Mining meaning of videos on YouTube: Unraveling latent content from digital influencers and their engagement

Eliane Francisco-Maffezzolli
Pontifical Catholic University of Paraná, Curitiba Campus (PUCPR)
Ana Cristina Munaro
Pontifícia Universidade Católica do Paraná PUCPR
João Pedro Santos Rodrigues
Pontifícia Universidade Católica do Paraná (PUCPR)
Emerson Cabrera Paraiso
Pontifícia Universidade Católica do Paraná (PUCPR)

Acknowledgements:
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

Cite as:
Francisco-Maffezzolli Eliane, Munaro Ana Cristina, Santos Rodrigues João Pedro, Cabrera Paraiso Emerson (2021), Mining meaning of videos on YouTube: Unraveling latent content from digital influencers and their engagement. *Proceedings of the European Marketing Academy*, 50th, (93144)

Paper from the 50th Annual EMAC Conference, Madrid, May 25-28, 2021

Mining meaning of videos on YouTube: Unraveling latent content from digital influencers and their engagement

Abstract: Unstructured data plays a key role in the consumer decision-making process. Advanced techniques for linguistic analysis allow extracting meaning from the content provided by digital influencers. In this paper, we identify the key dimensions of video content on YouTube using a data mining approach, Latent Dirichlet Analysis (LDA). The data set includes 38,427 videos transcript for 103 digital influencers channels over more than 10 years. LDA uncovers 19 content dimensions that remain stable over the past few years on YouTube, highlighting 6 content categories with greater digital engagement: Culture and Entertainment; Family; People, Behavior and Lifestyle; Education; Beauty, and Gastronomy. Dimensions that are key for content creators and professionals to strategically manage their digital engagement with the audience.

Keywords: *digital influencer content, digital engagement, latent Dirichlet allocation (LDA)*

Paper track: *Digital Marketing & Social Media*

1 Introduction

Visual information is becoming more and more prevalent in online markets, and companies are relying more than ever on online videos to introduce, promote, and advertise their products and services (Li, Shi, & Wang, 2019). Consumers, in turn, are overwhelmed by the proliferation of online content, and it seems clear that marketers will not succeed without engineering this content for their audience (Lee, Hosanagar, & Nair, 2018).

One of the most important factors affecting consumer engagement is the content that brands disseminate through social media (Lee et al., 2018). Brand-generated communication on social networks generates traceable attention, affects consumers' attitude toward branded content, engages consumers cognitively and emotionally, and can drive consumers to advocate for the brand (Gavilanes, Flatten, & Brettel, 2018). Also, can rise the customers' trust beliefs in the integrity, ability and benevolence of salespeople (Yaghtin, Safarzadeh, & Zand, 2020).

Understanding the strategy of brand content on social networks is crucial. High-quality content (organic and relevant) is imperative for brand communication (Voorveld, 2019). The content is elastic and is much broader than advertising (van Noort, Himelboim, Martin, & Collinger, 2020). And we still do not understand precisely what kinds of content work better for which firms and in what ways (Lee et al., 2018). Thus far, academic research focuses mainly on online behavioral advertising and social media advertising (van Noort et al., 2020).

Besides, while text information has been widely studied and used, academic research has lagged in analyzing visual information and provides little guidance on how to design an effective online video (Li et al., 2019; Ma & Sun, 2020). Further, the high dimensionality, massive volume and unstructured data make machine learning methods more efficient than human analyses (Liu, Burns, & Hou, 2017; Ma & Sun, 2020). Thereat, we adopted a method for knowledge extraction from videos through audio transcriptions (Rodrigues & Paraiso, 2020).

The goal of the study is to investigate what content-related topics do digital influencers discuss on YouTube. Then, we propose a unified framework for (1) extracting the latent dimensions of content from digital influencers channels on YouTube; (2) ascertaining the labels, dynamics, and heterogeneity of those dimensions; and (3) using those dimensions for strategy analysis. We use the brand-generated content and video features on YouTube as theoretical framework, considering digital influencers as content creators.

