RBGN REVISTA BRASILEIRA DE GESTÃO DE NEGÓCIOS
© FECAP

ISSN 1806-4892
e-ISSN 1983-0807

OPEN ACCESS

ARTICLE

Beyond the screen: a creative exploration of content that engages on YouTube discussed by social media influencers

Ana Cristina Munaro[1]
Eliane Cristine Francisco Maffezzolli[1]
João Pedro Santos Rodrigues[2]
Emerson Cabrera Paraiso[2]

Abstract

Purpose – The study aims to investigate the most popular content discussed by social media influencers on YouTube and its associated valence, to delineate the content categories favored by top Brazilian influencers, and to assess their impact on consumer digital engagement.

Theoretical framework – This study draws upon influencer marketing, social media influencer (SMI) literature, and digital engagement.

Design/methodology/approach – A data mining approach was used. The methodology includes the collection of video post characteristics, engagement metrics, and audio transcriptions from 34,563 videos on 103 YouTube channels. After textual preprocessing, a topic modeling stage is performed using the Latent Dirichlet Allocation (LDA) algorithm and sentiment analysis.

Findings – The study identified 19 critical dimensions of video content on YouTube. The top 3 content categories with the highest user digital engagement are: Family, Entertainment/General, and Culture & Entertainment. The sentiment analysis shows that content about Beauty, Gastronomy, and Economics, Entrepreneurship & Business have the highest proportional positive valence. Politics, Economy & News, Entertainment/General, and Gaming have high percentages of negative valence.

Practical & social implications of research – The results provide a deep understanding of YouTube's popular content and digital engagement rates. This is essential for companies and SMIs looking to maximize their reach, resonate with their target audience, and stay competitive in the dynamic digital landscape. It allows for more effective communication, content creation, and strategic decision-making.

Originality/value – Understanding the content on YouTube can provide valuable insights for businesses, marketers, and content creators to optimize their communication strategies.

Keywords: Social media influencer, YouTube, video content, digital engagement, topic modeling.

1. *Pontifícia Universidade Católica do Paraná, Escola de Negócios, Programa de Pós-graduação em Administração, Curitiba, PR, Brasil*
2. *Pontifícia Universidade Católica do Paraná, Programa de Pós-graduação em Informática, Curitiba, PR, Brasil*

How to cite:
Munaro, A. C., Francisco Maffezzolli, E. C., Rodrigues, J. P. S., Paraiso, E. C. (2024). Beyond the screen: a creative exploration of content that engages on YouTube discussed by social media influencers. *Revista Brasileira de Gestão de Negócios*, *26*(3), e20240002. https://doi.org/10.7819/rbgn.v26i03.4275

Received on:
Jan/11/2024
Approved on:
Sept/10/2024

Responsible editor:
Prof. Dr. Arnold Japutra

Reviewers:
Wilian Ramalho Feitosa;
Domingo Ribeiro-Soriano

Evaluation process:
Double Blind Review

This article is open data

Revista Brasileira de Gestão de Negócios

https://doi.org/10.7819/rbgn.v26i03.4275

1 Introduction

Brands dedicate a substantial portion of their budgets to influencer marketing, while social media companies focus on creating top-notch platforms for social media influencers (SMIs, hereafter influencers). The market grew from $1.7 billion in 2016 to $13.8 billion in 2021. According to Goldman Sachs, the global influencer market is expected to be worth around $480 billion by 2027 (Agência Brasil, 2024), indicating strong growth. Additionally, 93% of marketers have integrated influencer marketing into their campaigns (Santora, 2022). Video influencers are particularly noted for their ability to attract followers and influence purchasing decisions (Chen et al., 2023).

Influencers have gained widespread popularity across various domains, including marketing strategy. These skilled content creators offer businesses a significant advantage by engaging directly with targeted consumers (Ata et al., 2022). For brands, collaborating with influencers who resonate with their target audience is crucial for their customer acquisition efforts and campaign success (Digital Marketing Institute, 2024). According to the Digital Marketing Institute, businesses can earn an average of $5.78 for every dollar spent on influencers, with some seeing as much as $18 (Digital Marketing Institute, 2024). However, partnering with influencers can be risky due to the potential association with controversial electronic word-of-mouth (eWOM), making influencer selection a challenging task (Chung & Cho, 2017).

Implementing and evaluating influencer marketing requires substantial resources and effort, highlighting the importance of identifying effective decision criteria (Leung et al., 2022). For example, increasing video views is an important goal for those seeking fame, a large audience, or higher revenue (Yoon & Lee, 2022). However, the choice of influencers or YouTube channels should consider not only the number of views and subscribers, but also how influencers communicate and engage with their audience (Munaro et al., 2021; Berger et al., 2023). Therefore, although academic research has focused on online behavioral and social media advertising (van Noort et al., 2020), little is known about the factors that drive the success of online engagement with influencers (Hughes et al., 2019), and regarding the choice of a particular influencer as a brand endorser.

The success of influencer marketing hinges on the type of content shared, with certain posts more effectively driving consumer engagement and boosting sales (Zhang et al., 2017; Lee et al., 2018). Online videos have become increasingly important for promoting and advertising products and services, with visual content and video influencers playing a pivotal role in the digital marketplace (Li et al., 2019; Chen et al., 2023). However, given the overwhelming volume of content consumers face, marketers need to precisely tailor their content to resonate with their target audience (Lee et al., 2018).

High-quality content is imperative for brand communication (Voorveld, 2019). Content is elastic and much broader than advertising itself (van Noort et al., 2020), so understanding the strategy of content created on social networks is crucial. However, we still need to uncover the optimal content strategies tailored to specific companies and their unique effectiveness (Lee et al., 2018). Traditional research analyzes videos using controlled lab experiments, which are costly, time-consuming, and limited to small scales (Li et al., 2019). Moreover, while textual information has been widely studied and used, academic research needs more guidance on how to design compelling online videos (Li et al., 2019; Ma & Sun, 2020).

Our study seeks to thoroughly understand the types of content that are most well received by audiences on YouTube, in order to help companies select the most appropriate influencers based on user engagement. To address gaps in the existing literature, we investigate the most popular content and the emotional tone of social media influencer posts on YouTube. Specifically, the research identifies common content categories among top Brazilian influencers, analyzes the emotional tone of these topics, and assesses their effectiveness in terms of consumer digital engagement.

As contributions, first, our study broadens the scope of influencer marketing research, addressing a gap in the study of engagement predictors by emphasizing the significance of content based on textual analysis. Second, we advance the understanding of how content shapes consumer responses to marketing communications. Analyzing emerging content topics is essential for maximizing returns on digital platforms. Brands can craft effective messages with the right tone, reinforcing influencer-follower relationships and enhancing overall influence (Zhang et al., 2017; Jacobson et al., 2022). Third, while many studies emphasize the consumer's viewpoint and user-generated content (UGC) through reviews, tweets, and comments (e.g., Tirunillai & Tellis, 2014; Büschken & Allenby, 2016; Guo et al., 2017; Liu et al., 2017), our

study explores the dynamics of video content on social media from the perspective of influencers as creators of brand-generated content. Lastly, our study offers practical insights for marketers, influencers, and stakeholders seeking to enhance content design. Through real-world data analysis, we unveil how minor tweaks to influencers' content can significantly impact digital engagement.

2 Social media influencers

Social media influencers have become a global phenomenon. Brands invest a considerable portion of their budgets in influencer marketing, while social media companies invest in building the best digital platforms for influencers. Meanwhile, the number of people who want to become influencers has increased. An influencer is a person who consistently creates content for a specific audience on a social media platform, establishes a relationship with their audience, stands out as an opinion leader in a community, and ultimately influences the attitudes or behavior of other individuals (Sette & Brito, 2020).

