Journal of Business Research xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Business Research

Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach

Sérgio Moro [a,b,*], Paulo Rita [a], Bernardo Vala [c,1]

[a] Business Research Unit, ISCTE–University Institute of Lisbon, Portugal
[b] ALGORITMI Research Centre, University of Minho, Portugal
[c] ISCTE Business School, ISCTE–University Institute of Lisbon, Portugal

ARTICLE INFO

Article history:
Received 29 September 2015
Received in revised form 11 February 2016
Accepted 15 February 2016
Available online xxxx

Keywords:
Social networks
Social media
Data mining
Knowledge extraction
Sensitivity analysis
Brand building

ABSTRACT

This study presents a research approach using data mining for predicting the performance metrics of posts published in brands' Facebook pages. Twelve posts' performance metrics extracted from a cosmetic company's page including 790 publications were modeled, with the two best results achieving a mean absolute percentage error of around 27%. One of them, the "Lifetime Post Consumers" model, was assessed using sensitivity analysis to understand how each of the seven input features influenced it (category, page total likes, type, month, hour, weekday, paid). The type of content was considered the most relevant feature for the model, with a relevance of 36%. A status post captures around twice the attention of the remaining three types (link, photo, video). We have drawn a decision process flow from the "Lifetime Post Consumers" model, which by complementing the sensitivity analysis information may be used to support manager's decisions on whether to publish a post.

1. Introduction

The worldwide dissemination of social media was triggered by the exponential growth of Internet users, leading to a completely new environment for customers to exchange ideas and feedback about products and services (Kaplan and Haenlein, 2010). According to Statista Dossier (2014), the number of social network users will increase from 0.97 billion to 2.44 billion users in 2018, predicting an increase around 300% in 8 years. Considering its rapid development, social media may become the most important media channel for brands to reach their clients in the near future (Mangold & Faulds, 2009; Korschun and Du, 2013).

Companies soon realized the potential of using Internet-based social networks to influence customers, incorporating social media marketing communication in their strategies for leveraging their businesses. Measuring the impact of advertisement is an important issue to be included in a global social media strategy (Lariscy et al., 2009). Several studies focused on finding the relationships between online publications on social networks and the impact of such publications measured by users' interactions (e.g., Cvijikj et al., 2011). However, fewer studies devoted attention to research for implementing predictive systems that can effectively be used to predict the evolution of a post prior to its publication. A system able to predict the impact of individual published posts can provide a valuable advantage when deciding to communicate through social media, tailoring the promotion of products and services. Advertising managers could make judged decisions on the receptiveness of the posts published, thus aligning strategies toward optimizing the impact of posts, benefiting from the predictions made. Also, it has been shown that social media publications are highly related to brand building (Edosomwan et al., 2011). Therefore, the predictive tool outlined in this paper could leverage managerial decisions to improve brand recognition.

Data mining provides an interesting approach for extracting predictive knowledge from raw data (Turban et al., 2011). Its application to social media has been studied, especially for evaluating market trends from users' inputs (e.g., Trainor et al., 2014). However, most of the studies focused on a reactive evaluation of what users are saying through the network, with an emphasis on gathering information from different network groups or even personal posts (e.g., Bianchi and Andrews, 2015). We focused on predicting the impact of publishing individual posts on a social media network company's page. The impact is measured through several available metrics related to customer visualizations and interactions. The predictive knowledge found enables to support manager's decisions on whether to publish each post.

For validating the taken procedure, we addressed a worldwide cosmetics company with a renowned brand, including 790 posts published by this company in the year of 2014 in its Facebook social network

* Corresponding author at: Business Research Unit, ISCTE–University Institute of Lisbon, Lisbon, Portugal. Tel.: +351 21 790 30 24, +351 25 351 03 09.
E-mail address: scmoro@gmail.com (S. Moro).
[1] Tel.: +351 21 790 30 24

http://dx.doi.org/10.1016/j.jbusres.2016.02.010
0148-2963/

Table 1
Features from the compiled data set

Feature	Type of information	Source	Data type
Posted	Identification	Facebook	Date/time
Permanent link	Identification	Facebook	Text
Post ID			
Post message	Content	Facebook	Text
Type	Categorization	Facebook	Factor: {Link, Photo, Status, Video }
Category	Categorization	Facebook page managers	Factor: {action, product, inspiration }
Paid	Categorization	Facebook	Factor: {yes, no }
Page total likes	Performance	Facebook	Numeric
Lifetime post total reach			
Lifetime post total impressions			
Lifetime engaged users			
Lifetime post consumers			
Lifetime post consumptions			
Lifetime post impressions by people who have liked your page			
Lifetime post reach by people who like your page			
Lifetime people who have liked your page and engaged with your post			
Comments	Performance	Facebook	Numeric
Likes			
Shares			
Total interactions	Performance	Computed	Numeric

brand page. Therefore, this data set of posts is used as an input to the data mining procedure.

The main goals of this study are as follows:

- Implementing a model that predicts the impact of posts using their characteristics
- Measuring the predictive value of the model when applied to several output metric features, i.e., by evaluating the difference between the value predicted by the model and the real metric value
- Assessing the knowledge provided by the model in terms of which input features affect the impact metrics and how these input features influence each post, and hence supporting managers' decisions
- Defining a causal relation between the knowledge found and brand building by relating the influence of the input features and the impact of the posts on customers, and hypothesizing on how such metrics can effectively contribute to brand recognition

The next section describes the materials used (e.g., the input data set) as well as the methods chosen for the experiments. Section 3 is focused on providing specific background on the technical aspects of the data mining procedure, including prediction modeling and knowledge extraction. Such section also highlights the main motivations for evaluating post impact and how it affects brand building based on current theories. Sections 4 and 5 exhibit the results achieved and discuss them in the light of brand building through improving the value created by each post. Finally, the last section draws the main conclusions of this investigation.

2. Materials and methods

2.1. Data set

The proposed approach includes a data mining experimental method at its core, resulting in a data-driven procedure. Therefore, we needed to collect a representative data set of published posts. All the posts published between the 1st of January and the 31th of December of 2014 in the Facebook's page of a worldwide renowned cosmetic brand were included. As a result, the data set contained a total of 790 posts published. It should be noted that Facebook is the most used social network with an average of 1.28 billion monthly active users in 2014, followed by Youtube with 1 billion and Google + with 540 million (Insights, 2014).

The data set compiled contained four types of features:

- Identification—features that allow identifying each individual post
- Content—the textual content of the post
- Categorization—features that characterize the post
- Performance—metrics for measuring the impact of the post (or the impact of the page, in the case of "Page total likes")

Table 1 displays each of the features collected in the data set. Most of the information was exported directly from the company's Facebook page. The exceptions were "total interactions" and "category." The former represents a column computed based on the performance metrics exported from Facebook: it is the sum of the number of comments, likes, and shares of the post. The latter is the only column created manually by the Facebook page managers. This categorization was a request from the company's senior marketing managers as it relates to the types of campaigns performed by this specific cosmetic company. It provides a manual categorization according to the campaign to which the content posted is associated. For minimizing the risk of misclassification due to typing error for being a manual procedure, another experienced professional in social media within the company validated this categorization for all the 790 posts.

