ISCTE Business School
Instituto Universitário de Lisboa

The impact of Social Media in Brand Building

Bernardo Varela Vala

Project submitted as a partial requirement for the conferral of

Master of Science in Business Administration

Supervisor:

Paulo Rita, PhD, Full Professor

ISCTE – University Institute of Lisbon

Co-supervisor:

Sérgio Moro, PhD, Researcher

Business Research Unit (ISCTE-IUL)/ALGORITMI Centre (Univ. Minho)

September 2015

A journey of a thousand miles
begins with a single step.

Lao Tzu

This work results from various contributions and not only from my effort. Therefore I have to be thankful to a lot of people who are co-responsible for my success on delivering this document.

First of all I must congratulate Professor Paulo Rita and Sergio Moro for being capable of never letting the right path fade away and providing me with guidance and enlightenment during this period.

All the data used was retrieved while I was an intern at Live Content in Lisbon, and so I must thank to everyone who worked there with me and provided me with priceless knowledge regarding digital marketing and "general real-life business", which I will use during the rest of my life. I want to address some special words of appreciation to Sérgio and Domingos, as the face of Live Content, for making me feel at home and providing me with all the help I needed during those times.

Not less important was my family, nagging me about this "thesis writing", in the persons of Ana, Rui, Mafalda, Henriqueta and José, without you this might had happen but never without one or two postpones.

I must never forget my friends you walked alongside me during this time, not only regarding "thesis purposes", but also.

A special thanks to everyone whom I came across during this period and talked about my thesis as you provided me with new perspectives, questions and approaches which definitely contributed to make this a work that I am proud of.

Resumo

As publicações nas redes sociais influenciam de forma direta a formação das marcas na medida em que afectam a percepção que os consumidores têm da marca. Este estudo apresenta uma abordagem alavancada através de data mining que prevê o impacto de publicações numa página de Facebook de uma marca. Foram modeladas 791 publicações através de 12 métricas de performance, sendo que os dois melhores resultados atingiram um erro médio de cerca de 27%. Uma dessas variáveis, "Lifetime Post Consumers", foi analisada através de uma análise de sensibilidade para perceber de que forma é que cada uma das sete variáveis de input a influenciam. O tipo de conteúdo foi considerado a mais relevante com uma relevância de 36%. Uma publicação do tipo "Status" capta o dobro da atenção dos consumidores, quando comparado com os outros dois tipos de publicação. Foi também verificada elevada sazonalidade, de acordo com o mês da publicação.

Este tipo de conclusões são interessantes para ajudar os gestores nas suas decisões sobre fazer ou não uma publicação e em que moldes.

Palavras chave: Redes sociais; Data mining; Brand building; Facebook.

Sistema de classificação JEL:
M31 Marketing;
M37 Advertising.

Abstract

The impact of publications on social networks directly affects brand building through customers' perceptions on the brand. This research presents a data mining approach for predicting the impact of posts published on a Facebook page. Twelve posts' performance metrics extracted from a cosmetic company's page including 791 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. 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. Also, seasonality was observed regarding the month of the publication. Such knowledge is valuable for content managers' to make informed decisions on whether to publish or not a post.

Keywords: Social Media; Data mining; Brand building; Facebook.

JEL classification system:

M31 Marketing;

M37 Advertising.

Index

RESUMO IV

ABSTRACT V

1. INTRODUCTION 1

2. THEORY 3

2.1. SOCIAL MEDIA IMPACT ON COMPANIES 3

2.2. DATA MINING 3

3. MATERIAL AND METHODS 5

3.1 DATASET 5

3.2 DATA MINING 7

4. EXPERIMENTS AND RESULTS 11

4.1. PREDICTION 11

4.2. KNOWLEDGE EXTRACTION 14

5. CONCLUSIONS 21

6. REFERENCES 22

ANNEXES 26

ANNEX A - DEFINITIONS 26

Tables index

Table 1 - Features from the dataset compiled……5

Table 2 - List of input features used for modeling…….....8

Table 3 - List of output features to be modeled……....8

Table 4 - Posts information and model evaluation for "Lifetime Post Consumers"……12

Table 5 - Results for performance metrics predictions……13

Figures Index

Figure 1 - Data mining procedure……10

Figure 2 - Relevance of the input features for "Lifetime Post Consumers" ……15

Figure 3 - Influence of "Type" on "Lifetime Post Consumers"……...16

Figure 4 – Influence of "Page total likes" on "Lifetime Post Consumers"……..17

Figure 5 - Influence of "Month" on "Lifetime Post Consumers"……....18

Figure 6 - Influence of "Category" on "Lifetime Post Consumers"……19

Figure 7 - Influence of "Hour" on "Lifetime Post Consumers"……...19

Figure 8 - Influence of "Weekday" on "Lifetime Post Consumers"……....20

Figure 9 - Influence of "Paid" on "Lifetime Post Consumers"……20

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 & Haenlein, 2010). According to Statista Dossier (2014), the number of social networks 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 & 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 business. Measuring the impact of advertisement is an important issue to be included in a global social media strategy (Lariscy et al., 2009). A system that could predict the impact of each of their advertising posts in a social media would provide a valuable advantage when deciding to communicate through social media, helping to promote products and services, thus supporting brand building and there are no studies supporting the direct relation between post form and performance.