Currently, influencers exert a more substantial influence on brand attitudes and purchasing behavior compared to traditional celebrities (Nunes et al., 2018; Schouten et al., 2020; Chen et al., 2023). This is especially true because social networks have become the most important source of consumer insights, where consumers have access to quality, useful, and credible eWOM information, which in turn influences their attitudes (Nyagadza et al., 2023).

According to Ata et al. (2022), the influencer's trustworthiness, expertise, and attractiveness positively affect the attitude toward the advertisement. Moreover, this phenomenon is attributed to the consumers' heightened sense of similarity and identification with the influencer, driven by desired proximity, trust, and perceived similarity (Aleti et al., 2019; Schouten et al., 2020). Since influencers present the content generated about a product in their natural life, the given message is much more critical and increases trust. By establishing a connection with the brand, the influencer presents the brand to the consumer with their naturalness and clear language (Ata et al., 2022).

Notably, the literature on influencers highlights the pivotal role of specific content characteristics in shaping consumers' perceptions of information credibility and usefulness, consequently influencing brand attitudes and purchase intentions (Nunes et al., 2018; Hughes et al., 2019; Munaro et al., 2021). Casaló et al. (2020), for instance, underscore the significance of content originality and uniqueness in shaping an influencer's perceptions. Accordingly, when consumers compare the source of an advertisement, the influencer is perceived as a more credible source (Ata et al., 2022).

Furthermore, while no single content attribute exclusively explains social influence, insights from the influencer marketing literature provide clues to various elements that contribute to its success. Jacobson et al. (2022) emphasize that an influencer's strategies, content choices, and social presence collectively influence their relationships and level of influence with followers. Therefore, an exploration of the impact of social media content, including the type of content posted and interactions with followers, warrants deeper analysis (Shahbaznezhad et al., 2020).

2.1 Video content on YouTube

Video is quickly becoming a vital tool for business owners to help them leverage their brand, build loyalty, and add new customers. TikTok dances, YouTube celebrities, and influencers are prompting many small business owners to implement video marketing to reach customers on their mobile devices (Plummer, 2022). Videos are also an educational tool, and even in medicine, we see 3D surgical videos to record surgeries, provide a complete medical record, and allow surgeons to review their surgical actions (Patel et al., 2022, 2023).

The video format encourages active engagement by prompting users to share their opinions and comments (Shahbaznezhad et al., 2020), and how a video is interacted with and commented on is often a result the nature of its content (Cheng et al., 2012). For instance, the comments for a music video will differ from those for a comedy video (Yew & Shamma, 2011). To increase rebroadcasting activity, organizations should tailor their content to match the audience's interests (Zhang et al., 2017), mainly because different video genre categories have different patterns and signatures of contextual interactions surrounding them (Yew & Shamma, 2011). According to Munaro et al. (2021), managers should prioritize sponsoring or associating their ads with categories such as entertainment, gaming, and food over others such as fashion and fitness. However, increased engagement with these videos may also result in more dislikes.

According to Berger et al. (2023), maintaining audience attention depends on content that is easy to read,

employs simple syntax, and utilizes familiar or concrete language. For example, replacing abstract words with concrete alternatives and incorporating more familiar synonyms for less familiar terms can enhance audience engagement. Consequently, even when consumers aim to acquire knowledge or delve into personal interests through YouTube content, it is essential to maintain a personal, informal tone that is closer to the consumer's reality (Munaro et al., 2021).

However, studies are needed to elucidate the discovery and extraction of hidden semantic structures from textual data within YouTube videos, employing topic modeling to represent semantic information. With the increased competition for brand partnerships and the new opportunities for content creation on online platforms such as YouTube, TikTok, and Instagram, influencers must have a good understanding of how to generate content that followers will perceive as helpful and entertaining (Munaro et al., 2024). Given the centrality of text-based language in social media marketing communications, it is imperative to understand the language aspects that drive engagement (Pezzuti et al., 2021). Additionally, further research is warranted on the influence of social media content strategy on engagement behavior (Shahbaznezhad et al., 2020).

2.1.1 *Sentiment analysis*

Emotional valence, which refers to the degree of positive or negative emotion expressed, introduces a nuanced dynamic with context-dependent perceptions (Chen & Farn, 2020). Shahbaznezhad et al. (2020) suggest that video formats, with their higher media richness, elicit greater active engagement compared to photos when conveying emotional content. Positive emotional states, which include emotions such as amusement, excitement, joy, warmth, inspiration, and pride, not only cultivate a positive attitude toward shared content, but also increase opportunities for self-improvement and future reciprocity (Tellis et al., 2019). In the realm of social sharing, content that evokes positive emotions tends to outperform information-focused content (Tellis et al., 2019), consistent with the notion that people socialize more with those who bring them joy (Aleti et al., 2019; Tellis et al., 2019).

Tellis et al. (2019) suggest that intense drama, surprise, and the inclusion of celebrities, babies, and animals effectively evoke emotions, leading to increased social sharing. Similarly, Berger et al. (2023) affirm the impact of emotional language in sustaining attention, emphasizing its effectiveness in relation to specific emotions that induce uncertainty and arousal. For instance, language associated with anxiety (high arousal and uncertainty) is likely to capture attention, while sadness (low arousal) may have a dampening effect. Anger, which is characterized by high arousal and low uncertainty, operates on a spectrum and depends on the interplay of these aspects in a given situation (Berger et al., 2023).

Conversely, negative messages often contain more diagnostic features related to the product or service, making them more informative (Chen & Farn, 2020). An examination of consumer sentiment toward brands, based on Liu et al.'s (2017) analysis of 1.7 million tweets, reveals a notable imbalance. The proportion of negative tweets significantly exceeds positive tweets across all brands, suggesting that dissatisfied customers are approximately three times more likely to engage in negative eWOM compared to satisfied customers engaging in positive eWOM (Liu et al., 2017). This aligns with Shahbaznezhad et al.'s (2020) observation that fans are more inclined to express their opinions and articulate negative comments.

2.2 Digital engagement with influencers

Digital consumer engagement refers to consumer interactions with brands or influencers in a digital environment. It reinforces consumers' investment in the sponsored brand at different levels and phases and produces traceable reactions such as impressions, clicks, likes, comments, and shares (Gavilanes et al., 2018). Associating posted videos with their respective results of digital engagement is a strategy for understanding the specificities of the public. Practitioners and academics have suggested using likes and comments as essential metrics of consumer engagement behavior in social media (Oh et al., 2017).

Anticipating different reactions to different content and considering the impact of a video's characteristics on engagement outcomes is crucial (Gavilanes et al., 2018). A comprehensive understanding of digital engagement with influencers requires evaluating the effects of an influencer's post on various user behaviors, particularly in the context of YouTube videos, where metrics such as views, likes, dislikes, and comments serve as indicators of online popularity and viewer satisfaction with the influencer's content (Munaro et al., 2021). These metrics not only reflect the online standing

of the influencer and the video, but also offer insights into the potential success of the featured product or service in the market (Aggrawal et al., 2018). Digital marketers and professionals need to manage consumers on social media, engaging them strategically and dividing them according to different promotional strategies based on their behavioral segments. This will improve conversion rates from the positive impact of social media eWOM to brand-loyal, non-switching consumers (Nyagadza et al., 2023).

Central to consumer engagement is the content that brands disseminate on social media (Lee et al., 2018; Hughes et al., 2019). As posited by Lee et al. (2018), brand content, especially emotional and humorous content, has a positive association with increased engagement. Conversely, directly informative content tends to yield lower social media engagement, although certain types of informative content can prompt higher click-through rates. Consequently, selecting specific content attributes, including word choice and communication style, is pivotal in shaping various forms of consumer engagement.

