Visual electronic Word of Mouth: a multimodal brand approach and case study

Rietveld, R. ; Mazloom, M.; Van Dolen, W.; Worring, M.

Publication date
2016
Document Version
Final published version

Visual electronic Word of Mouth: a multimodal brand approach and case study

Abstract

An emerging activity on internet is to create and share visual content, our understanding of this activity and its impact however is limited. In this paper we aim to define and operationalize the concept of visual eWom and embed it in the current eWom literature. Different from existing eWom research which relies on textual information only, we conceptualize visual eWom as a multi-modal construct consisting of textual and visual concepts. We test our operationalization on a dataset of 6435 consumer posts crawled from Instagram. We apply state-of-the-art machine learning techniques, convolutional neural networks, for detecting the visual concepts in images posted on Instagram. We use OLS regression to test the impact of textual and visual concepts on image popularity. The results in our case study show that textual and visual concepts provide complementary information and have a different impact on image popularity.

Keywords: visual eWom, deep learning, social media

Track: Online marketing and social media

1. Introduction

The availability of data and potential impact of eWom on consumer behavior have led to a considerable amount of academic attention, resulting in over 100 studies over the past 15 years (Babić, Sotgiu, de Valck, & Bijmolt, 2015). While this extensive body of research provides insight into the antecedents and consequences of textual eWom, little is known about visual eWom (King et al. (2014). This is surprising since the popularity of visual based social media such as Instagram has grown faster than textual base social media platforms (e.g. Twitter, Facebook); Instagram has more active users (> 400 million) than Twitter (Kharpal, Images, & Origin, 2015) who upload on average 58 million images every day ("Instagram Company Statistics | Statistic Brain," 2015). Given the widespread adoption of visual based social media, managers are faced with the challenge to manage their brand and integrate consumer-generated visual media and textual social media into their brand strategy (Gensler, Völckner, Liu-Thompkins, & Wiertz, 2013). In particular visual eWom seems to be relevant for brand strategy as the impact of images are perceptually more distinct than words (Childers & Houston, 1984) and visual processing of images is rapidly and almost automatically. In this paper we answer to the call of King et al. (2014) to overcome the gap in our understanding of visual eWom by answering three research questions: What is visual eWom? How do visual and textual elements of visual eWom differ? How do visual and textual elements impact brand image popularity? In this paper we aim to bridge this gap by conceptualizing visual eWom as a multi-modal construct. Based on advertising literature we decompose visual eWom into brand, text and pictorial elements to be able to capture both textual information provided by the sender and visual information contained in the image. We apply machine learning methods, in particular convolutional neural networks, to provide a scalable and testable way to analyze the content of textual and visual elements. We test our methods on a dataset from Instagram and demonstrate the different impact of textual and visual elements on image popularity. The contribution of our research is threefold: First, to our knowledge we are the first to provide a formal conceptualization of visual eWom in a marketing context. Second,we demonstrate that visual eWom consists of both textual and visual elements each providing different information. We illustrate this by modelling the popularity of images by using both image and textual data from the same dataset. Third, while we do not develop new to the world text or visual analysis methods, we do integrate methods from machine learning to analyse text and visual information in a marketing context. To our knowledge we are the first to adopt convolutional neural networks to analyse visual eWom.

2. Theory

WOM has a long standing tradition within the marketing discipline dating back to the 60's (Arndt, 1967). We continue in this vein and introduce the concept of visual eWom to describe the widespread creation, sharing and consumption of brand related images on the internet by consumers. We argue that visual eWom constitutes a new form of eWom based on the communication modality, vision, it introduces. 2 Surprisingly however none of the studies, or any of the studies in the recent meta analysis on eWom (Babić et al., n.d.; You, Vadakkepatt, & Joshi, 2015), explicitly consider visual eWom.

