ASSOCIATION FOR CONSUMER RESEARCH

Labovitz School of Business & Economics, University of Minnesota Duluth, 11 E. Superior Street, Suite 210, Duluth, MN 55802

Black-Box Emotion Detection: on the Variability and Predictive Accuracy of Automated Emotion Detection Algorithms

Francesc Busquet, University of St.Gallen

Christian Hildebrand, University of St. Gallen

The current research demonstrates considerable variability in predictive accuracy across major emotion detection systems (such as Google ML or Microsoft Cognitive Services) with lower (higher) classification accuracy for negative (positive) discrete emotions. We provide two modelling strategies to improve prediction accuracy by either combining feature sets or using ensemble methods.

[to cite]:

Francesc Busquet and Christian Hildebrand (2020) ,"Black-Box Emotion Detection: on the Variability and Predictive Accuracy of Automated Emotion Detection Algorithms", in NA - Advances in Consumer Research Volume 48, eds. Jennifer Argo, Tina M. Lowrey, and Hope Jensen Schau, Duluth, MN : Association for Consumer Research, Pages: 831-835.

[url]:

http://www.acrwebsite.org/volumes/2661730/volumes/v48/NA-48

Consumer Behavior Meets Machine Learning

Chairs: Siham El Kihal, Frankfurt School of Finance & Management, Germany
A. Selin Atalay, Frankfurt School of Finance & Management, Germany

SESSION OVERVIEW

In a recent paper, Proserpio and colleagues (2019) highlight how advances in machine learning research help marketing researchers estimate models that predict extremely well and that are able to leverage the power of new data sources such as images, text, audio, and video. At the same time, they raise how machine learning research needs a "soul" which should capture competitors' reactions in complex markets as well as consumer theories to be able to capture the complexity of consumer behaviors, their needs and preferences.

The papers in this special session "Consumer Behavior Meets Machine Learning" aim at improving our understanding of consumer behavior and attitudes by combining advanced machine learning techniques with behavioral theories. The papers contribute to answering the following high-level questions:

- How can machine learning enhance our understanding of consumer attitudes and behaviors?
- How can behavioral theories help machine learning researchers open the prediction black-box and generate more interpretable features?

Using Natural Language Processing to Investigate the Role of Syntactic Structure in Persuasive Communication

EXTENDED ABSTRACT

How to use language, that is, how to formulate marketing communications so that they are persuasive, is a decision problem most marketers face. Language, spoken and written alike, is a fundamental element of marketing messages. Language is used in various contexts such as political debates, product reviews, product descriptions and brand communications with the goal of persuading the recipient. Today, the use of email, social media posts, live chats and blogs provides unprecedented opportunities to reach consumers with verbal messages in addition to existing channels such as TV, radio or newspaper ads.

Extant literature in marketing suggests that the language used influences the persuasive outcomes in various marketing contexts. For instance, language used in news articles reflects the emotions and legitimacy of the article, and influences consumers' attitudes toward the article as well as their likelihood of reading or sharing the article (Berger and Milkman 2012; Humphreys and LaTour 2013; Berger, Moe, and Schweidel 2019). Language used in movie scripts reflects the theme, scenes, and emotion in the movie and thus predicts the box office success (Eliashberg, Hui, and Zhang 2007). Language used in songs impacts the songs' success on the market as the typicality of the lyrics impacts the popularity of a song (Berger and Packard 2018). Elaborateness and the emotionality of tweets about political debates influences which topics get consumers' attention and are shared with others (Berman et al. 2019). Even though a large stream of recent literature is dedicated to understanding the role of language used in marketing contexts (see Berger et al. 2020 for an overview), most of this research is dedicated to the content of the message communicated (i.e., choice of words). Tools for understanding what makes the language used in marketing messages persuasive independently of the content are still missing.