3 Data and method

Our research begins with the collection of audio transcriptions from videos using meticulous textual preprocessing techniques. This is followed by a pivotal stage of topic modeling using the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003; Ma & Sun, 2020) coupled with text analysis. Text analysis methods such as topic modeling and sentiment analysis are commonly employed by researchers to delve into themes, sentiments, viewpoints, and other pertinent factors. These analyses are facilitated by the implementation of machine learning algorithms (Rouhani & Mozaffari, 2022; Cano-Marin et al., 2023).

Figure 1 provides a concise overview of our methodological journey, encompassing three key phases: (1) Data Acquisition: Involving the collection of metadata and transcripts from videos; (2) Preprocessing: Encompassing the removal of stopwords, text normalization, lemmatization, and the generation of n-grams to improve text quality; (3) Creation of Multiple LDA Models: Developing various LDA models with different numbers of topics (k) to comprehensively explore and identify latent themes within the data. This methodical process ensures a thorough investigation, unveiling insights into the intricate structures and patterns embedded in the video transcripts.

In summary, our study introduces a comprehensive framework aimed at extracting latent content-related topics from influencer channels on YouTube. This framework includes the identification of labels, valence, and heterogeneity within these dimensions. The extracted dimensions serve as valuable inputs for strategy analysis, aligning with digital engagement measures, as outlined by Rodrigues and Paraiso (2020).

3.1 Data acquisition

The data collected include the number of views, likes, dislikes, comments, topic content, and other video post characteristics of 34,563 videos posted on YouTube among 103 different channels in 26 content categories between January 2008 and October 2020 (see full database in Supplementary Data 3- Full Video database - sentiment analysis). We identified the top Brazilian influencer channels through the *"Prêmio Influenciadores Digitais"* list of 2019-2020,which ranks influencers based on their relevance, popularity, and engagement in each influencer's area of expertise.

We relied on the platform's application programming interfaces (APIs) to extract the data. Data collection was performed using the Python programming language, and data persistence was performed using a MySQL relational database. We used the YouTube API (YouTube Data, 2024)

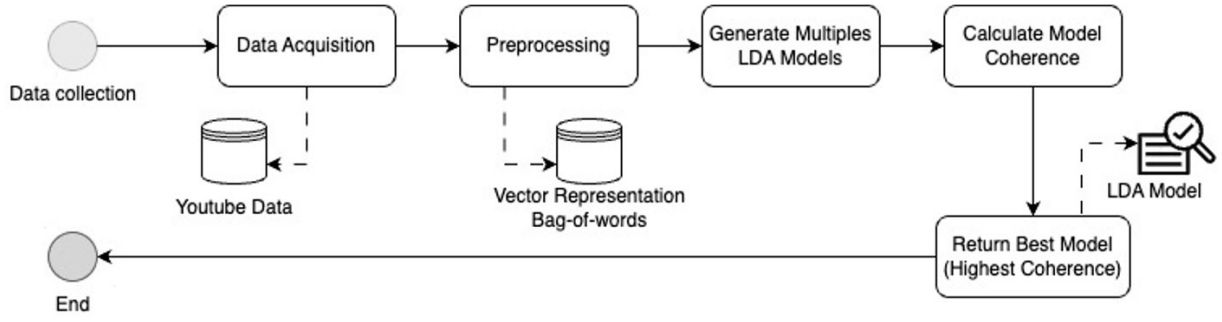

Figure 1. Research Extraction Model

for the extraction process. To collect video transcriptions on YouTube, we used automatic speech recognition (ASR) technology, which converts the spoken words into text. A named "bag of words" can be assembled from this data, which can establish a structure to be analyzed and tell the story of an emerging technology (Daniel & Dutta, 2018). We used an open-source Python API tool to extract all available auto-generated captions from the videos (Depoix, 2023); it converts audio data into textual data (JSON file).

3.2 Data preprocessing

Next, we implemented text preprocessing using modules from the Natural Language Toolkit (Guo et al., 2017). Preprocessing is an integral stage in the Natural Language Processing (NLP) pipeline, as it entails cleaning and preparing raw text data for subsequent analysis (Cano-Marin et al., 2023). For the n-gram (a sequence of n words) stage, we used the Gensim library (Řehůřek & Sojka, 2010). The text preprocessing steps were similar to those adopted in prior studies (e.g., Tirunillai & Tellis, 2014; Guo et al., 2017; Debortoli et al., 2016).

The first step was to normalize the text by removing stopwords, converting it to lowercase, and applying lemmatization. Stopwords refer to frequently occurring words that are likely to reduce the interpretability of the results (Dantu et al., 2020). Researchers often use lemmatization, i.e., reduction to a base dictionary form or lemma, e.g., "reviews" and "reviewing" to "review" (Debortoli et al., 2016; Dantu et al., 2020). We use lemmatization to facilitate human interpretation because it is more advanced, e.g., lemmatization looks at the synonyms of a word, resulting in more relevant documents (Balakrishnan & Lloyd-Yemoh, 2014). These tasks were performed to minimize the noise in the textual data, resulting in more accurate results and facilitating easier processing by NLP algorithms (Cano-Marin et al., 2023).

We applied part-of-speech (POS) tagging to retain only words that are adjectives, nouns, or adverbs (Debortoli et al., 2016; Tirunillai & Tellis, 2014). Bigrams and trigrams, two- and three-word sequences of words that frequently appear together in the input corpus, respectively, were identified, but only those appearing more than five times in the documents. This task is essential for obtaining the n-grams and allows for a reduction of the dimensionality of the terms while increasing their representativeness. N-gram analysis can be used for a variety of NLP tasks, including language modeling, text classification, and information retrieval. This technique is useful for generating insights from a text corpus because it can capture the context of words or phrases and can be handled flexibly (Cano-Marin et al., 2023). This improves the performance of the model (Blei et al., 2003). Finally, the bag-of-words representation was generated from the count of the present terms.

3.3 Generating multiple LDA models

Topic modeling is a technique for uncovering the underlying structure of a corpus, defined as an extensive collection of text data. Topic modeling can be used to organize and summarize large amounts of text data and to discover hidden patterns and relationships in the data (Cano-Marin et al., 2023). The Latent Dirichlet Allocation (LDA) algorithm is a Bayesian probabilistic model that generates a distribution of topics populated with words over documents and is well documented in the academic literature in terms of its application and automated topic generation from data sources (Daniel & Dutta, 2018; Feng et al., 2021). LDA makes several assumptions, the most important of which is the exchangeability of words (Wallach, 2006). Exchangeability assumes that the word order is irrelevant and only the presence or absence of words matters. Thus, the document can be expressed only by the frequency of the words it contains (Wallach, 2006; Feng et al., 2021). Words with higher probabilities in a particular topic reflect the subject matter of that topic (Feng et al., 2021).

In LDA, it is necessary to specify the number of topics to be extracted before running the analysis and, with the help of a specialist, analyze the results obtained until a satisfactory result is found (Souza & Souza, 2019). This is a crucial parameter of the analysis and is often not easy to determine (Dantu et al., 2020). Thus, it was necessary to develop an approach to find the best number of k in a corpus. The method generates a set of N models to increase the number of k topics. In this way, 20 different models were trained, starting with five topics and stopping at 100, incremented by 5 in 5. After generating each model, its coherence is calculated, and the next step is to select the best model with the most appropriate number of topics representative of the corpus. Coherence refers to the semantic relationship between the most frequently used words of a single topic; it is necessary for topics to be interpretable and for finding meaningful topic labels

(Büschken & Allenby, 2016). The model with the most significant coherence available was selected; it included 50 topics and 44% coherence (see Supplementary Data 2 – Topic modeling results).

