strength with each and every member of $SN(u_i)$. The number of interactions

considered in this set can be fixed (e.g.: the latest 1000) or can be dependent on a timely factor (e.g.: all interaction in the last 3 months). The larger the set of interactions, the more accurate the Tie-Strength but also the higher the risk of neglecting decaying Tie-Strengths (user that used to have a very close interaction-rich relationship in the past but no longer at present). Similarly and as explained in the Section 3.4, the set of exposed users is computed with the formula (14).

If CARESOME is configured in speed modus, the approximations explained in the Subsection 3.6.2. Depending on the methods configured for the approximation, CARESOME applies the weighted distribution explained in the Eq. (21) for Tie-Strength and in the Eq. (22) for the Exposure. Both length and weight need to be configured in the system. If the decision is to keep *memory* on previously calculated Tie-Strength values between users, the system provides the lookup table to support this process.

Once the individual Impact has been computed, the Metrics Generator module computes the overall Impact the aggregation as explaining in the formula (19).

4.4.3. Output category mapper

To make the impact values per communication purpose category actionable, CARESOME provides a dynamic mapping of values to categories, whose min and max values automatically adjust based on the values distribution. The number of categories is configurable, as well as the time granularity the impact value is provided for (as explained in the Subsection 3.5). In the next section, we show concrete examples of this mapping applied (see also Figs. 17 and 15).

5. Analysis of performance and discussion

To analyze the performance of our system in action we chose 2 well-known locations with a high volume of visitors and where people are likely to have time and therefore prone to create Social Media interactions: the biggest two airports in the city of London, namely Gatwick and Heathrow. We set 2 harvesters centered in the middle point of both airports with a radius big enough (5 km) to capture all activity happening at both airports (see Fig. 9).

To increase the number of interactions retrieved by the geo-location query, we also ran 2 harvesters configured to just gather tweets with the words *Gatwick* or *Heathrow* present. Thus, we were able to gather an additional set of interactions from those users that did not have the geo-location functionality enable but referred to one of these airports.

Between the 24th of November 2013 and 23rd of January 2014, 852,319 SM interactions have been gathered. During this period of time there were severe weather conditions, spreading the chaos all over the country with strong winds and flooding episodes, which impacted the quality of all transportation services in UK. Thousands of passengers were affected and the Social Media platforms filled with users' statements on how well the different carriers handled the incident.

For our show case we took as entities a subset of the airlines operating in these airports and gathered the identifying terms (see Fig. 10).

As Communication Purpose Categories, we worked with the standard ones: *Complaints and Criticism (c)*, *Praise and Positive Feedback (p)*, *Information Request and Customer Care (ir)* and a forth one for the rest called *Neutral (n)*. The creation of the definition files for these categories has been performed by enhancing a pre-defined default file with typical terms for each category with the most frequent terms in manually flagged airline specific Tweets. Fig. 11 shows the terms per category sorted by frequency over all airlines.

For Tie-Strength and Exposure, CARESOME was configured to rely on the weighted distribution presented in the Fig. 6.

Fig. 12 shows the result of the classifier per airline and per harvester. As expected, airlines just operating from one airport present a much higher amount of interactions in this airport (e.g.: *Easyjet* showing just a few interactions in Heathrow compared with Gatwick). An interesting exception is *British Airways*, as it is used as reference in opposition to low-cost carriers in Gatwick, even if no BA flights departures from or lands there.

The adverse weather conditions on the 24th and the 25th of December left thousands of passengers stranded in the Gatwick Airport due to power problems.[2] Countless flights were canceled or suffered severe delays.[3] In this emergency situation, a blame game between Gatwick airport and the airline Easyjet started.[4] In Fig. 13, we can see the daily values for the single impact metric components, both intrinsic and extrinsic for the entity *Easyjet* and the communication purpose category *Complaints* measured by the Gatwick harvester. Especially on the 24th we observe a peak over all submetrics, motivated by the increase of SM interactions (297 different users) criticizing the way *Easyjet* handled the emergency situation. These results produced by CARESOME would have given *Easyjet* enough quantified evidence to trigger some sort of reaction and the corresponding communication back to the SM channel to palliate the incident effect. After the potential airline reaction, CARSOME can then measure the SM community response. In general, CARESOME's role is providing enough insights for a company to steer the SM dialog in all fronts.

On the Christmas eve, due to the disruptions in the railway transportation, many passengers were about to miss their flights departing from Gatwick. *[...]Ryanair uses the South Terminal, but decided to delay its services by an hour to Cork, Shannon and Dublin by an hour "To ensure all those affected by rail delays at Gatwick get home [...]".*[5] CARESOME reported a peak in the impact created by SM interactions (more than 85 in total) talking about it in the communication *Praise* purpose, with messages such as: *"I must say @Ryanair handled everything really well yesterday at gatwick #gladtobehome"*. Fig. 14 shows the peak at 9:00 am the 24th, consequence of all praise-related interactions. Ryanair delayed its flights to allow people get home for the Christmas eve and such a small decision had a huge impact over the SM channels as reported by our system, outperforming even the bad press related to service disruption. This example shows very well when finer time granularity (hourly instead of daily) makes sense and how the impact measured by CARESOME delivers meaningful results aligned with one event that happened in the real world and got reflected in the SM channels.

CARESOME can also help understanding and measuring those small things that might not be considered by the company as relevant for its customers but perceived by those as such. A captain successfully landing a plane after complicated maneuvering with adverse weather conditions might be seen as part of his job, but might also trigger a set wave of SM interactions praising the action[6] (which also contributed to the increase in the category *Praise* on the 24th Dec. in Heathrow for *British Airways* as we can see in Fig. 19). As a potential take to action, British Airways might have well created and launched a campaign to reinforce the idea of security in extreme conditions. With CARESOME, the impact of this

[2] http://www.bbc.com/news/uk-england-sussex-25503513.

[3] http://www.itv.com/news/story/2013-12-25/gatwick-airport-christmas-travel-disruption-cancellations/.

[4] http://www.dailymail.co.uk/travel/article-2535822/Blame-game-Gatwick-easyJet-clash-responsibility-Christmas-Eve-chaos.html.