Data mining provides an interesting approach for extracting predictive knowledge from raw data (Turban et al., 2011). Its application to social media has been extensively 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, posing legal issues (e.g., Bianchi & Andrews, 2015). In the present article, the focus of research is on the prediction of the impact a given post published by a specific company will have on customers prior to its publication, providing a previous evaluation that enables to support company's social media managers decisions whether to publish it or not.

For validating the procedure undertaken, the case of a worldwide cosmetic company with a renowned brand is used, including 790 posts published by this company in the year of 2014 in the Facebook social network. Therefore, this dataset of posts is used as an input to the data mining procedure.

The main goals of this work are:

- 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, therefore contributing to brand building through the support of managers' decisions.

Next section describes the materials used (e.g., the input dataset) as well as the methods chosen for the experiments. Section 2.2 is focused on providing specific background on the technical aspects of the data mining procedure, including prediction modeling and knowledge extraction. Sections 3 and 4 exhibit the results achieved and discuss them in the light of brand building through improving the value created by each post. Finally, in the last section conclusions are drawn.

2. Theory

2.1. Social media impact on companies

Laroche et al. (2012) demonstrated the effects of brand communities established on the platform of social media on the underlying elements and practices in communities as well as on brand trust and brand loyalty. According to "Navigating the New Digital Divide", a report from the consulting firm Deloitte, based on a survey of over 3,000 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 are increasing their direct impact on companies revenues (Deloitte Digital, 2015).

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, to creating 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).

Data mining has the potential of 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 published posts about a certain brand acts as an enabler of brand building, 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.

2.2. Data mining

The current research adopted a data mining approach for modeling 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 of this a regression problem.

For the experiments, the support vector machine-modeling algorithm was adopted. Support vector machines emerged in the nineties to become one of the most widespread advanced machine learning techniques. Support vector machine fits input data by finding the best linear separating hyper plane, related to a set of support vector points, in the feature space (Steinwart & Christmann, 2008). A nonlinear kernel function is used for building the feature space. 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 as proved to be an effective method for extracting useful knowledge from black-box models such as the support vector machines (Cortez & 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 & Embrechts (2013) and selects a sample from the input data used to train the model for assessing model sensitivity to several inputs variation at the same time. This method has been extensively used in several distinct domains such as bank marketing (Moro et al., 2015) and geotechnical engineering (Tinoco et al., 2014).

For all the experiments, the R statistical tool was adopted. It provides a free and open source environment easy to use for data manipulation. Furthermore, it has a worldwide community of users that make available a huge number of packages at the CRAN repository (Comprehensive R Archive Network). Moreover, the “rminer” package was chosen for it has implemented a short set of functions that can be easily parameterized for conducting a wide range of data mining experiments (Cortez, 2010). Such functions include modeling using support vector machines and sensitivity analysis, both methods applied in this research.

3. Material and methods

3.1 Dataset

The proposed approach was tested using the case of a worldwide renowned cosmetic brand. A dataset was collected including 790 posts published by this company between the 1st of January and the 31th of December 2014 in its Facebook page. 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 (Digital Insights, 2014).

The dataset compiled contains information referring to the identification of the publication, categorization and performance, as shown on Table 1.

Table 1 – Features from the dataset compiled.

Feature	Type of information	Source	Data type
Posted	Identification	Facebook	Date/time
Permanent link	Identification	Facebook	Text
Post ID	Identification	Facebook	Text
Post message	Identification	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	Performance	Facebook	Numeric
Lifetime Post Total Impressions	Performance	Facebook	Numeric
Lifetime Engaged Users	Performance	Facebook	Numeric
Lifetime Post Consumers	Performance	Facebook	Numeric
Lifetime Post Consumptions	Performance	Facebook	Numeric
Lifetime Post Impressions by people who have liked your Page	Performance	Facebook	Numeric
Lifetime Post reach by people who like your Page	Performance	Facebook	Numeric
Lifetime People who have liked your Page and engaged with your post	Performance	Facebook	Numeric
Comments	Performance	Facebook	Numeric
Likes	Performance	Facebook	Numeric
Shares	Performance	Facebook	Numeric
Total interactions	Performance	Computed	Numeric

Most of the information is exported directly from the company's Facebook page. The exceptions are the "total interactions" and the "category". The former represents a column computed based on the performance metrics exported from Facebook. The latter is the only column created manually by the Facebook page managers. It provides a manual categorization according to the content. An experienced professional in social media

validated this categorization for all the 790 posts included to minimize the risk of this procedure.

3.2 Data mining

A data mining approach typically includes phases such as data understanding, data preparation, modeling and evaluation (Han et al., 2011). The dataset described in Table 1 includes twelve features (eleven exported from Facebook plus the computed total interactions) that may be used to measure performance, thus any of them can be used as an output to predict. Therefore, the procedure included modeling each of those twelve features to assess which ones can be better predicted. The meaning of each of those features is detailed in table 3**.** The seven remaining features are known prior to the post publication and can be used as an input. However, three of them are unique per post: the permanent link, the post ID and the post message itself; therefore, such features are of no value to modeling, considering that those do not represent any type of relationship between posts. One could argue that the posted date is also 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 previous three (excluding the Posted date-time value since it is distinct for each post) provides a dataset with seven distinct input features for feeding the model (table 2).

