THE ROLE OF DATA MININIG IN CREATING SOCIAL MEDIA MARKETING STRATEGY

Sara Gračić[1]

Abstract: Social media are an important part of state-of-the-art marketing strategy, because of higher profits, positive reputation via Word of Mouth marketing and abundance of user-generated data. Therefore, implementing data mining enables acquisition of new knowledge regarding customers and their behavioral patterns. To determine influence of different algorithms on creating marketing strategy, data set containing Facebook posts was used for creating predictive models in Weka 3.7. All classifiers that work with categorical and numerical input and output attributes were used. Regardless of variations of different parameters and test methods, all models had relative absolute error of 57% and up. The "best worst" results were achieved using SMOreg classifier. These models cannot be used for creating marketing strategy, but if access to all data was possible and if SVM had been used, this prediction error could have been lower and then the built models would be applicable in practice.

Key words: social media, data mining, classification, strategy, knowledge, marketing

1. INTRODUCTION

Social media is the new "Wild West" of marketing where organizations "attack" individuals in order to get into the public's focus and build their communities in virtual space. From platforms to services, social media influences people one on one and creates P2P (person-to-person) communication that affects people's awareness, acceptance and behavior. As a powerful tactics and a communication tool, social networks can and should play an important role in branding, preservation and business protection strategies, according to Weinstein (2011).

Social media are becoming an important part of state-of-the-art marketing strategy, because of multiple benefits for a company, e.g. increased product sales and, consequently, higher profits, as well as positive reputation via social media users and Word of Mouth marketing. However, if companies do not know how to do business on this market "created" by social media users, they can suffer great loses.

On the other hand, social media is also an indispensable source of information about customers and their habits, so implementation of data mining can enable acquisition of new knowledge regarding customers and their behavioral patterns. Therefore, it is very important for managers to adopt this approach and make decisions in accordance with the guidelines provided by data mining results. By analyzing data related to the effects of content posted on social media profiles, companies and marketing agencies can become familiar with the preferred content of their visitors. This enables them to get to know their visitors and optimize their content on social media according to their visitors' preferences and predict the effects of optimized content for longer period of time.

After the Introduction, the rest of the paper is structured in the following order. Section 2 gives research background regarding social media in business; Section 3 presents focus of this research and its objective, as well as Facebook performance metrics data set; Section 4 gives insight in created models and compares the obtained results with findings of the authors Moro, Rita & Vala in 2016. Section 5 discusses built models and Section 6 concludes the paper.

2. RESEARCH BACKGROUND

Although the number of social media users is increasing, thus providing companies with the opportunity to enlarge their virtual communities, many companies, despite their increased investments

[1] PhD student, University of Novi Sad, Faculty of Economics, Subotica, Segedinski put 9-11, email:saritta4u@gmail.com

in social media efforts, do not make any return on investment as Fertik (2014) points out. This indicates that a major problem lies in not understanding how social media function (Gračić, 2017).

On the other hand, this market type provides large amount of data that companies should carefully analyze, because of valuable first-hand information from their consumers. Previously, these analyses were very problematic, but with the development of analytics software, artificial intelligence and hardware, knowledge discovery can be performed without difficulties.

Culnan, McHugh & Zubillaga (2010) emphasize that social media enable creation of virtual customer environments (VCEs) around companies, brands or products, which companies need to take into consideration. Otherwise, they risk losing many opportunities offered by social media.

Gensler, Völckner, Egger, Fischbach & Schoder (2015) point out that consumer-generated product reviews have created colossal amount of data and gave the opportunity for companies to "listen" to their consumers. Their approach of combining text mining and network analysis can reveal strengths and weaknesses of the brand image, so managers could take necessary actions to use the strengths and minimize or eliminate weaknesses.

Based on a small-scale UK study performed by Griffiths & McLean in 2015, the authors concluded that, although some companies understand the importance of "real customer conversations", only a few have actually adopted the "human brand" approach, and even fewer are focused on strategic communication on social media.

Habibi, Laroche & Richard conducted a research in 2014, which showed that 3 out of 4 relationships positively influence brand trust, while customer-to-other-consumer relationships have a negative influence on brand trust. Community engagement amplifies the strength of relationships that consumers have with the brand community elements and there is a translation of this relationship on brand trust.

Hudson, Huang, Roth & Madden (2015) point out in their research that companies allocate more resources from marketing budgets for social media campaigns. Their 3 studies have shown that the use of social media is positively related with the brand relationship quality and the effect was more pronounced with high perceptions of anthropomorphism.