The performance metrics collected characterized posts' performance in several aspects. Some of them were intuitively derived from interactions with posts, such as the number of comments, likes, and shares of the post. The "page total likes" measures the number of likes the page had when the post was published. The remaining metrics are not so intuitive. Fig. 1 shows a concept map for understanding the concepts underlined in each performance metric. These can be logically divided in visualizations and interactions. The former, named "impressions," are based on counting the number of times the post was loaded onto the user's browser, whether directly (organic reach) or through another user's interaction (viral reach). The latter, "engagements," account for all the types and origins of clicks on the post. Considering that engagements define explicit user actions on the post, these constitute a stronger measure for user feedback on the post when compared to impressions, since loading the contents on the browser does not actually means the user has paid attention to it. The only research published we found using these Facebook performance metrics is the study by Smyser et al. (2014) using impressions and interactions for measuring the performance of a campaign for tobacco control. However, such paper did not include any predictive computation for unveiling the

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010

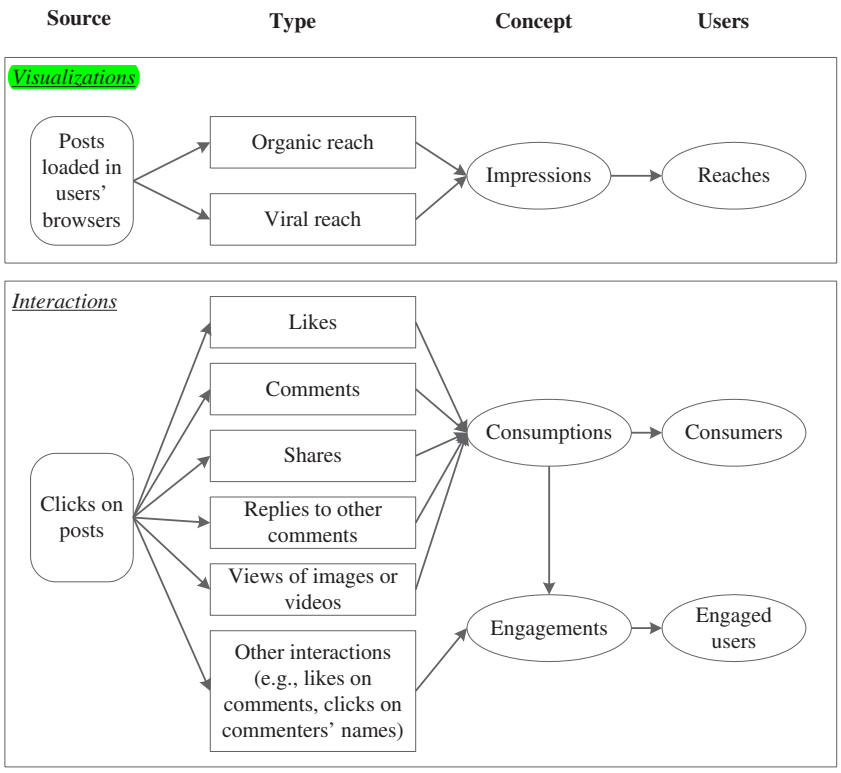

Fig. 1. Conceptual map on Facebook's performance metrics.
More detailed information can be obtained from:

https://developers.facebook.com/docs/graph-api/reference/v2.5/insights
http://www.agorapulse.com/blog/facebook-post-consumers-and-post-consumption

performance of posts. While each concept accounts for every visualization or interaction, Facebook also makes available metrics on a per-user basis, taking into account that each user may visualize or interact more than once per post.

2.2. Data mining

A data mining approach typically includes phases such as data understanding, data preparation, modeling, and evaluation (Han et al., 2011). The data set described in Table 1 includes twelve features (eleven exported from Facebook plus the computed total interactions) that may be used to measure posts' performance. Thus, any of the features can be used as an output to predict. It should be stressed that the "Page total likes" feature is not linked to any post, but instead to the page's performance. Hence, by considering it may affect the impact of publishing each post, we included it as an input feature. Therefore, the procedure included modeling each of those twelve features related to post's performance for assessing which ones could be better predicted. The meaning of each of those features is detailed in Table 2, based on the concepts illustrated in Fig. 1.

The seven remaining features are known prior to the post publication and can be used as an input. However, two of them are unique per post: the permanent link and the post ID. Thus, such features are of no value to modeling, considering that these do not represent any type of relationship between posts. Also, the post message itself is unique per post. Nevertheless, it may potentially conceal valuable knowledge in the unstructured textual message. However, one would need to include a text mining procedure in this approach for unveiling such knowledge, which was out-of-scope for the present study (it is suggested as future research in Section 5). One could also argue that the posted date is unique per post. Nevertheless, a few characteristics may be extracted from the date: the month, the weekday, and the hour. Adding these three computed features to the remaining four (excluding the posted date–time value since it is distinct for each post, the permanent link, the post ID, and the post message) provides a

Table 2
List of output features to be modeled

Feature	Description[a]
Lifetime post total reach	The number of people who saw a page post (unique users).
Lifetime post total impressions	Impressions are the number of times a post from a page is displayed, whether the post is clicked or not. People may see multiple impressions of the same post. For example, someone might see a Page update in News Feed once, and then a second time if a friend shares it.
Lifetime engaged users	The number of people who clicked anywhere in a post (unique users).
Lifetime post consumers	The number of people who clicked anywhere in a post.
Lifetime post consumptions	The number of clicks anywhere in a post.
Lifetime post impressions by people who have liked a page	Total number of impressions just from people who have liked a page.
Lifetime post reach by people who like a page	The number of people who saw a page post because they have liked that page (unique users).
Lifetime people who have liked a page and engaged with a post	The number of people who have liked a Page and clicked anywhere in a post (Unique users).
Comments	Number of comments on the publication.
Likes	Number of "Likes" on the publication.
Shares	Number of times the publication was shared.
Total interactions	The sum of "likes," "comments," and "shares" of the post.

[a] Descriptions extracted from:

- http://www.agorapulse.com/blog/facebook-reach-metrics-ultimate-guide
- https://www.facebook.com/help/274400362581037

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010

Table 3
List of input features used for modeling

Feature	Description
Category	Manual content characterization: action (special offers and contests), product (direct advertisement, explicit brand content), and inspiration (non-explicit brand related content).
Page total likes	Number of people who have liked the company's page.
Type	Type of content (Link, Photo, Status, Video).
Post month	Month the post was published (January, February, March, …, December).
Post hour	Hour the post was published (0, 1, 2, 3, 4, …, 23).
Post weekday	Weekday the post was published (Sunday, Monday, …, Saturday).
Paid	If the company paid to Facebook for advertising (yes, no).

data set with seven distinct input features for feeding the model (Table 3).

Fig. 2 exhibits the data mining procedure undertaken for implementing the model, for validating the results, and for extracting useful knowledge for leveraging post publications decisions. Such procedure was executed twelve times, for evaluating the predictive performance of every output feature available. First, the data mining algorithm chosen was fed with the seven input features from Table 3, resulting in a model. Such model was then tested to obtain the values predicted for the output performance metric of the post. The differences between the real performance metrics and the predicted values were compared to assess model performance. The performance metric that could better be modeled, i.e., in which the model predictions showed less differences to the real values, was then assessed to understand how input features influenced this performance metric. In Section 3.2, further details are provided on the specific data mining technique employed.