[5] http://www.independent.co.uk/travel/news-and-advice/passengers-stranded-at-gatwick-airport-as-flooding-causes-power-outages-9023990.html.

[6] Video recorded by a passenger showing the captains' heroic landing https://www.youtube.com/watch?v=MPT3bdEr_VM.

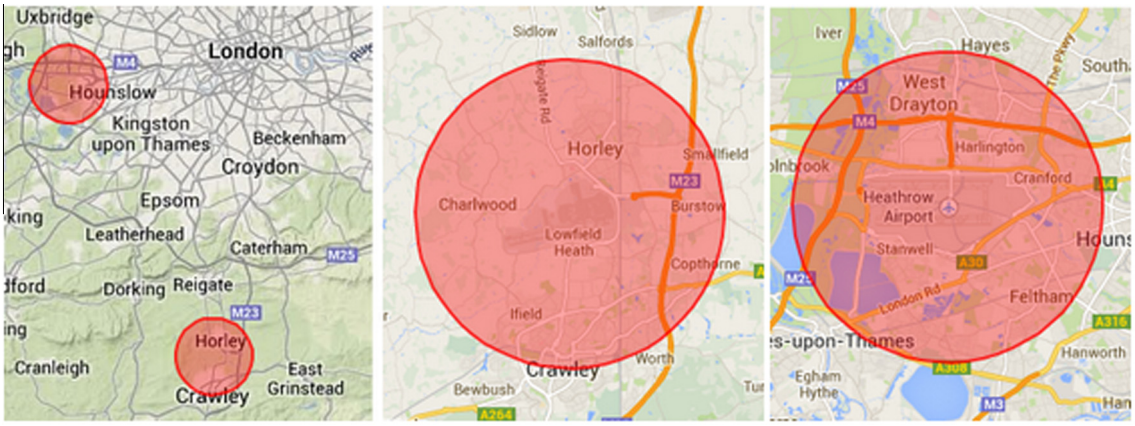

Fig. 9. Harvesters overview.

aegean	air india	british airways	etihad airways	Kenya Airways	qatar	Thomson Airways
aegean	air india	british airways	etihad airways	Kenya Airways	qatar	Thomson Airways
aegeanairlines	air lingus	British_Airways	EtihadAirways	KenyaAirways	qatarairways	transavia
aeroflot	AerLingus	BritishAirways	finnair	klm	Ryanair	transavia
Aeroflot	lingus	brussels airlines	finnair	klm	Ryanair	tunisair
aeroflot	Air New Zealand	brussels airline:	flybe	KLM_EIR	SAS	tunisair
aeromexico	Air New Zealand	FlyingBrussels	Flybe	KLM_SE	SAS	Turkish Airlines
aeromexico	FlyAirNZ	cathay pacific	germanwings	KLM_UK	singapore	turkish
air asia	alaska air	cathay pacific	german wings	KLM_US	singapore	TurkishAirlines
AirAsia	alaska air	cathaypacific	germanwings	klmfan	Singapore Air	united
air baltic	AlaskaAir	cathaypacificUS	Hawaiian Air	korean air	SingaporeAir	united airlines
airBaltic	alitalia	czech airlines	Hawaiian Air	Korean Air	south african	UnitedAirlines
air berlin	Alitalia	Czech Airlines	HawaiianAir	KoreanAir	south african	us airways
air berlin	american air	CzechAirlines	Iberia	KoreanAir_KE	swiss airlines	us airways
airberlin	American Air	delta airlines	Iberia	KoreanAir_KE	FlySWISS	USAirways
airberlin_com	AmericanAir	Delta	Iberia_en	lufthansa	TAM Airlines	virgin
airberlin_US	american airlines	DeltaBlog	Icelandair	Lufthansa_DE	tam airlines	virgin
air canada	american airlines	DeltaNewsroom	Icelandair	Lufthansa_USA	TAMAirlines	VirginAmerica
Air Canada	asiana	DeltaSkyBonus	japan airlines	Midwest Airlines	tap	VirginAtlantic
AirCanada	asiana	DeltaTechOps	japan airlines	Midwest Airlines	tap airlines	VirginAustralia
air europa	AsianaAirlines	easyjet	jetairways	MidwestAirlines	Thai Airways	vueling
Air Europa	Flyasiana	easyjet	jetairways	monarch	Thai Airways	vueling
AirEuropa	atlantic airways	easyjetservice	JetBlue	Monarch	ThaiAirwaysAust	
air france	Atlantic Airways	emirates	JetBlue	norwegian	ThaiAirwaysIT	
air france	AtlanticAir	emirates	Jetstar	Fly_Norwegian	ThaiAviation	
AirFranceFR	AtlanticJet	estonian air	Jetstar_Japan	Norwegian	Thomas Cook	
AirFranceIE	austrian airlines	estonian air	JetstarAirways	qantas	Thomas Cook	
AirFranceUK	austrian air	Estonian_Air		qantas	ThomasCookUK	
AirFranceUS				QantasAirways		
				QantasUSA		

Fig. 10. SM accounts and hashed tags used to identify the major Gatwick/Hethrow airlines.

campaign could also be measured or any other public relationship action in an ongoing basis.

Fig. 16 shows the system cockpit for *Easyjet* in Gatwick with the time granularity set to one week (in this case the week starting 22nd December).[7] In addition to the impact values for each category, the change from the previous week in percentage helps understanding whether something exceptionally changed which requires some kind of reaction by the brand. As our impact metric can deliver pretty high numbers, a heatmap-like visualization over time units allows for a quick visual identification of high-impact increases (see Fig. 15 displaying hourly values for complaints and praise over 10 days). Clicking on a particular square for a given day and a given hour displays the interactions flagged for the communication purpose category that took place there. To better makes sense of the impact values, the dashboard offers as well a calendar heat-map showing the number of interactions per hour (see Fig. 17). Apart from tailoring retention campaigns on locations where for example the impact of complaints substantially increases or keeps increasing, CARESOME can be used to monitor the perception of direct competitors in a given location to spot acquisition opportunities. Fig. 18 is a snapshot from the system front-end showing the category share per competitor for *Easyjet* in Gatwick over the 51th week of the year.