Hutter, Hautz, Dennhardt & Füller analyzed the activities on a Facebook page of a car manufacturer in 2013. The results showed a positive effect of user engagement on fan-page on consumers' brand awareness, word of mouth (VOM) and purchase intentions. On the other hand, irritation with fan-page, due to information overload, leads to negative effects on fan-page commitment and reduced VOM activities.

Laroche, Habibi, Richard & Sankaranarayanan were investigating brand communities in 2012. The results showed that brand communities established on social media have positive effects on community markers, which have positive influence on value creation practices, thus improving brand loyalty.

3. MATERIALS AND METHODS

It is clear that social media, as a newly-created market by consumers, represent a virtually inexhaustible source of information regarding customers and their habits.

3.1. Research focus and objective

The focus of this research is to examine how companies use data mining to become familiar with their consumers' preferences and optimize their content on social media according to consumers' needs as a part of their social media marketing strategy. The objective of this paper is to determine how different data mining algorithms behave when analyzing user-generated data on social media and

whether developed models can be used for giving recommendation during social media marketing strategy development. To determine how different data mining algorithms influence on creation of social media marketing strategy, the author of this paper has downloaded Facebook performance metrics data set. Data analyses were performed using Excel 2010. Algorithms in Weka 3.7 were used for creation of models for giving recommendation during social media marketing strategy development. Results obtained by the author of this paper will also be compared with the results obtained by Moro et al. (2016).

3.2. Data set in focus of this research

The Facebook performance metrics data set is in focus of this research. It was created by Moro et al. (2016). User-generated data include 790 posts from a cosmetic company's Facebook page from January 1st to December 31st, 2014. However, the authors of this data set made only 500 posts publicly available to eliminate privacy issues. Out of 500 posts, one post has a missing value, related to add payment to Facebook – it is unclear whether the ad was paid or not. Moro et al. (2016) used this data set for building models for giving recommendations for branding cosmetic products. Data set can be downloaded from https://archive.ics.uci.edu/ml/datasets/Facebook+metrics. The data set has 7 input and 12 output attributes.

The input attributes include:

- **Category** (action, product and inspiration),
- **Page total likes**,
- **Type** (status, photo, video and link),
- **Post month** (from January to December),
- **Post weekday** (from Monday to Sunday),
- **Post hour** (all hours except midnight) and
- **Paid** (yes and no).

These input attributes affect each of the twelve output attributes:

- **Lifetime post consumers,**
- **Lifetime post total reach,**
- **Lifetime post total impressions,**
- **Lifetime engaged users,**
- **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,**
- **Comment,**
- **Like,**
- **Share** and
- **Total interactions.**

All observed attributes are numeric, except **Type of publication**, which is categorical. For details regarding meaning of the above stated attributes, consult the work of Moro et al. (2016), which is available at: https://www.sciencedirect.com/science/article/pii/S0148296316000813.

4. DATA ANALYSIS, MODEL CREATION AND RESULTS COMPARISON

The purpose of this analysis of the data set is to compare the obtained results with the results presented in the paper by Moro et al. (2016), in order to determine how applying different algorithms influences on research results and development of social media marketing strategy, that is will

recommendations given by Moro et al. (2016) be similar to recommendations created after this research.

4.1. Comparison of the influence on attribute Lifetime post consumers

In order to compare the influence of 7 input attributes on Lifetime post consumers (LPC) output attribute, Excel 2010 was used. Influence I of an input attribute X, on a output attribut Y is calculated in the same way as Moro et al. (2016):

$$I = \frac{\sum \text{LFT post consumer (Y) for an input attribute X}}{\text{Count on the value of an input attribute X}} \quad (1)$$

The influence of the attribute **Paid** on LPC shows that paid ads have greater impact than unpaid ads, which is identical to the result obtained by Moro et al. (2016).

The influence of **Category** on LPC shows that inspiration has the smallest impact. Action has greater impact, while products have the biggest influence on consumers. However, Moro et al. (2016) determined that consumers show the highest interest in actions, i.e. Facebook posts that promote action sales of cosmetics.

When **Type** of publication is considered, by far, the highest influence on LPC have statuses posted by the company, followed by videos, but with drastic decline in impact, and then photos and links. Rankings are identical in results presented by Moro et al. (2016).

When **Post month** is investigated, the biggest influence have posts published in February, probably because of Valentine's Day. In research conducted by Moro et al. (2016) however, highest numbers appear during summer months, due to holiday season, followed by influence decline in February. In March LPC is the lowest. It is possible that flowers and chocolates are bought more than cosmetics, but this statement needs further research.