3. Theory

3.1. Social media impact on brand building

Laroche et al. (2012) demonstrated the effects of brand communities established in social media platforms on the underlying elements and practices in communities as well as on brand trust and brand loyalty. According to Deloitte Digital (2015), based on a survey of over 3000 US consumers, digital interactions are expected to influence 64 cents of every dollar spent in retail stores by the end of 2015, up from 14 cents in 2012, meaning that social media is increasing its direct impact on companies' revenues. The creation of virtual customer environments may be triggered by social media networks such as Twitter and Facebook, providing an emergent interest around specific firms, brands, and products. Therefore, in order to create business value, organizations need to incorporate community building as part of the implementation of social media (Culnan et al., 2010). Brand communities established on social media enhance feelings of community among members and contribute to creating value for both members and companies (Laroche et al., 2012).

Table 4
Studies on social media impact on branding

Reference	Materials	Problem addressed
Smyser et al. (2014)	Facebook's metrics	Tobacco control campaign
Jansen et al. (2009)	Tweets (from Twitter)	Brand sentiment analysis on 5 brands (Banana Republic, SMART For-Two, Wii Fit, Google, and Forever Stamp)
Shen and Bissell (2013)	Facebook's likes and comments	Branding on six cosmetic companies

Several empirical studies were published in the last few years for assessing the impact of social media on brand building. Hudson et al. (2015) conducted three surveys to explore the relationship between social media usage and customer–brand relationships. Their results showed that engaging customers via social media was associated with higher consumer–brand relationships. The study by Hutter et al. (2013) was focused on the case of the MINI car brand through a survey published in MINI's Facebook page for assessing impact on brand awareness and purchase intentions. The conclusions supported previous research insights in that social media content influences the economic outcome of brands. Another empirical paper evaluated the impact of social media adoption on a small-scale UK-based company, concluding that social media advertising had a positive effect on brand image (Griffiths and McLean, 2015). Generically, the literature seems to support an existing influence of social media on brand building. Nevertheless, research in this domain is still scarce. Adding up to the novelty and interest of social media, research is expected to flourish in the next few years, hopefully unveiling novel studies for filling this research gap (Okazaki and Taylor, 2013).

Most of the empirical research found in the literature and cited above adopted survey studies. While survey studies are a valuable method for obtaining data, we were more interested in articles that passively collected relevant information from social media networks and used it for empirical research. Such method is better aligned with our focus on using performance metrics for assessing the impact on brand building. We selected three distinctive studies using different social media metrics (detailed in Table 4), including the only one we found using Facebook's performance metrics, for drawing a model in a structure adapted from a concept map format (Kinchin et al., 2000) that summarizes the impact on brand building (Fig. 3). Such model has four layers: the source metrics' types, the data effectively used for the experiments, the references, and the branding effect levels. For

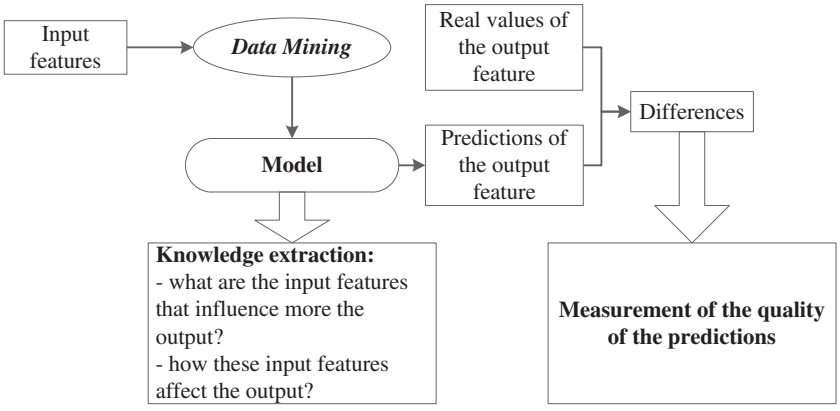

Fig. 2. Data mining procedure.

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010

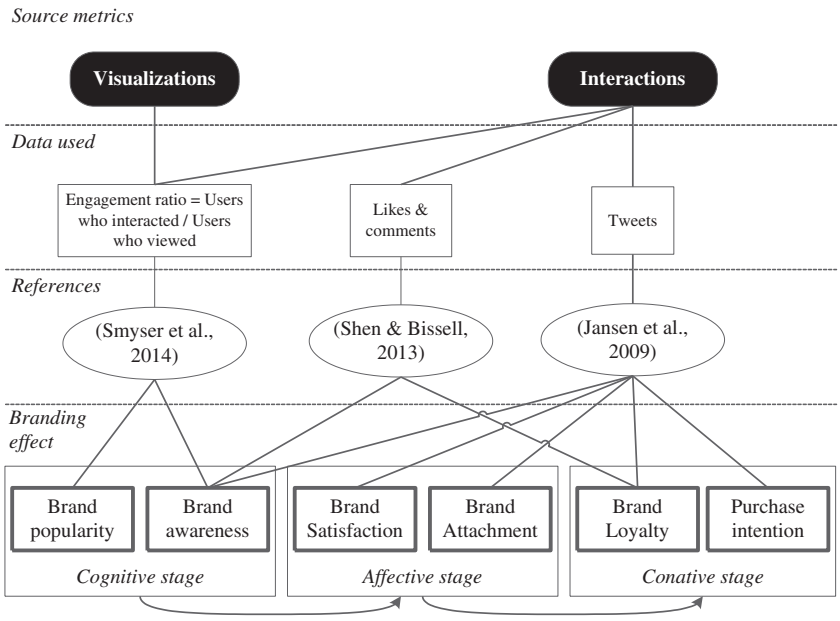

Fig. 3. Social media impact on branding.

defining the branding stages, we adapted the model described by Smith et al. (2008). Visualization metrics appear to be closer related to the cognitive stage, influencing awareness, while interactions affect all branding stages.

While engagement in social media has been shown to influence every stages of branding, other types of relationships should be also accounted for. Habibi et al. (2014) defined a model for brand communities based on social media which included five relationships that affected brand trust: customer–product, customer–brand, customer–company, customer–other customers, and brand community engagement. Their results supported that brand community engagement amplifies the impacts of customers' relationships with both the brand and the product. Therefore, while numerous aspects may influence brand building, literature supports that engagement also plays a significant role. Thus, a predictive system that anticipates posts' engagement in social media can provide a valuable tool to support managers' decisions on publishing posts.

Data mining has the potential for discovering valuable trends and insights concealed in social networks (Gupta et al., 2014). The interactions between customers about a brand in online social networks are powerful mindset enablers that can have a huge impact in brand building (Gensler et al., 2015). By using the predictive potential of data mining to understand how each of the posts published about a certain brand acts as an enabler of brand building in its different stages (Fig. 3), social media managers could make solid-grounded decisions on whether to publish a certain post. Such premise is the main driver of the current research.

3.2. Data mining

Data mining enables to identify coherent patterns of information from where to extract useful knowledge (Turban et al., 2011). Its roots include both traditional statistical analysis and artificial intelligence/machine learning sciences, aiming to benefit from both. We adopted data mining for modeling the twelve numeric metrics related to the performance of posts published in a social network, enumerated in Table 1. Since the algorithm tries to fit the input data to model a numeric variable, it makes this a regression problem.