2 Brand-generated Content

Content marketing seeks to develop content that better engages targeted users and drives the desired goals of the marketer (Lee et al., 2018). Rancati and Gordini (2014) define content marketing as being a tool to share content, but also to create value and high returns along with the financial means of customer distribution, attraction, involvement, acquisition, and retention. We focused on brand-generated content (BGC), specifically, content created by digital influencers.

BGC entails content that is deliberately planned and distributed by the brand (van Noort et al., 2020). Generally, the content must: 1) be able to generate interest, involving, but also informing and educating the customer; 2) express all those values that identify the firm in terms of uniqueness, consistency, quality and relevance; 3) be pro-active, that can evolve over time (Rancati & Gordini, 2014).

Yaghtin et al. (2020) categorized four main content classes including task-oriented (e.g., corporates' products/services/resources information), emotion-oriented (content that influences the audiences and makes them comprehend deeply the concept), interaction-oriented (content not directly related to the brand or corporates' products/services) and advertising-oriented (content with the main purpose of directly promoting the corporates' brands or the products).

BGC classifications in the literature focus on the purpose of the content and neglect underlying factors of that content regarding "how" it is delivered to the public, for instance, the most used terms, and relevant keywords. Analyzing the emerging topics of content on digital platforms is a determinant facet to achieve effective returns. Firms could identify the appropriate content to disseminate to their unique audience and develop a message by using words that have a high probability of being associated with that content topic (as determined by LDA) (Zhang, Moe, & Schweidel, 2017). They could also use influentials and disseminate messages with words that are associated with the topics generally broadcasted by them (Zhang et al., 2017).

2.1 Video content on YouTube

Li et al. (2019) presented measures automatically obtained from videos by a machine learning algorithm, such as hue, brightness, and saturation. However, it is missing studies presenting the discovery and extraction of hidden semantic structures from textual data from

videos on YouTube, with the semantic information represented using topics as topic modeling proposes. Since text-based language is a central component of marketing communications on social media, understanding aspects of language that drive engagement is imperative (Pezzuti, Leonhardt, & Warren, 2021).

Users are more likely to rebroadcast content that matches their interests. This implies that organizations can tailor content to match the audience's interests to increase rebroadcasting activity from them rather than simply disseminate viral content (Zhang et al., 2017). When the reactions of audiences are available, it is helpful to incorporate individual heterogeneity into the analysis of videos, this development could also be important for practitioners: with the ability to personalize recommendations, firms could deliver different video content to different users (Li et al., 2019). Therefore, associating videos posted to their respective results of digital engagement (number of views, likes, comments) is a strategy to understand the idiosyncrasies of the public.

3. Method

The process used in our study comprises collect audio transcriptions from videos, after executing a text preprocessing, perform the Latent Dirichlet Allocation algorithm (Blei, Ng, & Jordan, 2003), and text analysis. Figure 1 summarizes the methodological process.

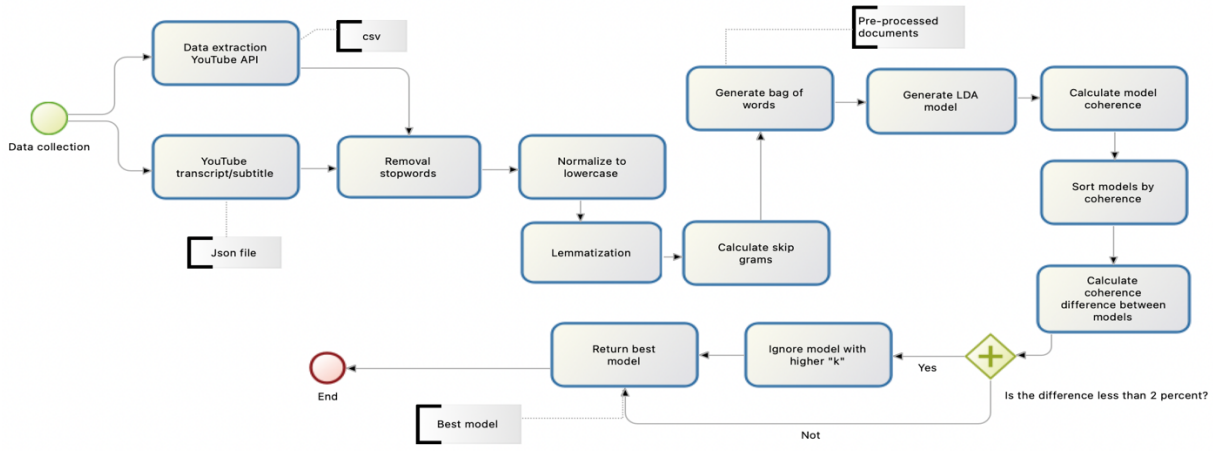

Figure 1. Methodological model

3.1 Data collection and data processing

The study collected data on the number of views, likes, dislikes, and comments, topics content and other video post characteristics, from 38,427 videos posted on YouTube among 103

