Quality Indicators for Social Business Intelligence

Conference Paper · October 2019

DOI: 10.1109/SNAMS.2019.8931862

Quality Indicators for Social Business Intelligence

Rafael Berlanga
Department de Llenguatges i Sistemes Informàtics
Universitat Jaume I
Castelló de la Plana, Spain
berlanga@uji.es

Indira Lanza-Cruz
Department de Llenguatges i Sistemes Informàtics
Universitat Jaume I
Castelló de la Plana, Spain
lanza@uji.es

María José Aramburu
Department d'Enginyeria i Ciència dels Computadors
Universitat Jaume I
Castelló de la Plana, Spain
aramburu@uji.es

***Abstract*—The main purpose of Social Business Intelligence is to help companies in making decisions by performing multidimensional analysis of the relevant information disseminated on social networks. Although data quality is a general issue in SBI, few approaches have aimed at assessing it for any data collection, being this a context dependent task. In this paper, we define a quality indicator as a metric that serves to assess the overall quality of a collection, and that integrates the measures obtained by several quality criteria applied to filter the posts relevant for a SBI project. The selection of the best quality criteria to include in each quality indicator is a complex task that requires a deep understanding of both the context and objectives of analysis. In this paper, we propose a new methodology to design quality indicators for SBI projects whose quality criteria consider contents coherence and data provenance. Thus, for the context defined by an objective of analysis, this methodology helps users to find the quality criteria that best suit both the users and the available data, and then integrate them into a valid quality indicator.**

Keywords—data quality, business intelligence, social media, business indicators

I. Introduction

Social networks have become a new source of useful information for companies, helping them, among others, to know the opinions of their customers, to analyse the trends of their market, and to discover new business opportunities [1] [2]. Social media constitutes a fundamental part of the information ecosystem, and there has been a growing interest in the development of solutions for social data analysis from the commercial and scientific perspectives.

The main purpose of Social Business Intelligence (SBI) is to help companies in making decisions by performing multidimensional analysis of the relevant information disseminated on social networks. However, today businesses mainly use social networks to produce a group of social metrics [3] [4] that are analysed in an isolated way. Rarely, companies integrate social metrics with other business measures to calculate and analyse key performance indicators. The fundamental cause of this underutilization is the lack of trust that companies have in this kind of data, since, coming from social networks, they do not have control over their origin, contents and quality.

In an environment as open and out of control as Twitter, or any other social network, it is difficult to find the right data for a SBI project. As in any other Business Intelligence application, the subject of analysis is described, as exactly as possible, by using natural language expressions. Then, a proper group of keywords are chosen to program the retrieval mechanisms. However, as experience shows, the resulting collection use to be noisy, incomplete or biased. Most times, it will include posts generated with very different purposes, as well as posts without a true relation with the subject of analysis. Their specific contents may even be counterproductive and produce false conclusions, due to the misinformation and the noise they generate. This is a data quality problem that requires executing data cleansing operations on posts retrieved from social networks before extracting any social metrics from them. Thus, the execution of a SBI project must start by assessing the quality of the available data and, when needed, improving it until the established standard in reached [2].

Although data quality is a general issue in SBI [2] [5], few approaches have aimed at assessing this data aspect for large collections of posts. Most approaches just apply a series of ad-hoc quality rules over posts (e.g., tweets with more than three retweets, users with more than 100 followers, and so on) in order to filter those that will be analysed. In addition, many works aim at analysing concrete events such as a catastrophe or a terrorist attack where the main issue is to score tweets credibility [6] [7].

Evaluating the quality of social data is a context dependent task [8], and each SBI project will require the definition of the best quality indicators for the source data. We define a quality indicator as a metric that serves to assess the overall quality of a collection and that integrates the measures obtained by the various quality criteria. With this approach, as in [8], several quality metrics will serve to filter the posts that are relevant for a SBI project. However, the selection of the best quality criteria and metrics, and how to combine them into a quality indicator, is a complex task that requires a deep understanding of both the context and objectives of analysis.

In this paper, we propose a new methodology to design quality indicators for SBI projects. After analysing the quality attributes of tweets, we have identified that the main quality criteria for SBI analysis are contents coherence and data provenance, two aspects not treated in the literature. Thus, our methodology relies on two foundations. Firstly, considering the collection of tweets of an analysis context, it is possible to model the language of the context and measure the contents coherence of each tweet, as well as each user profile description. These measures, plus other useful metrics from users and tweets, such as the number of favourites, replies, repeats, and followers, are the main quality criteria to consider by our methodology. Secondly, a classification of the different user profile categories that participate in the context of analysis helps to identify the best quality criteria for each kind of user. In this way, given an objective of analysis in a specific context, with this methodology it is possible to find the quality criteria that best suit both its users and the available data, and then integrate them into a quality indicator valid for that SBI project.