Several data mining techniques can be used to model numeric variables, such as linear regression, support vector machines, and neural networks (Cortez, 2010). We adopted the support vector machines for conducting the experiments. Support vector machines emerged

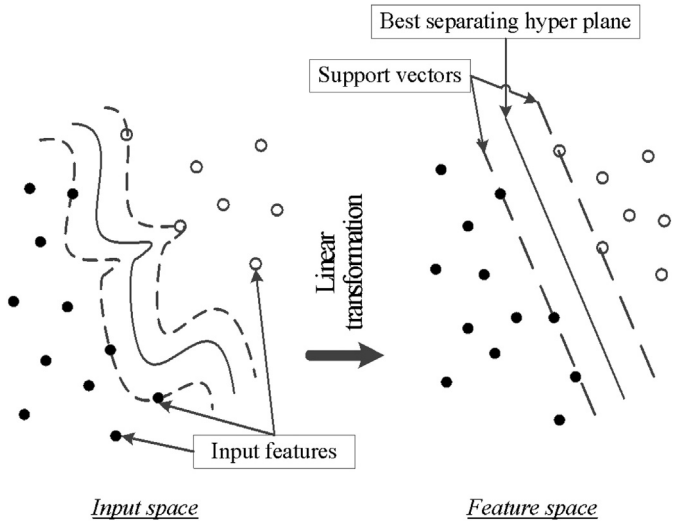

Fig. 4. Support vector machines.

Table 5
Results for performance metrics predictions

Performance metric	Mean absolute percentage error	Source of metric
Lifetime people who have liked your page and engaged with your post	26.9	Interactions
Lifetime post consumers	27.2	
Lifetime engaged users	28.8	
Lifetime post consumptions	33.1	
Shares	35.8	
Lifetime post reach by people who like your page	37.5	Visualizations
Likes	41.2	Interactions
Lifetime post impressions by people who have liked your page	47.8	Visualizations
Lifetime post total reach	49.6	
Comments	63.9	Interactions
Lifetime post total impressions	69.3	Visualizations

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010

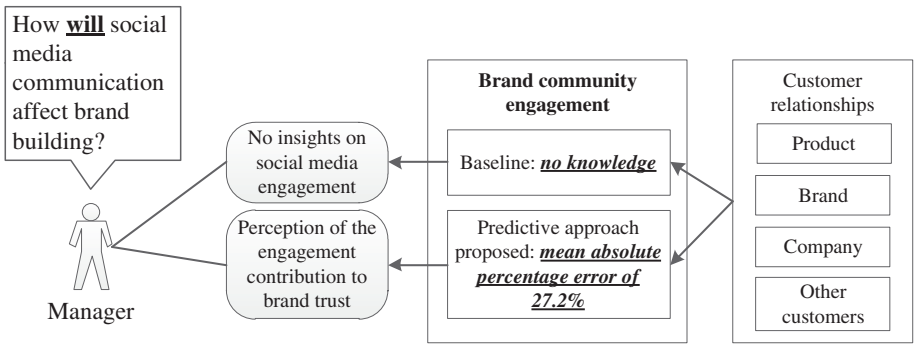

Fig. 5. Application of the model for "Lifetime Post Consumers" (adapted from Habibi et al., 2014).

in the nineties to become one of the most widespread advanced machine learning techniques. Support vector machine transforms the input $x \in \Re^M$ space into a high m-dimensional feature space by using a nonlinear mapping that depends on a kernel. Then the algorithm finds the best linear separating hyper plane, related to a set of support vector points, which define the support vectors in the feature space, as shown in Fig. 4 (Steinwart and Christmann, 2008). For the experiments, the popular Gaussian kernel was chosen, which has the advantage of having less parameters than other kernel functions (Hastie et al., 2005).

The support vector machine provides a high accuracy performance model, although it has the disadvantage of being difficult to understand by humans, in contrast to traditional methods such as linear regression or decision trees, from which the rules comprising these can be directly read. A sensitivity analysis has proved to be an effective method for extracting useful knowledge from black box models such as the support vector machines (Cortez and Embrechts, 2013). Such method consists in assessing model sensitivity to changes in the inputs by evaluating how the output predicted value changes when varying the input features through their range of values. Moreover, the data-based sensitivity analysis was proposed by Cortez and Embrechts (2013) and selects a sample from the input data used to train the model for assessing model sensitivity to several inputs' variations at the same time. This method has been extensively used in several distinct domains (e.g., Moro et al., 2015). All the experiments described in this article were conducted using the R statistical tool.

4. Experiments and results

4.1. Prediction

As stated in Section 2, we used the seven input features from Table 3 to predict each of the twelve performance metric features described in Table 2. In order to prepare the 790 rows containing the information about the posts published on this cosmetics company's Facebook page, outliers were analyzed for each of the performance features. We adopted the Shapiro–Wilk test to assess if each of the output columns for the features to be predicted followed a normal distribution (Razali and Wah, 2011). Such a validation provided the ground needed to discard the 5% posts from which the performance metric value deviated the most, leaving 751 of the posts for building the model.

After generating the model for each of the twelve performance metrics, we evaluated the results by comparing the real value for that metric with the value predicted by the model. A good model implies fitting all the input data in a way that the predicted values are as close as possible to the real values. Hence, we first computed the absolute difference between the predicted values and the real values for each post. We also computed the difference in percentage to assess the relative deviation of the predictions. Finally, we calculated the mean absolute percentage error for each model, which is a metric widely used to evaluate regression model's performance based on the relative difference (Hyndman and Koehler, 2006). The results are shown in Table 5.

Table 5 indicates also which are the source types for the metrics, as described in Fig. 1: visualization metrics or interaction metrics (grayed rows). The results show that the models built for predicting interaction-based metrics were more accurate than for visualizations. Nevertheless, the specific Facebook interactions, shares, likes, and comments, were predicted with a larger difference when compared to the Facebook report interaction metrics, which ranked in the top four best predicted metrics. However, comments were by far poorly predicted. Such result may be derived from the fact that comments may hold either a positive or a negative connotation, which has been a subject of several studies (e.g., Ballantine et al., 2015). Thus, for evaluating comments, one would need to consider the textual message contained in each comment.

Visualization metrics were more difficult to model, according to Table 5. The visualization-based performance metric that achieved the best result was the "Lifetime Post reach by people who like a Page", with 37.5%, more than 10% of difference to the best interaction-based metrics. Nevertheless, it concealed an interaction with the page since it considered only users who liked the page. Likewise, the second best predicted visualization-based metric also considered only users who liked the page. The two purely visualization-based metrics were poorly predicted. Such result may be derived that visualizations are more subjected to randomness since any user may get the post contents loaded onto its browser for numerous reasons. It would be interesting in a future study to consider feature enrichment strategies for improving the accuracy in predicting visualizations, as the viral reach is becoming more relevant in brand awareness (Moro et al., 2016).

