Online Consumer Behaviour in Social Media Post Types: A Data Mining Approach

Dimitrios Gkikas
University of Patras
Theodoros Theodoridis
University of Salford
Prokopis Theodoridis
University of Patras
Androniki Kavoura
Prof. University of West Attica

Cite as:
Gkikas Dimitrios, Theodoridis Theodoros, Theodoridis Prokopis, Kavoura Androniki (2020), Online Consumer Behaviour in Social Media Post Types: A Data Mining Approach. *Proceedings of the European Marketing Academy*, 49th, (63455)

Paper from the 49th Annual EMAC Conference, Budapest, May 26-29, 2020.

Online Consumer Behaviour in Social Media Post Types: A Data Mining Approach

Abstract.

Our research focuses in justifying the performance of different types of social media posts extracted from real posts in fashion and cosmetics Facebook business pages after a live video was introduced as a new posting type. The data we used include posts of a different nature like video, photos, statuses, and links. User engagement metrics consist of comments, shares, and reactions. The dataset is analysed through a study of the averages of the different engagement metrics for different timeframes. We applied machine learning and data mining classification techniques on benchmarked dataset using the WEKA platform to highlight a variety of reactions on different status posts. Finally, we present the classified posts performances upon several status posts and users' reactions. We hope that our research will reveal to decision makers, marketers and managers valuable information incorporating new social media strategy for leveraging their fashion businesses.

Keywords: consumer behavior, Facebook metrics, decision trees

Track: Digital Marketing & Social Media

1. Introduction

Nowadays social media platforms play an essential role to people's lives. People express their feelings by sharing, reacting and commenting their or other people's experiences and behaviors on social media. Apart from people, businesses also take advantage of these mediums for marketing purposes. With 2.45 billion monthly active users, as of the third quarter of 2019, Facebook daily generates more than 4 petabytes of data (Statista, 2019). Under these circumstances data scientists and marketers try to gather and analyse all that information in order to extract marketing rules that will benefit companies and consumers. Due to this highly internet accessibility of companies worldwide, customers manage to purchase products or services from more than one seller at the lowest price, exchanging ideas and reviewing products and services. Thus, companies struggle to compete each other and eventually survive (Kaplan & Haenlein, 2010; Kim, Kim, & Lee, 2000). This new internet era made companies start thinking the potential of using this new information coming from using social media platforms to promote and increase customers and profits. Among others, social media marketing gave the opportunity to marketers reach many potential customers, study the relationships occurred during this processes and extract marketing rules. However, due to high volume of data and correlations generated, data scientists and marketing managers focused on statistically studying the relations between social media posts and the impact on the online users' behaviour. Although, there is still lack of substantial efforts to create sophisticated customer behaviour prediction systems (Kaur et al., 2019). Since branding is highly affected by the proper use of different types of social media posts marketing managers could be benefited from the use of such research and be able to highly predict any potential increase of sales or customers (Edosomwan et al., 2011). In our paper an artificial intelligence through machine learning and data mining techniques in online customer behavior prediction for online sellers is presented. We provide information about which social media statuses generate more reactions from the users after Facebook Live was introduced. Our goal is to help the marketers to understand the behavior of potential and current clients to regenerate a social media strategy based on different type of post (Chen et al. 2012).

2. Related Work

There is an interesting number of researches, which point out the major principles in social media consumer behavior providing prediction evidences mostly based on text mining and sentiment analysis. Text data mining refer to extraction of useful information out of textual databases using machine learning techniques such as natural language processing. Text mining

has been used social media networks to predict consumer behavior based on likes, shares, comments and reactions. Comments refer to written texts thus sentiment analysis converts it into “emotions/opinions” (Akaichi, 2013; Ortigosa et al., 2014; Pavaloaia et al., 2019; Clos et al., 2019). Likes, shares and reactions refer to quantity facts measured and processed such as positives, negatives or neutral behaviors (Kaur et al., 2019; Kim & Yang, 2017; Zell & Moeller, 2018). One significant research approach using data mining for predicting the performance metrics of posts published in brands' Facebook pages generated a decision process flow model, which may be used to support manager's decisions on whether to publish a post (Moro et al., 2016). Facebook user behavior has also been examined and classified based on content relevancy in order to avoid seeing irrelevant advertisements (Forouzandeh et al., 2014). The variability of consumer engagement was also analysed highlighting the changes induced by the use of Facebook Live videos revealing higher emotional and cognitive engagement with the live videos compared to other types of posts (Dehouche & Wongkitrungrueng, 2018). However, the number of articles using decision trees combined with social media remain scarce.