5.1. Extreme cases analysis

Analyzing how the metrics perform in extreme cases helps understanding both sensibility and suitability for real-world business scenarios. Let's assume following setup for a given user u_i over a period of time Δt:

[7] Pictures for logos and airline data have been taken directly from Twitter for the purpose of this research.

complaint			praise			info request		
term	#	share	term	#	share	term	#	share
delay	74	7,64%	great	44	14,29%	can you	35	50,72%
cancel	63	6,51%	good	38	12,34%	which	19	27,54%
no info	52	5,37%	love	29	9,42%	where	4	5,80%
late	39	4,03%	thank you	29	9,42%	do i	3	4,35%
bad	34	3,51%	nice	22	7,14%	do you	3	4,35%
miss	29	3,00%	best	21	6,82%	how do	2	2,90%
stranded	29	3,00%	lovely	16	5,19%	can i	1	1,45%
ruin	28	2,89%	impressed	15	4,87%	can u	1	1,45%
stuck	28	2,89%	better	13	4,22%	do u	1	1,45%
poor	27	2,79%	amazing	12	3,90%			
chaos	24	2,48%	awesome	12	3,90%			
fail	23	2,38%	well done	11	3,57%			
lose	18	1,86%	excellent	9	2,92%			
break	17	1,76%	cool	7	2,27%			
no staff	17	1,76%	lucky	6	1,95%			
worst	17	1,76%	favourite	4	1,30%			
problem	16	1,65%	loving	4	1,30%			
other	433	44,73%	other	16	5,19%			

Fig. 11. Top 30 terms for purposes identification.

- *All parameters maximized*: which means:
  - *DPF* is (close to) 1: the only complaint the SM user posted was about the Entity and otherwise, the user does not post any complaint.
  - The *EEI* is also 1: the user just posts about the Entity and nothing else.
  - The *Tie Strength* between the user u_i and any user in $SN(u_i)$ is also 1: all Foreign Interactions of any of the users in $SN(u_i)$ have been done by user u_i. In other words, u_i got the full attention of any user in $SN(u_i)$.
  - All users in $SN(u_i)$ has been exposed to the interactions of u_i.

In this case, the value of or Impact metric is equal to the size of the Social Network of the user u_i. $Impact(u_i, E, P, \Delta t) = \#SN(u_i)$

- *Low user entity engagement*: overly active users posting a lot of content to their SM Networks and just engaging once with an Entity E to express a complaint, ask a question, etc. are penalized over rather passive users who turn active to engage with E.
- *Overly complaining users*: when most of the interactions of user u_i belong the same Communication Purpose Category (e.g.: always complaining, always expressing a "kudos" or a "well done", his/her network tends to lower the perceived impact of the interaction). The Differential Perception Factor helps

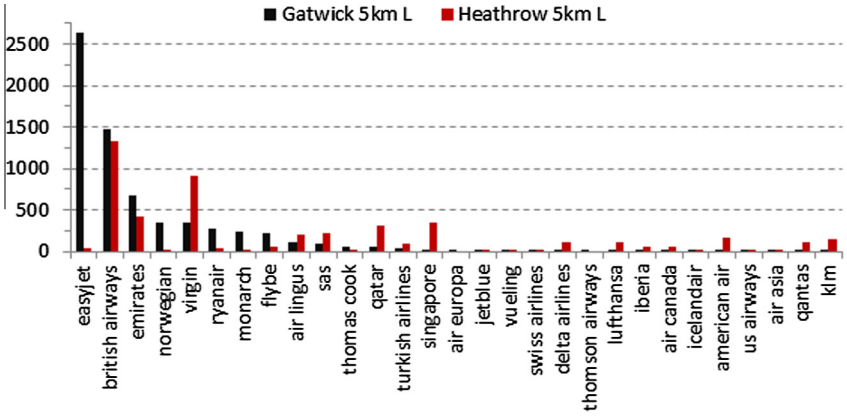

Fig. 12. Number of flagged interactions per airline and location for the top 30 airlines.

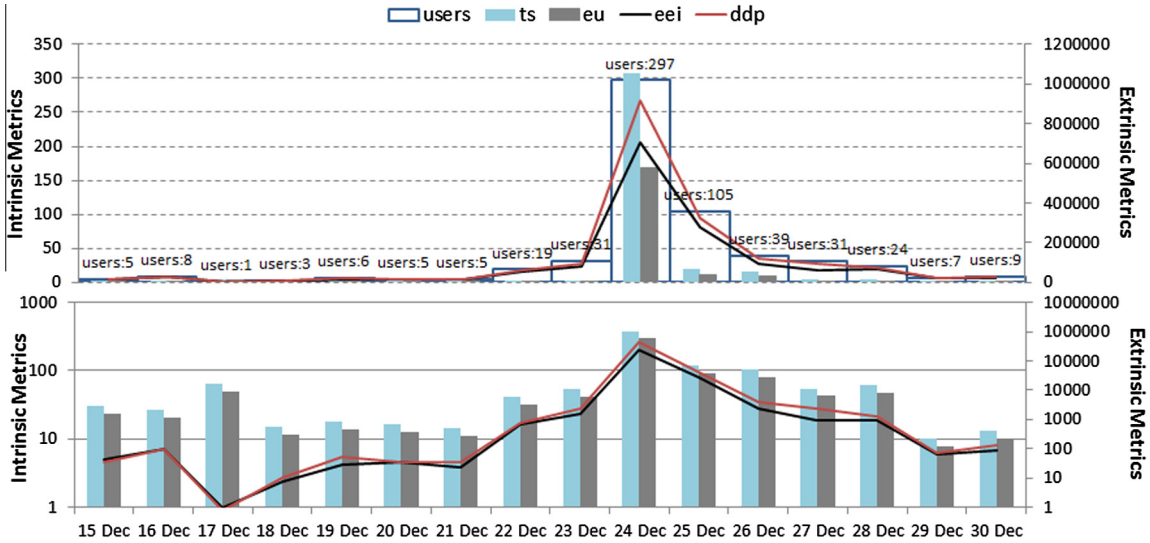

Fig. 13. Metrics components (TS, EU, EEI and DDP) as defined in Section 3.5 for Easyjet in Gatwick over 2 weeks. Natural (above) and logaritmic (below) scale.