different YouTubers channels in 26 categories between January 2008 and October 2020. We identified the top digital influencers channels by "*Prêmio Influenciadores Digitais*" list of 2019-2020. Data collection was conducted via the Python programming language. We used the YouTube API (application programming interface) for the extraction process, also an open-source tool from Python API to extract all available auto-generated captions for the videos.

Before performing LDA, we implemented text pre-processing by using modules of the Natural Language Toolkit (www.nltk.org) (Guo et al., 2017). Text pre-processing used steps very similar to those adopted in prior studies (e.g., Tirunillai & Tellis, 2014; Guo et al., 2017; Debortoli, Müller, Junglas, & vom Brocke, 2016). The first step was to normalize the text, were removed stopwords, converting it to lower case, lemmatization is also applied. We apply part-of-speech (POS) tagging to retain only words that are adjectives, nouns, or adverbs (Debortoli et al., 2016; Tirunillai & Tellis, 2014). Also, all bi-grams and tri-grams present in the corpus are identified. Finally, the bag-of-words was generated from the count of the present terms.

LDA requires us to specify the number of topics to be extracted prior to running the analysis. Thus, it was necessary to develop an approach that would allow finding the best number of k in a corpus. The method generates a set of N models to increase the number of k topics. After creating each model, its coherence is calculated, the next step is to select the best model, with the most appropriate number of topics representative of the corpus. The model selected was that one with 50 topics and a 45,4% consistency.

4. Data Analysis

We assigned a label to the given dimension such that it reflects the topic of discussion being evaluated across all the videos expressing the dimension, the words as well as its weights that are important for a given dimension determine its label or provide direction to its labeling (Tirunillai & Tellis, 2014; Liu et al., 2017; Li & Ma, 2020). As with manually coding texts, following Debortoli et al., (2016), two independent researchers interpreted and labeled the topics. Moreover, was analyzed channels and previous content categories related to each topic.

The naming of 50 topics resulted in 19 different content labels, which are Beauty, Culture & Entertainment, Decoration, Organization & DIY, Education, Economics, Entrepreneurship & Business, Entertainment/general, Family, Fashion/Lifestyle, Gaming, Gardening, Gastronomy,

Health & healthy lifestyle, Military, People, Behavior & Lifestyle, Pets & Animals, Politics, Economy & News, Sports, Tech, and Travel, learnings & curiosities. In Table 1, we present the 15 topics with the highest percentage of videos to illustrate the topics, top words, and labeling.