3. Theory and Data Analysis

3.1. Decision Trees

Data mining is a process of valuable information extraction using a set of analytical tools and algorithms, which provide data correlations very useful for decision making and predictions. Most commonly used techniques in data mining are decision trees (Theodoridis et al., 2010), genetic algorithms (Agapitos et al., 2011), neural networks, association rules, clusters and logistic regression. Data mining usage serves the need of discovering new knowledge generating new data connections and correlations and revealing interesting new patterns (Karim & Rahman, 2013; Ling & Li, 1998). A decision tree is used for data classification as follows: it reads data, separates classes, and sets values to each class. Usually, decision trees follow the rules of the sequence if – then – else. Datasets are sets of attributes and each attribute can have properties and several instances. Attributes types of data can be numeric and nominal. A decision tree consists of nodes, lines, branches and leaves. The nodes represent the attributes of the dataset; the branches represent attributes values. The first node is on the top is a super-class, the leave nodes are sub-classes. The examples are divided into the training, the validation, and the testing sets. After training the algorithm with the biggest set of examples (training set) the hypothesis will be generated, and the percentage of validations sets’ correctly classified examples is calculated. This procedure is repeated for as long as the size of

the training set changes. The testing set is used to validate the outcome of the precious procedure with completely new data. Overfitting occurs when the training of the tree reaches a point where it has no data to continue the classification it starts guessing values, it is also possible when going deep noise will occur which will affect the classification process; thus, tree pruning techniques are used. In each run of the algorithm a different small fraction of data from the entire example will be selected to validate the results thus is called the validation set and these number of tests are called folds. In our research we used ten-fold cross-validation and the dataset is randomly split into ten equal sized subsets. Then ten separate experiments will be run each of them using one subset for testing and the other nine subsets for training. In order to calculate the success rate of the data mining method, one should then calculate the average success rate of the ten experiments (Kohavi, 1995; Quinlan, 1986; Riasi and Wang, 2016). The Shannon function reflects the possibility of being less gained information when it is already known that some outcomes are more likely to occur than others. We list some of the C4.5 (the successor of ID3) decision trees mathematical formulas:

$$Information \quad = \sum_{i=1}^{N} ln_2(p_i) bits \tag{1}$$

$$Gain(X, A) = I(X) - \sum_{n \epsilon values(A)} \quad \frac{|X_v|}{|X|} I(X_v) \tag{2}$$

$$Splitlnf(X, A) = -\sum_{n \epsilon values(A)} \quad \frac{|X_v|}{|X|} log_2 \frac{|X_v|}{|X|} \tag{3}$$

$$GainRatio(X, A) = \frac{Gain(X,A)}{Splitlnf(X,A)} \tag{4}$$

3.2. Dataset Analysis

The dataset processed in this article is retrieved from the UCI machine learning repository and it consists of ten fashion and cosmetics retail sellers' Fakebook business pages (Dehouche & Wongkitrungrueng, 2018). The dataset provides data for different Facebook status posts including video, photos, statuses, and links. The engagement metrics consists of comments, shares, and reactions. The dataset consists of 7051 instances, 9 numeric attributes useful for performance analysis, and 3 categorical attributes useful for classification and prediction. In this section, we provide the preprocess analysis of the dataset which took place using the

Preprocess WEKA tool. Table 1 demonstrates the “Attributes” preprocess features. Figure 1 demonstrates the attribute types by a visualisation of occurrences.

Attributes	Type	Labels	Mean	Standard Deviation
status_id	Nominal	6997, Unique 6944-98%	-	-
status_type	Nominal	Video, Photo, Link, Status	-	-
status_published	Nominal	6913, Unique 6777-96%	-	-
num_reactions	Numeric	-	230.117	462.625
num_comments	Numeric	-	224.356	889.637
num_shares	Numeric	-	40.023	131.6
num_likes	Numeric	-	215.043	449.472
num_loves	Numeric	-	12.729	39.973
num_wows	Numeric	-	1.289	8.72
num_hahas	Numeric	-	0.696	3.957
num_sads	Numeric	-	0.244	1.597
num_angrys	Numeric	-	0.113	0.727

Table 1. Dataset Preprocess Features.