The main idea behind this methodology is that the users with a Twitter profile description in high correlation to the subject of analysis and showing a coherent activity in their posts will have higher quality than the rest of users. These users will constitute the group of relevant users of a SBI project and will serve as reference to build quality indicators. By finding the quality criteria that produce good measures for the relevant users of an analysis context it is possible to identify the best quality metrics for that SBI project. Our method also helps to integrate these metrics in order to build effective quality indicators.

The rest of the paper is organised as follows. Section II reviews SBI approaches and methodologies. Section III specifies the analysis dimensions that are involved in the assessment of social data quality. A methodology for measuring quality in SBI projects is described in Section IV, and some results and conclusions are presented in sections V and VI respectively.

II. Social Business Intelligence Approaches

SBI is an emerging discipline that combines corporate data with content generated by users on social media with the aim of improving decision making in the company [2] [4]. Social networks can be analysed from different user perspectives such as contents, relationships and behaviour, becoming an abundant source of information on opinions, interests, needs and attitudes of users. The challenge lies in the efficient management of information from social networks considered as Big Data, characterized by an immense amount and variety of unstructured and noisy data, which change at a high speed.

This section provides a review of the current state of art of Social Business Intelligence that distinguishes between methodologies for the development of new applications and the quality assessment of social media data. It also depicts the main issues and novelties of our approach.

A. SBI Methodologies

Despite the growing interest in the development of SBI applications in the industry, in the scientific literature there have been very few approaches that establish a methodology for their design and implementation. In general, SBI requires highly integrated multidisciplinary research [9] and here we highlight the main approaches of methodologies and architectures.

The work in [10] proposes an iterative methodology for the design and maintenance of SBI applications, establishing an ordering for the common tasks executed during social media analysis. It emphasizes the need for agile and effective support during the maintenance of the infrastructure, due to the dynamism of the user generated contents and the changes in the environment. The main tasks proposed are iterative and can be optimized, they are organized as follows: macro-analysis (definition of the scope of the project and the questions it will solve), ontology design, source selection, semantic enrichment, crawling design, ETL and OLAP design, and finally, the execution and testing. In turn, the authors propose the development of an architecture where the information resulting from the analysis is stored within a data mart in the form of multidimensional cubes that can be exploited by OLAP techniques. The basic problem is that all the information is stored in a historical repository that requires large storage resources. The task of analysing the relevance of the contents generated during the crawling process relies on a manual labelling that distinguishes between in-topic and off-topic clips (textual data). In a later stage, the recovered documents are filtered based on search criteria that do not guarantee the credibility and quality of the source. On the other hand, when analysing large volumes of data this process can be unsustainable.

More in accordance with the current needs for SBI and Big Data processing, a new approach focuses on the speed and immediacy of information, processing data in streaming and incorporating batch analysis processes to obtain knowledge models. With the learned models it is possible to apply inductive processing algorithms on the data stream and to favour the semantic enrichment of the data [11] [12]. Only the information elements that are needed for the knowledge models are stored, optimizing memory usage. However, so far, there are few solutions for SBI, and analysis tasks are mainly oriented towards event detection [13] and recommendation systems [12].

The SLOD-BI project [14] is an infrastructure for linked open data that lays the foundation for the implementation of SBI. It offers mechanisms for the extraction, linking and publication of social data in the form of RDF triplets modelled as multidimensional stars. Providing access to large open knowledge resources and multidimensional analytical models to define efficient methods of data extraction and analysis, the project proposes the combination of cognitive models with statistical language models to infer useful information from the texts generated by the users. A semantic meta-structure of multidimensional analytical patterns (user facts, social facts and dimensions) is proposed to model social data in a way that facilitates the integration with corporate data stored in traditional repositories and data warehouses. In [15] SLOD-BI is extended with both a methodology and a framework oriented towards the formalization of social indicators and performance measures to support decision-making. This proposal allows social measures exploration and aggregation over dynamic multidimensional contexts for on-demand objectives.

Currently, the main challenges that SBI projects have to face are the high dynamicity of both the elements implied in the analytics and the analytical requests, as well as, the high percentage of noisy data. In this sense, [16] proposes an architecture and a methodology to model analytical streams for SBI that relies on both linked data and multidimensional modelling. The architecture eases the cleaning and semantic enrichment of data, whereas the methodology serves to shape the data for analysis purposes. The adoption of semantics facilitates the development, validation and follow-up of workflows. Thus, instead of storing the semantically enriched facts, they can be generated and processed on the fly. The solution corresponds to a Kappa streaming architecture that consists of two stages: a long-term stage for keeping recent historical information as a data stream of long duration (to collect data for inductive processing tasks) and a short-term stage with some workflows for generating the required analysis data in real-time.