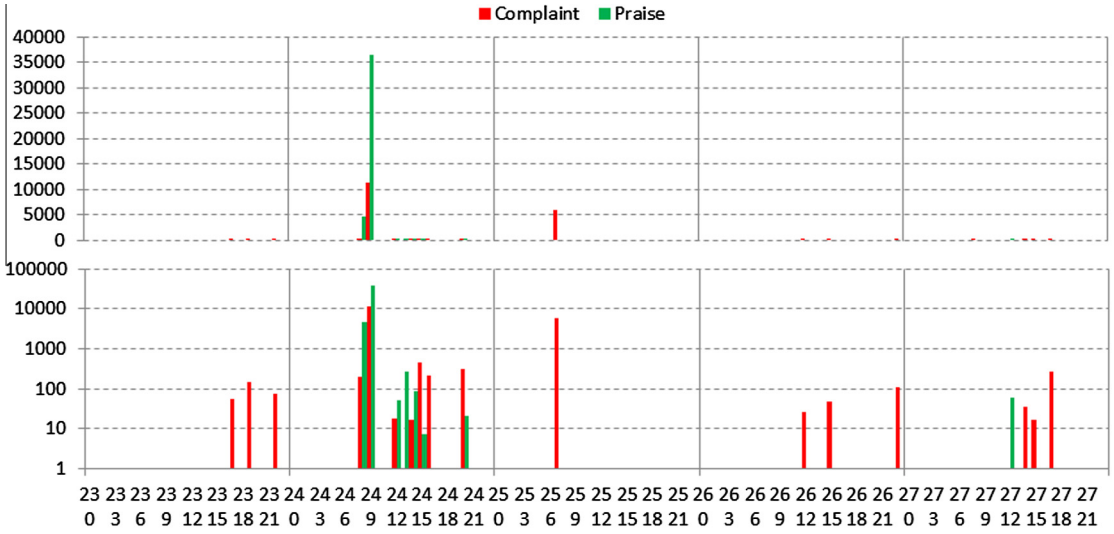

Fig. 14. Hourly impact for Ryanair in Gatwick between the 23rd and 27th December 2013.

modulating the impact metric based on the user's SM behavior; the impact of an overly complaining user posting a complaint about an Entity is lowered down according to the DPF.

- *User loosely tied to his/her network*: When all interactions where the user u_i directly mentions other users but the share of mentions in the Foreign Interactions set of all their SM Network users is very low, the posts authored by u_i are not very likely to have a great impact on his/her Network.
- *Low number of followers being rarely online*: Our impact metric rewards the users with a large network of active followers. If $\#SN(u_i)$ is low or $(ExposedUsers(u_i, E, \Delta t) \cap SN(u_i)^C$ is pretty high, the Impact metric is going to be low as well, as the number of users that can be impacted remains low.
- *Low social media activity location*: When the number of users in a location engaging with an Entity E is rather low, the impact metric can generate volatile results. In this circumstances it's advisable to extend the geographical coverage of the harvester (e.g.: increasing the radius).
- *High-activity SM location*: The design of the impact metric is not resolving overlappings in the *ExposedUsers* sets of the users behind the impact on an particular Entity E in a location. If the same SM user is part of different ExposedUsers sets, the contributions of the impacting users rather add up, which intents to reflect the combined effect on the user being exposed.

5.2. Design decisions and performance evaluation

To define our metrics and implement our system, several design decisions have been taken (e.g.: points we intentionally left unaddressed for further research, deliberate decisions against other approaches to solve punctual problems for the sake of simplicity, decisions where we opted for the most complete solution trading off simplicity for accuracy, etc.). In this subsection, we go through each decision explaining the rationale behind it and pointing out alternatives for future research.

5.2.1. Data gathering

- *Users geo-location*Our system relies on the geo-location capabilities of the Twitter Search API to periodically retrieve all the interactions of any kind created over SM channels in the specified area. A limitation is that some transactions created by users in the area are not geo-localized and cannot therefore be retrieved by a geo-query. To overcome this problem, a potential solution would be implementing a *user-place stickiness factor*, which computes based on the user's history of interactions, the likelihood of a particular interaction to be located in the area under analysis. Implementing such an approach would improve the data gathering recall.

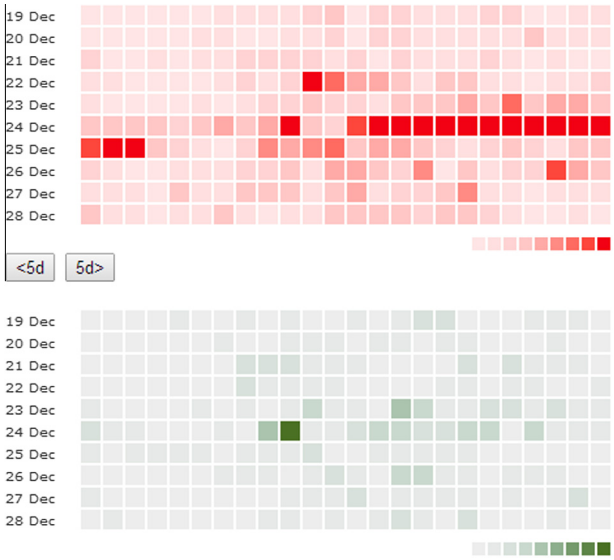

Fig. 15. 10-days hourly heatmap for all airlines for categories *Complaint* and *Praise* for the Gatwick Harvester.