2.1 Definition of visual eWom

Central to our definition of visual eWom is the multi-modality of the construct. We follow Pieters et al. (2004) who divide the content of print advertisements into brand, textual and pictorial content. The brand element entails the visual brand identity cues such as the brand name, trademark and logo of the brand (Keller 2003).The presence of a brand element has been considered to be the primary difference between eWom and user generated content (Lovett, Peres, & Shachar, 2013). The text element comprises all textual information of the message, excluding all references of the brand name. The pictorial element comprises all nontextual information of the image, excluding all incidences of the brand trademark and logo. Visual eWom consists of brand, text, and pictorial

elements with the primary carrier of information being visual, which is a combination of brand and pictorial elements. By applying this multi-modal approach to the definition provided by Hennig-Thurau (2004) we define visual eWom as: *any message where the main carrier of information is visual (brand + pictorial elements), created by potential, actual, or former customers about a product, brand or company, which is made available to a multitude of people and institutions via the Internet.* Our definition contains several key elements. First, a brand element must be present in the message in order to qualify as visual eWom. Second, the main carrier of information is visual.The visual aspect (pictorial and brand elements) of the message must make up more than half of the message surface. The message in an eWom context is the unit of the venue format for example blogpost, forum entry or tweet in case of a microblogs. Visual eWom is therefore not dependent upon the venue format or platform on which it is posted. Third, the sender of the image is a consumer. Hence firm generated content is not visual eWom. Sharing content created by others however also qualifies as visual eWom. The sender does not have to be the creator of the message. Fourth, the image must be publically available on the Internet for other people to view with minimal effort.

2.2 How do visual and textual elements differ?

We argue that the visual and textual elements contain different information in terms of the subjective and objective information provided. Users who post brand-related content have the opportunity to augment their images with textual information (for example a caption or hashtag) to convey additional information or meaning to the receiver. In content based information retrieval the process of adding text to images is known as image annotation (Momeni, Cardie, & Ott, 2013). Image annotation is a way to share and categorize images which enables users to express their thoughts, perceptions, and feelings with respect to diverse concepts.The most common approach to image annotation is tagging, which allows the users to annotate images with a chosen set of keywords ("tags") from a controlled or uncontrolled vocabulary (Yan, Natsev, & Campbell, 2007). The ability to annotate images from an uncontrolled vocabulary provides users with a means to invent new hashtags which do not exist in our formal language. We focus on tags as our textual element since tags represent a high semantic value where users abstract away words which are less relevant (Nam & Kannan, 2014). When assigning tags users also make decisions to exclude information which could be relevant in describing an image. This leads to three sources of information discrepancies between visual and textual elements; 1) Users may invent concepts to describe their subjective experience, which are not related to the visual objective information the image provides; 2) User generated tags contain subjective information which is not directly related to the concepts represented in the image; 3) User generated tags omit concepts present in the image (for various reasons).

Proposition 1: Visual eWom messages contain both visual and textual elements which provide complementary information.

2.3 Does the difference matter?

We next propose that visual and textual elements have a different impact on image popularity. Processing of brand related information is an antecedent in the brand attitude formation process (Maclnnis & Jaworski, 1989). The different processing of text and visual content is well established as images have a universal quality in that it is genotypically much older than language representation (Öhman, Flykt, & Esteves, 2001). These different cognitive processes on the receiver part play a role in their subsequent actions. Prior research on Instagram images for example, showed that the presence of a face in an image has a positive impact on the popularity of the image (Bakhshi, Shamma, & Gilbert, 2014). Based on the impact of both tags and image content we hypothesize that both text and visual elements of eWom provide complementary information and have an impact on the receiver's action. In line with de Vries (2012 et al.) we

tested our hypothesis by explaining image popularity (likes) using both visual and text elements. We expect that the inclusion of each element helps to explain part of the heterogeneity observed in likes per image.

Proposition 2: Text and visual elements from visual eWom have a different impact on image popularity.

3. Data

We selected Instagram as a platform to illustrate application of the visual eWom concept. Instagram is a social network site designed around photo and video-sharing. Following our criteria for visual eWom, we collected images from a single brand. We choose McDonald's as a case study which has been subject in prior research of brand associations (John, Loken, Kim, & Monga, 2006). Images contain metadata consisting of geolocation, author information, hashtags, caption, amount of likes, and amount of comments. We collected data on all images by querying for the hashtag 'mcdonalds' (spelled in small caps on Instagram) between september and october 2015 which had geolocation data (latitude and longitude) indicating the image was uploaded in the US . We excluded images which originated from instagram accounts which are owned by or affiliated with McDonald's. To restrict the set to 10.000 image we took the most recent images. We consider this dataset to fulfill the requirements of being considered visual eWom conform the requirements set in section 2.1 since, the visual elements are the main carrier of information, consumers have posted the content on a publicly available social network and it is brand related as the hashtag mcdonalds is present.