All these approaches (i.e., [10] [11] [12] [14]) have in common the semantic enrichment of the social data in order to increase the processing capabilities of the collection. Entity resolution and semantic enrichment do not only provide a good context for understanding post contents, but also transform them into meaningful data with a representation easier to process by analysis tools. In this sense, semantic enrichment processes serve to enhance the quality of social data collections. However, they cannot measure the quality and filter the posts that a SBI project require.

B. *Social Data Quality Assessment*

When implementing a SBI project, it is critical to assess the quantity and quality of the available data, since without an amount large enough of valuable data, any implementation of a BI-oriented application will fail to [17] [18]. However, as shown by a recent review paper [5], few approaches to business social media analytics depict pre-analytics processing activities, and the task of assessing data quality is left out of consideration in all revised works.

In [8], the profiling and supervision of the quality of the data are considered as primary concerns of the Big Data processing cycle. With this methodology, the user is in charge of analysing the quality metrics provided by the system. In their solution, a quality management module includes an interface that enables the end-user to configure the quality policies of a company. The module also includes a quality evaluator for visualizing the values for the metrics of the quality attributes of each police. Similarly, to the quality indicators of our approach, quality policies combine several metrics to measure the quality of a collection. However, the work in [8] does not address the problem of finding and integrating the best quality metrics to assess data quality in a specific SBI project.

In spite of the great interest that the analysis of the quality of social data arouses in the scientific community, there are very few related works within the scope of SBI. The evaluation of the quality of contents published on microblogging platforms has focused mainly on post retrieval operations. Searching for posts related to a topic [19] [20]; filtering posts based on their credibility and quality [21] [22]; detection of events and disasters [13] [23] [24] [25]; analysis of feelings, political and consumer opinions [14] [26] [27]; and detection of influencers [28] [29], are some example applications. Other applications aimed at the detection of spammers, bots and advertising campaigns, have proposed intelligent analysis techniques for social metrics [30] [31] [32].

In the literature, quality measures are defined at post and user level. At post level, there are a large number of metrics covering the characteristics of the text (e.g., grammar, contents and semantics), together with the metrics specific to microblogs that reflect their social impact (e.g., number of comments and retweets). On the other hand, at user level, there are activity metrics to assess the relevance (e.g., account age and number of posts) and popularity (e.g., number of followers, likes and mentions) of issuers.

Most of the quality attributes applicable for social media data classified in [8] can easily be measured by means of the contents and metadata of Twitter posts. Among them, relevance (measured as post content keywords), believability (measured with author attributes like followers count or registration age), popularity (measured as number of readers or re-tweets) and timeliness (i.e., tweet date). Corroboration (i.e., the number of data sets where the issue has been recognized) and validity (i.e., likelihood of data validity for its purpose) are other quality attributes included in this classification. Although, during data extraction, these two attributes cannot be measured for a single data element, they are considered important when evaluating several data sets in the data analysis phase. The paper does not provide any methods to measure them.

C. *Our Approach*

In contrast to previous approaches, in our approach, we consider contents coherence together with the origin of the social data (i.e., data provenance) as fundamental quality attributes. So far, the relevance of a post was ensured when it contained one or more retrieval keywords. As it is recognised, keywords are imprecise, and using them to build a complete collection produces the retrieval of many useless posts. With our approach, by measuring the coherence of the posts contents with respect to the language model of the analysis context, it is possible to filter noisy posts and to improve the quality of the collection.

On the other hand, Twitter user profiles can constitute a source of valuable information about posts and issuers that previous approaches have not taken into consideration. From our point of view, measuring the coherence of user profile descriptions with respect to the language model of the analysis context also helps to identify the relevant posts of an SBI project. The challenge lies in developing efficient natural language processing mechanisms to identify semantic and syntactic patterns within the texts of a fully open environment such as the Twitter social network. Here, tweets are composed of very short texts, written with a style that is informal and full of hashtags and abbreviations, and produced by a large range of different kinds of users, including the noise produced by fake users and bots.

These two new quality attributes can be applied together with other commonly used metrics to assess the relevance, believability and popularity of each post and its issuer. However, the idea of integrating several quality measures into a single quality indicator is completely new. Up to our knowledge, there are not previous works that have proposed a method to define quality indicators for social data [33]. The quality policies proposed by [8] allow users to describe the relevant quality attributes and their importance for a SBI project. However, this work does not propose a method to build indicators from a set of metrics with the purpose of assessing and analysing the overall quality of a collection of posts from different perspectives (e.g. types of social network users, temporal evolution or geographical distribution). In the next sections, we present a semi-automatic method that, using a ranking of relevant users as a reference, helps to create quality indicators and social media data collections.