Table 1. The 15 most representative topics in the sample

Topic	Proportion (total videos)	Top 10 words	Labeling	Examples of associated channels
1	8.50 (2,937)	power, guys, fight, time, hero, new, powerful, strong, picture, world	Culture & Entertainment	Ei Nerd, Bibi, Meteoro Brasil, Whindersson Nunes
2	5.38 (1,861)	cake, good, recipe, chocolate, dough, milk, love, little, form, minute	Gastronomy	TPM por Ju Ferraz, Receitas da Cris, Receitas de Pai, Dani Noce
3	4.60 (1,589)	govern, bolsonaro, president, politician, brazil, lula, public, be, country, leave	Politics, Economy & News	Kim Kataguiri, TV Afiada, Mamaefalei, Nando Moura
4	3.94 (1,361)	hair, makeup, foundation, little, product, shadow, face, look, good, tone	Beauty	Mariana Saad, Mari Maria, NiinaSecrets, Bianca Andrade
5	3.83 (1,323)	device, camera, screen, photo, good, hit, best, cellphone, samsung, iphone	Tech	TudoCelular, Canaltech, Dudu Rocha, Be!Tech
6	3.77 (1,303)	cool, white, see, blue, red, landmark, time, pool, nice, new	Family	Brancoala, Flavia Calina, resendeevil, T3ddy
7	3.47 (1,199)	god, good, love, photo, friend, kiss, happy, life, world, day	People, Behavior & Lifestyle	Taciele Alcolea, Central de fãs de Luisa Mell, Graciele Lacerda dia a dia, Evelyn Regly
8	3.15 (1,088)	cut, side up, paper, paint, paste, ready, down, line, piece	Decoration, Organization & DIY	Dany Martines, Paula Stephânia, Diycore com Karla Amadori, Manual do Mundo
9	2.72 (941)	good, tasty, food, water, little, meat, coffee, dish, chicken, cheese	Gastronomy	Tastemade Brasil, Dani Noce, Receitas de Pai, Sal de Flor
10	2.59 (894)	wall, bedroom, bathroom, door, cooking, space, room, wood, table, bed	Decoration, Organization & DIY	Doma Arquitetura, Diycore com Karla Amadori, Organize sem Frescuras!, Maurício Arruda
11	2.42 (837)	money, real, year, month, account, bank, investment, value, tax, person	Economics, Entrepreneurship & Business	Me poupe!, O Primo Rico, Bruno Perini, Tiago Fonseca
12	2.35 (812)	dog, liven, cat, animal, creature, species, fish, cat, big, huge	Pets & Animals	Richard Rasmussen, Estopinha & Alexandre rossi, Central de fãs de Luisa Mell, Você Sabia?
13	2.33 (805)	travel, place, hotel, cool, hour, plane, city, world, dollar, day	Travel, learnings & curiosities	Estevam Pelo Mundo, Melhores Destinos, Prefiro Viajar, Viajo logo existo
14	2.23 (772)	ball, play, goal, team, challenge, football, first, cup, fred, good	Sports	Desimpedidos, Raquel Freestyle, Jogo Aberto, Denílson Show
15	2.21 (763)	clothes, cool, beautiful, store, lovely, box, pretty, wonderful, gift, purse	Fashion/ Lifestyle	Organize sem frescuras!, Taciele Alcolea, NiinaSecrets, Flavia Pavanelli

Note. Most frequent words (top 10) from selected topics using the LDA model.

The top 5 topics represent more than 26% of the total sample. Topic 13 had the highest number of related videos and has appeared in 95 different channels of digital influencers. Topic 26 showed up in 43 different channels, its keywords representing cooking ingredients,

preparation methods, and adjectives from the gastronomy universe. Topic 7 was presented in 31 channels, the most likely words referring to politics and the Brazilian political/economic scenario. Topic 0 comprised 43 channels, its keywords refer to makeup and aesthetics products, body parts, and related adjectives. Topic 46 revealed in 15 channels, it is interesting to note that it is one of the top 5 topics in the corpus with less distribution among channels.

4.1 YouTube content overview

Based on the study's results, 'Education', 'Culture & Entertainment', 'People, Behavior & Lifestyle' and 'Gastronomy' content categories are the most popular among the digital influencers on YouTube. Education is the most prevalent content among the more than 38 thousand videos analyzed. It covers 8 topics in 3,555 videos, representing 10.3% of the total. This result is consistent with the main reasons for the audience watching YouTube. According to more than 12,000 people worldwide (Google, 2019), the best reasons given by people to watch YouTube include the opportunity to learn something new and to dig deeper into one's interests.

Culture & Entertainment ranks second on the most created content on YouTube (10.13% of total). It consists of two topics corresponding to 3,501 videos. It covers content that also aims to supply the curiosities and needs of the public combined with entertainment common to other social networks. People, Behavior & Lifestyle is the third most representative content among Brazilian influencers (8.5% of total), composed of 5 topics in 2,929 videos. It is a style of content that reflects human behavior, self-learning, reflections and lifestyles. Thus, it is a topic that permeates most channels in greater or lesser amounts.