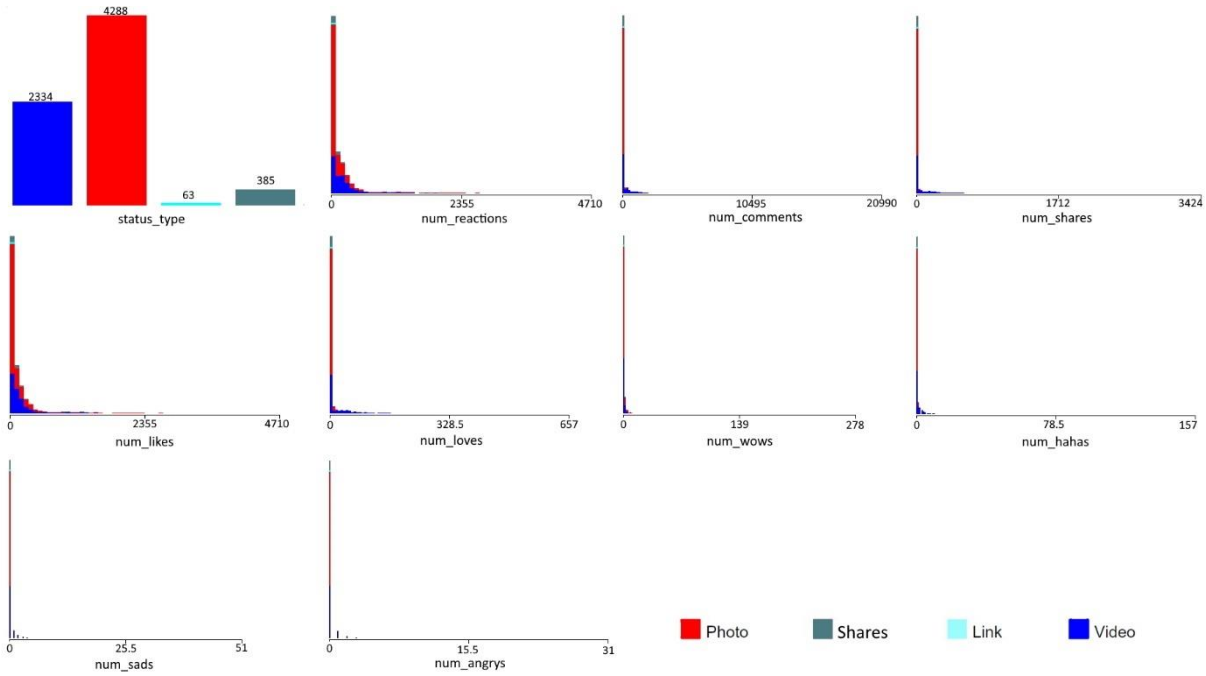

Figure 1. Attributes Preprocess Visualisation. Blue: refers to 2334 video posts, Red: refers to 4288 photo posts, Cyan: refers to 63 link posts, and Green: refers to 365 shares.

4. Results

The conducted experiments used WEKA 3 workbench as a mining tool and the decision trees were constructed by using pruned weka.classifiers.trees.J48 -C 0.25 -M 2. WEKA software, which is provided by the Waikato machine learning group from the departments of Computer Science at University of Waikato in New Zealand. The datasets are provided in ARFF format (Weka, 2019). Table 2 demonstrates the overall classification summary.

Correctly Classified Instances	**76.72**% (5409)
Incorrectly Classified Instances	**23.27**% (1641)
Kappa Statistic	0.4923
Mean absolute error	0.178
Root mean squared error	0.2993
Relative absolute error	68.73%
Root relative squared error	83.19%
Total Number of Instances	7050

Table 2. Classification Summary.

During our experiments and in order to avoid noise and overfitting, we used the pruned decision tree. Thus, due to low possibility of occurrence some of the status types were pruned and not displayed in the outcome. In each pair of brackets, the first number indicates the number of correctly classified instances; and the second the number of incorrectly classified instances.

num_loves <= 16
| num_likes <= 4: video (575.0/221.0) → 72.24% Accuracy
| num_likes > 4: photo (5426.0/1402.0) → 79.47% Accuracy
num_loves > 16
| num_comments <= 124
| | num_reactions <= 409
| | | num_comments <= 13: photo (2.0/1.0) → 67% Accuracy
| | | num_comments > 13: video (29.0/1.0) → 96.7% Accuracy
| | num_reactions > 409: photo (45.0/3.0) → 94% Accuracy
| num_comments > 124
| | num_reactions <= 1732: video (948.0/1.0) → 99.99% Accuracy
| | num_reactions > 1732
| | | num_loves <= 137: photo (5.0/1.0) → 83% Accuracy
| | | num_loves > 137: video (20.0) → 100% Accuracy
Number of Leaves: 8
Size of the tree: 15

Table 3. C4.5 (J48) Pruned Tree.