5.2.2. Significance for the author

- *Quantifying engagement*We defined the *Entity Engagement Index* as a share of *Entity Related Interactions* over the overall number of interactions to measure the relevance of the *Entity Related Interaction* within the set of all interactions authored by the user.
  Additionally, the system could separate interactions initiated by the user him/herself from forwarding behaviors (re-tweets,

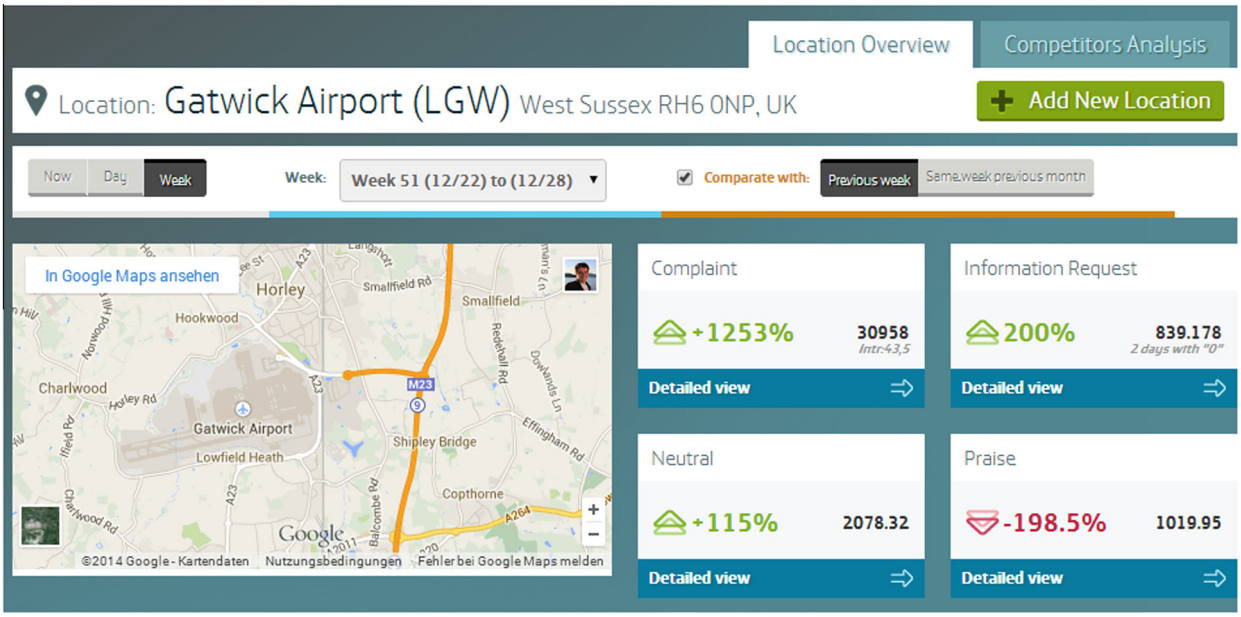

Fig. 16. Weekly dashboard view for *Easyjet* in the Gatwick location.

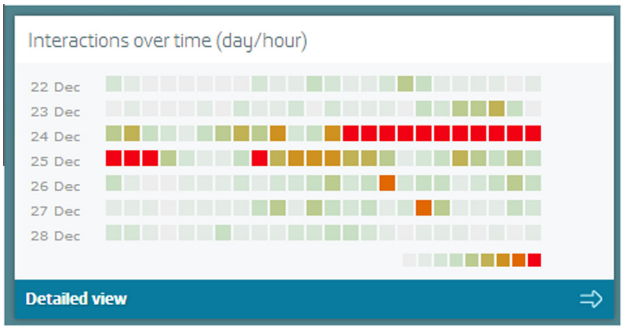

Fig. 17. Weekly social media interactions overview for Gatwick airport.

shares, etc, depending on the SM platform) to define a weighting schema based on the intensity—e.g.: a re-tweet would have lower weights than an interaction where the author is also the initiator of the conversational thread—.

A point worth researching would also be understanding the effect of taking a share over the number of SM interactions in the same industry only (e.g.: Transportation).

- *Modeling the differential perception factor*

CARESOME considers all interactions in a category to have the same weight. Enhancing the purpose tagging with tonality—e.g.: based on sentiment analysis—and comparing it to the baseline tonality for the particular author can also be use to modulate the perception factor in future works.

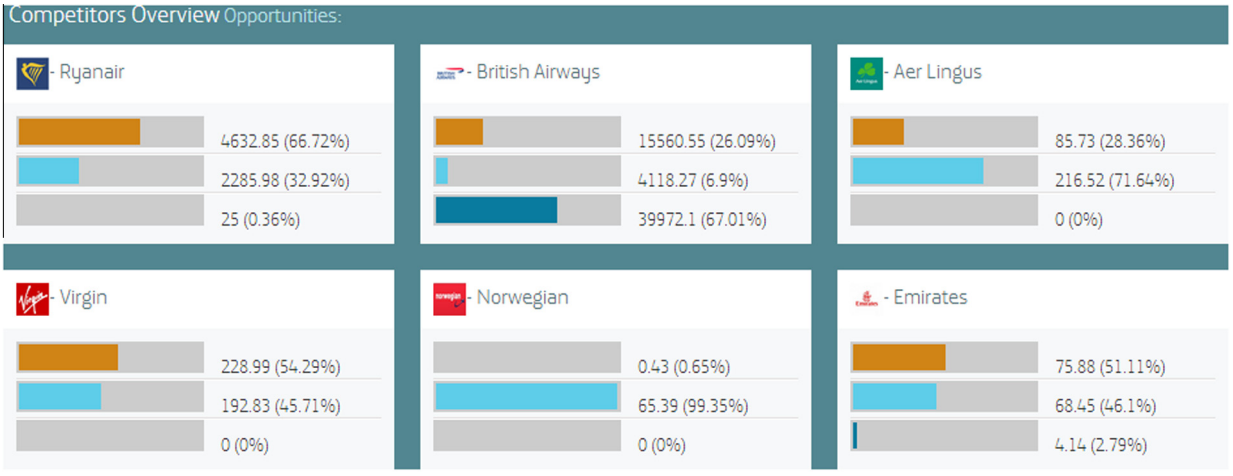

Fig. 18. Weekly Easyjet competitors' overview for Gatwick airport: C (orange), P (light blue), IR (dark blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