4. Methodology

In order to analyze the multi-modality of visual eWom we use methods from machine learning. To analyze textual and image content in the same framework we needed to map hashtags and visual concepts onto a common taxonomy. To that end we choose to use WordNet and ImageNet respectively (Deng et al., 2009; Miller, 1995). Doing so allowed for a comparable set of concepts from text and image modalities. WordNet is a large

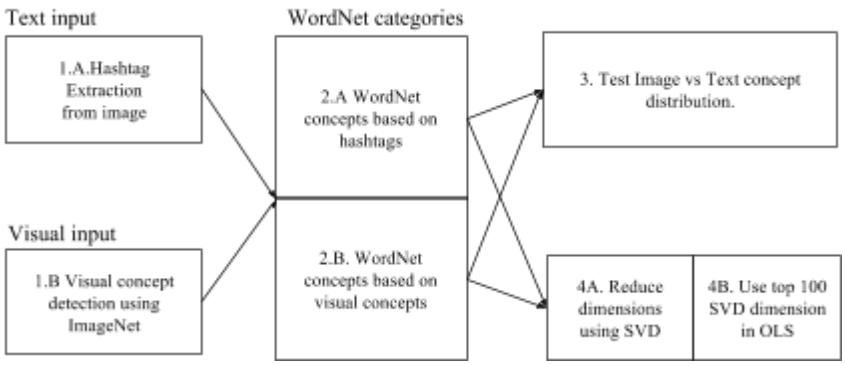

Figure 1

lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept (Miller, 1995). WordNet contains 117.791 different synsets denoting the most commonly used words and concepts in the English language. ImageNet is an image dataset organized according to the subset of the concepts in the WordNet hierarchy which can be visually represented. 14.197.122 images have been manually classified in 21.841 hierarchical categories. Recent advances in employing convolutional neural networks (CNN) trained on the imagenet dataset (Hinton, Osindero, & Teh, 2006) have proven very effective in correctly identifying concepts in images (Deng et al., 2009). Figure 1 summarizes our approach in mapping textual and visual concepts to the common taxonomy following the following steps:

Step 1.A Extraction of hashtags: Each collected image on Instagram has one or more hashtags. Hashtags are user generated combinations of characters without spaces, for example one of the most popular hashtags on Instagram is instagood (421 million images have this hashtag).

Step 1.B Visual ImageNet concept detection: For extracting visual concepts per image we used deep learning features from an existing convolutional neural network (Mazloom, Li, & Snoek, 2015). Similar to Krizhevsky et al. (2012), we used a pre-trained convolutional neural network for mapping every image to a 15.293-dimensional vector which is the output of the soft-max layer of the convolutional neural network. The network was pre-trained on all the 15.293 concept categories in the ImageNet dataset, for which at least 50 positive examples are available. This collection includes semantic concepts related to objects such as car, food, table and different scenes such as indoor scenes and outdoor scenes. The output of each concept demonstrate the probability of existing concept in image. The top 10 concepts per image in terms of highest probability were selected to be included for further analysis.

Step 2.A. Mapping hash tag onto WordNet: We only included hashtags which have a direct equivalent in WordNet in order ensure a comparable sample with concepts extracted from images. 6435 images are included for further analysis,since one or more hashtags appears in WordNet.

Step 3. Assess concept distribution: In order to assess the difference between text and visual concepts, we aggregated the concept occurrence for text and images. Next, we tested whether the distribution of concepts originated from the same distribution using the Kolmorov Smirov (KS) two sample test. We selected concepts which occur more than 10 times, resulting in 1659 distinct concepts for images and 1322 distinct concepts for text. We indexed the concept occurrences to reflect the proportion of the concept occurrences within the text and image respectively which sum to 1.