The influence of **Post hour** attribute on LPC shows that the lowest impact has content posted at 6, 8, 18 and 23 o'clock, and the highest, content posted at 21 o'clock. However, Moro et al. (2016) observed the relative uniformity of LPC per hour and it ranges from 600 to 700 people. The highest impact is achieved at 5, 10, 14 and 23 o'clock, while the lowest is achieved at 19 o'clock. It is high likely that the cosmetic company operates on several continents, so different time zones should be taken into consideration when developing social media strategy.

The influence of **Post weekday** attribute on LPC indicates that the lowest impact have posts created on Tuesdays and the greatest those created on Wednesdays and Mondays. In research of Moro et al. (2016), LPC was the highest on Fridays and the lowest on Saturdays and Sundays. However, a study conducted by Cvijikj, Spiegler & Michahelles in 2011 showed that the day of the week has nothing to do with performance metrics.

4.2. Comparison of the constructed classification models

Moro et al. (2016) analyzed the outliers for each of the 12 output attributes. They used the Shapiro-Wilk test to evaluate whether each of the output attributes follows a normal distribution of values. On this basis, 5% of posts from the original data set were excluded and 751 instances were used to build classification models. Models were built for all 12 output attributes using data mining technique support vector machine (SVM). Out of all 12 models, 2 with the lowest prediction error of around 27% were chosen for managers. These models refered to output attributes Lifetime post consumers and Lifetime people who have liked a page and engaged with a post. These results mean

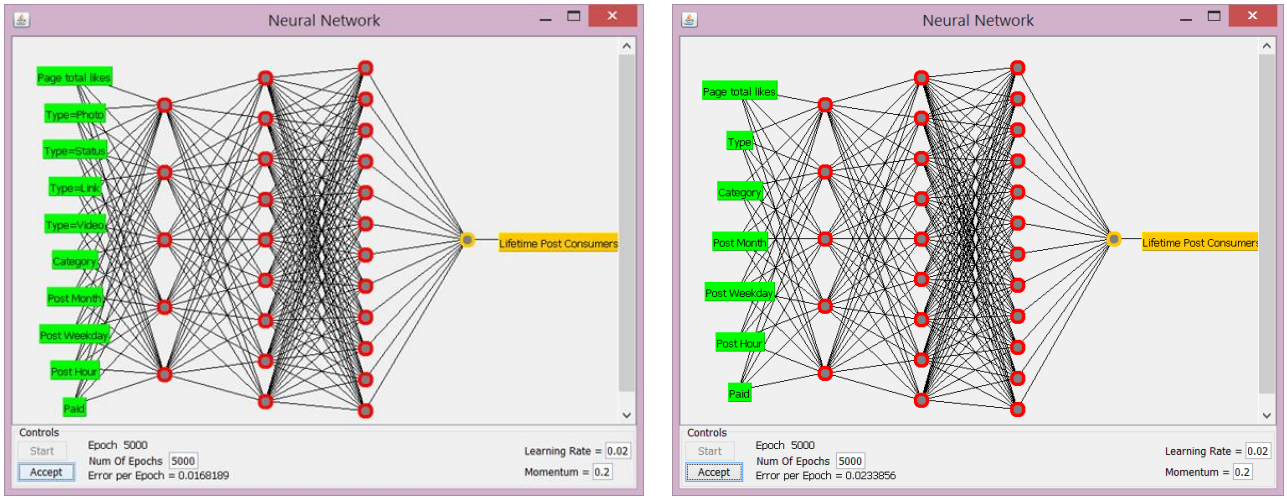
that a company's manager, when making a decision to post content on Facebook page, can predict the impact of that content with a 27% error.

The author of this paper used Weka 3.7 for rebuilding models based on the data set published by Moro et al. (2016). All classifiers that can handle categorical and numeric input and output attributes were used, while retaining their predefined values. The percentage split test method was selected, with the 80% -20% ratio (representing the number of instances for training and testing, respectively). All 500 instances were used and models were built for all 12 output attributes, using all 7 input attributes. All classifiers showed an extremely high classification error.

After applying all classifiers for creating the model for output attribute **Lifetime Post Total Reach**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 67.7871%. After applying all classifiers for creating the model for output attribute **Lifetime Post Total Impressions**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 62.4485%. After applying all classifiers for creating the model for output attribute **Lifetime Engaged Users**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 68.6134%. After applying all classifiers for creating the model for output attribute **Lifetime Post Consumptions**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 73.559%. After applying all classifiers for creating the model for output attribute **Lifetime Post Impressions by people who have liked your Page**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 58.3089%. After applying all classifiers for creating the model for output attribute **Lifetime Post reach by people who like your Page**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 67.5557%.