Both the models for the "Lifetime People who have liked a Page and engaged with a post" and the "Lifetime Post Consumers" features achieved an average difference of around 27% to the real values (with 26.9% for the former and 27.2% for the latter features). Such results

Table 6
Posts information and model evaluation for "Lifetime Post Consumers"

Category	Page total likes	Type	Month	Hour	Weekday	Paid	Real	Predicted	Absolute difference	% difference
Product	139,441	Photo	Dec	3	Thu	No	134	228	94	70%
Action	136,642	Photo	Oct	13	Sat	No	356	346	10	3%
Inspiration	135,617	Photo	Sep	10	Fri	No	614	520	94	15%
Product	139,441	Status	Dec	3	Sat	No	1407	1502	95	7%

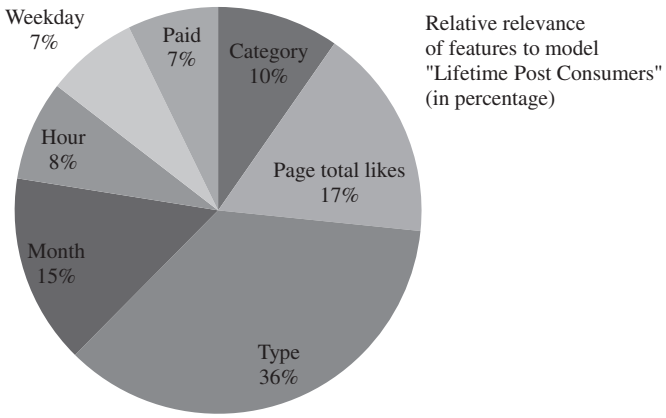

Fig. 6. Relevance of the input features for "Lifetime Post Consumers."

mean that one can know prior to post publication the results on these two metrics with an error of around 27%. Hence, managers can decide to use both models for having a perception on the impact a given post may have. Although the models still hold a difference of 27%, these provide a better judged decision than not having an educated guess at all. Considering the theoretical analysis on Section 3.1, these models may be useful for assessing social media user engagement, which influences brand trust (Habibi et al., 2014). Also, the conceptual model shown in Fig. 3 references that interaction in social media affects all branding stages.

For a deeper model analysis, there was a focus on the "Lifetime Post Consumers" performance metric. It achieved 27.2% of mean absolute percentage error, slightly higher than "Lifetime People who liked a Page and engaged with a post" (26.9%). However, the latter performance metric is dependent on users liking the page. Therefore, the analysis will focus on the "Lifetime Post Consumers." Fig. 5 exemplifies how the predictive approach proposed could help answering manager's question of obtaining insights about the impact that social media communication will have on brand building prior to post publication.

Table 6 shows four randomly selected examples of the information for four of the posts used to feed the model at the seven left columns, while "Real" provides the "Lifetime Post Consumers" real value. The three last columns are used to validate the model: "Predicted" shows the value predicted by the model, "Absolute difference" shows the absolute difference to the real value while the last column shows the percentage difference.

The examples provided in Table 6 for "Lifetime Post Consumers" illustrate how a manager can understand the impact of the posts: the two bottom rows even though showing a difference around a hundred in terms of absolute values can provide a glimpse of the order of magnitude of the real values. Furthermore, it should be stressed that such result is achieved by only using seven input features, with three of them being related to the date and time the post was published ("Month,"

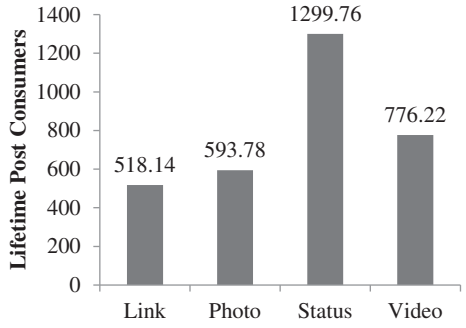

Fig. 7. Influence of "Type" on "Lifetime Post Consumers."

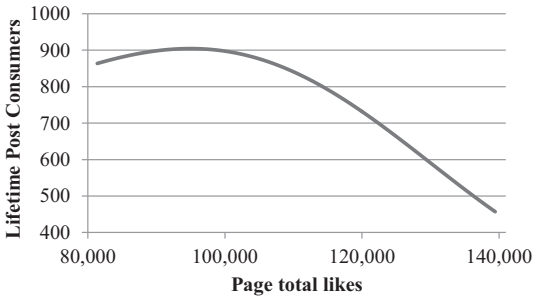

Fig. 8. Influence of "Page total likes" on "Lifetime Post Consumers."

"Weekday," and "Hour"), three directly obtained from Facebook ("Type," "Page total likes," and "Paid") and only one fed by the page content manager ("Category"). Usually, content managers have a richer set of features for characterizing each of the contents they are publishing (e.g., specificities about a product being advertised, if the product or service has any associated promotion). Enriching the data set with such features may result in an increase in model's accuracy (Moro et al., 2016).

4.2. Knowledge extraction

In Section 4.1, two models for two distinctive performance metrics achieved an average difference of around 27% to the real values, namely, the "Lifetime People who have liked a Page and engaged with a post" and the "Lifetime Post Consumers." The "Lifetime Post Consumers" provided a more interesting metric for the decision of publishing the post, as it focused solely on the impact of the post, while the "Lifetime People who have liked a Page and engaged with a post" contained an inner relation to liking the page besides interacting with the post. Therefore, the "Lifetime Post Consumers" was chosen for analysis.

For extracting knowledge from the "Lifetime Post Consumers" implemented model, a data-based sensitivity analysis was performed via two complementary approaches: first, the model was assessed to understand which of the input features affected more the outcome of the studied metric; second, all input features from the most to the least relevant for the model were assessed to discern how each of them influenced the outcome.

Fig. 6 shows the contribution of each input feature for the model of "Lifetime Post Consumers." The relevance of the "Type" of content published was remarkable since it accounted for 36% of relevance to the model. This finding is aligned with the results reported by Cvijikj et al. (2011), which analyzed fourteen sponsored brand pages using statistical analysis for assessing the correlation between "Type" and the number of likes and comments. First, building a predictive model with several features is more complex than just finding correlation between two variables since the machine learning technique needs to

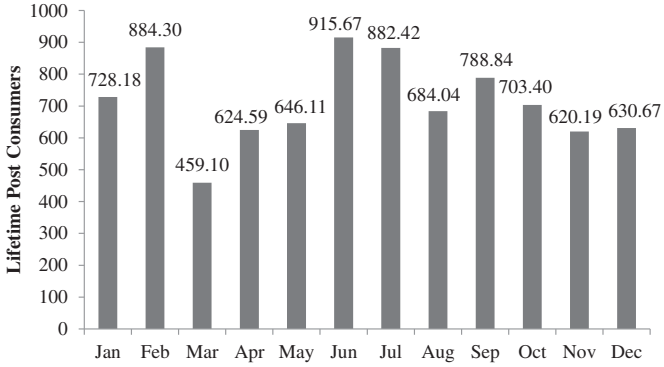

Fig. 9. Influence of "Month" on "Lifetime Post Consumers."

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010

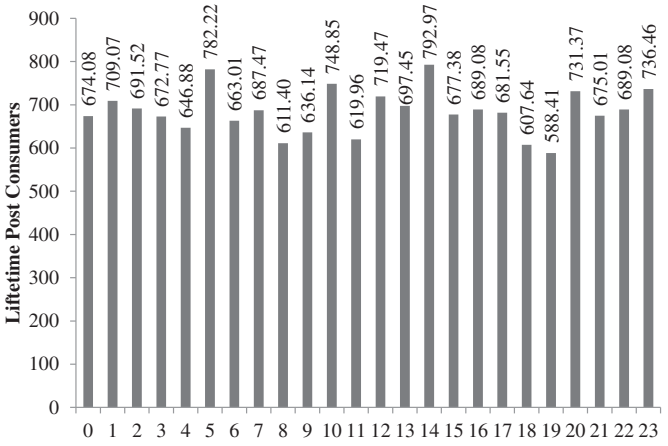

Fig. 10. Influence of "Hour" on "Lifetime Post Consumers."