Sentiment Analysis: Sentiment analysis, a subfield of natural language processing, is a means of automatically classifying texts by valence (Liu et al., 2017). Sentiment analysis can identify the overall sentiment of a piece of text, such as whether it is positive, negative, or neutral, as well as analyze the sentiment of specific words or phrases within the text (Cano-Marin et al., 2023). Positive emotions can be captured by the frequency of words such as happy, excited, and thrilled, whereas negative emotions are associated with words such as anxious, tragic, and selfish (Aleti et al., 2019). The output values for valence ranged from -1 to +1, with +1 being an extremely positive text and -1 being an extremely negative one.

For sentiment extraction, we used the LeIA (Lexicon for Adapted Inference) algorithm (Almeida, 2018). LeIA is a Brazilian Portuguese adaptation of VADER (Valence Aware Dictionary and Sentiment Reasoner), a traditional sentiment extraction algorithm. We assume that the message in each video explicitly expresses the writer's opinion on aspects of the content (see Liu et al., 2017). We analyzed the videos and counted the number of negative, positive, and neutral videos for each topic. Our criteria for categorizing the videos were as follows: any value below -0.2 was considered negative, any value above 0.2 was considered positive, and any value between -0.2 and 0.2 was considered neutral. We then calculated the overall sentiment for each topic by determining the proportion of negative, positive, and neutral videos relative to the total number of videos for that topic (see sentiment analysis database in Appendix D. Supplementary Data 3 – Full Video database - sentiment analysis).

4 Data analysis

Latent topics were extracted according to their probability of occurrence. Based on the terms of each topic, the semantic or representative label for that topic was inferred by domain knowledge (Rouhani & Mozaffari, 2022). 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 their 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). Following the methodology proposed by Debortoli et al. (2016), two researchers worked independently to interpret and label the topics, considering their semantic qualities: the topics had to be meaningful, interpretable, coherent, and useful. The channels and previous content categories related to each topic were also analyzed.

We named the 50 topics and grouped some of them based on semantic similarities, video content, and channel participation. In this way, the naming of the 50 topics resulted in 19 different content labels: 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, Learning & Curiosities. Supplementary Data 4 - Appendix A shows the names, groupings, and video percentages of the 50 topics. It also provides a short description of each of the 19 content categories. After labeling the generated topics, a valence analysis and statistical analyses were performed. The dashboard data can be viewed at LokerStudio (2023).

5 Results

The words that occur most frequently together in a particular topic are the ones with higher probabilities. Usually, the top 10 or 15 words are used to semantically interpret and label the topics. The top 10 most representative words for each topic are shown in Appendix B and are arranged in order of the probability that each term is assigned to the topic (Supplementary Data 5 - Appendix B). Table 1 shows the 15 topics with the highest percentage of videos illustrating the topics, the top words, and the labeling process. The second column indicates the percentage of each topic. Column 3 contains the top 10 essential words; column 4 shows the descriptive labels used; column 5 shows examples of SMI channels.

The top 5 topics (numbers 13, 26, 7, 0, 46) represent more than 26% of the sample. Topic 13 had the highest number of related videos and appeared on 95 influencer channels. The wide distribution across almost all evaluated channels (n = 103) contributes to its popularity, which is associated with more generic keywords such as power, fight, time, hero, and world. The primary content categories for the topic on YouTube are Culture & Entertainment, Gaming, Humor, and Politics, Economy

Table 1
The 15 most representative topics

Topic	Percentage (total videos)	Top 10 words	Labeling	Examples of associated channels
13	8.50 (2,937)	power, guys, fight, time, hero, new, powerful, strong, picture, world	Culture & Entertainment	Ei Nerd, Bibi, Meteoro Brasil, Whindersson Nunes
26	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
7	4.60 (1,589)	govern, Bolsonaro, president, politician, brazil, Lula, public, be, country, leave	Politics, Economy & News	Kim Kataguiri, TV Afiada, Mamaefalei, Nando Moura
0	3.94 (1,361)	hair, makeup, foundation, little, product, shadow, face, look, good, tone	Beauty	Mariana Saad, Mari Maria, NiinaSecrets, Bianca Andrade
46	3.83 (1,323)	device, camera, screen, photo, good, hit, best, cellphone, Samsung, iPhone	Tech	TudoCelular, Canaltech, Dudu Rocha, Be!Tech
40	3.77 (1,303)	cool, white, see, blue, red, landmark, time, pool, nice, new	Family	Brancoala, Flavia Calina, resendeevil, T3ddy
27	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
17	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
42	2.72 (941)	good, tasty, food, water, little, meat, coffee, dish, chicken, cheese	Gastronomy	Tastemade Brasil, Dani Noce, Receitas de Pai, Sal de Flor
31	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
34	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
44	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?,
15	2.33 (805)	travel, place, hotel, cool, hour, plane, city, world, dollar, day	Travel, Learning & 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
3	2.21 (763)	clothes, cool, beautiful, store, lovely, box, pretty, wonderful, gift, purse	Fashion/Lifestyle	Organize sem frescuras!, Taciele Alcolea, NiinaSecrets, Flavia Pavanelli

& News. The most representative channels are Ei Nerd, Bibi, Meteoro Brasil, and Whindersson Nunes, with the topic representing 34%, 89%, 31%, and 39% of the channels' content, respectively.

Topic 26 appeared in 43 different channels; its keywords represented cooking ingredients, preparation methods, and adjectives related to the field of gastronomy. The main categories of the topic are Gastronomy, Fitness, and Decoration, Organization & DIY. The representative channels are (topic share in parentheses): TPM por Ju Ferraz (82%), Receitas da Cris (87%), Receitas de Pai (79%), Danielle Noce (58%), and Tastemade Brasil (50%). Topic 7 was presented in 31 channels, with the most likely words related to politics and the Brazilian political and economic scenario. The main categories of the topic are Politics, Economy & News, Humor, and History. The representative channels of the topic are Kim Kataguiri (86%), TV Afiada (69%), Mamaefalei (50%), and Gabriela Prioli (59%). Topic 0 comprised 43 channels; its keywords relate to makeup and aesthetics products, body parts, and related adjectives. The YouTube video categories for this topic include Beauty, Fashion, Behavior & Lifestyle, and Fitness. Examples of channels with greater representation are Mariana Saad (73%), Mari Maria (72%), NiinaSecrets (56%), and Rodrigo Cintra (52%). Topic 46 was 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. The keywords revolve

around attributes and parts of electronic devices, brands, and performance adjectives. The video categories include Digital Technology and Travel and Tourism. Examples of the most representative channels are TudoCelular (91%), Canaltech (62%), Dudu Rocha (48%), and Be!Tech (43%).

5.1 YouTube content overview

Based on the results of the study, the content categories Education, Culture & Entertainment, and People, Behavior & Lifestyle are the most popular in terms of quantity, i.e., they have the most significant number of videos among the influencers on YouTube (see Supplementary Data 4 - Appendix A). Education is the most popular type of content, with 3,555 videos covering eight different topics, representing 10.3% of the total. This result is consistent with the main reasons people watch YouTube – to learn something new and to further explore their interests. While tutorial videos have always been popular on YouTube, the global watch time of how-to videos has increased more than 50% yearly, an opportunity to learn something new and dig deeper into one's interests. (Google, 2020). In this scenario, many video influencers have emerged as key opinion leaders in their areas of interest (Chen et al., 2023). As a result, YouTube has become a go-to platform for troubleshooting and informational videos.

Culture & Entertainment is the second most created content on YouTube in terms of quantity (10.13%). It consists of two distinct topics corresponding to 3,501 videos. It covers content that also aims to satisfy 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%), comprising five topics in 2,929 videos. It is content that reflects human behavior, self-learning, reflections, and lifestyles. Thus, it is a topic that permeates most channels to a greater or lesser extent.

Gastronomy (2 topics – 8.1%) and Tech (4 topics – 7.9%) are the contents that complete the top 5 most popular in terms of quantity on YouTube, 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 topics – 5.4%), which complete the list of the top 10 most created content among SMIs on YouTube.