After applying all classifiers for creating the model for output attribute **Lifetime People who have liked your Page and engaged with your post**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 58.0834%. After applying all classifiers for creating the model for output attribute **Comment**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 57.7214%. After applying all classifiers for creating the model for output attribute **Like**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 60.6259%. After applying all classifiers for creating the model for output attribute **Share**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 67.6568%. After applying all classifiers for creating the model for output attribute **Total interactions**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 63.1041%. After applying all classifiers for creating the model for output attribute **Lifetime Post Consumers**, SMOreg classifier showed the best performance, with classification error (relative absolute error) of 67.5695%.

All classifiers were first applied on the original data set available on the Internet. After that, input attribute **Type** was transformed from nominal to binary, to determine whether this change would decrease classification error. 4 new dummy variables were created (Picture 1). When classifiers were applied, no improvements were achieved; classification error was still over 50%. Applying another test method cross-validation and changing the parameters did not result in any improvements either. After all experiments were finished, SMOreg classifier in Weka 3.7 showed the "best worst" results for all 12 output attributes and model Comment had the lowest relative absolute error of 57,7214%. Unfortunately, none of the created models can be applied in practice due to high relative absolute error.



Picture 1: Neural Network with dummy variables (left) and without dummy variables (right)

5. DISCUSSION

Classification models built by Moro et al. (2016) and the models built and presented in this paper vary greatly. All algorithms in Weka 3.7, which can perform classification with numeric and categorical input and output attributes, showed extremely poor performance for all 12 output attributes, as their relative absolute error reached extremely high percentages, over 50%. This means that managers of the cosmetic company could not use these models (due to high classification errors) for more reliable decision-making when creating marketing strategy, because they would have more than 50% chance of failing in the process.

"The best of the worst" classification models were created with SMOreg classifier, whose relative absolute error span was from 57.7214% for the output attribute Comment up to 73.559% for the output attribute Lifetime Post Consumptions.

Differences in the scope of the used data set (66.58% of the initial data set used by Moro et al. (2016)), differences in attribute values, unclear meaning of certain values (e.g. does 1 mean that an add was paid or does it mean that an add was not paid) and the differences in the applied algorithms have resulted not only in the drastic difference in the relative absolute errors, but have also influenced on model choice: the model based on the output attribute Comment (with classification error of 57.72%) was chosen as the best-worst, as opposed to the model based on the output attribute LPC (with classification error of 27%) chosen by the authors Moro et al. (2016) as the most suitable for classification.

6. CONCLUSION

Social media are an important part of state-of-the-art marketing strategy, so, implementing data mining enables acquisition of new knowledge regarding customers and their behavioral patterns. To determine influence of different algorithms on creating marketing strategy, data set containing Facebook posts was used for creating predictive models in Weka 3.7. All classifiers that work with categorical and numerical input and output attributes were used. Regardless of variations of different parameters and test methods, all models had relative absolute error of 57% and up. The "best worst" results were achieved using SMOreg classifier.

Built models cannot be used for creating marketing strategy, but if access to all data was possible and if SVM had been used, this prediction error could have been lower and then the built models

would be applicable in practice, thus matching the pace of customers and the pace of the company, to get the best of efforts invested in social media marketing strategy.

7. REFERENCES

[1] Cohen, H. (2011, May 11). Social Media Definitions. Retrieved April 5, 2017, from http://heidicohen.com/social-media-definition/

[2] 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.

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

[4] Fertik, M. (2014, August 21). Is Social Media Worth It For Small Businesses? Retrieved April 5, 2017, from http://www.forbes.com/sites/michaelfertik/2014/08/21/is-social-media-worth-it-for-small-businesses/#a0efe34cef48

[5] 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.

[6] Gračić, S. (2017). *Kreiranje zajednica u virtuelnom prostoru „divljeg zapada“ radi postizanja svesnosti, prihvatanja, ponašanja i profita. XXII International Scientific Conference SM 2017, Strategic Management and Decision Support Systems in Strategic Management* (pages. 577-585). Subotica: Faculty of Economics

[7] 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.

[8] 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.

[9] 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, 33*, 27-41.

[10] 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.

[11] 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.

[12] Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research 69*, 3341-3351