Step 4.A. Reducing dimensions: To analyse the impact of both text and image concepts we created an image versus concept incidence matrix with images as rows and concepts as columns. Each cell has a binary value indicating the occurrence of a concept in an image (i.e. an entry for each of the 10 top scoring concepts). Due to the large number of dimensions (1659 and 1322 respectively) and sparse nature of the data we used singular value decomposition(SVD) to reduce the dimensionality. Following Kosinski et al. (2013), who studied facebook likes, we used the top 100 factors from the SVD analysis to be included in the subsequent regression analysis.

Step 4.B. Regression analysis: Likes are a continuous count variable following a Poisson distribution (Strawderman, 1999). Similar to de Vries (2012), we modelled the data using the log of likes and OLS regression using the top 100 dimensions from the SVD analysis. We used two user properties as control variables, the amount of followers on instagram and the amount of images posted. A user with a large following is more likely to generate exposure for posted content, since more users on Instagram have the opportunity to view the image (Bakhshi et al., 2014).

5. Results

We performed two sample Kolmogorov-Smirnov tests, the first on the full dataset including concepts which had 0 occurrences in either text or image. Only 4% (116 concepts) of the concepts have occurrences in both image and text modalities. The KS-test for difference distribution was significant ($p < 000.1$), which confirmed the limited overlap between information from text and images. Next, we performed the same test but only on the 4% of the concepts which overlap. We reindexed the data and tested the sample again using two sided KS-test. Also in this case the distribution on concepts between text and images differ significantly ($p = 0.001$). Also within concepts that occur in both image and text the distribution of those topics was different. Therefore, we conclude that proposition 1 holds for our dataset.

The regression model on the log of the number of likes from image concepts was significant as a whole (F-value = 6.756, p-value 0.01) and explains the variance of the dependent variable reasonably well (R^2= 9.81%, adj. R^2= 8.35%). The control variables of user followers was significant ($p < 0.01$), however the amount of images showed no effect ($p=0.36$). Next we added

the textual concepts and ran the regression on log of likes again. The regression model was significant as a whole (F-value = 9.84, p-value 0.01) and explained the variance of the dependent variable reasonably well (R^2= 13.73%, adj. R^2= 12.34%). The same effects hold in this model for the control variables of user followers was significant ($p < 0.01$), however again the amount of images showed no effect (p=0.25). Therefore, we conclude that proposition 2 is holds true for our dataset since the second model explained more variance in terms of image popularity.

6. Conclusions

We define visual eWom and provide a scalable approach to analyse this data and provide case based evidence that visual information differs from textual information. Not only do the type of topics differ in the case of McDonald's dataset, also the matching concepts distribution is different for visual vs textual elements. Both results indicate that visual eWom should be considered as a multi-modal construct. Our approach provides a methodology for further research and we demonstrate through a case based study that the identified visual concepts have impact on image popularity, an important measure for marketing managers. Furthermore, we show that within the context of our case study the impact on outcome measures such as image popularity differs for text and visual content. We find that the image popularity can be explained reasonably well by the image annotation generated by users, but also by the content of the image itself. These results suggest further study is required into the workings of the communication modality and the message receiver outcome of the WOM process.

7.Limitations and Further Research

The article provides a first step in exploring the topic of visual eWom. Our study is limited in a number of ways. First, although our methods could be implemented on a larger scale we only selected a limited amount of images from a single brand. Extending the scope and depth of the data could yield category effects. Second, our way of mapping text based hashtags onto WordNet taxonomy is rather crude by eliminating hashtags (and images without associated hashtags) which do not have an entry in WordNet.

Further research is needed to understand the visual eWom type on a consumer perspective and a management perspective. Antecedents as to why consumer select images over other communication modalities would advance our knowledge. Also what would be the effect of (repeated) visual eWom on the cognitive processes which involve learning and updating the network of brand associations? As social media venues are opening up more of their platforms to brand advertising, our method could be helpful in understanding which content drivers contribute to image popularity. The availability of data and advances in visual pattern recognition using deep learning models provide ample opportunity to understand antecedents and consequences of visual eWom better.

References

Arndt, J. (1967). Word of mouth advertising: a review of the literature.