On average, there are around 20 topics per channel, indicating that channels have a greater dispersion of content. However, we found that in 52 channels, a single topic accounts for more than half of the video content, ranging from 51.6% to 97.1%. Additionally, most channels focus on only one or two topics, indicating that social media influencers stick to a specific script when creating content. Appendix C lists 20 individual channels among Brazilian influencers on YouTube with the most likes received in the sample (Supplementary Data 6 - Appendix C).

Table 2 shows the top 20 topics with the highest number of views, likes, comments, and dislikes, and analyzes their relationship with user digital engagement. Topic 13 (Culture & Entertainment) ranks second in number of views, has the highest number of likes and comments, and is second in number of dislikes. Topic 40 (Family) has the highest number of views, the second highest number of dislikes, the sixth highest number of comments, and the third highest number of likes. Topic 27 (People, Behavior & Lifestyle) ranks fifth in number of views and comments, third in likes, and seventh in dislikes.

Figure 2 shows the proportion of valence identified in each content topic. Topics are identified by their numbers on the x-axis. Regarding the positive sentiment analysis, the top 10 topics with the highest positive valence were 46 - Tech (91%), 26 - Gastronomy (89.9%), 34 - Economics, Entrepreneurship & Business (89.2%), 0 - Beauty (89.1%), 10 - Education (87.9%), 3 - Fashion/Lifestyle (84.5%), 15 - Travel, Learning & Curiosities (84.4%), 21 - Tech (84.4%), 31 - Decoration, Organization & DIY (83.4%), and 28 - Economics, Entrepreneurship & Business (83.1%). These topics had more than 83% positive content in their related videos. Three also have the highest digital engagement (topics 26, 34, and 0 – see Table 2).

Upon analyzing the negative valence, we found that the topic Entertainment/General had the highest negative valence percentage with 90.6%. This was followed by the Politics, Economy & News category with topics 11 (87.6%) and 7 (77.1%), respectively. The Education category had four topics with high negative valence percentages: 33 (75.6%), 32 (71.3%), 24 (71.2%), and 29 (65.7%). The Gaming category had a negative valence percentage of 69.5% (topic 6), while the Health & Healthy Lifestyle category had 64% (topic 1), and the Travel, Learning & Curiosities category had 59.2% (topic 12). It is important to note that six of these topics are among those with the

Table 2
Topics associated with greater digital engagement on YouTube

Topic	Associated label	Number channels	Total videos	Nº views	Nº likes	Nº dislikes	Nº comments
13	Culture & Entertainment	95	2,937	2,837,621,147	295,615,053	3,709,281	9,964,773
40	Family	74	1,303	3,400,544,604	103,450,628	3,796,103	2,266,477
27	People, Behavior & Lifestyle	69	1,199	846,274,668	89,509,236	930,214	2,753,016
8	Entertainment/General	23	372	837,121,596	75,437,368	813,837	1,660,115
29	Education	35	517	776,596,493	73,958,484	1,106,737	2,906,724
0	Beauty	43	1,361	714,310,073	68,455,565	767,523	2,871,953
26	Gastronomy	43	1,861	1,071,900,602	66,547,497	1,024,933	1,964,037
44	Pets & Animals	38	812	696,201,768	56,566,211	757,986	1,452,640
14	Sports	39	772	720,392,877	54,159,710	766,397	1,468,965
36	Education	26	529	533,635,953	53,074,079	672,553	2,158,053
7	Politics, Economy & News	31	1,589	446,821,941	52,750,412	2,923,144	3,816,130
12	Travel, Learning & Curiosities	65	393	619,938,100	52,432,747	738,320	1,274,099
17	Decoration, Organization & DIY	39	1,088	924,291,357	49,230,238	759,413	1,705,115
38	Culture & Entertainment	33	564	527,061,542	48,050,369	704,981	1,545,639
43	People, Behavior & Lifestyle	49	529	404,529,577	46,779,720	495,255	1,727,293
4	Family	56	553	547,394,390	43,847,063	484,000	1,043,645
6	Gaming	40	663	456,882,188	42,278,282	503,941	1,458,319
11	Politics, Economy & News	35	629	336,688,921	38,548,484	1,166,129	1,890,595
45	People, Behavior & Lifestyle	48	533	408,030,571	33,715,477	351,538	961,774
34	Economics, Entrepreneurship & Business	39	837	356,237,568	31,723,586	450,462	1,011,786

Note. As several topics can form a content category, there may be repetitions of some content labels.

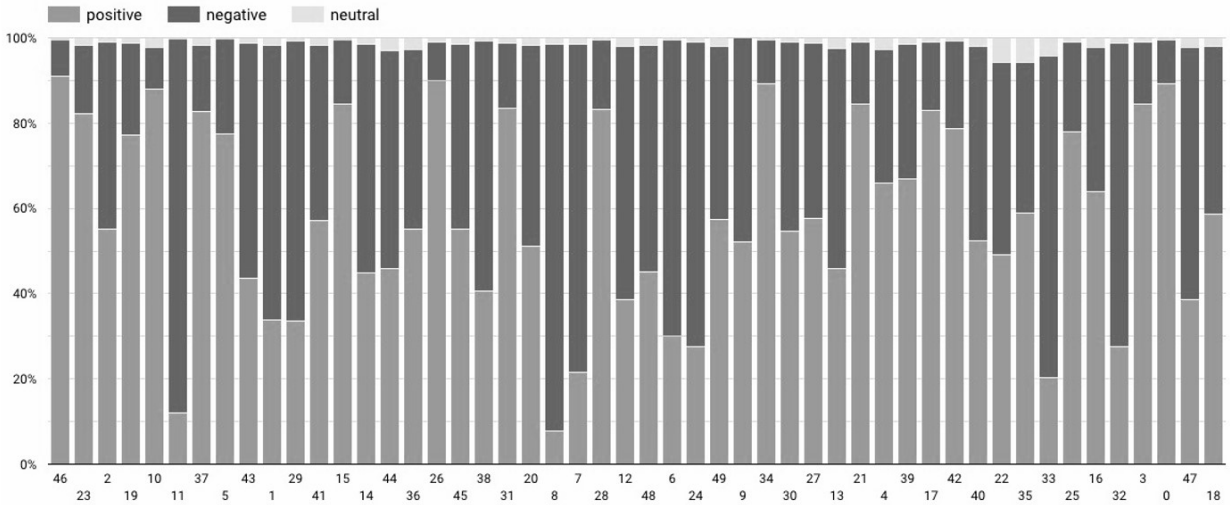

Figure 2. Sentiment analysis result for each topic

greatest digital engagement (topics 8, 11, 7, 6, 29, and 12 – see Table 2), suggesting that content that deals with controversial issues and evokes negative emotions in the audience has the potential to be more engaging.