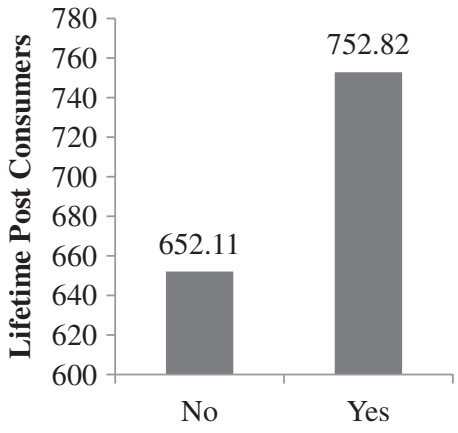

Fig. 12. Influence of "Paid" on "Lifetime Post Consumers."

find the best fit possible for all features. Also, the Pearson correlations found by Cvijikj et al. (2011) were weak ($r = 0.11$) to moderate ($r = 0.62$); thus, no strong relationship was found for likes and comments (Rumsey, 2011). This weak to moderate correlation may be reflected in the 36% relative relevance of type for the "Lifetime Post Consumers" model. Nevertheless, more studies on different data would be needed to confirm such hypothesis. The "Page total likes" and the "Month" the post was published appear in the second and third positions with 17% and 15% of relevance, respectively. The fact that the "Month" has 15% of relevance suggests seasonality which could be associated with the type of industry of this particular company. It should be stressed that the three most relevant features accounted for almost 70% of the relevance to the model. For comparison, we performed a similar analysis (see Figs. 6 and 15 in the Appendix) to the model obtained for "Lifetime People who liked a Page and engaged with a post". While the order of the features for relevance does not change much, the relative relevance has some variations which are discussed on the Appendix.

The "Category" set by the Page manager had 10% of relevance. This result showed a weak relative relevance, especially when compared to the study of Cvijikj et al. (2011), in which a weak to moderate correlation was found between category and likes/comments. However, post categorization is dependent on the company's strategy. Furthermore, such research used data from fourteen sponsored brand and consumer goods pages, with a specific categorization assigned manually (seven categories: information, designed question, statements, advertisements, competitions, questionnaires, and announcements). Therefore, the results cannot be directly compared. The "Hour" and the "Weekday" the post was published account for just 8% and 7% of relevance, respectively. Cvijikj et al. (2011) found no correlation between weekday and likes/comments in their study. This is a subject that should demand more research, considering that people are more available in weekends, thus intuitively one would expect to see more engagement on these days. Finally, the feature that indicates the company paid for the page to be specifically advertised appears with just 7% of relevance, the same of weekdays. This is an interesting result and suggests that paying for the specific post to boost the reachability does not compensate as many as focusing on publishing on the right month, for example.

After understanding the importance of the content "Type" to the impact of the post as translated by "Lifetime Post Consumers" measured from the model, it was intended to observe how each of the possible types influenced this output metric. Fig. 7 illustrates this influence and shows that "Status" posts have clearly the largest impact on the performance of the post, more than twice the values for "Photo" and "Link," and 60% more than "Video." This result is aligned with the findings of Cvijikj et al. (2011), which found that "Status" posts caused the greatest number of comments, "Videos" caused the most likes, while "Photos" and "Links" had the least number of interactions. Moreover, the study published by Kwok and Yu (2013), while achieving a similar conclusion for "Status" posts, came to a different conclusion for "Photos," stating that these received more likes and comments than "Links" and "Videos." However, while Cvijikj et al. (2011) analyzed fourteen sponsored brand pages selected from Fan Page web page, which ranks the Facebook pages, Kwok and Yu (2013) focused specifically on restaurant pages, making their case more specific than the former cited study.

The second most relevant features, "Page total likes," is by far less relevant than "Type," although still showing an influence of 17%. This input feature relates to the likes the company's page where the post is being published had at the moment of publishing the post. Fig. 8

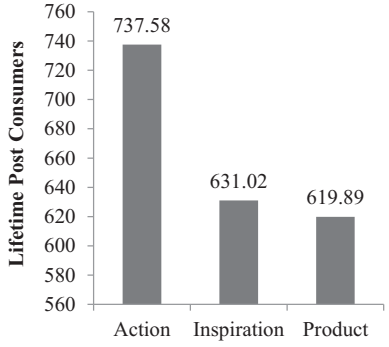

Fig. 11. Influence of "Category" on "Lifetime Post Consumers."

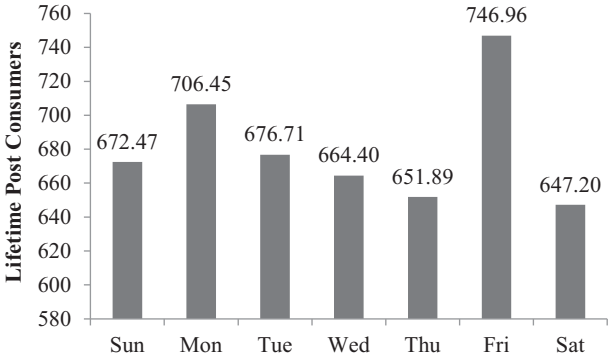

Fig. 13. Influence of "Weekday" on "Lifetime Post Consumers."

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010

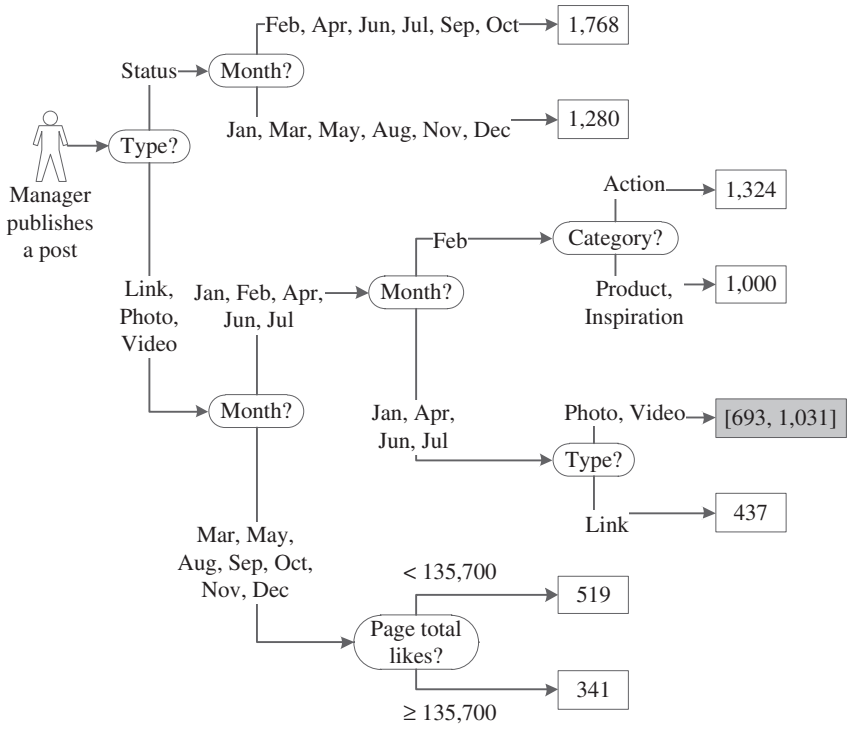

Fig. 14. Rules extracted from the support vector machine model.

shows that "Lifetime Post Consumers" decreased after reaching a peak of around 95,000 page likes. Page likes usually increase overtime as brand satisfaction is translated into social media interaction with the company's page. However, unsatisfied users may explicitly press the dislike button, removing the like previously counted. Such actions directly affect page total likes. As page total likes increases overtime, more users are receiving feedback on the posts published in the page they previously liked. However, an observation of Fig. 8 reveals that, while page likes increased, post consumers were expected to decrease, i.e., users are not so keen to engage with posts being published. Such issue may disclose some erosion of the company's Facebook page, since users are seeing the contents published, but are not interacting with it.