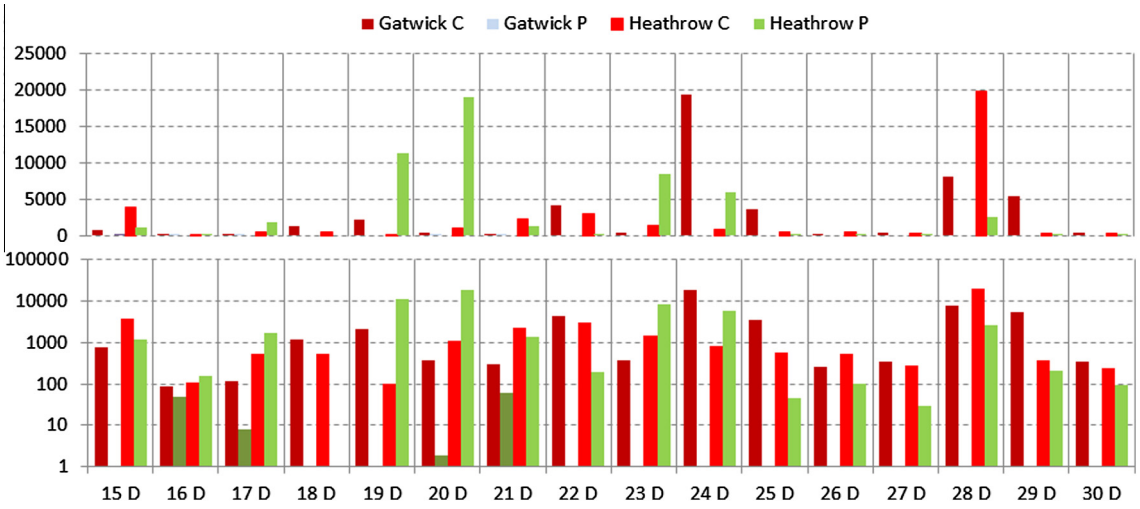

Fig. 19. 15 days monitoring of British Airways impact for complaints and praise in Gatwick and Heathrow – natural and logaritmic scale.

5.2.3. Communication category share analysis

- *Categories prevalence*
  CARESOME flags the set of harvested tweets in several waves (one per purpose category). If some terms from category A and other terms from category B have positive matching with the tweet content, a prevalence rule is triggered to decide which communication purpose category the tweet is assigned to (e.g.: the purpose with higher number of matching terms, or in case of a draw, the one defined as more dominant). The prevalence rules need to be defined and apply in the same way to all terms. In order to make these rules more effective, terms can be given a weight indicated how representative it is for the category (e.g.: for complaints, the bi-gram "no info" is less representative for "complaints" than "sucks"). Working with weights would help defining better which category the tweet should be flagged with.
- *Categories overlapping*
  Our implementation do not foresee that a given SM interaction can be flagged in 2 communication purpose categories (e.g.: "You lost my suitcase, now what? your service sucks"—*Information Request* and *Complaints*—). As the overlapping of purposes usual is, future research could implement a mechanism to address that.
- *Irony and Brand comparison*
  CARESOME does not implement any mechanism to handle irony. Taking into account all interactions from the author with the brand might help uncovering outliers (e.g.: many complaints and a punctual praise). Also situations where in a the same tweet 2 competing entities are mentioned are not handled by CARESOME at present (e.g.: "After being in a @[Entity A] flight I cannot fly @[EntityB] anymore because the service sucks!").

5.2.4. Impact on the SM network

- *Defining exposure*We worked with probability windows referred to the points of time where the user authored the interaction to engage with the Entity and the evidence that the user whose exposure is being checked has been active within this probability window. Our approach does not take into account users' activity patterns of any kind, likelihood of reading based on the time of the day where the interaction was created, etc. Exposure modeling is certainly a research line that can provide promising results with respect to certainly simplified way we have implemented it. In our system, we took the decision of defining a window, whereas more fuzzy-oriented implementations could have also been analyzed, like a decay-gradient function instead of the crisp simplification we implemented.
- *Defining tie-strength*
  Similarly, the interpretation for Tie-Strength we have implemented in our system might look simplistic. We opted for merely measuring the direct interactions—what we believe is the most defining factor—but keeping it bidirectional. Other factors might be thrown into the mix in further studies, like SM networks overlapping, Tie-Strength with common first degree connections, size of the SM network, etc.
- *Differential influence*
  Putting Tie-Strength aside, the importance of the user within the social network of the follower has not been implemented for simplicity reasons. The impact caused by the interaction of certain user on another one depends to certain extent on how important the first one is within the SM network of the later. Ranking users within a SM network requires complex modeling which led us to postpone this aspect to further analysis.
  Another possible improvement would be introducing a reputation index for SM authors, which can be taken into account in the impact computation on the SM network. Another improvement could be achieved by modeling the quality of the interaction, approach which has delivered good results in the recommender system domain [66–68].

5.2.5. Optimizing for time-2-results

- *Computing exposure*
  A good compromise between computing the Exposure at user level and applying a method to approximate it would be keeping memory creating a long-term index at user level. This long term index might also vary per user and per time of the day/day of the week, which adjusts much better to the interaction patterns in the SM world. Future work could significantly improve the approximation to the Exposure computation.
- *Tie-strength computing*
  Our suggestion to model the Tie-Strength is pretty simplistic and works reasonably well in scenarios where prompt decision taking is required. For those use cases where precision has priority over speed, the Tie-Strength can be redefined taking into consideration other factors like overlapping of SM Networks, interactions where both users are mentioned, etc.

In general, the approximation of both Tie-Strength and Exposure as explained in Section 3.6.2 is per se a research area needing exploration.

6. Conclusions

In this paper we introduced CARESOME, a system that leverages geo-located SM insights to support both customer retention and acquisition activities. CARESOME turns the SM channels into a sensor that companies can use to understanding the impact of the unfiltered feedback given by their customers and prospect customers, but also to uncover competitors' weak spots and engineer acquisition strategies targeting them. Our system relies on a framework of metrics intended to quantify what we defined as intrinsic and extrinsic impact, where we modeled the contribution of all potential factors playing a role in the impact perception, such as author's engagement with the topic, the underlying communication purpose per interaction and how the authors of these interactions are connected to other SM users.