III. Analysis Dimensions for Quality Assessment

Some of the most powerful mechanisms for modelling analysis dimensions are the categories of customer segmentation, which vary depending on the business context. For example, demographic data and behavioural styles are usually applied to model the categories of the customer dimension. In the case of social media data, the user that writes a post can represent an individual person or a company, and it can play a role that depends on the context of analysis. Thus, our approach models the different business roles that the social media users of an SBI project can take as a way of providing a full range of perspectives of analysis.

More specifically, we propose four disjoint categories for the “user business role” dimension, named as follows: Professional, Journalist, Public Service and Lovers & Fans. The first three categories fall within the services sector, whereas the category Lovers & Fans comprises general users who use social networks by their own interest and could generate posts relevant to the context of analysis. Below, we describe these four categories and give examples of Twitter user profile descriptions associated with each category in the context of an automotive-related SBI project.

A. Professional

The "Professional" category is related to a subset of activities in the services sector. In turn, we divide this class into two subcategories.

"Professional on domain", which covers official companies, small dealers or professional groups within the analysis domain that use Twitter to promote their services. The following is an example of a professional in the automotive sector:

```
since 1921, pep boys has been the top
automotive aftermarket chain w/ quality
auto repair & car parts in 800 locations
across 35 states & puerto rico
```

"Professional others" represents professionals who do not belong to the domain of analysis. The following example shows an instance of this category:

```
sales   director   @adeogroup   providing
ecommerse  and  web  solutions  including
responsive web design to a diverse range
of clients visit {lnk}
```

B. Journalist

This category includes users dedicated to the publication of news, such us journalists, newspapers and magazines.

```
wews newschannel5 is on your side with
breaking  news  &  weather  updates  from
northeast ohio
```

C. Public service

This category represents organizations dedicated to public services such as government agencies, emergency and security services.

```
york regional police official twitter . in
case of emergency dial 911/non-emergency
call  1-866-876-5423  .  account  is  not
monitored 24/7
```

D. Lovers & Fans

This category represents people or groups that do not use social networks with a professional interest, but as a means of learning, fun and entertainment. They are usually consumers of information instead of emitters. This category represents the highest percentage of individual users in social networks and constitutes the main source of noise for many analysis contexts.

```
book  lover  ,  gamer  ,  cat  wrangler  ,
autoimmune fighter . tweets : {lnk}"
```

IV. A Methodology for Measuring Quality in SBI Projects

As in our previous work, we rely on a linked open data infrastructure where social network data is continuously stored as social BI facts, called SLOD-BI [14]. In this paper, we aim at directly deriving the quality indicators from the metrics associated to users and posts within the infrastructure. These quality indicators will be stored and used by the own infrastructure to processes and filter social data before their analysis. Figure 1 shows an example of a quality indicator derived from the post and user facts, and relying on the business role dimension.

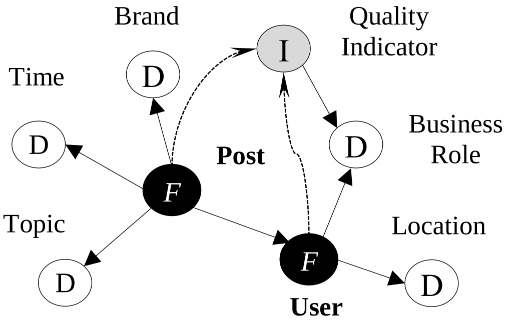

Fig. 1. Quality indicator (I) derived from Social BI facts. As in traditional BI, social facts (F) consist of metrics and dimensions (D). Metrics have been omitted in the figure.

Quality indicators must be adapted to the needs of the analysis at hand. As we adopt multi-dimensional models for social BI analysis, quality indicators can be also associated to some specific analysis dimensions. For example, in Fig. 1, the quality indicator is associated to the business role dimension, so that we can get different quality criteria depending on this perspective.

A quality indicator is formally expressed as a formula over fact metrics, which calculates the quality score of each fact according to the selected perspective. This quality score can be then applied to filter the data to be analysed.

The proposed method can be summarised in the following steps:

1) First, we establish a reference dataset of relevant users associated to the dimension of the quality indicator we aim at (e.g., business role in Figure 1).