The "Month" is the third most relevant feature, with 15% of influence. Fig. 9 displays some seasonality, with a large increase starting in April and reaching a peak in June of almost twice the value of April. From November to February there is also a steady although not so steep increase. Golder et al. (2007) analyzed a data set of messages sent through Facebook between February 2004 and March 2006, a period of 26 months, also identifying a large increase between March and June, while another increase appears between September and January, in a time frame displaced by one/two months in relation to the November–February increase observed in Fig. 9. The results are not directly comparable due to different contexts, namely, the mentioned study used private messages, it focused on an academic community, and it considered the period when Facebook was in its infancy. Nevertheless, Golder et al. (2007) also found a seasonality effect in an early stage of Facebook, aligned with current findings in a more mature stage of the same social network.

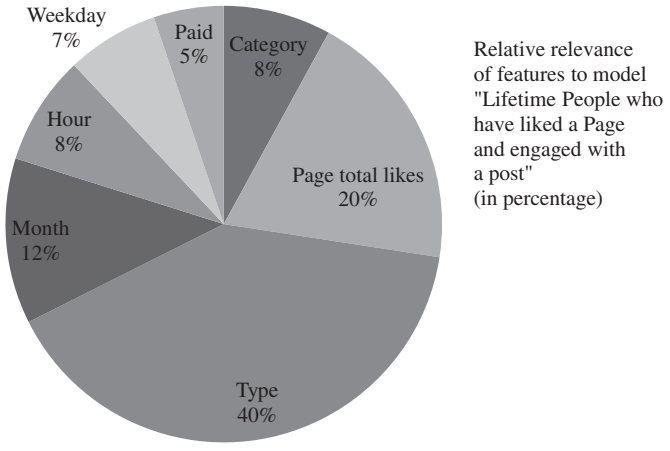

Fig. 15. Relevance of input features to the model of "Lifetime People who have liked a Page and engaged with a post".

The influence of the remaining four least relevant features which conceal 32% of the model's knowledge is displayed in Fig. 10 for "Hour," Fig. 11 for the "Category," Fig. 12 for "Paid," and finally Fig. 13 for "Weekday." Regarding "Category," it is remarkable the influence that "Action" had when compared to the remaining two features. This "Actions" category stands for special offers and contests, clearly gathering more attention than "Products" and non-explicit brand related contents ("Inspiration"). Campaigns have proven to be a valuable asset for brand awareness (Hanna et al., 2011). The company's Facebook page managers may use this type of "Actions" for increasing social media engagement, thus contributing to an increase in brand building. The "Hour" influence graphic appears to show that no trends associated with the hour of publication exists, although some peaks can be observed. The "Weekday" shows that "Monday" has a local maximum of impact, decreasing along the week until "Friday," when the global maximum of impact occurs. The study of Cvijikj et al. (2011) also resulted in a global maximum on "Friday," although they did not report a trend for the "Monday" local maximum and then decrease observed in Fig. 13. It was expected more impact on weekend days, considering users tend to be more available in this period. This is an interesting result to explore in future studies with additional data from when users interact with posts (e.g., by analyzing the hours in each day of the week in which the users engage the most). The result shown for "Paid" is expected: a post for which the company paid for advertising has a larger impact than a post not paid. Nevertheless, this is one of

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010

the least relevant input features for the defined model, with just 7% of relevance.

For providing the "big picture" of the decision process flow defined by the "Lifetime Post Consumers" model, we performed a rule extraction procedure over the model, by applying a decision tree modeling over the predicted values of the support vector machine built (Fig. 14). This technique has been shown to be valuable in providing a readable image of a black box model such as support vector machines, complementing the sensitivity analysis conducted from Figs. 6 to 13 (Moro et al., 2014).

The concise results shown in Fig. 14 provide a decision process path until a leaf node is reached (in squares), where the predicted "Lifetime Post Consumers" value lies. The leaf in gray represents a subset of the decision process path ranging from 693 to 1,031 of "Lifetime Post Consumers," which was omitted due to page space constraints. Nevertheless, the more a decision node is to the left, the more relevant that feature is. In fact, the "Type" of post appears as the first decision node, aligned with the results of Fig. 6 ("Type" was the top ranked feature in terms of relevance). However, the "Month" comes next in terms of relevance, while "Page total likes" is third, on the contrary of Fig. 6 ("Month" got a relevance of 15% while "Page total likes" achieved 17%). Such result may be due to the fact that decision tree's modeling does not apprehend the complex nonlinear mappings such as more advanced black box techniques (e.g., neural networks or support vector machines).

The results and analyses presented in this paper are based on the experiments of a specific case of posts published by a cosmetic company's Facebook page during the year of 2014. Models built using data mining are purely data driven; thus, these rely on the patterns of knowledge hidden in data (Turban et al., 2011). If data sources change, models need to be updated, which may happen not only if one considers another case study, but also during the natural evolution of the context surrounding the company and the users. Also, unexpected events may have a huge impact on the predicting capabilities of models. Hence, data mining models need constant updating to incorporate these changes. Therefore, the results presented cannot be generalized. However, the experimental approach conducted can be applied to another company and period, unveiling potential useful knowledge.

5. Conclusions

This research focused in modeling performance metrics extracted from posts published in a company's Facebook page through the usage of data mining. Moreover, the support vector machine technique was employed by feeding it with seven input features, all provided by Facebook's page, except a content specific categorization provided by the page's manager. Twelve performance metrics were modeled with these input features, from which the two models achieving the best performance modeled the "Lifetime Post Consumers" and the "Lifetime People who have liked a Page and engaged with a post" output features, with a mean absolute percentage error of 27.2% and of 26.9%, respectively.

Based on the "Lifetime Post Consumers" model, this study showed how it could benefit through its predictions brand building by providing insights on social media engagement. The advantages of using the model were also linked to all the stages of branding (cognitive, affective, and cognative stages). A data-based sensitivity analysis was then applied for extracting valuable knowledge from the model of "Lifetime Post Consumers." The "Type" of the content published was considered by far the most relevant input feature for the model. Posts from the "Status type" are likely to result in twice the impact of the remaining "Types." Also, seasonality was found regarding the "Month" the post was published. Publications related to special offers and contests are likely to produce posts with greater impact than "Product" and other non-explicit brand related contents. We also produced a decision flow process based on rules extraction from the model. Facebook page managers can use this knowledge to make informed decisions on the posts they publish, enhancing their impact, thus contributing for brand building.

Several ideas arise from this study for future research. First, the model may be enriched with other context features (e.g., if the product is being advertised elsewhere) for tuning its performance. Also, text mining methods could be employed to the content for extracting additional knowledge. Finally, text mining the comments of each post for user sentiment analysis could reveal the feelings each post is generating.

Acknowledgments

We would like to thank the two anonymous reviewers for their valuable recommendations, which highly enhanced the value of the final manuscript.