Finally, to answer the question of which content categories receive the most digital engagement and which is the predominant valence, we compiled the 19 identified and labeled content categories in Table 3. The choice of predominant content changes depending on the strategic marketing goal. Overall, Family, Entertainment/General, and Culture & Entertainment were the contents with the best engagement indicators (number of views, likes, comments). Regarding the number of comments, we found that Culture & Entertainment content had the highest

Table 3
Summary of 19 content categories with digital engagement and valence data

Label	Number of videos (%)	Nº views	Nº likes	Nº dislikes	Nº comments	Positive valence	Negative valence
Beauty	1361 (3.9)	524,842.08	50,297.99	563.94	2,134.40	89.13%	10.43%
Culture & Entertainment	3501 (10.1)	961,063.61	98,162.07	1,260.86	3,299.16	44.93%	52.90%
Decoration, Organization & DIY	1982 (5.7)	612,701.28	34,527.17	511.97	1,170.84	83.15%	15.64%
Economics, Entrepreneurship & Business	1905 (5.5)	284,987.25	27,012.64	391.53	949.99	85.62%	13.65%
Education	3555 (10.3)	641,230.50	55,268.27	821.25	2,106.61	54.26%	43.77%
Entertainment/General	750 (2.2)	1,432,471.91	125,524.70	1,507.26	2,957.33	31.33%	64.93%
Family	1856 (5.4)	2,127,122.30	79,363.52	2,306.63	1,986.39	56.30%	41.54%
Fashion/Lifestyle	1292 (3.7)	371,831.14	22,680.94	348.42	550.49	81.58%	17.34%
Gaming	1398 (4)	683,878.34	66,073.60	939.52	2,582.22	39.77%	59.23%
Gardening	391 (1.1)	203,085.99	13,236.55	154.70	361.64	82.61%	15.60%
Gastronomy	2802 (8.1)	550,532.95	33,972.93	526.09	996.26	86.19%	12.92%
Health & Healthy Lifestyle	2036 (5.9)	329,021.96	24,475.54	327.87	907.98	51.96%	46.32%
Military	648 (1.9)	278,865.59	31,622.47	283.95	1,435.84	51.23%	47.07%
People, Behavior & Lifestyle	2929 (8.4)	685,388.35	70,121.77	765.17	2,354.42	55.21%	43.43%
Pets & Animals	812 (2.3)	857,391.34	69,662.82	933.48	1,790.20	45.94%	50.99%
Politics, Economy & News	2218 (6.4)	353,251.06	41,166.32	1,847.28	2,603.09	18.80%	80.07%
Sports	1193 (3.4)	740,539.15	51,673.53	768.69	1,438.12	44.93%	53.56%
Tech	2736 (7.9)	220,264.81	13,741.35	262.51	777.21	81.54%	17.91%
Travel, Learning & Curiosities	1198 (3.4)	612,750.54	50,825.90	719.61	1,358.29	69.45%	29.47%

Notes. Digital engagement values correspond to the average of the sum of videos in each content category. Each content category's average share (%) is shown for valence.

average of all (3,299.16), followed by Entertainment/General (2,957.33), Politics, Economy & News (2,603.09), and Gaming (2,582.22).

The most popular content on YouTube can vary depending on the region and time, but certain categories tend to have consistently high numbers of videos and digital engagement. Entertainment content, such as comedy sketches, TV show clips, movie trailers, reviews, and reactions to memes, challenges, and humorous situations, is very popular on YouTube. Many YouTubers have built massive audiences by creating original content in this category. Moreover, YouTube has become a platform where parents can find information on a variety of parenting topics. From advice on pregnancy and childbirth to tips on raising children of all ages, a vast amount of parenting content is available on the platform. Family content can offer a sense of community and connection with other parents who may be going through similar experiences.

6 Discussion

Currently, unstructured data have a significant impact on consumer decision-making (Li et al., 2019). Mining and analyzing these unstructured data present numerous challenges and diverse applications, resulting in extensive research and literature studies. This framework adeptly utilizes both LDA and sentiment analysis, two widely recognized machine learning methods tailored to address the complexities of big data in textual formats (Liu et al., 2017). Identifying latent topics becomes a crucial guide for marketing managers, helping to track, evaluate, and integrate results into automated marketing planning (Li & Ma, 2020). This knowledge is pivotal in fostering and sustaining consumer engagement across various channels and online communities.

We identified 19 critical dimensions of YouTube video content through data mining in a dataset of 34,563 video transcripts from 103 influencer channels. The study reveals that the top 3 content categories that drive higher user digital engagement are Family, Entertainment/General, and Culture & Entertainment. These categories differ from the most popular ones in terms of quantity, which include Education, Culture & Entertainment, and People, Behavior & Lifestyle. Moreover, the sentiment analysis indicates that content related to Beauty, Gastronomy, and Economics, Entrepreneurship & Business has the highest proportional positive valence. Conversely, Politics, Economy & News, Entertainment/General, and Gaming

have higher percentages of negative valence, suggesting topics that are more contradictory, controversial, or evoke negative emotions in consumers.

The study identified prevalent topics in influencer discussions or promotions, providing valuable insights for influencer marketing. This information enhances the understanding of influencer branding and messaging. For businesses or marketers considering collaborations with these influencers, leveraging this knowledge ensures alignment between their product or service and the influencer's messaging and values. We also offer detailed steps for uncovering and interpreting latent topics derived from LDA from brand-related content. This topic modeling approach unveils keywords and tags linked to popular content, introducing a data-driven dimension to decision-making. Companies can employ this information to optimize their video titles, descriptions, and tags to improve discoverability and search engine optimization.

Second, our study results help identify the interests and preferences of the influencer's audience, offering valuable insights for crafting marketing strategies or products that resonate with the influencer's followers. By understanding successful content on the platform, companies can optimize their own content for enhanced visibility, search rankings, and suggested video placements. The study's insights also reveal trends and patterns in social media content, encompassing frequently discussed topics, commonly used words, and sentiment analysis. This information enriches marketing strategies, identifies content gaps, and enhances social media engagement.

Third, a sentiment analysis of video content topics is presented. The propensity of negative content to go viral on YouTube can be attributed to its attention-grabbing nature and its ability to evoke strong emotional responses, leading to increased engagement and sharing. Shocking, controversial, or unexpected content, including topics related to Politics, Economy & News, or Education, can also go viral. Social identity also plays a role, with individuals sharing negative content that aligns with their beliefs and values.

Fourth, beyond identifying current trends, cultural movements, and potential content gaps in influencer channels, such as environmental issues, gardening, and the animal world, the study results provide actionable insights. By using state-of-the-art technologies such as NLP, data can be processed to generate inferences, making it possible to make proactive decisions based on vast amounts of information and prevent potential future situations (Cano-Marin et al., 2023). Companies can leverage this information to fill content gaps, attract a wider audience, or increase engagement with existing followers by staying relevant and incorporating popular (or unpopular) topics into their marketing strategies. When determining which influencer to use as a product endorser in an advertisement, the influencer should be selected based on which feature should be emphasized (see Ata et al., 2022).

Finally, our study helps bridge a gap in the literature on the dynamics of video content on social media from the perspective of influencers as brand-generated content. The influence of influencers on brand attitudes and purchase behaviors surpasses that of traditional celebrities (Schouten et al., 2020). In summary, a deep understanding of YouTube's popular content and digital engagement rates is essential for companies aiming to maximize reach, resonate with their target audience, and remain competitive in the dynamic digital landscape. This knowledge facilitates more effective communication, content creation, and strategic decision-making.

7 Limitations and further research

There are several limitations to this study. First, we employed a computerized technique to extract all available autogenerated captions from the videos in our sample as textual data. However, relying solely on these captions has its drawbacks. The captions are generated through the automatic conversion of audio into text, and the quality of this process is contingent on various factors, such as semantic errors, accents, dialects, background noise, and technical jargon. Consequently, errors or omissions in the captions may impact the accuracy of our findings.

Second, the study did not account for the personality traits of social media influencers or the number of subscribers to each channel. Future research could delve into analyzing identified topics and their relationship with the personal aspects of SMIs and their channels. Third, the study did not consider influencers' audience data, including demographic and/or psychographic data. The fit between message content and audience interest is a significant driver of rebroadcasting behavior (Zhang et al., 2017). Fourth, we did not analyze infrequent words in the long tail of the distribution. These words could reflect emerging content topics, which could be invaluable for understanding latent content on YouTube. Each of these

limitations represents a promising avenue for further research.

Many interesting questions remain unexplored. The study examines the content produced by leading Brazilian YouTube influencers. For future research, it is essential to conduct a more in-depth analysis of the video content to better understand the dynamics of digital engagement and the factors that contribute to a video's popularity. One promising approach is to analyze the linguistic elements in video transcripts. Additionally, it is important to identify videos with sponsored content and evaluate potential differences in language use between these and non-sponsored videos. For instance, are sponsored videos of higher quality compared to non-sponsored ones?