Table 2 – 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).

Table 3 – List of output features to be modeled.

Feature	Description [1]
Lifetime Post Total Reach	The number of people who saw your Page post. (Unique users).
Lifetime Post Total	Impressions are the number of times a post from your Page

[1] Descriptions extracted from:
- http://www.agorapulse.com/blog/facebook-reach-metrics-ultimate-guide
- https://www.facebook.com/help/274400362581037

Impressions	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 their friend shares it.
Lifetime Engaged Users	The number of people who clicked anywhere in your posts. (Unique users).
Lifetime Post Consumers	The number of people who clicked anywhere in your post.
Lifetime Post Consumptions	The number of clicks anywhere in your post.
Lifetime Post Impressions by people who have liked your Page	Total number of impressions just from people who have liked your page.
Lifetime Post reach by people who like your Page	The number of people who saw you Page post because they have liked your page (Unique users).
Lifetime People who have liked your Page and engaged with your post	The number of people who have liked your Page and clicked anywhere in your posts (Unique users).
Comments	Number of commentaries 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" on the post.

Figure 1 exhibits the data mining procedure undertaken for implementing the model, for validating the results, and for extracting useful knowledge for leveraging posts publications decisions. Such procedure was executed twelve times, for evaluating the predictive performance of every output feature available.

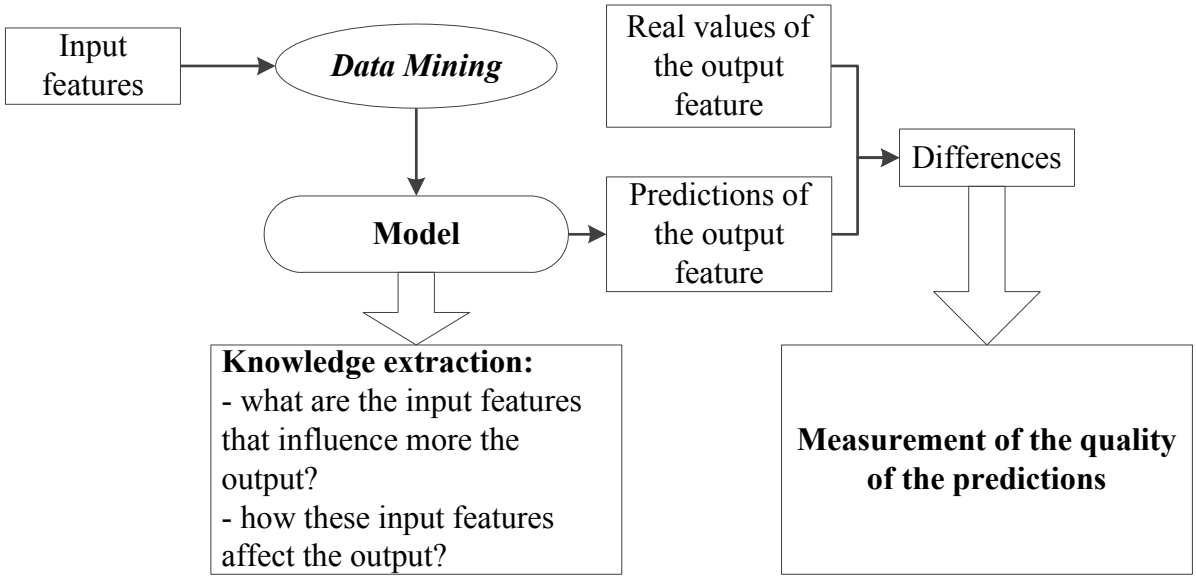

Figure 1 – Data mining procedure

First, the data mining algorithm chosen is fed with the seven input features from Table 2, resulting in a model. Such model is 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 are compared to assess model performance. The model of the performance metric that can better be modeled, i.e., in which the model predictions show less differences to the real values is then assessed to understand how input features influence this performance metric. In Section 3.2., further details are provided on the specific data mining techniques employed.

4. Experiments and results

4.1. Prediction

As stated in Section 2, the seven input features from table 2 were used for predicting each of the twelve performance metric features described in table 3. 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. The Shapiro-Wilk test was used for assessing that each of the columns for the features to be predicted followed a normal distribution (Razali & Wah, 2011). Such a validation provided the ground needed to remove 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 a given performance metric, the results were evaluated by comparing the real values for that metric with the value predicted by the model. A good model implies that it fits all the input data in a way that the predicted values are as close as possible to the real values. Table 4 shows four randomly selected examples of the information of four of the posts used to feed the model at the seven left columns, while the "Real" provides the "Lifetime Post Consumers" real value. The three last columns are used to validate the model: the "Predicted" shows the value predicted by the model, the "Absolute difference" shows the difference to the real value while the last column shows the percentage difference.

Table 4 – 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	Mon	No	356	346	10	3%
Inspiration	135,617	Photo	Sep	10	Wed	No	614	520	94	15%
Product	139,441	Status	Dec	3	Thu	No	1,407	1502	95	7%

The four examples shown in Table 4 attest that the model is valuable as a predictive tool to support the company's Facebook page manager's decisions whether or not to publish a given post, considering the manager can have a perception of the impact that the post will have on the future.