CARESOME is designed to produce actionable insights supporting the customer facing departments of any service company. Thus, in addition to the suggested approach to compute the impact metrics, a speed modus is available, which trades accuracy against time-to-results. To make the generated insights more actionable and enable a prompter decision making, CARESOME also implements a mapping of the results to categories so that the system users do not have to deal with large, hard to compare numbers, but with simple shaded impact categories over time.

To discuss the system performance we presented a real case scenario from the travel industry and engaged into a discussion about the design decisions, indicating potential limitations and pointing at further research lines to contribute to the evolution of CARESOME, especially in the impact modeling area.

Acknowledgements

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

References

[1] T. Hennig-Thurau, C. Wiertz, F. Feldhaus, Exploring the twitter effect: an investigation of the impact of microblogging word of mouth on consumers' early adoption of new products, SSRN 2016548, 2012, pp. 1–32.
[2] A.D. Stein, M.F. Smith, R.A. Lancioni, The development and diffusion of customer relationship management (crm) intelligence in business-to-business environments, Indust. Market. Manag. 42 (6) (2013) 855–861.
[3] M. Meadows, S. Dibb, Progress in customer relationship management adoption: a cross-sector study, J. Strategic Market. 20 (4) (2012) 323–344.
[4] A. Sen, A.P. Sinha, It alignment strategies for customer relationship management, Decis. Support Syst. 51 (3) (2011) 609–619.
[5] Y. Chen, J. Xie, Online consumer review: word-of-mouth as a new element of marketing communication mix, Manag. Sci. 54 (3) (2008) 477–491.
[6] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 137–146.
[7] D. Yates, S. Paquette, Emergency knowledge management and social media technologies: a case study of the 2010 haitian earthquake, Int. J. Inform. Manag. 31 (1) (2011) 6–13.
[8] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp. 851–860.
[9] H. Gao, G. Barbier, R. Goolsby, Harnessing the crowdsourcing power of social media for disaster relief, IEEE Intell. Syst. 26 (3) (2011) 10–14.
[10] C.D. Corley, D.J. Cook, A.R. Mikler, K.P. Singh, Text and structural data mining of influenza mentions in web and social media, Int. J. Environ. Res. Public Health 7 (2) (2010) 596–615.
[11] M. Thomson, F. Doblas-Reyes, S. Mason, R. Hagedorn, S. Connor, T. Phindela, A. Morse, T. Palmer, Malaria early warnings based on seasonal climate forecasts from multi-model ensembles, Nature 439 (7076) (2006) 576–579.
[12] R. Basher, Global early warning systems for natural hazards: systematic and people-centred, Philos. Trans. Roy. Soc. A: Math. Phys. Eng. Sci. 3645 (184) (2006) 2167–2182.
[13] X. Zhang, H. Fuehres, P.A. Gloor, Predicting stock market indicators through twitter i hope it is not as bad as i fear, Proc.–Soc. Behav. Sci. 26 (2011) 55–62.
[14] B. Candelon, E.-I. Dumitrescu, C. Hurlin, How to evaluate an early-warning system: toward a unified statistical framework for assessing financial crises forecasting methods, IMF Econ. Rev. 60 (1) (2012) 75–113.
[15] J. Smailović, M. Grčar, N. Lavrač, M. Žnidaršič, Stream-based active learning for sentiment analysis in the financial domain, Inform. Sci. (2014).
[16] M. Bussière, M. Fratzscher, Towards a new early warning system of financial crises, J. Int. Money Finan. 25 (6) (2006) 953–973.
[17] C. Jiang, K. Liang, H. Chen, Y. Ding, Analyzing market performance via social media: a case study of a banking industry crisis, Sci. China Informat. Sci. 57 (5) (2014) 1–18.
[18] M. Scheffer, J. Bascompte, W.A. Brock, V. Brovkin, S.R. Carpenter, V. Dakos, H. Held, E.H. Van Nes, M. Rietkerk, G. Sugihara, Early-warning signals for critical transitions, Nature 461 (7260) (2009) 53–59.
[19] T. Hennig-Thurau, E.C. Malthouse, C. Friege, S. Gensler, L. Lobschat, A. Rangaswamy, B. Skiera, The impact of new media on customer relationships, J. Service Res. 13 (3) (2010) 311–330.
[20] C.L. Philip Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data, Inform. Sci. 275 (2014) 314–347.
[21] A.M. Kaplan, M. Haenlein, Users of the world, unite! the challenges and opportunities of social media, Business Horizons 53 (1) (2010) 59–68.
[22] W.G. Mangold, D.J. Faulds, Social media: the new hybrid element of the promotion mix, Business Horizons 52 (4) (2009) 357–365.
[23] B.J. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: tweets as electronic word of mouth, J. Am. Soc. Inform. Sci. Technol. 60 (11) (2009) 2169–2188.
[24] T. Li, G. Berens, M. de Maertelaere, Corporate twitter channels: the impact of engagement and informedness on corporate reputation, Int. J. Electron. Commerce 18 (2) (2013) 97–126.
[25] A. Java, X. Song, T. Finin, B. Tseng, Why we twitter: understanding microblogging usage and communities, in: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, ACM, 2007, pp. 56–65.
[26] Y. Kim, M.A. Ahmad, Trust, distrust and lack of confidence of users in online social media-sharing communities, Knowl.-Based Syst. 37 (2013) 438–450.
[27] J. Marés, V. Torra, On the protection of social networks users information, Knowl.-Based Syst. 49 (2013) 134–144.
[28] M.J. Culnan, P.J. McHugh, J.I. Zubillaga, How large us companies can use twitter and other social media to gain business value, MIS Quart. Exec. 9 (4) (2010) 112–115.
[29] J.M. Chan, R. Yazdanifard, How social media marketing can influence the profitability of an online company from a consumer point of view, J. Res. Market. 2 (2) (2014) 157–160.
[30] A. Rapp, L. Beitelspacher, D. Grewal, D. Hughes, Understanding social media effects across seller, retailer, and consumer interactions, J. Acad. Market. Sci. 41 (5) (2013) 547–566.
[31] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: online book reviews, J. Market. Res. 43 (3) (2006) 345–354.
[32] J. Villanueva, S. Yoo, D.M. Hanssens, The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth, J. Market. Res. 45 (1) (2008) 48–59.
[33] R.N. Bolton, A dynamic model of the duration of the customer's relationship with a continuous service provider: the role of satisfaction, Market. Sci. 17 (1) (1998) 45–65.
[34] R. Rishika, A. Kumar, R. Janakiraman, R. Bezawada, The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation, Inform. Syst. Res. 24 (1) (2013) 108–127.
[35] N. Naveed, T. Gottron, J. Kunegis, A.C. Alhadi, Bad news travel fast: a content-based analysis of interestingness on twitter, in: Proceedings of the 3rd International Web Science Conference, ACM, 2011, p. 8.
[36] J. Park, M. Cha, H. Kim, J. Jeong, Managing bad news in social media: a case study on domino's pizza crisis, in: ICWSM, 2012.
[37] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 851–860.
[38] A. Culotta, Towards detecting influenza epidemics by analyzing twitter messages, in: Proceedings of the First Workshop on Social Media Analytics, ACM, 2010, pp. 115–122.
[39] S. Middleton, L. Middleton, S. Modafferi, Real-time crisis mapping of natural disasters using social media, IEEE Intell. Syst. 99 (2014) 1.
[40] J. Yin, A. Lampert, M. Cameron, B. Robinson, R. Power, Using social media to enhance emergency situation awareness, IEEE Intell. Syst. 27 (6) (2012) 52–59.
[41] R. Colbaugh, K. Glass, Early warning analysis for social diffusion events, Security Inform. 1 (1) (2012) 1–26.
[42] N.R. Adam, B. Shafiq, R. Staffin, Spatial computing and social media in the context of disaster management, IEEE Intell. Syst. 27 (6) (2012) 90–96.
[43] C. Dellarocas, The digitization of word of mouth: promise and challenges of online feedback mechanisms, Manag. Sci. 49 (10) (2003) 1407–1424.
[44] Y. Liu, Word of mouth for movies: its dynamics and impact on box office revenue, J. Market. 70 (3) (2006) 74–89.