In the current research, we seek to understand what makes marketing messages persuasive independently of the message content. In that, our focus is on the impact of the syntax of the language used (i.e., grammar) on the persuasiveness of the message communicated. Following a natural language processing (NLP) approach, we posit that the syntactic structure of a message influences its persuasiveness (Gibson 1998). We develop and test a novel approach to measuring syntactic structure in a series of three studies, and identify the specific syntactic elements, referred to as dependencies, which impact the persuasiveness of a marketing message. In our main study, using a dataset consisting of 134 debates with 129,480 sentences, we develop an NLP-based measure of syntactic structure. In this study, we show that the syntactic structure of the language used in the debates impacts the attitude change caused by the debates. Using our measure, we identify the specific syntactic elements that affect the persuasiveness of a marketing message. Our measure captures syntactic structure more systematically and better in terms of predicting the persuasiveness of a message than past measures of syntax.

We validate our measure of syntactic structure and the impact of the specific syntactic elements that affect the persuasiveness of a marketing message in two experiments that follow the main study: First, we conduct a laboratory experiment where we manipulate the syntactic structure of a marketing message using the specific syntactic elements identified in our main study. The context is a persuasive message about the benefits of wearing a bike helmet. We show that when the syntactic structure of the message is changed according to our findings, the persuasiveness of the message is increased. Second, we conduct a field experiment on Facebook, where we improve the syntactic structure in the copy of an advertisement designed by a marketing agency using the specific syntactic elements identified in our main study, and measure click through rates for the ad. We show that when the syntactic structure of the ad language is improved using the specific syntactic elements we identified, the persuasiveness

of the ad is increased: The click-through rate for the ad increased significantly by 40% compared to the baseline ad version, which was designed by the marketing agency. This experiment also demonstrates how marketers can use our method to improve their marketing messages to make them more persuasive.

Our work contributes to marketing research by developing a novel approach of measuring and modifying syntactic structure, allowing for a more precise way of formulating persuasive marketing communications. Additionally, our work contributes to marketing practice by providing a tool that marketers can use to efficiently improve the persuasiveness of their messages as well as any other communication that uses language as a medium.

Scalable Content Curation: Learning from Human Effort

EXTENDED ABSTRACT

Many firms rely on the crowdsourcing platforms to generate new product and advertising ideas. For example, Colgate-Palmolive, a large consumer products firm, collaborated with a digital creative agency to generate ideas for one of its most visible advertising campaigns. The firm organized an ideation contest - members of the agency's innovation platform submitted ad ideas, and the winning ideas were used to develop an ad that was shown to over 109 million viewers of the Super Bowl - one of the world's most popular television events. Major advertisers such as P&G and Unilever generate ideas for new products and advertising campaigns either using internal crowdsourcing platforms or through specialized creative agencies. Crowdsourced ideas can be more creative than internal solutions. The firm may uncover an unexpected extreme-value solution thanks to a large number of opinions from diverse participants. Crowds can also identify niche consumer interests which can lead to valuable innovations that the firm would not have considered otherwise.

Crowdsourcing campaigns can yield thousands of submissions that firms cannot exhaustively and carefully evaluate. For example, our data provider attracted 74,436 ideas over 153 ideation contests. The scale of these campaigns raises the important question of how to identify the best ideas efficiently. Current approaches involve using either the crowd itself or a set of internal employees to prescreen ideas. However, the crowd-based approach is infeasible if the firm prefers to keep submissions confidential. In this research, we focus on the case where the firm uses internal evaluators to maintain submission confidentiality and thereby abide with data protection laws or withhold the submissions it receives as a competitive advantage. In our field setting, each idea was evaluated once by several employees of the crowdsourcing platform in a prescreening stage, which resulted in a total of 186,111 evaluations and consumed about 14 workdays per month of human effort in the median month.

Our research aims to develop an efficient approach to identify the best ideas from crowdsourcing contests using data on the behavior of internal evaluators, whom we refer to as "voters", idea characteristics and text, and the selection of winning ideas by the clients, whom we refer to as "sponsors". We develop a model to reconcile the difference between the sponsor's selection of finalist ideas and the evaluations of voters who prescreen the ideas. In our field setting, the platform sorts ideas based on their average voter score before presenting them to the sponsor. A model that transforms voter scores and incorporates idea characteristics and text can yield a better score to sort ideas, thereby reducing the number of ideas that a sponsor must review to identify all finalists by 7-18%, depending on the specific evaluation approach. Equivalently, the platform can reduce the workload of its employees by 8-25% while maintaining the same quality of the idea ranking for the sponsor.