The differences for the 751 posts were obtained similarly to the Table 4 for each of the twelve performance features. For evaluating the overall prediction results, the mean absolute percentage error was computed for every of the twelve models (Hyndman & Koehler, 2006). These results are shown on table 5, with the two best results (lowest mean absolute percentage error) highlighted in gray.

Table 5 – Results for performance metrics predictions.

Performance metric	Mean absolute percentage error
Lifetime Post Total Reach	49.6%
Lifetime Post Total Impressions	69.3%
Lifetime Engaged Users	28.8%
Lifetime Post Consumers	27.2%
Lifetime Post Consumptions	33.1%
Lifetime Post Impressions by people who have liked your Page	47.8%
Lifetime Post reach by people who like your Page	37.5%
Lifetime People who have liked your Page and engaged with your post	26.9%
Comments	63.9%
Likes	41.2%
Shares	35.8%

Both the models for the "Lifetime People who have liked your Page and engaged with your 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). This is a low difference, implying that both models can provide a good perception of the impact a given post can have, supporting page manager's decision. The examples provided in Table 4 for "Lifetime Post Consumers" illustrate how the 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", "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, etc.). The precursory work of Moro et al. (2014) on context features suggests that enriching the input features' set with these additional context features may possibly have a huge impact in the model's performance, improving the results achieved.

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 your Page and engaged with your post" and the "Lifetime Post Consumers". The "Lifetime Post Consumers" provides a more interesting metric for the decision of publishing the post, as it focus solely on the impact of the post, while the "Lifetime People who have liked your Page and engaged with your post" contains 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, the data-based sensitivity analysis was adopted in two complementary approaches: first, the model is assessed to understand which of the input features affect more the outcome of the studied metric; finally, all input features from the most to the least relevant for the model are assessed to discern how each of them influences the outcome.

Figure 2 shows the contribution of each input feature for the model of "Lifetime Post Consumers". The relevance of the "Type" of content published is notorious: it accounts 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. 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 account for almost 70% of the relevance to the model.

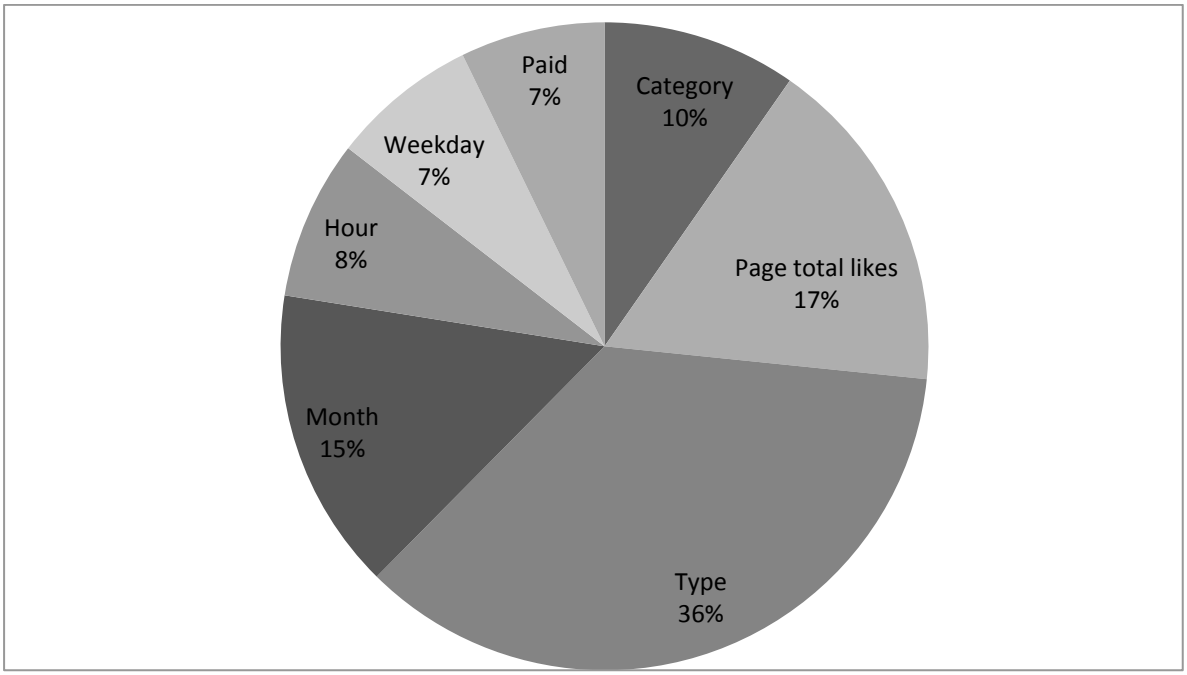

Figure 2 – Relevance of the input features for "Lifetime Post Consumers".

The "Category" set by the Page manager has 10% of relevance while the "Hour" and "Weekday" the post was published account for 8% and 7% of relevance, respectively. Finally, the feature that indicates the company paid for the page to be specifically advertised appears with just 7% of relevance. This an interesting result, for it 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 a model considered with good prediction performance, it is interesting to observe how each of the types influences this output metric. Figure 3 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". Interestingly, 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. On the other hand, the work published by Kwok & Yu (2013) while achieving a

similar conclusion for "Status" posts, achieved a different conclusion for "Photos", stating that these received more likes and comments than "Links" and "Videos".