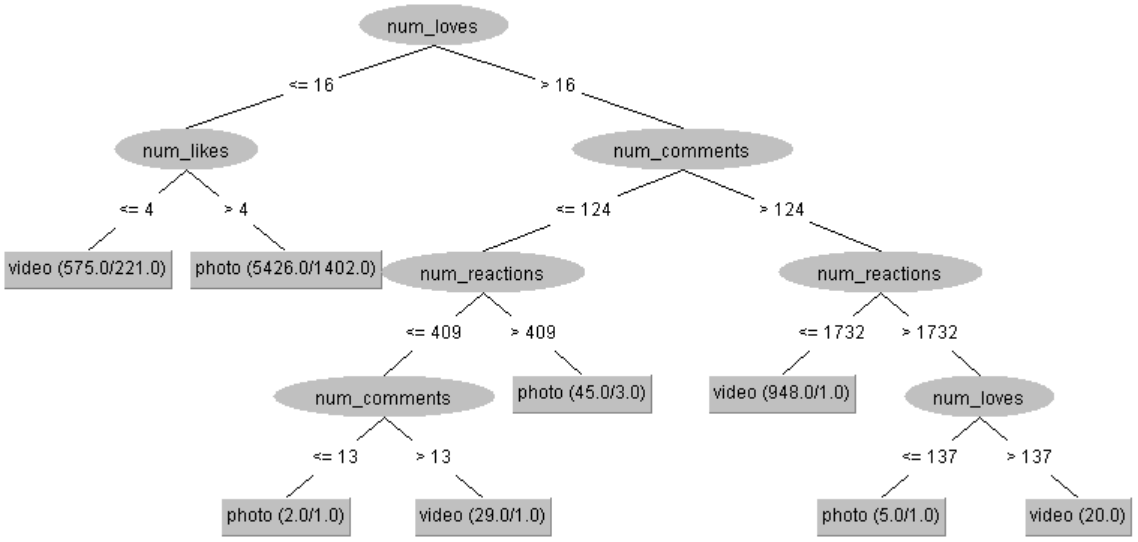

Figure 2. Decision Tree induced with J48, pruned Decision Tree visualisation.

5. Discussion

Our research results confirm the preferences of Facebook users on several types of posts. A previous conclusion showed that live videos increase the number of like reactions (Dehouche & Wongkitrungrueng, 2018). In this work, using data mining techniques, we found that live videos and photos cause the biggest number of reactions. Especially live videos posts correlate with a larger number of loves reactions. These results reveal the power that live video, and photo posts have over the other post types. Even though the dataset precisely describes live video performance, we must point out that there is a status type missing from the dataset and this is simple video posts uploads unless it is considered that link status type includes video posts, but we cannot do such an assumption. It would be interesting if we could measure and compare the performance of live videos versus the simple video posts. Due to low occurrence within the generated tree surprisingly, link and status post types are pruned. Hence, we can assume that this kind of posts do not contribute to fashion and cosmetics social media marketing purposes. Since Facebook gives different weight to different behaviors to determine what to show in user's screen our research proves to be essential when it comes to social media marketing decision making (Kim & Yang, 2017).

6. Conclusions

This study addresses an important question: Which types of Facebook posts generate bigger user engagement. It is proven that different types of Facebook posts trigger different users' reactions. Thus, marketers and decision makers across the globe need to understand the

way that social media work and apply tactics according to what attracts online users' attention. This study accomplished to predict online user behavior providing a detailed way of formulating the differences among Facebook post types. The dataset was tested using decision trees classification algorithms using Weka. Several experiments were conducted, but relatively to what we could accomplish there are steps which need to be made forward. The goal of this research was to track the number of likes, comments, shares and reactions on Facebook posts. Furthermore, our objective is to provide solutions that will meet the needs of marketers and customers generating between them a win-win relationship and create a track of experiments of how we get to such a position. However, limitations exist basically due to the lack of sufficient behavioural data volume across the industry sectors and secondly due to technology evolution and thirdly due to GDPR. The current dataset size, compared to the entire examples of retail population in a multimedium, the multichannel perspective is relatively small. Therefore, due to the uniqueness of human beings, different cultures, countries etc., we could assume that extracting specific rules out this research could be inappropriate or unethical. Probably, in a larger scale experiment processing millions of instances and customers' behaviour, we might be in a position where we could extract generalised rules including specifications. Nevertheless, due to highly accurate results in certain industry sectors within a specific country, we could assume that our prediction should be taken under consideration from marketing experts. A more sophisticated algorithm combining the advantages of a decision tree classifier along with the precision of a genetic algorithm approach might produce better results. This study is part of a much larger research effort on consumer behavior data analysis.