Appendix

Considering the novelty of the proposed approach, the knowledge hidden in the "Lifetime People who have liked a Page and engaged with a post" was also evaluated, which achieved a mean absolute percentage error of 26.9%. Nevertheless, such performance metric was influenced by considering only users who have liked the page, as argued in Section 4.1. Therefore, this Appendix shows in Fig. 15 the relative relevance of each input feature to the model (similar to the exercise displayed in Fig. 6).

The results are aligned with those for the model of "Lifetime Post Consumers," even though the "Type" is now more relevant (40%) than for the latter model (36%). Also, "Page total likes" are more relevant, while "Month" is less. Further studies would be required for a deeper analysis of the differences. Moreover, these studies would require additional data for differentiating the engagements of users who have liked from those that didn't but also engaged.

References

Ballantine, P. W., Lin, Y., & Veer, E. (2015). The influence of user comments on perceptions of Facebook relationship status updates. *Computers in Human Behavior*, *49*, 50–55.

Bianchi, C., & Andrews, L. (2015). Investigating marketing managers' perspectives on social media in Chile. *Journal of Business Research* (Available online 20 June 2015, in press).

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. *Advances in data mining. Applications and theoretical aspects* (pp. 572–583). Berlin Heidelberg: Springer.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1–17. http://dx.doi.org/10.1016/j.ins.2012.10.039.

Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large US companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive*, *9*(4), 243–259.

Cvijikj, I. P., Spiegler, E. D., & Michahelles, F. (2011). The effect of post type, category and posting day on user interaction level on Facebook. *In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 810–813). IEEE. http://dx.doi.org/10.1109/PASSAT/SocialCom.2011.21.

Digital, Deloitte (2015). Navigating the new digital divide—capitalizing on digital influence in retail. Retrieved September 12, 2015, from http://www2.deloitte.com/content/dam/Deloitte/us/Documents/consumer-business/us-cb-navigating-the-new-digital-divide-v2-051315.pdf

Dossier, Statista (2014). Social media & user-generated content—Number of global social network users 2010–2018—Statista Dossier 2014. Retrieved September 10, 2015, from http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Edosomwan, S., Prakasan, S. K., Kouame, D., Watson, J., & Seymour, T. (2011). The history of social media and its impact on business. *Journal of Applied Management and Entrepreneurship*, *16*(3), 79–91.

Gensler, S., Völckner, F., Egger, M., Fischbach, K., & Schoder, D. (2015). Listen to your customers: Insights into brand image using online consumer-generated product reviews. *International Journal of Electronic Commerce*, *20*(1), 112–141. http://dx.doi.org/10.1080/10864415.2016.1061792.

Golder, S. A., Wilkinson, D. M., & Huberman, B. A. (2007). Rhythms of social interaction: Messaging within a massive online network. *Communities and technologies 2007* (pp. 41–66). London: Springer. http://dx.doi.org/10.1007/978–1-84628-905-7_3.

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010

Griffiths, M., & McLean, R. (2015). Unleashing corporate communications via social media: A UK study of brand management and conversations with customers. *Journal of Customer Behaviour*, *14*(2), 147–162.
Gupta, S., Hanssens, D., Hauser, J. R., Lehmann, D., & Schmitt, B. (2014). Introduction to theory and practice in marketing conference special section of marketing science. *Marketing Science*, *33*(1), 1–5. http://dx.doi.org/10.1287/mksc.2013.0830.
Habibi, M. R., Laroche, M., & Richard, M. O. (2014). The roles of brand community and community engagement in building brand trust on social media. *Computers in Human Behavior*, *37*, 152–161.
Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques: concepts and techniques*. Elsevier.
Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business Horizons*, *54*(3), 265–273.
Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85. http://dx.doi.org/10.1007/BF02985802.
Hudson, S., Huang, L., Roth, M. S., & Madden, T. J. (2015). The influence of social media interactions on consumer–brand relationships: A three-country study of brand perceptions and marketing behaviors. *International Journal of Research in Marketing*.
Hutter, K., Hautz, J., Dennhardt, S., & Füller, J. (2013). The impact of user interactions in social media on brand awareness and purchase intention: The case of MINI on Facebook. *Journal of Product & Brand Management*, *22*(5/6), 342–351.
Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001.
Insights, Digital (2014). Social Media 2014 Statistics. Retrieved September 10, 2015, from http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html
Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188.
Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. http://dx.doi.org/10.1016/j.bushor.2009.09.003.
Kinchin, I. M., Hay, D. B., & Adams, A. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research*, *42*(1), 43–57.
Korschun, D., & Du, S. (2013). How virtual corporate social responsibility dialogs generate value: A framework and propositions. *Journal of Business Research*, *66*(9), 1494–1504. http://dx.doi.org/10.1016/j.jbusres.2012.09.011.
Kwok, L., & Yu, B. (2013). Spreading social media messages on Facebook: An analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly*, *54*(1), 84–94. http://dx.doi.org/10.1177/1938965512458360.
Lariscy, R. W., Avery, E. J., Sweetser, K. D., & Howes, P. (2009). Monitoring public opinion in cyberspace: How corporate public relations is facing the challenge. *Public Relations Journal*, *3*(4), 1–17.
Laroche, M., Habibi, M. R., Richard, M. O., & Sankaranarayanan, R. (2012). The effects of social media based brand communities on brand community markers, value creation practices, brand trust and brand loyalty. *Computers in Human Behavior*, *28*(5), 1755–1767. http://dx.doi.org/10.1016/j.chb.2012.04.016.
Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons*, *52*(4), 357–365. http://dx.doi.org/10.1016/j.bushor.2009.03.002.
Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. http://dx.doi.org/10.1016/j.dss.2014.03.001.
Moro, S., Cortez, P., & Rita, P. (2015). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing and Applications*, *26*(1), 131–139. http://dx.doi.org/10.1007/s00521-014-1703-0.
Moro, S., Cortez, P., & Rita, P. (2016). A framework for increasing the value of predictive data-driven models by enriching problem domain characterization with novel features. *Neural Computing and Applications*, *1-9*. http://dx.doi.org/10.1007/s00521-015-2157-8 (in press).
Okazaki, S., & Taylor, C. R. (2013). Social media and international advertising: Theoretical challenges and future directions. *International Marketing Review*, *30*(1), 56–71.
Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.
Rumsey, D. J. (2011). *Statistics for dummies*. John Wiley & Sons.
Shen, B., & Bissell, K. (2013). Social media, social me: A content analysis of beauty companies' use of Facebook in marketing and branding. *Journal of Promotion Management*, *19*(5), 629–651.
Smith, R. E., Chen, J., & Yang, X. (2008). The impact of advertising creativity on the hierarchy of effects. *Journal of Advertising*, *37*(4), 47–62.
Smyser, J., Novotny, T., Dayan, S., & Rodwell, T. (2014). "Toxic Butts": Key performance indicators from a California statewide social media campaign for tobacco control. *British Journal of Medicine and Medical Research*, *4*(25), 4341.
Steinwart, I., & Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.
Trainor, K. J., Andzulis, J. M., Rapp, A., & Agnihotri, R. (2014). Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM. *Journal of Business Research*, *67*(6), 1201–1208. http://dx.doi.org/10.1016/j.jbusres.2013.05.002.
Turban, E., Sharda, R., Delen, D., & Efraim, T. (2011). *Decision support and business intelligence systems* (9th ed.). Pearson.

Please cite this article as: Moro, S., et al., Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.02.010