The second most relevant features, "Page total likes", is by far less relevant than "Type"; nevertheless, it still has an influence of 17%. This input feature relates to the likes the company's page where the post is being published has at the moment of publishing the post. Therefore, the conclusions drawn should account that this is a variable that is always increasing along time. Nevertheless, Figure 4 shows that "Lifetime Post Consumers" decreased after reaching a peak of around 95000 page likes, perhaps disclosing some erosion of the company's Facebook page.

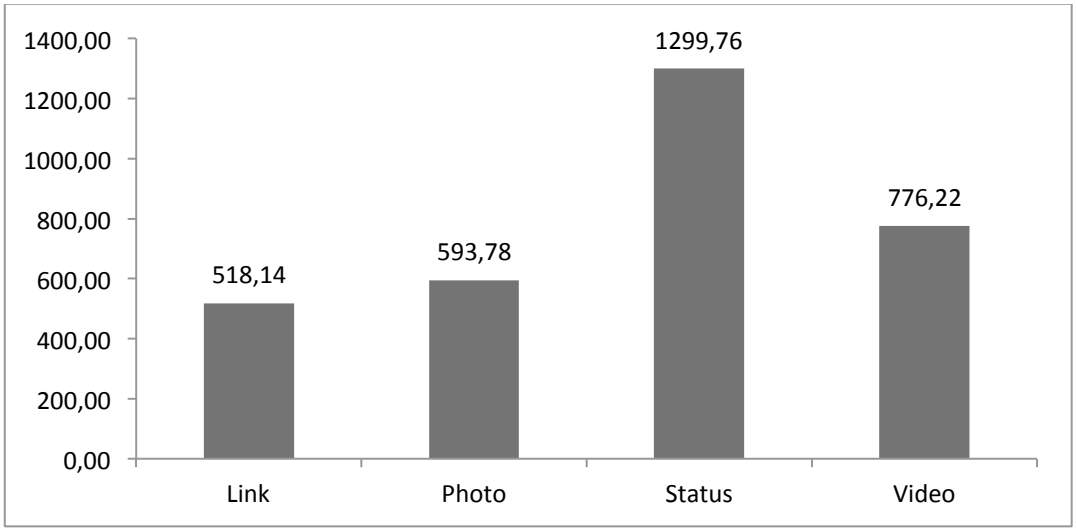

Figure 3 – Influence of "Type" on "Lifetime Post Consumers".

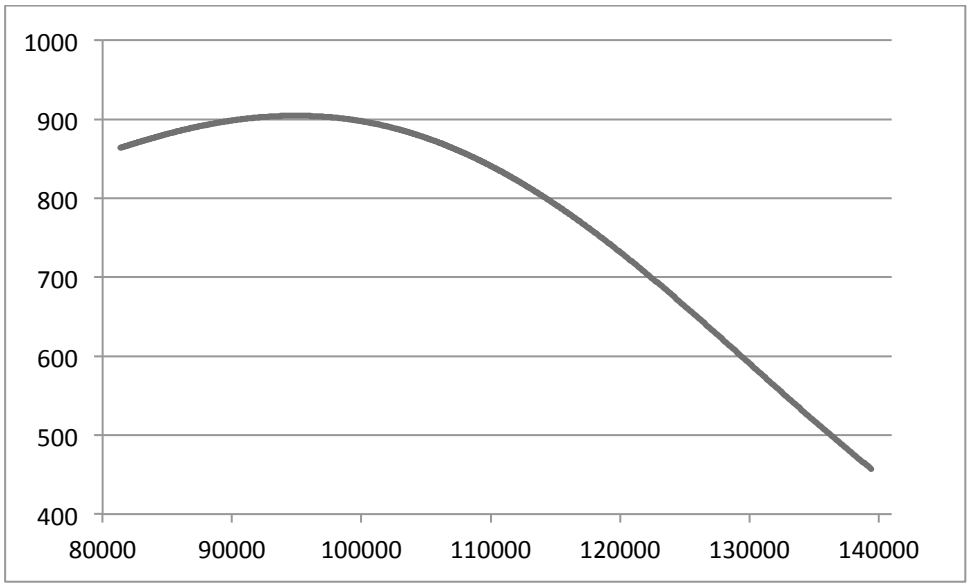

Figure 4 – Influence of "Page total likes" on "Lifetime Post Consumers".

The "Month" is the third most relevant feature, with 15% of influence. Figure 5 displays some seasonality, with a large increase starting in April and reaching a peak in June of almost the twice the value of April. From November to February there is also a steady although not so steep increase. While the present study analyzed only a year, previous work of Golder et al. (2007), which analyzed Facebook messages sent between February 2004 and March 2006, a period of 26 months, confirms the trends in Figure 5: the March to June large increase is also observed, 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 Figure 5.