We find that voters exhibit persistent differences in their "importance", or the extent to which their votes help identify finalists. As a result, a model trained on one set of contests can extend to future contests even if ideation topics differ across contests. We define a metric for measuring the importance of a voter as the deterioration in the quality of the idea ranking when we remove this voter. Interestingly, voter importance may differ depending on whether a model is used or not. We call this phenomenon "predictable inaccuracy" – a voter may be more predictive of the sponsor's choice if her votes are properly transformed. Moreover, two voters who are similarly important when the platform uses average score to rank ideas may differ significantly in their importance when a model is used, suggesting that different voters can exhibit different levels of predictable inaccuracy. Voter "redundancy", or the extent to which voters are correlated with each other, can also lead to varying levels of predictably inaccuracy. Our results suggest that a firm interested in identifying its most valuable voters may arrive at different conclusions depending on whether it assesses voter performance jointly with a statistical model or independently of a model based on the initial ranking criteria.

We complement our findings with a survey based on a product design contest organized by an open innovation platform called Lego Ideas. The survey allows us to design simulations to study the generalizability of our findings, recover ground-truth measures of idea quality that are not available in the field data, and study how voter characteristics may relate to voter importance. We capture voter heterogeneity through differences in brand appreciation and generate a sponsor score from a subset of the available votes. We find that even if voters do not influence sponsor choices, and sponsor preferences are representative of voter preferences, certain types of voters still emerge as more important than others. A statistical model with text data yields improvements in the survey setting as well, but there is only subtle evidence of predictable inaccuracy in the particular simulation we study because of limited nonlinearities in the relationship between voter and sponsor scores. The survey confirms that a statistical model can exploit varying voter importance and redundancy, thereby reducing evaluation effort in this alternative setting, even when there are no significant preference differences between voters and the sponsor.

Overall, our findings suggest that firms can enhance their content curation processes, even if they crowdsource ideas on different topics over time. In assessing the performance of different voters, firms must consider how the voters perform in conjunction with a model as opposed to independently. The "predictable inaccuracy" of certain voters, if it is present, can alter their relative importance.

The Power of Brand Selfies in Consumer-Generated Brand Images

EXTENDED ABSTRACT

Every day more than 5 billion images are shared on social media networks such as Facebook, Twitter, or Instagram. Of particular interest to marketers are images that feature brands and consumption experiences that contribute approximately 1% of all social media images, resulting in 50 million social media images featured with brand logos daily. At the same time, one of the biggest trends introduced by smartphone cameras and social media are selfies. Today, more than 383 million images with the hashtag #selfie exist on a single photo-sharing platform such as Instagram. The emergence of selfies merits

the question of how brands appear in selfie images and how observers respond to this type of social media brand images.

In fact, recent marketing practice attempts to capitalize on the selfie phenomenon, mainly by actively encouraging consumers to post selfies of their product encounters. Brand images are also of interest to managers when passively listening in to social media posts. Among other things, companies track brand logo presence on social media to understand social media popularity, rank consumer-generated images on their social media brand page, or use such images as part of their own marketing campaigns.

However, not all brand logo appearances are created equal. Some may generate more valuable consumer-brand engagement than others. Accordingly, the objective of this research is to investigate how brands appear in consumer-generated images and to examine the effectiveness of different types of social media brand images on generating engagement among consumers. Specifically, we investigate both sender engagement objectives in terms of how brand images may generate image engagement (i.e., likes or comments) and brand engagement objectives in terms of how brand images may generate brand engagement (i.e., brand-related comments such as expressed purchase intent).

Academic research provides ample evidence for the effectiveness of images in advertising (e.g., Hanssens and Weitz 1980; Xiao and Ding 2014). Recent studies have also explored the motivations to share content and take photos and how this affects subsequent sender behavior (e.g., Barasch, Zauberman and Diehl 2018; Grewal, Stephen and Coleman 2019). However, little is known about how observers respond to brand images in social media.