Gastronomy (2 topics – 8.1%) and tech (4 topics – 7.9%) are contents that complete the 'top 5' among digital influencers. Followed by Politics, Economy & News (2 topics – 6.4%), Health and healthy lifestyle (4 topics – 5.9%), Decoration, Organization & DIY (2 topics – 5.7%), Economics, Entrepreneurship & Business (3 topics – 5.5%) and Family (2 topis – 5.4%) that complete the list of 10 most created/disseminated content among digital influencers.

LDA can be used to visualize the temporal development of topics. Overall, the contents of the topics remain stable. From 2014-2015, the topics started to gain relevance in a number of videos, years that match with the beginning of the professional career of digital influencer in Brazil. The most prominent topics are related to Fashion/Lifestyle, Culture & Entertainment, and

Family in continuous growth. Some topics showed some decline such as those relates to Gaming, and Entertainment. One wonders if these are terms that are in decline and if the topics have a life cycle, for example, in the case of games, one may be discontinued or lose the public's attention.

Analyzing the distribution of topics between channels, an average of 20 topics per channel is perceived, however, 52 channels present a single topic representing more than 50% of the content of the channel's videos (range 51.6% to 97.1%). Besides, for the most part, one or two topics represent much of the content of the channels. Basically, two topics summarize the digital influencer's content creation script.

Analyzing the relationship between the content and the results of digital engagement, 6 topics are highlighted, which are among the top 10 channels with the highest number of views, number of likes, comments, and dislikes (see Table 2). Topic 40 (Family labeled) is the largest in number of views, and in number of dislikes, the second topic in number of likes, and the sixth in number of comments. Topic 13 (Culture & Entertainment), the second in the ranking in number of views, the highest number of likes and comments, and the second in the number of dislikes.

Topic 26 (Gastronomy), third in number of views, seventh in likes, sixth in dislikes and eighth in number of comments. Followed by topic 27 (People, Behavior and Lifestyle), fifth in number of views and comments, third in number of likes and seventh in dislikes. Rounding out the top 6 list, topic 29 (Education), and topic 0 (Beauty).

Table 2. Topics associated with greater digital engagement

Label	Number channels	Topic	Total Videos	Nº Views	Nº Likes	Nº Dislikes	Nº Comments
Culture & Entertainment	95	13	2,937	**2,837,621,147**	**295,615,053**	**3,709,281**	**9,964,773**
Family	74	40	1,303	**3,400,544,604**	**103,450,628**	**3,796,103**	**2,266,477**
People, Behavior & Lifestyle	69	27	1,199	**846,274,668**	**89,509,236**	**930,214**	**2,753,016**
Entertainment/general	23	8	372	837,121,596	75,437,368	813,837	1,660,115
Education	35	29	517	**776,596,493**	**73,958,484**	**1,106,737**	**2,906,724**
Beauty	43	0	1,361	**714,310,073**	**68,455,565**	**767,523**	**2,871,953**
Gastronomy	43	26	1,861	**1,071,900,602**	**66,547,497**	**1,024,933**	**1,964,037**
Pets & Animals	38	44	812	696,201,768	56,566,211	757,986	1,452,640
Sports	39	14	772	720,392,877	54,159,710	766,397	1,468,965
Education	26	36	529	533,635,953	53,074,079	672,553	2,158,053
Politics, Economy & News	31	7	1,589	446,821,941	52,750,412	2,923,144	3,816,130
Travel, learnings&curiosities	65	12	393	619,938,100	52,432,747	738,320	1,274,099
Decoration, Organization & DIY	39	17	1,088	924,291,357	49,230,238	759,413	1,705,115
Culture & Entertainment	33	38	564	527,061,542	48,050,369	704,981	1,545,639

People, Behavior&Lifestyle	49	43	529	404,529,577	46,779,720	495,255	1,727,293
Family	56	4	553	547,394,390	43,847,063	484,000	1,043,645
Gaming	40	6	663	456,882,188	42,278,282	503,941	1,458,319
Politics, Economy & News	35	11	629	336,688,921	38,548,484	1,166,129	1,890,595
People, Behavior&Lifestyle	48	45	533	408,030,571	33,715,477	351,538	961,774
Economics, Entrepreneurship & Business	39	34	837	356,237,568	31,723,586	450,462	1,011,786

Note. The 20 most popular channels ranked by the number of likes.