2) Then, we select a series of metrics associated to user and post facts that are deemed as relevant indicators of quality. Some of these metrics are directly taken from the infrastructure (e.g., followers, retweets, etc.) whereas others are derived by processing both the post contents (e.g., length of messages) and the user descriptions (e.g., size of the description).

3) A quality score is calculated for each selected fact metric and each dimension perspective.

4) The formula of the quality indicator is derived by a simple linear combination of the fact metrics. The weights of the linear combination are directly obtained from the quality scores of Step 3.

5) Finally, the obtained quality indicator is applied over each incoming post/user fact to decide whether it is included in the analysis or not.

We describe in detail these steps in turn.

Step 1: Reference dataset of relevant users

In this work, we consider relevant users those that publish reliable messages according to the analysis task at hand. Useful analyses should rely on high quality data, which mainly comes from reliable users. In this way, we assume that the provenance of posts is a primary quality criterion. Another relevant criterion for quality is the coherence of the published contents and their sources. Since we cannot know all relevant users publishing in a data stream, we should rely on a small reference set of relevant users to assess the impact of existing metrics on their quality level. The main idea is then to predict a quality formula that can rank incoming user/post facts similarly to the reference set of relevant users.

For selecting relevant users we could make use of existing resources such as analytical platforms, like Socialbakers®, or existing datasets, like RepLab 2014 [34]. However, these resources are limited to a few vertical domains and mainly aim at identifying influencers from different perspectives. Indeed, users included in these resources usually have a very large number of followers. For this reason, in this paper we propose a different way to obtain reference relevant users according to their descriptions.

The proposed method consists in manually labelling the top most frequent word bigrams of the user profile descriptions. In Twitter these descriptions are included along with the tweet metadata, so this information can be easily obtained from the incoming data stream. Bigrams are labelled with the dimension values associated to the quality indicator. For example, following the schema of Figure 1, bigrams are labelled with the values of the business roles (e.g., journalist, professional and so on). Finally, user descriptions containing labelled bigrams with the same label (i.e. they have associated a unique dimension value) are considered relevant users for that label. Table I shows an example of labelled bigrams and the number of relevant users selected for each label.

Step 2: Selection of fact metrics

After analysing the different methods proposed in the literature for social data quality, we selected those metrics widely adopted for filtering posts. Then, we defined some new metrics to measure the coherence of texts in the posts and user descriptions with respect to the overall vocabulary of the stream. These metrics are organized into two levels: users and posts. For example, user quality is usually associated to metrics like number of followers, account age, and interaction metrics with other users. Posts metrics are related to metrics associated to the users' interactions with the post (e.g., likes, retweets, etc.) as well as metrics derived from the post text like sentiment indicators, semantics, and different lexical features.

Tables II and III show the selected metrics for users and posts respectively in our preliminary experiments. Concerning the user-level metrics (Table II), "interactions to post" is the number of interactions over posts performed by the user (e.g., retweets, likes, etc.), whereas "interactions from users" is the number of interactions received from other users. Description coherence is the entropy of the language model of the user description with respect to the overall language model of the stream. The lower the entropy the better correlated is the description with the domain. The account age is directly associated to the user identifier number (i.e., the lower the older). "Posts on domain" is the number of tweets published by the user in the analytical data stream. The rest of user-level metrics are provided as metadata by the incoming data stream.

With respect to the post-level metrics (Table III), "repetitions" is the number of times the same message has been published in the stream without being retweets. "Text coherence" is the entropy of the language model derived from the post text with respect to the overall language model of the domain. The next metrics are directly taken from the incoming data stream. Finally, the last four metrics are facts and sentiment indicators extracted from the post text [14].

Step 3: Impact of metrics

To measure the impact of a metric in the data quality, we propose a novel method based on Information Retrieval evaluation. Our hypothesis is that a metric will have a high impact in data quality if the ranking of items produced by that metric promotes the relevant users associated to a given dimension.

A direct way to measure this impact is to use de Mean Average Precision (MAP) metric [35]. A high MAP value indicates that relevant items are mostly placed at the top positions of the ranking. However, it heavily depends on the number of relevant items (the larger this number the higher the scores). As we use different datasets of relevant users for each dimension, we need to normalize MAP scores. In this work we propose to use the relative change with respect to a uniform distribution of relevant items.

$$MAP = \frac{\sum_{k=0}^{N} pre(k) \cdot rel(k)}{R}$$

$$MAP_{relative} = \frac{(MAP - N/R)}{MAP}$$

Where *N* is the total number of items in the ranking and *R* is the number of relevant items. Function *pre* returns the precision at position *k* in the ranking. Function *rel* returns 1 if the item at position *k* is relevant and 0 otherwise.