This study used the interpretability of topics generated using LDA, representing a class of probabilistic topic models. However, future studies could use large language models (LLMs) to compare different topic modeling techniques. The recent launch of ChatGPT, an artificial intelligence chatbot developed using LLMs, has garnered considerable attention, highlighting the profound impact of these models on academia, industry, and society. LLMs could improve generalization, topic coherence, and diversity compared to LDA. This approach eliminates the need for manual parameter tuning. It would improve the quality of the extracted topics, mainly because the LDA labeling process was very challenging given the number of words in each topic and the overlapping topic labels in the same topic.

Finally, marketers are increasingly using virtual influencers and computer-generated image personas to promote products and services. Do the observed effects of interaction and content evaluation in human influencer posts extend to virtual influencers? Additionally, the significance of language in discussing a topic, which is as important as the topic itself for capturing and maintaining audience attention (Berger et al., 2023), provides a positive insight for communicators working on less engaging topics. Further research in this area is essential to delve deeper into this finding.

References

Agência Brasil. (2024). *Pesquisa aponta pulverização no mercado de influenciadores*. https://agenciabrasil.ebc.com.br/geral/noticia/2024-06/pesquisa-aponta-pulverizacao-no-mercado-de-influenciadores-digitais#

Aggrawal, N., Arora, A., Anand, A., & Irshad, M. S. (2018). View-count based modeling for YouTube videos and weighted criteria–based ranking. In M. Ram & J. Paulo Davim (Eds.), *Advanced Mathematical Techniques in Engineering Sciences (*pp. 149-160). CRC Press. http://doi.org/10.1201/b22440-8.

Aleti, T., Pallant, J. I., Tuan, A., & van Laer, T. (2019). Tweeting with the stars: Automated text analysis of the effect of celebrity social media communications on consumer word of mouth. *Journal of Interactive Marketing*, *48*, 17-32. http://doi.org/10.1016/j.intmar.2019.03.003.

Almeida, R. J. A. (2018). *Leia-léxico para inferência adaptada*. https://github.com/rafjaa/LeIA.

Ata, S., Arslan, H. M., Baydaş, A., & Pazvant, E. (2022). The effect of social media influencers' credibility on consumer's purchase intentions through attitude toward advertisement. *ESIC Market*, *53*(1), e280-e280. http://doi.org/10.7200/esicm.53.280.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering*, *2*(3), 174-179. http://doi.org/10.7763/LNSE.2014.V2.134.

Berger, J., Moe, W. W., & Schweidel, D. A. (2023). What holds attention? Linguistic drivers of engagement. *Journal of Marketing*, *87*(5), 793-809. http://doi.org/10.1177/00222429231152880.

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. http://doi.org/10.1287/mksc.2016.0993.

Cano-Marin, E., Ribeiro-Soriano, D., Mardani, A., & Gonzalez-Tejero, C. B. (2023). Exploring the challenges of the COVID-19 vaccine supply chain using social media analytics: A global perspective. *Sustainable Technology and Entrepreneurship*, *2*(3), 100047. http://doi.org/10.1016/j.stae.2023.100047.

Casaló, L. V., Flavián, C., & Ibáñez-Sánchez, S. (2020). Influencers on Instagram: Antecedents and consequences

of opinion leadership. *Journal of Business Research*, *117*, 510-519. http://doi.org/10.1016/j.jbusres.2018.07.005.

Chen, L., Yan, Y., & Smith, A. N. (2023). What drives digital engagement with sponsored videos? An investigation of video influencers' authenticity management strategies. *Journal of the Academy of Marketing Science*, *51*(1), 198-221. http://doi.org/10.1007/s11747-022-00887-2.

Chen, M. J., & Farn, C.-K. (2020). Examining the influence of emotional expressions in online consumer reviews on perceived helpfulness. *Information Processing & Management*, *57*(6), 1-15. http://doi.org/10.1016/j.ipm.2020.102266.

Cheng, Y., Xie, Y., Zhang, K., Agrawal, A., & Choudhary, A. (2012). How online content is received by users in social media: A case study on Facebook.com posts. In *2nd Social Media Analytics Workshop, ACM, ACM SIGKDD, 8/1/12*.

Chung, S., & Cho, H. (2017). Fostering Parasocial Relationships with Celebrities on social media: Implications for Celebrity Endorsement. *Psychology and Marketing*, *34*(4), 481-495. http://doi.org/10.1002/mar.21001.

Daniel, C., & Dutta, K. (2018). Automated generation of latent topics on emerging technologies from YouTube Video content. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 1762-1770). ScholarSpace. http://doi.org/10.24251/HICSS.2018.222.

Dantu, R., Dissanayake, I., & Nerur, S. (2020). Exploratory analysis of internet of things (IoT) in healthcare: A topic modelling & co-citation approaches. *Information Systems Management*, *38*(1), 62-78. http://doi.org/10.1080/10580530.2020.1746982.

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. http://doi.org/10.17705/1CAIS.03907.

Depoix, J. (2023). *YouTube Transcript API. v. 0.6.2*. GitHub. https://github.com/jdepoix/youtube-transcript-api

Digital Marketing Institute. (2024). *20 Surprising Influencer Marketing Statistics*. https://digitalmarketinginstitute.com/blog/20-influencer-marketing-statistics-that-will-surprise-you.

Feng, J., Mu, X., Wang, W., & Xu, Y. (2021). A topic analysis method based on a three-dimensional strategic diagram. *Journal of Information Science*, *47*(6), 770-782. http://doi.org/10.1177/0165551520930907.

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. http://doi.org/10.1080/00913367.2017.1405751.

Google (2020). *3 ways people are using YouTube to learn at home during the coronavirus pandemic. Consumer Insights, Think with Google*. https://www.thinkwithgoogle.com/intl/en-emea/consumer-insights/consumer-trends/how-people-use-youtube-for-learning/.

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. http://doi.org/10.1016/j.tourman.2016.09.009.

Hughes, C., Swaminathan, V., & Brooks, G. (2019). Driving brand engagement through online social influencers: An empirical investigation of sponsored blogging campaigns. *Journal of Marketing*, *83*(5), 78-96. http://doi.org/10.1177/0022242919854374.

Jacobson, J., Hodson, J., & Mittelman, R. (2022). Pupularity contest: The advertising practices of popular animal influencers on Instagram. *Technological Forecasting and Social Change*, *174*, 121226. http://doi.org/10.1016/j.techfore.2021.121226.

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. http://doi.org/10.1287/mnsc.2017.2902.

Leung, F. F., Gu, F. F., Li, Y., Zhang, J. Z., & Palmatier, R. W. (2022). Influencer marketing effectiveness. *Journal of Marketing*, *86*(6), 93-115. http://doi.org/10.1177/00222429221102889.

Li, H., & Ma, L. (2020). Charting the path to purchase using topic models. *JMR, Journal of Marketing Research*, *57*(6), 1019-1036. http://doi.org/10.1177/0022243720954376.

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. http://doi.org/10.1016/j.ijresmar.2019.02.004.

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. http://doi.org/10.1080/00913367.2017.1297273.

LokerStudio. (2023). *What video engages the most? an analysis of social media influencers' content on YouTube*. LokerStudio. https://lookerstudio.google.com/reporting/5bee2ab6-31b3-447b-875d-a3af12dd3417/page/p_zfzvikc3cd.

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481-504. http://doi.org/10.1016/j.ijresmar.2020.04.005.

Munaro, A. C., Barcelos, R. H., Francisco-Maffezzolli, E. C. F., Rodrigues, J. P. S., & Paraiso, E. C. (2021). To engage or not engage? The features of video content on YouTube affecting digital consumer engagement. *Journal of Consumer Behaviour*, *20*(5), 1336-1352. http://doi.org/10.1002/cb.1939.