[45] N.I. Bruce, N.Z. Foutz, C. Kolsarici, Dynamic effectiveness of advertising and word of mouth in sequential distribution of new products, J. Market. Res. 49 (4) (2012) 469–486.
[46] E. Bothos, D. Apostolou, G. Mentzas, Using social media to predict future events with agent-based markets, IEEE Intell. Syst. 25 (6) (2010) 50–58.
[47] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp 591–600.
[48] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: bringing order to the web, in: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 1998, pp. 161–172.
[49] S. Ye, S.F. Wu, Measuring message propagation and social influence on twitter.com, in: Proceedings of the Second International Conference on Social Informatics, SocInfo'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 216–231.
[50] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring user influence in twitter: the million follower fallacy, in: ICWSM10: Proceedings of International AAAI Conference on Weblogs and Social, 2010.
[51] D.M. Romero, B. Meeder, J. Kleinberg, 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, ACM, New York, NY, USA, 2011, pp. 695–704.
[52] C.C. Yang, X. Tang, Estimating user influence in the medhelp social network, IEEE Intell. Syst. 27 (5) (2012) 44–50.
[53] P.V. Marsden, K.E. Campbell, Measuring tie strength, Social Forces 63 (2) (1984) 482–501.
[54] M. Granovetter, The strength of weak ties, Am. J. Sociol. 78 (6) (1973) l.
[55] E. Gilbert, K. Karahalios, Predicting tie strength with social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2009, pp. 211–220.
[56] C. Haythornthwaite, Tie strength and the impact of new media, in: Proceedings of the Annual Hawaii International Conference on System Sciences, 2001, p. 11.
[57] P.A. Grabowicz, J.J. Ramasco, E. Moro, J.M. Pujol, V.M. Eguiluz, Social features of online networks: the strength of intermediary ties in online social media, PloS (Pub. Library Sci.) One 7 (1) (2012) 1–27.
[58] R.K. Pan, J. Saramäki, The strength of strong ties in scientific collaboration networks, EPL (Euro Phys. Lett.) 97 (1) (2012) 18007.
[59] H. Shin, J. Lee, Impact and degree of user sociability in social media, Inform. Sci. 196 (2012) 28–46.
[60] R. Likert, A technique for the measurement of attitudes, Arch. Psychol. 22 (140) (1932) 1–55.
[61] Q.X. Yang, S.S. Yuan, L. Zhao, L. Chun, S. Peng, Faster algorithm of string comparison, Pattern Anal. Appl. 6 (2) (2003) 122–133.
[62] M. Krieger, D. Ahn, Tweetmotif: exploratory search and topic summarization for twitter, 2010.
[63] T. Kiss, J. Strunk, Unsupervised multilingual sentence boundary detection, Comput. Linguist. 32 (4) (2006) 485–525.
[64] X. Wang, A. McCallum, X. Wei, Topical n-grams: phrase and topic discovery, with an application to information retrieval, in: Seventh IEEE International Conference on Data Mining, 2007 (ICDM 2007), IEEE, 2007, pp. 697–702.
[65] B.C. Gencosman, H.C. Ozmutlu, S. Ozmutlu, Character n-gram application for automatic new topic identification, Inform. Process. Manag. 50 (6) (2014) 821–856.
[66] J. Serrano-Guerrero, E. Herrera-Viedma, J.A. Olivas, A. Cerezo, F.P. Romero, A google wave-based fuzzy recommender system to disseminate information in university digital libraries 2.0, Inform. Sci. 181 (9) (2011) 1503–1516.
[67] C. Porcel, A. Tejeda-Lorente, M. Martnez, E. Herrera-Viedma, A hybrid recommender system for the selective dissemination of research resources in a technology transfer office, Inform. Sci. 184 (1) (2012) 1–19.
[68] A. Tejeda-Lorente, C. Porcel, E. Peis, R. Sanz, E. Herrera-Viedma, A quality based recommender system to disseminate information in a university digital library, Inform. Sci. 261 (2014) 52–69.