Using both manual annotations of a sample of images and convolutional neural networks (CNN) for automated image classification of nearly half a million brand images related to 185 different brands that were posted on Twitter and Instagram, we identify three types of user-generated brand images that differ in terms of human and facial presence. Consumers post either images of products in isolation or holding products themselves (i.e., selfies). We find that these brand-related selfies exist in two forms with consumers either visible (their face) or invisible to the viewer (e.g., first-person point of view of the product). We name the former consumer selfie and the latter brand selfie to indicate the focus in brand selfies is exclusively on the product and differentiate it from consumer selfies where the face of the sender is visible[1]. This results in the following typology of brand images:

1. Brand Selfies: branded products held by an invisible consumer,
2. Consumer Selfies: visible consumer faces together with a branded product,
3. Packshots: standalone images of branded products.

The CNN algorithm accurately classified images into these three categories with a hold-out accuracy level of > 80%. In both datasets, we find that consumer selfies have the lowest fraction of all types, suggesting that consumers are reluctant to post photos of themselves with the brand on their own accord. At the same time, consumer selfies are the image type most often encouraged by corporate communication campaigns, and face images, akin to consumer selfies, are ubiquitous in print advertising (Xiao and Ding 2014). Whether consumer selfies are well suited as a user-generated media content is an open question, which we attempt to answer in this research.

Analyzing consumers' response to the different types of brand images in terms of likes and comments, we find that consumer selfies, in which a person appears with the brand, generate the highest level of engagement towards the image or the sender in terms of the number of likes and comments on the image. However, these simple engagement measures, while encouraging for brands in terms of user potential, may be misleading. Examining the content of the user comments, using both dictionary-based and machine learning text mining tools, we find that consumer selfies generate fewer self-brand mentions and stated purchase intentions of receivers in response to the original image post. These results are consistent with research from traditional advertising and information systems, which indicate that, on the one hand, images with faces catch more attention than those without faces (Bakhshi, Shamma and Gilbert 2014; Xiao and Ding 2014), but, on the other hand, may detract attention away from the brand itself (Erfgen, Zenker and Sattler 2015).

These results are consistent for both the Twitter and Instagram dataset, suggesting that the effects are not driven by different consumer motives of the respective platforms. We complement these results with a lab experiment, which allows us to control for the prominence of the brand in the image and collect purchase intent ratings, as well as to

test the underlying potential psychological mechanism that may trigger consumers' varying reaction to different brand image types. The results of the experiment further suggest that brand selfies have a superior impact on perceived purchase intent compared to consumer selfies. In addition, it suggests the differential impact of brand selfies is related to easier and more accessible self-reference and mental simulation offered by these image types.

Black-Box Emotion Detection: On the Variability and Predictive Accuracy of Automated Emotion Detection Algorithms

EXTENDED ABSTRACT

Automated emotion detection from facial expressions refers to the use of algorithms that detect facial landmarks in pictures to classify people's discrete emotions (Liu et al. 2014; Wood et al. 2016). These automated emotion detection systems classify discrete positive emotions such as happiness or surprise to negative emotions such as anger and fear. A computer vision algorithm first identifies facial landmarks in a picture and then assigns each picture a discrete emotion label based on the features or composition of the existing facial landmarks (Fox et al. 2000; Pantic and Rothkrantz 2000) development of an automated system that accomplishes this task is rather difficult. There are several related problems: detection of an image segment as a face, extraction of the facial expression information, and classification of the expression (e.g., in emotion categories. Companies such as Microsoft, Google, or GfK provide platforms to perform such automated emotion detection with recent industry reports suggesting a CAGR of 32.7% and a market size of 25 billion by 2020, highlighting the importance and dominance of these AI-powered technologies for the future of marketing. Despite the increasing availability of automated emotion detection systems, fundamental methodological questions arise. Are automated emotion detection systems valid? Is the same picture classified correctly across emotion detection systems? These questions are important as emotion detection systems are using pre-trained algorithms to classify discrete emotions from facial expressions in a way that is unknown to

1 Note, we distinguish brand and consumer selfies based on the visibility of consumer faces. Conceptually, the term selfie suggests the person on the image took the photo herself. Yet, in some cases, a third person may have photographed the sender (e.g., < 3% of brand selfies and < 39% of consumer selfies in our Twitter data). This can be difficult to distinguish empirically for both the human eye and an automated image classifier.

the user of these systems. Thus, it is likely that the same picture is assigned a different discrete emotion conditional on the type of emotion detection system being used. To the best of our knowledge, both a formal test of this hypothesis is non-existent as is a formal analysis that unravels under which conditions emotion detection algorithms would increase in predictive accuracy. The current paper fills this gap by providing a tightly controlled validation study comparing the effectiveness (i.e., in terms of predictive accuracy) of major automated emotion detection systems across discrete emotions, dependencies on the modelling technique being used, the stimuli set, and ways to improve predictive accuracy by combining feature sets or using ensemble methods.