Babić, A., Sotgiu, F., de Valck, K., & Bijmolt, T. H. A. The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. *JMR, Journal of Marketing Research*, *forthcoming*. http://doi.org/10.1509/jmr.14.0380

Bakhshi, S., Shamma, D. A., & Gilbert, E. (2014). Faces engage us: Photos with faces attract more likes and comments on instagram. *Proceedings of the International ACM SIGCHI Conference on Supporting Group Work.*

Berger, J., & Iyengar, R. (2013). Communication Channels and Word of Mouth: How the Medium Shapes the Message. *The Journal of Consumer Research*, *40*(3), 567–579.

Childers, T. L., & Houston, M. J. (1984). Conditions for a Picture-Superiority Effect on Consumer Memory. *The Journal of Consumer Research*, *11*(2), 643–654.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*

(pp. 248–255).
De Vries, L., Gensler, S., & Leeflang, P. S. H. (2012). Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *Journal of Interactive Marketing*, *26*(2), 83–91.
Gensler, S., Völckner, F., Liu-Thompkins, Y., & Wiertz, C. (2013). Managing Brands in the Social Media Environment. *Journal of Interactive Marketing*, *27*(4), 242–256.
Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52.
Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.
Instagram Company Statistics | Statistic Brain. (2015, September 11). Retrieved December 1, 2015, from http://www.statisticbrain.com/instagram-company-statistics/
John, D. R., Loken, B., Kim, K., & Monga, A. B. (2006). Brand Concept Maps: A Methodology for Identifying Brand Association Networks. *JMR, Journal of Marketing Research*, *43*(4), 549–563.
Kharpal, A., Images, G., & Origin, S. B. (2015, September 23). Facebook's Instagram hits 400M users, beats Twitter. Retrieved November 25, 2015, from http://www.cnbc.com/2015/09/23/instagram-hits-400-million-users-beating-twitter.html
King, R. A., Racherla, P., & Bush, V. D. (2014). What We Know and Don't Know About Online Word-of-Mouth: A Review and Synthesis of the Literature. *Journal of Interactive Marketing*, *28*(3), 167–183.
Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc.
Lovett, M. J., Peres, R., & Shachar, R. (2013). On Brands and Word of Mouth. *JMR, Journal of Marketing Research*, *50*(4), 427–444.
Maclnnis, D. J., & Jaworski, B. J. (1989). Information Processing from Advertisements: Toward an Integrative Framework. *Journal of Marketing*, *53*(4), 1–23.
Mazloom, M., Li, X., & Snoek, C. G. M. (2015, October 10). *TagBook: A Semantic Video Representation without Supervision for Event Detection*. *arXiv [cs.CV]*.
Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39–41.
Momeni, E., Cardie, C., & Ott, M. (2013). Properties, Prediction, and Prevalence of Useful User-Generated Comments for Descriptive Annotation of Social Media Objects. In *ICWSM*. myleott.com.
Nam, H., & Kannan, P. K. (2014). The informational value of social tagging networks. *Journal of Marketing*. Retrieved from http://journals.ama.org/doi/abs/10.1509/jm.12.0151
Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology. General*, *130*(3), 466.
Pieters, R., & Wedel, M. (2004). Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects. *Journal of Marketing*, *68*(2), 36–50.
Schweidel, D. A., & Moe, W. W. (2014). Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice. *JMR, Journal of Marketing Research*, *51*(4), 387–402.
Smith, A. N., Fischer, E., & Yongjian, C. (2012). How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter? *Journal of Interactive Marketing*, *26*(2), 102–113.
Strawderman, R. L. (1999). Regression Analysis of Count Data. *Journal of the American Statistical Association*, *94*(447), 984–986.
Yan, R., Natsev, A., & Campbell, M. (2007). An Efficient Manual Image Annotation Approach Based on Tagging and Browsing. In *Workshop on Multimedia Information Retrieval on The Many Faces of Multimedia Semantics* (pp. 13–20). New York, NY, USA: ACM.
You, Y., Vadakkepatt, G. G., & Joshi, A. M. (2015). A Meta-Analysis of Electronic Word-of-Mouth Elasticity. *Journal of Marketing*, *79*(2), 19–39.