Notice that a value of $MAP_{relative}$ near to 0 indicates a low impact of the metric to rank relevant items. Notice also that large negative values indicate that the ranking should be reversed to have a positive impact in the promotion of relevant items.

Since metrics can be associated to either users or posts, depending on the target metric we will rank either users or posts to calculate the corresponding MAP score. Additionally, a MAP score can be calculated for each dimension value (labels) by considering only the corresponding subset of relevant items associated to it.

Step 4: Quality Indicators

The last step consists in applying the quality scores of each metric to obtain the corresponding quality indicators. In this paper we propose a simple linear combination of the metrics where impact scores act as weights. More specifically, given a fact *F* with metrics (m_1, ... , m_M), the quality indicator for dimension value *Di* is the following formula:

$$I_{Di}(F) = \sum_{k=0}^{M} \alpha_k^{Di} \cdot norm(F[m_k])$$

Here, the quality scores of each metric and dimension value are represented by the coefficients α_k^{Di}. To properly combine the metrics, we normalize and scale them with the function *norm* as described in turn.

As previously mentioned, when the quality factor of a metric is negative, the ranking must be inverted. For this purpose, we apply the complement with respect to the maximum value of the metric (i.e., max(m_k) - F[m_k]). Additionally, we must multiply by -1 to make the corresponding factor positive.

The function *norm* also applies a logarithm transformation to all the metrics that counts things (e.g., retweets, statuses, etc.), since all these metrics follow a power law distribution. Finally, all metrics are normalized in the range [0, 100].

Step 5: Quality Data Filtering

Quality indicators can be applied in different ways to filter the incoming posts before analysing them. The most straightforward way is to firstly apply a threshold over the indicators at user-level, and then a second threshold over the indicators at post-level. Additionally, we can select one of the perspectives of the dimension associated to the quality indicator (e.g., the business role to be focused on).

V. RESULTS

For demonstrating the usefulness of the previous metrics, we have chosen a long-term stream of tweets related to the automotive domain. This stream has been active from 2015 until now and has served as basis of several studies about Social BI [14] [16]. The stream is generated with a series of keywords representing different car models and brands. It currently contains 1.930.617 tweets, written in both Spanish (456.059) and English (1.474.558).

Table I shows the different categories contained in the user dimension for performing different analysis tasks. The second row in this table indicates the number of relevant users for each class. An external indicator of the relevance of these users is the ratio of verified accounts, which is of 0.9% for the whole stream and 3.3% for the selected relevant users.

TABLE I. EXAMPLES OF TOP FREQUENT BIGRAMS ASSIGNED TO THE DIFFERENT USER CATEGORIES

Journalists (J)	·news information ·motoring news ·auto news
5173	
Lovers & Fans (L&F)	·sports fanatic ·love cars ·auto enthusiast
4572	
Professional on domain (P.D)	·used cars ·cars service ·car parts
4441	
Professional others (P.O)	·community manager ·project manager ·writer photographer
3286	
Public Service (PS)	·call emergency ·crime call ·report call
193	

Table II shows the impact weights obtained for the different user categories by applying the MAP scores and the relevant users associated to them. In this table, MAP scores have been transformed into normalized weights. Cells marked with asterisks correspond to negative MAP scores (see Step 4), and therefore their metrics should be inverted when contributing to the indicator. As can be noticed, different categories show quite different weights which leads to different quality indicators. Notice also that most categories promote the coherence of the user description with respect to the domain.

TABLE II. USER-LEVEL QUALITY SCORES

METRICS	**User business roles**				
	J	***L & F***	***P.D***	***P.O***	***PS***
Interactions to posts	0.01	0.09	0.09	0.05	0.04
Interactions from users	0.09	0.07*	0.12	0.04	**0.26**
Description coherence	0.16	**0.23**	**0.24**	**0.29**	0.11
Account age	0.05	0.13*	0.18	0.02	0.04
Posts on domain	0.19	0.08	0.07*	0.10	0.22
Followers	0.12	0.18	0.04*	0.17	0.04*
Friends	**0.24**	0.17	0.07*	0.24	0.20
Listed users	0.13	0.06	0.18*	0.10*	0.08*
Published posts	0.01	0.09	0.09	0.05	0.04

Table III shows the impact weights for the post-level metrics. We use the same conventions than in Table II. As for the user-level metrics, the different categories show different weight configurations. It can be noticed that the most relevant metrics correspond to relevant users of each perspective. Thus, journalist and public services promote tweets with a high number of interactions, whereas professional categories promote tweets with sentiment data. Notice also that automotive-related professionals (P.D) have a greater weight in the text coherence metric than other professionals (P.O).