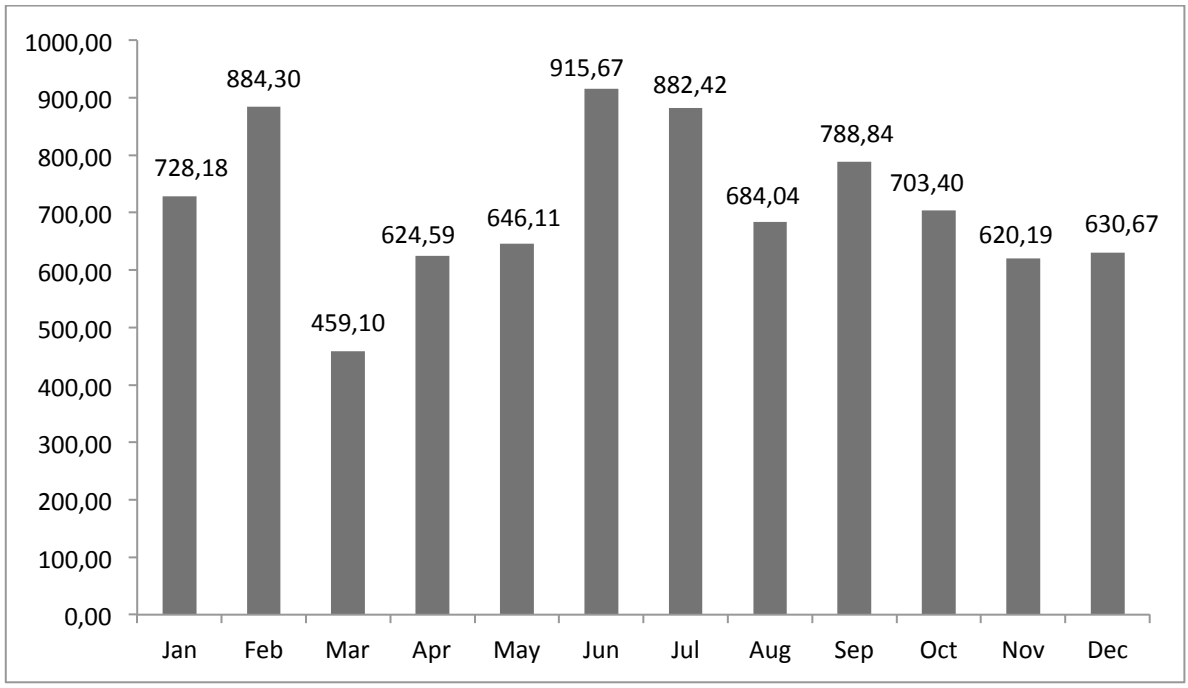

Figure 5 – Influence of "Month" on "Lifetime Post Consumers".

The influence of the remaining four least relevant features which conceal 32% of the model's knowledge is displayed in Figure 6 for the "Category", Figure 7 for "Hour", Figure 8 for "Weekday", and finally Figure 9 for "Paid". Regarding "Category", it is notorious the influence that "Action" has when compared to the remaining two. This "Actions" category stands for special offers and contests, clearly gathering more attention than "Products" and non explicit brand related contents ("Inspiration"). The "Hour" influence graph appears to show that it does not exist any trend associated with the hour that the post is published, 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 do not report a trend for the "Monday" local maximum and then decrease observed in Figure 8. The result shown for "Paid" is expected: a post for which the company paid for advertising has a larger impact that a post not paid. Nevertheless, this is one of the least relevant input features for the model defined, with just 7% of relevance.

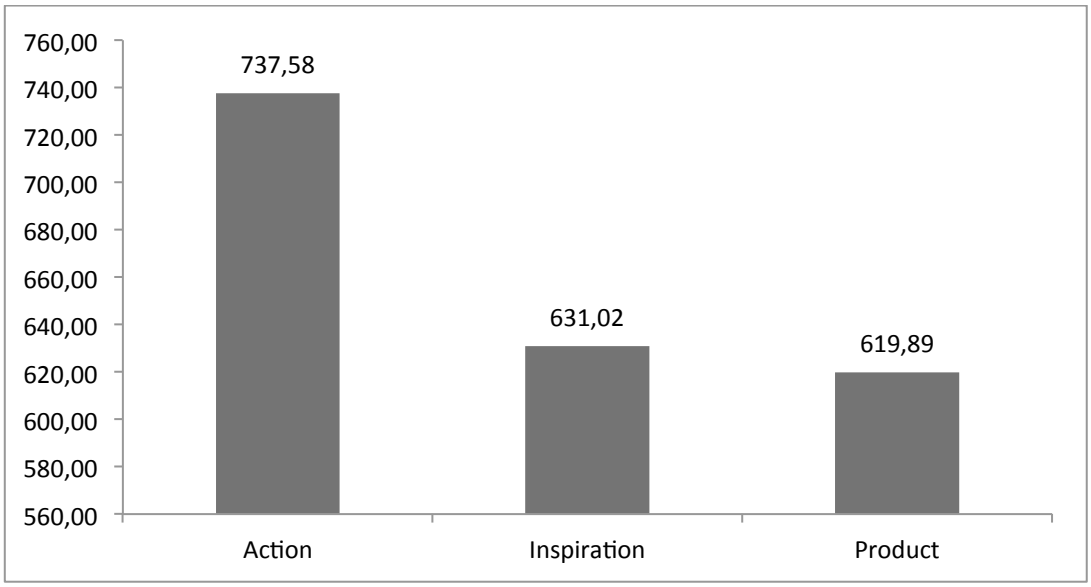

Figure 6 – Influence of "Category" on "Lifetime Post Consumers".