We used a standardized picture set with objective ground truth, i.e. knowledge about the discrete emotion that is displayed by actors in a standardized picture set. Specifically, we used the Chicago face database (Ma, Correll, and Wittenbrink 2015) to evaluate automated emotion detection systems. The Chicago face database contains photos from 597 male and female targets of varying ethnicity between 18 and 40 years under standardized conditions. For a subset of 158 targets, images display either a neutral, angry, fearful, or two positive emotional states (happy face with either an open or closed mouth). Each of these images of the 158 subjects displaying various emotions was classified using five major automated emotion detection systems via an Application Programming Interface (API): Microsoft Cognitive Services, Google ML, Sightcorp, Kairos, and the GfK EmoScan. The data was normalized across emotion detection systems to provide meaningful comparisons in the following analyses.

To set a baseline, we evaluated the effectiveness of the different emotion detection systems by predicting each image emotion label via the corresponding emotion measure (i.e., predicting happy faces using the happiness feature of each emotion detection system). A multinomial logit model revealed substantial variation across emotion detection systems with the highest, cross-validated (CV) prediction accuracy for Google ML with a test accuracy of 58.94% and a test Kappa of 40.25%, compared to the lowest prediction accuracy for GfK EmoScan with a test sample accuracy of 49.81% and a test Kappa of 36.88%. Positive emotions were predicted consistently better compared to the substantial variation across negative emotion labels (variability across emotion detection systems: $\sigma_{\text{happy open}} = 3.81$, $\sigma_{\text{neutral}} = 6.89$, $\sigma_{\text{fear}} = 14.98$, $\sigma_{\text{anger}} = 30.9$) and overall lower predictive power for negative emotions than for positive and neutral emotions (average balanced accuracy for the different emotion classes: $\mu_{\text{happy open}} = 78.97$, $\mu_{\text{neutral}} = 79.97$, $\mu_{\text{fear}} = 65.45$, $\mu_{\text{anger}} = 44.3$). These findings were robust even after controlling for all other emotion features.

Next, we aimed at further improving prediction accuracy in two ways. First, by increasing the feature space through combining the entire set of features across all emotion detection systems. Second, by using more flexible model estimation and machine learning techniques. Specifically, recent advances in deep learning might further improve predictive accuracy by increasing the number of layers and / or boosting a model by aiming at predicting misclassified cases. In short, increasing the feature space resulted in a greater test accuracy with an averaged accuracy of 85.73% and a Kappa of 81.02%.

Next, we aimed at further improving prediction accuracy by using more flexible machine learning techniques. Parameters were selected through grid search in the parameter space. Lowest performance was achieved by Support Vector Machines with radial basis kernel resulting in a test accuracy of 84.41% and a test Kappa of 79.33%. Highest performance was achieved by using Random Forests and LogitBoost achieving 92.02% test accuracy, 89.25% test kappa and 93.60% accuracy, 91.34% kappa, respectively. These changes led to a significant increase in predictive power also for negative discrete emotions and substantially reduced variation across positive, neutral and also negative emotion labels ($\sigma_{\text{happy open}} = 2.71$, $\sigma_{\text{neutral}} = 1.36$, $\sigma_{\text{fear}} = 5.32$, $\sigma_{\text{anger}} = 4.61$).

To the best of our knowledge, this is the first systematic study demonstrating the striking variability in automated emotion detection systems across discrete emotions and we provide two easy to implement modelling strategies to improve prediction accuracy across automated emotion detection systems.

REFERENCES

Bakhshi, Saeideh, David A. Shamma and Eric Gilbert (2014), "Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Toronto, Canada: ACM, 965–974.