7. References

Agapitos, A., O'Neill, M., Brabazon, A., & Theodoridis, T. (2011). Maximum Margin Decision Surfaces for Increased Generalisation in Evolutionary Decision Tree Learning, 14th European Conference on Genetic Programming (EUROGP'11), (pp. 61-72). https://doi.org/10.1007/978-3-642-20407-4_6

Akaichi, J. (2013). Social Networks' Facebook' Statutes Updates Mining for Sentiment Classification. *International Conference on Social Computing.* 886 – 891. https://doi.org/10.1109/SocialCom.2013.135

Riasi, A., Wang, D. (2016). Comparing the Performance of Different Data Mining Techniques in Evaluating Loan Applications. DOI:10.5539/ibr.v9n7p164

Chen, D., Sain, S., & Guo, K. J. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management, 19*(3), 197-208. https://doi.org/10.1057/dbm.2012.17

Clos, J., Bandhakavi, A., Wiratunga, N., & Cabanac, G. (2017). Predicting Emotional Reaction in Social Networks. In: Jose J. et al. (eds), *Advances in Information Retrieval. Lecture Notes in Computer Science, 10193*. https://doi.org/10.1007/978-3-319-56608-5_44

Dehouche, N., & Wongkitrungrueng, A. (2018). Facebook Live as a Direct Selling Channel. Proceedings of ANZMAC 2018: *The 20th Conference of the Australian and New Zealand Marketing Academy*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand. (Last accessed: November 27, 2019).

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.

Forouzandeh, S., Soltanpanah, H., & Sheikhahmadi, A. (2014). Content marketing through data mining on Facebook social network. Webology, 11(1). Retrieved from http://www.webology.org/2014/v11n1/a118.pdf. (Last accessed: November 27, 2019).

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.

Karim, M., & Rahman, R. M. (2013). Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *Journal of Software Engineering and Applications, 6*, 196-206. http://dx.doi.org/10.4236/jsea.2013.64025

Kaur, W., Balakrishnan, V., Ranab, O., & Sinniah, A. (2019). Liking, sharing, commenting and reacting on Facebook: User behaviors' impact on sentiment intensity. *Telematics and Informatics, 39*, 25-36. https://doi.org/10.1016/j.tele.2018.12.005

Kim, C., & Yang, S. (2017). Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review 43*(2), 441-449. https://doi.org/10.1016/j.pubrev.2017.02.006

Kim, E., Kim, W., & Lee, Y. (2001). Purchase propensity prediction of EC customer by combining multiple classifier based on GA. *International Conference on Electronic Commerce*, 274–280.

Kohavi, R. (1995). The power of decision tables. In Lavrac, N., Wrobel, S. (eds). *Machine Learning: ECML-95. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), 912* (pp. 174-189). https://doi.org/10.1007/3-540-59286-5_57

Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. *Proceedings of International Conference on Knowledge Discovery from Data,* 73-79.

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. http://dx.doi.org/10.1016/j.jbusres.2016.02.010

Ortigosa, A., Martín, J. M., Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computer in Human Behavior, 31,* 527-541.

Panigrahi, R., & Borah, S. (2019). Classification and Analysis of Facebook Metrics Dataset Using Supervised Classifiers. *Social Network Analytics: Computational Research Methods and Techniques 1*, 1-19. DOI:10.1016/b978-0-12-815458-8.00001-3

Pavaloaia, V. D., Teodor, H. M., Fotache, D., & Danilet, M. (2019). Opinion Mining on Social Media Data: Sentiment Analysis of User Preferences. *Sustainability 11*(16), 4459. https://doi.org/10.3390/su11164459

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81–106. https://doi.org/10.1007/BF00116251

Statista. (2019). Social media & user-generated content— Facebook: number of monthly active users worldwide 2008-2019. Retrieved from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/. (Last accessed: November 27, 2019).

Theodoridis, T., Agapitos, A., Hu, H., & Lucas, S. M. (2010). A QA-TSK Fuzzy Model versus Evolutionary Decision Trees Towards Nonlinear Action Recognition. IEEE Int. Conference on Information and Automation (ICIA'10), (pp. 1813-1818). DOI: 10.1109/ICINFA.2010.5512225

Zell, A. L., & Moeller, L. (2018). Are you happy for me…on Facebook? The potential importance of "likes" and comments. *Computers in Human Behavior 78*, 26-33. https://doi.org/10.1016/j.chb.2017.08.050

Weka. (2019) Weka 3: Machine Learning Software in Java. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/downloading.html/. (Last accessed: November 27, 2019).