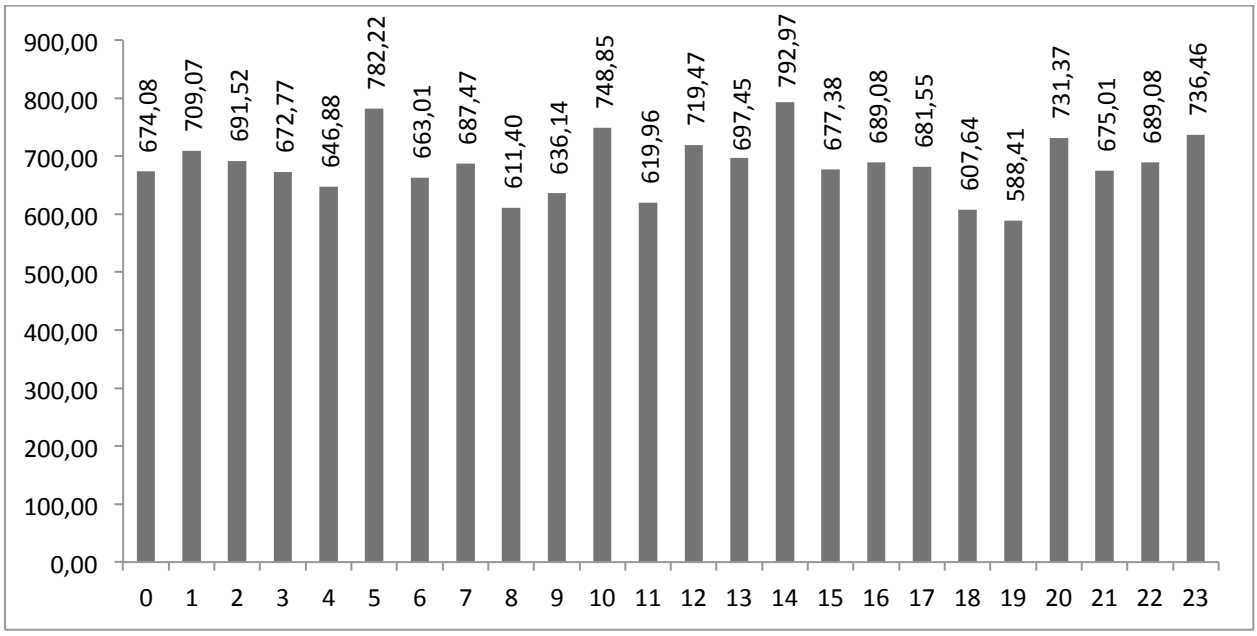

Figure 7 – Influence of "Hour" on "Lifetime Post Consumers".

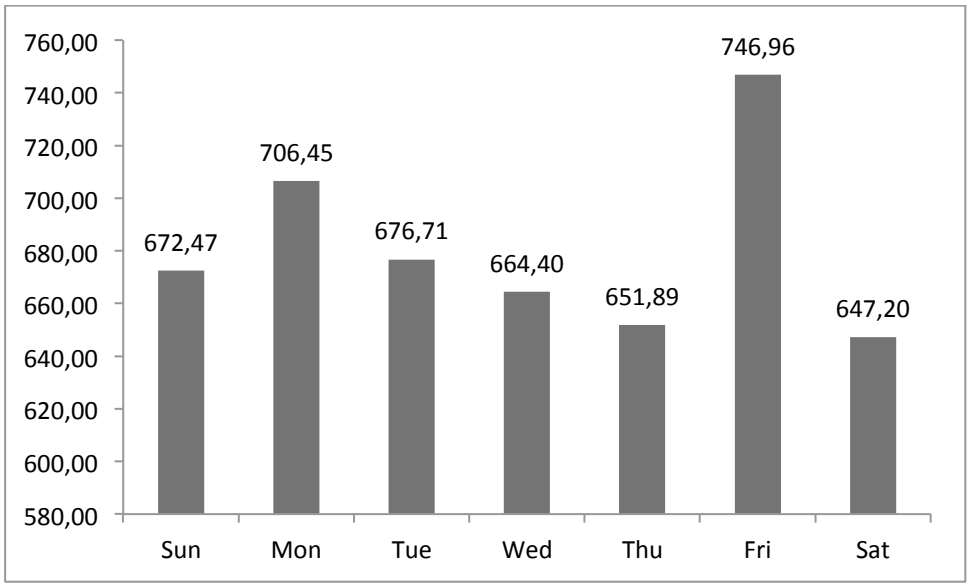

Figure 8 – Influence of "Weekday" on "Lifetime Post Consumers".