Barasch, Alixandra, Gal Zauberman and Kristin Diehl (2018), "How the Intention to Share Can Undermine Enjoyment: Photo-Taking Goals and Evaluation of Experiences," *Journal of Consumer Research*, 44 (6), 1220–1237.

Berger, Jonah, Ashlee Humphrees, Stephen Ludwig, Wendy Moe, Oded Netzer, and David Schweidel (2020), "Uniting the Tribes: Using Text for Marketing Insight," *Journal of Marketing*, 84 (January), forthcoming.

Berger, Jonah, and Grant Packard (2018), "Are Atypical Things More Popular?" *Psychological Science*, 29 (7), 1178–84.

Berger, Jonah, and Katherine L. Milkman (2012), "What Makes Online Content Viral?" *Journal of Marketing Research*, 49 (2), 192–205.

Berger, Jonah, Wendy W. Moe, and David A. Schweidel (2019), "What Makes Stories More Engaging? Continued Reading in Online Content," *Working Paper*.

Berman, Ron, Shiri Melumad, Colman Humphrey, and Robert Meyer (2019), "A Tale of Two Twitterspheres: Political Microblogging During and After the 2016 Primary and Presidential Debates," *Journal of Marketing Research*, 56(6), 895–917.

Eliashberg, Jehoshua, Sam K. Hui, and Z. John Zhang (2007), "Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach," *IEEE Transactions on Knowledge & Data Engineering*, 26 (11), 2639–48.

Erfgen, Carsten, Sebastian Zenker and Henrik Sattler (2015), "The Vampire Effect: When Do Celebrity Endorsers Harm Brand Recall?", *International Journal of Research in Marketing*, 32 (2), 155–163.

Fox, Elaine, Victoria Lester, Riccardo Russo, R. J. Bowles, Alessio Pichler, and Kevin Dutton (2000), "Facial expressions of emotion: Are angry faces detected more efficiently?," *Cognition and Emotion*.

Gibson, Edward (1998), "Linguistic Complexity: Locality of Syntactic Dependencies," *Cognition*, 68(1), 1–76.

Grewal, Lauren, Andrew T. Stephen and Nicole V. Coleman (2019), "When Posting About Products on Social Media Backfires: The Negative Effects of Consumer Identity Signaling on Product Interest," *Journal of Marketing Research*, forthcoming.

Hanssens, Dominique M. and Barton A. Weitz (1980), "The Effectiveness of Industrial Print Advertisements Across Product Categories," *Journal of Marketing Research*, 17 (3), 294–306.

Humphreys, Ashlee, and Kathryn A. LaTour (2013), "Framing the Game: Assessing the Impact of Cultural Representations on Consumer Perceptions of Legitimacy," *Journal of Consumer Research*, 40(4), 773–795.

Liu, Ping, Shizhong Han, Zibo Meng, and Yan Tong (2014), "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1805–12.

Ma, Debbie S, Joshua Correll, and Bernd Wittenbrink (2015), "The Chicago face database: A free stimulus set of faces and norming data," *Behavior research methods*, 47(4), 1122–35.

Pantic, Maja, and Leon Ü M. Rothkrantz (2000), "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Proserpio, Davide and Hauser, John R. and Liu, Xiao and Amano, Tomomichi and Burnap, Alex and Guo, Tong and Lee, Dokyun (DK) and Lewis, Randall A. and Misra, Kanishka and Schwartz, Eric M. and Timoshenko, Artem and Xu, Lilei and Yoganarasimhan, Hema, Soul and Machine (Learning) (September 16, 2019). Available at SSRN: https://ssrn.com/abstract=3454294 or http://dx.doi.org/10.2139/ssrn.3454294

Wood, Adrienne, Magdalena Rychlowska, Sebastian Korb, and Paula Niedenthal (2016), "Fashioning the Face: Sensorimotor Simulation Contributes to Facial Expression Recognition," *Trends in Cognitive Sciences*.

Xiao, Li and Min Ding (2014), "Just the Faces: Exploring the Effects of Facial Features in Print Advertising," *Marketing Science*, 33 (3), 338–352.