5. Discussion and Conclusions

We identify 19 key dimensions of video content on YouTube using a data mining approach, latent dirichlet analysis (LDA), in a data set includes 38,427 videos transcript for 103 digital influencers channels for more than ten years. The study highlights six content categories with greater digital engagement among digital influencers: Culture and Entertainment; Family; People, Behavior and Lifestyle; Education; Beauty, and Gastronomy. Besides, it can be seen that the topics have remained stable over the past few years on YouTube, which launches insights for content creators and brand managers on the consumption of content on the platform. Furthermore, the topic analysis opens paths for little content addressed or that can be better exploited as environmental issues, gardening, and the animal world.

Today, unstructured data plays a key role in the consumer decision-making process (Li et al., 2019). Latent topics and the dynamics of latent topics can serve as important indicators for marketing managers to track, evaluate, and incorporated the outcomes into the firm's automated marketing planning and allocation (Li & Ma, 2020). Several studies in the area focus on the consumer's view (e.g., Tirunillai & Tellis, 2014; Büschken & Allenby, 2016; Guo et al, 2017; Liu et al., 2017). Our study fills a gap in the literature on the dynamics of video content on social networks from the view of content creators (digital influencers) as BGC. Digital influencers have a more significant impact on brand attitudes and purchase behaviors than traditional celebrities, they may affect positively advertising effectiveness (Schouten, Janssen, & Verspaget, 2020).

This study has some important limitations. The study did not analyze the emotional polarity of the topics found and does not consider the personality traits of digital influencers. Second, the study does not consider audience data from influencers, such as demographic and/or psychographic data. Third, we do not analyze rare or infrequent words in the long tail of the distribution. Each of the above limitations could be rich avenues for further research.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953-975.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, *39*(1), 7.

Gavilanes, J. M., Flatten, T. C., & Brettel, M. (2018). Content strategies for digital consumer engagement in social networks: Why advertising is an antecedent of engagement. *Journal of Advertising*, *47*(1), 4-23.

Google (2019). Insight Strategy Group, Global, "Premium Is Personal" studies, AU, BR, CA, DE, IN, JP, KR, U.K., U.S. In: What the world watched in a day. Retrieved from https://www.thi nkwithgoogle.com/feature/youtube-video-data-watching-habits/. (Last accessed: Nov. 20, 2020).

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467-483.

Lee, D., Hosanagar, K., & Nair, H. S. (2018). Advertising content and consumer engagement on social media: evidence from Facebook. *Management Science*, *64*(11), 5105-5131.

Li, X., Shi, M., & Wang, X. S. (2019). Video mining: Measuring visual information using automatic methods. *International Journal of Research in Marketing*, *36*(2), 216-231.

Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on Twitter. *Journal of Advertising*, *46*(2), 236-247.

Pezzuti, T., Leonhardt, J. M., & Warren, C. (2021). Certainty in Language Increases Consumer Engagement on Social Media. *Journal of Interactive Marketing*, *53*, 32-46.

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing*.

Rancati, E., & Gordini, N. (2014). Content marketing metrics: Theoretical aspects and empirical evidence. *European Scientific Journal*, *10*(34).

Rodrigues, J. P., & Paraiso, E. (2020). From audio to information: Learning topics from audio transcripts. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*,121-128.

Schouten, A. P., Janssen, L., & Verspaget, M. (2020). Celebrity vs. Influencer endorsements in advertising: the role of identification, credibility, and Product-Endorser fit. International journal of advertising, 39(2), 258-281.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463-479.

van Noort, G., Himelboim, I., Martin, J., & Collinger, T. (2020). Introducing a model of automated brand-generated content in an era of computational advertising. *Journal of Advertising*, *49*(4), 411-427.

Voorveld, H. A. (2019). Brand communication in social media: a research agenda. *Journal of Advertising*, *48*(1), 14-26.

Yaghtin, S., Safarzadeh, H., & Zand, M. K. (2020). Planning a goal-oriented B2B content marketing strategy. *Marketing Intelligence & Planning*, 38(7), 1007-1020.

Zhang, Y., Moe, W. W., & Schweidel, D. A. (2017). Modeling the role of message content and influencers in social media rebroadcasting. *Int. Journal of Research in Marketing*, 34(1),100-119.