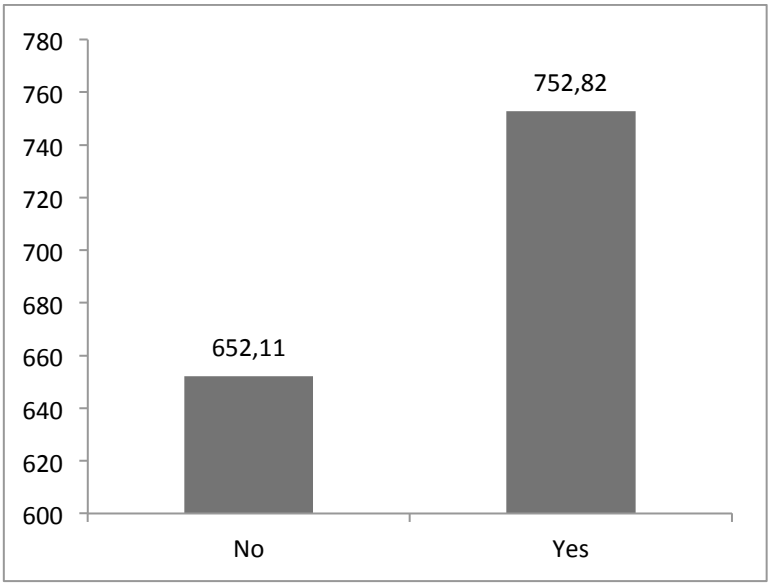

Figure 9 – Influence of "Paid" on "Lifetime Post Consumers".

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 your Page and engaged with your post" output features, with a mean absolute percentage error of 27.2% and of 26.9%, respectively. These differences can be considered low, implying that both models are useful for providing support to the page manager's decision on whether or not to publish a given post. The procedure adopted may potentially benefit other companies and pages, considering the models were built using only seven input features mostly derived from Facebook.

For extracting valuable knowledge from the model achieved for "Lifetime Post Consumers", the databased sensitivity analysis was applied. 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 is published. Both discoveries are aligned with previous research, now applied to the cosmetic industry, proving that these two attributes cross the type of industry. Publications related to special offers and contests are likely to produce posts with greater impact than "Product" and other non explicit brand related contents. 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 work 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. Furthermore, text mining the comments of each post for user sentiment analysis could reveal the feelings each post is generating.

6. References

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.

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. doi:10.1109/PASSAT/SocialCom.2011.21

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

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17. doi: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.

Deloitte Digital (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

Digital Insights (2014). Social Media 2014 Statistics. Retrieved September 10, 2015, from http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html

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. doi: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. In Communities and technologies 2007 (pp. 41-66). Springer London. doi:10.1007/978-1-84628-905-7_3

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. doi:10.1287/mksc.2013.0830

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

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. doi:10.1007/BF02985802

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688. doi:10.1016/j.ijforecast.2006.03.001

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 53(1), 59-68. doi:10.1016/j.bushor.2009.09.003

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. doi: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. doi: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. doi:10.1016/j.chb.2012.04.016

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31. doi: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. doi:10.1007/s00521-014-1703-0

Statista Dossier (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/.

Steinwart, I., & Christmann, A. (2008). Support vector machines. Springer Science & Business Media.

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.

Tinoco, J., Correia, A. G., & Cortez, P. (2014). Support vector machines applied to uniaxial compressive strength prediction of jet grouting columns. Computers and Geotechnics, 55, 132-140. doi:10.1016/j.compgeo.2013.08.010

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. doi:10.1016/j.jbusres.2013.05.002

Turban, E., Sharda, R., Delen, D., & Efraim, T. (2011). Decision support and business intelligence systems, 9th Edition. Pearson.

Annexes

Annex A - Definitions

Post ID – Numeric permanent identification of each publication.

Permanent Link – Direct and permanent link to the publication.

Post message - The written text that goes with each publication.

Type – The kind of publication, it might be of four types: Photo, link, status or video.

Category – The kind of publication. This can be divided in: Inspirational (non explicit brand related content), product (direct advertisement, explicit brand content) or action (special offers and contests).

Posted – Date when the content was posted.

Impressions - Impressions are the number of times a post from your Page is displayed, whether the post is clicked or not.

Lifetime post total reach – The number of people who saw your Page post. (Unique users)[2].

[2] http://www.agorapulse.com/blog/facebook-reach-metrics-ultimate-guide

Lifetime post total impressions – Impressions are the number of times a post from your 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 their friend shares it. [3]

Lifetime engaged users – The number of people who clicked anywhere in your posts. (Unique Users)

Lifetime post consumers – The number of people who clicked anywhere in your post. Clicks generating stories are included in "Other Clicks." (Unique Users)[4]

Lifetime post consumptions - The number of clicks anywhere in your post. Clicks generating stories are included in "Other Clicks." (Total Count) [5]

Lifetime post impressions by people who have liked your page– Total number of impressions just from people who have liked your page.

Lifetime post reach by people who like your page – The number of people who saw you Page post because they have liked your page. (Unique users)

Lifetime people who have liked your page and engaged with your post – The number of people who clicked anywhere in your posts, within the users who "like" the page. (Unique Users)

[3] https://www.facebook.com/help/274400362581037
[4] http://www.agorapulse.com/blog/facebook-post-consumers-and-post-consumption
[5] http://www.agorapulse.com/blog/facebook-post-consumers-and-post-consumption

Comments – Number of commentaries 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" on the post.

Reached consumers – All users that were exposed to the content.

Passive consumers – All users that just read the contents but do not interact with it generating content (liking, commenting or sharing).

Active consumers – All users that interacted with the content.