Munaro, A. C., Barcelos, R. H., Francisco-Maffezzolli, E. C., Rodrigues, J. P. S., & Paraiso, E. C. (2024). Does your style engage? Linguistic styles of influencers and digital consumer engagement on YouTube. *Computers in Human Behavior*, *156*, 108217. http://doi.org/10.1016/j.chb.2024.108217.

Nunes, R. H., Ferreira, J. B., Freitas, A. S. D., & Ramos, F. L. (2018). The effects of social media opinion leaders' recommendations on followers' intention to buy. *Revista Brasileira de Gestão de Negócios*, *20*(1), 57-73. http://doi.org/10.7819/rbgn.v20i1.3678.

Nyagadza, B., Mazuruse, G., Simango, K., Chikazhe, L., Tsokota, T., & Macheka, L. (2023). Examining the influence of social media eWOM on consumers' purchase intentions of commercialised indigenous fruits (IFs) products in FMCGs retailers. *Sustainable Technology and Entrepreneurship*, *2*(3), 100040. http://doi.org/10.1016/j.stae.2023.100040.

Oh, C., Roumani, Y., Nwankpa, J. K., & Hu, H. F. (2017). Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. *Information & Management*, *54*(1), 25-37. http://doi.org/10.1016/j.im.2016.03.004.

Patel, P. C., Parida, V., & Tran, P. K. (2022). Perceived risk and the need for trust as drivers of improved surgical skills in 3D surgical video technology. *Journal of Innovation & Knowledge*, *7*(4), 100269. http://doi.org/10.1016/j.jik.2022.100269.

Patel, P. C., Stenmark, M., Parida, V., & Tran, P. K. (2023). A socio-institutional perspective on the reluctance among the elderly concerning the commercialization of 3D surgical video technology in Sweden. *Journal of Innovation & Knowledge*, *8*(2), 100361. http://doi.org/10.1016/j.jik.2023.100361.

Pezzuti, T., Leonhardt, J. M., & Warren, C. (2021). Certainty in language increases consumer engagement on social media. *Journal of Interactive Marketing*, *53*(1), 32-46. http://doi.org/10.1016/j.intmar.2020.06.005.

Plummer, M. (2022). *Why video plays a key role in today's marketing scene*. https://www.forbes.com/sites/forbestechcouncil/2022/06/03/why-video-plays-a-key-role-in-todays-marketing-scene/.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modeling with large corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (pp. 46-50). Elra.

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

Rouhani, S., & Mozaffari, F. (2022). Sentiment analysis researches story narrated by topic modeling approach. *Social Sciences & Humanities Open*, *6*(1), 100309. http://doi.org/10.1016/j.ssaho.2022.100309.

Santora, J. (2022). *35 Influencer Marketing Statistics Shaping 2024. Influencer Marketing Hub.* https://

influencermarketinghub.com/influencer-marketing-statistics/amp/

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. http://doi.org/10.1080/02650487.2019.1634898.

Sette, G., & Brito, P. Q. (2020). To what extent are digital influencers creative? *Creativity and Innovation Management*, *29*(S1), 90-102. http://doi.org/10.1111/caim.12365.

Shahbaznezhad, H., Dolan, R., & Rashidirad, M. (2020). The role of social media content format and platform in users' engagement behavior. *Journal of Interactive Marketing*, *53*(1), 47-65. http://doi.org/10.1016/j.intmar.2020.05.001.

Souza, M., & Souza, R. R. (2019). Modelagem de tópicos: Resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina. *Múltiplos Olhares em Ciência da Informação*, 9(2), 1-11. https://periodicos.ufmg.br/index.php/moci/article/view/19138.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., & Zhang, Y. (2019). What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence. *Journal of Marketing*, *83*(4), 1-20. http://doi.org/10.1177/0022242919841034.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *JMR, Journal of Marketing Research*, *51*(4), 463-479. http://doi.org/10.1509/jmr.12.0106.

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. http://doi.org/10.1080/00913367.2020.1795954.

Voorveld, H. A. (2019). Brand communication in social media: A research agenda. *Journal of Advertising*, *48*(1), 14-26. http://doi.org/10.1080/00913367.2019.1588808.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine Learning* (pp. 977-984). ACM. http://doi.org/10.1145/1143844.1143967.

Yew, J., & Shamma, D. A. (2011). Know your data: Understanding implicit usage versus explicit action in video content classification. In *Multimedia on Mobile Devices 2011; and Multimedia Content Access: Algorithms and Systems V* (Vol. 7881, pp. 355-362). SPIE.

Yoon, S. H., & Lee, S. H. (2022). What likeability attributes attract people to watch online video advertisements? *Electronics (Basel)*, *11*(13), 1960. http://doi.org/10.3390/electronics11131960.

YouTube Data (2024). *Data API*. Google. https://developers.google.com/youtube/v3

Zhang, Y., Moe, W. W., & Schweidel, D. A. (2017). Modeling the role of message content and influencers in social media rebroadcasting. *International Journal of Research in Marketing*, *34*(1), 100-119. http://doi.org/10.1016/j.ijresmar.2016.07.003.

SUPPLEMENTARY MATERIAL

Supplementary Data 1 - Database - Appendix D
Supplementary Data 2 - Topic modeling results - database
Supplementary Data 3 - Full video database - sentiment analysis
Supplementary Data 4 - Appendix A
Supplementary Data 5 - Appendix B
Supplementary Data 6 - Appendix C
Supplementary Material can be found online at https://doi.org/10.7910/DVN/EETODI

Financial support:
The authors declare that no financial support was received.

Open science:
Ana Cristina Munaro; Eliane Cristine Francisco Maffezzolli; João Pedro Santos Rodrigues; Emerson Cabrera Paraiso, 2024, "Beyond the Screen: A Creative Exploration of Content that Engages on YouTube by Social Media Influencers", https://doi.org/10.7910/DVN/EETODI, Harvard Dataverse, V1

Conflicts of interest:
The authors have no conflicts of interest to declare.

Plagiarism analysis:
RBGN performs plagiarism analysis on all its articles at the time of submission and after approval of the manuscript using the iThenticate tool.

Authors:
1. Ana Cristina Munaro, Doutora em Administração, Pontifícia Universidade Católica do Paraná, Curitiba, Brasil.
E-mail: ana.munaro@pucpr.edu.br
2. Eliane Cristine Francisco Maffezzolli, Doutora em Administração, Professora do Programa de Pós-Graduação em Administração (PPAD), Pontifícia Universidade Católica do Paraná, Curitiba, Brasil.
E-mail: eliane.francisco@pucpr.br
3. João Pedro Santos Rodrigues, Mestre em Informática, Pontifícia Universidade Católica do Paraná, Curitiba, Brasil.
E-mail: jpsanr@gmail.com
4. Emerson Cabrera Paraiso, Doutor em Tecnologia em Sistema de Informação pela Université de Technologie de Compiègne, França, Professor do Programa de Pós-Graduação em Informática (PPGIa), Pontifícia Universidade Católica do Paraná, Curitiba, Brasil.
E-mail: paraiso@ppgia.pucpr.br

Authors' contributions:
1st author: definition of research problem; development of hypotheses or research questions (empirical studies); definition of methodological procedures; literature review; statistical analysis; analysis and interpretation of data; manuscript writing.
2nd author: definition of research problem; literature review; statistical analysis; analysis and interpretation of data; critical revision of the manuscript; manuscript writing.
3rd author: definition of research problem; development of hypotheses or research questions (empirical studies); definition of methodological procedures; data collection; analysis and interpretation of data.
4th author: definition of methodological procedures; data collection; analysis and interpretation of data; critical revision of the manuscript.