TABLE III. POST-LEVEL QUALITY SCORES

METRICS	**User business roles**				
	J	***L & F***	***P.D***	***P.O***	***PS***
Repetitions	0.06*	0.02*	0.15*	0.03*	0.12*
Text Coherence	0.06*	0.09*	**0.21**	0.03*	**0.16***
Replies	0.02	0.10	0.01*	0.09	0.01*
Retweets	**0.21**	0.08	0.02*	0.09	**0.18**
Favourites	**0.22**	0.11	0.05*	0.09	0.14
Sentiment Score	0.04	0.12	**0.21**	**0.24**	0.14*
Sentiment facts	0.13	0.15	0.05	0.15	0.10
Opinion expressions	0.14	**0.17**	0.16	**0.16**	0.10
Feature expressions	0.12	**0.16**	0.13	0.12	0.06

Finally, once quality indicators are built, we need to set the thresholds that will be applied for data quality filtering. If no perspective is selected, we assume that the thresholds are applied to the best-scored perspective at user level. As previously mentioned, metrics are normalized in the range [0, 100] and therefore final indicators will be also in that range.

When applying the best-scored perspective, the very dominant role is "Professional on domain", which mainly covers all the advertisements and promotions generated by agents of the automotive business. Threshold setting is performed by exhaustive exploration, looking for a trade-off between relevant user coverage and the size of the filtered dataset.

Alternatively, we can select the perspective that best fits our analysis. For example, we can select the journalist perspective if we want to analyse the events associated to the different brands. Threshold setting for each perspective can be different, as shown in Table IV. In this table, we fix the filter ratio to 20% to find out their corresponding thresholds. Journalist role gets the lowest thresholds, which indicates that quality scores generated for this perspective are usually lower than in the other ones because its profile is much more difficult to cope with.

TABLE IV. THRESOLD SETTINGS PER PERSPECTIVE

	J	L&F	P.D	P.O	PS
User-level threshold	>25	>25	>40	>25	>20
Post-level threshold	>5	>20	>20	>20	>20

[a.] filtering ratio fixed to 20%

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel methodology to semi-automatically define quality indicators for social BI. It relies on the selection of a ranking of relevant users associated to specific analysis dimensions. Taking this ranking as reference, the method automatically calculates the impact of each metric in the quality indicators applied to filter social data before analysis.

Preliminary results show the viability of the approach for a large stream of tweets in the automotive domain. The resulting quality indicators show that there is not a unique formula for all the analysis tasks but different configurations depending on the kind of contents we want to analyse.

Future work will be mainly focused on a comprehensive evaluation of existing social data metrics to make the obtained quality indicators much more effective. We will also analyse semantic-based methods to automate as much as possible the selection of relevant users for specific dimensions. Finally, further experiments must be done to measure the noise reduction produced by the proposed quality indicators.

ACKNOWLEDGMENT

This research has been funded by the Spanish Ministry of Industry and Commerce grant number TIN2017-88805-R and by the pre-doctoral grant of the Universitat Jaume I with reference PREDOC/2017/28.

REFERENCES

[1] Akter, S., Bhattacharya, M., Fosso Wamba, S. and Aditya, S. “How does Social Media Analytics Create Value?” in Journal of Organizational and End User Computing 28, 1-9, 2016.

[2] Ruhi, U. “Social Media Analytics as a BI Practice: Current Landscape and Future Prospects” in Journal of Internet Social Networking and Virtual Communities. 1-12, 2014.

[3] Keegan, B. and Rowley, J. “Evaluation and decision-making in social media marketing” in Management Decision. 55, pp. 15-31, 2017.

[4] Lee, I. “Social media analytics for enterprises: Typology, methods, and processes” in Business Horizons. 61.2. 199-210, , 2018.

[5] Holsapple, C., Hsiao, S. and Pakath, R. “Business social media analytics: Characterization and conceptual framework” in Decision Support Systems. 110, 2018.

[6] Gupta, A., Kumaraguru, P., Castillo, C. and Meier, P. “TweetCred: Real-Time Credibility Assessment of Content on Twitter” in Proceedings of the 6th International Conference on Social Informatics. 228-243, 2014.

[7] O’Donovan, J., Kang, B., Meyer, G., Höllerer, T. and Adalii, S. “Credibility in Context: An Analysis of Feature Distributions in Twitter” in International Conference on Privacy, Security, Risk and Trust, pp. 293-301, 2012.

[8] Immonen, A., Pääkkönen, P. and Ovaska, E. “Evaluating the Quality of Social Media Data in Big Data Architecture” in IEEE Access. 3, 2015.

[9] Zeng, D., Chen, H. and Lusch, R. “Social Media Analytics and Intelligence” in IEEE Intelligent Systems, Vol. 25, n. 6, pp. 13-16, 2010.

[10] Francia, M. Golfarelli, M.and Rizzi. S. “A methodology for social BI” in Proceedings of the 18th International Database Engineering & Applications Symposium, ACM, 2014.

[11] Barbieri D. et. Al., “Deductive and Inductive Stream Reasoning for Semantic Social Media Analytics” in IEEE Intelligent Systems, pp. 32-41, 2010.

[12] Nadal, S. et al. “A software reference architecture for semantic-aware Big Data systems” in Information & Software Technology 90, pp. 75-92, 2017.

[13] Zhang, C. et al. “GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams”, In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, 2016.

[14] Berlanga, R. et al. “SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence” in International Journal on Data Warehousing and Data Mining, vol. 11, 4, pp. 1-28, 2015.

[15] Lanza-Cruz, I. and Berlanga, R. “Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter” in the 10th International Joint Conference on Knowledge Discovery and Information Retrieval, pp. 219-226, 2018.

[16] Lanza-Cruz, I., Berlanga, R. and Aramburu, MJ. “Modeling Analytical Streams for Social Business Intelligence” in Informatics 5(3), 2018.

[17] Czernek, A. “Social Measurement Depends on Data Quantity and Quality” in Millward Brown Dynamic Logic, 2018. http://www.millwardbrown.com/docs/default-source/insight-documents/points-of-view/Millward_Brown_POV_Social_Measurement_Depends_on_Data_Quantity_and_Quality.pdf (accessed 5 September 2019).

[18] Inmon, B. “Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump”. Technics Publications, 2016.

[19] Massoudi, K., Tsagkias, M., de Rijke, M. and Weerkamp, W. “Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts” in Advances in Information Retrieval - 33rd European Conference on IR Research, pp. 18-21, 2011.

[20] Xie, W., Zhu, F., Jiang, . Lim, P. and Wang, K. “TopicSketch: Real-Time Bursty Topic Detection from Twitter” in IEEE Transactions on Knowledge and Data Engineering , pp. 2216 - 2229, 2016.

[21] Momeni , E., Tao K. and Haslhofer, B. “Identification of Useful User Comments in Social Media: A Case Study on Flickr Commons” in Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 1-10, 2013.

[22] Chen, W. et al. “A study on real-time low-quality content detection on Twitter from the users’perspective” in PLoS ONE, vol. 12, 8, https://doi.org/10.1371/journal.pone.0182487, 2017.

[23] Feng, W. et al. “STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream” in Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, pp. 13–17, 2015.

[24] Zhou, X. and Chen, L. “Event detection over twitter social media streams” in The VLDB Journal, pp. 381–400, 2014.

[25] Zubiaga, A., Spina, D., Martínez R. and Fresno, V. “Real - time classification of Twitter trends” in Journal of the Association for Information Science and Technology, pp. 462‑473, 2015.

[26] Liu, X. et al. “A Text Cube Approach to Human, Social and Cultural Behavior in the Twitter Stream,” in International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, 2013.

[27] Rosenthal, S. Farra, N. and Nakov, P. “Sentiment Analysis in Twitter” in Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), Vancouver, Canada, 2017.

[28] Rodríguez - Vidal, J., Gonzalo, J., Plaza, L. and Anaya, H. “ Automatic detection of influencers in social networks: Authority versus domain signals” in Journal of the Association for Information Science and Technology, vol. 70,7, pp. 675-684, 2019.

[29] Mahalakshmi, G. S., Koquilamballe, K. and Sendhilkumar, S. “Influential Detection in Twitter Using Tweet Quality Analysis” in Second International Conference on Recent Trends and Challenges in Computational Models, pp. 315-319, 2017.

[30] Miller, Z. et al. “Twitter spammer detection using data stream clustering” in Information Sciences, pp. 64-73, 2014.

[31] Varol, O. et al. “Online Human-Bot Interactions: Detection, Estimation, and Characterization” in Social and Information Networks, 2017.

[32] Li, H. et al. “Detecting Campaign Promoters on Twitter using Markov Random Fields” in theIEEE International Conference on Data Mining, Shenzhen, 2014.

[33] Taleb, I., Serhani, M. and Dssouli, R. “Big data quality: A survey” in Proc. IEEE Int. Congr. Big Data, pp. 166–173, 2018.

[34] Amigó, E. et al. “Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management” in Kanoulas E. et al. (eds) Information Access Evaluation. Multilinguality, Multimodality, and Interaction, 2014.

[35] Manning, C., Raghavan, P. and Schütze, H. “Introduction to Information Retrieval”. Cambridge University Press, 2008.