SemeAd 2023
XXVI Seminários em Administração

07, 08, 09 e 10 de novembro de 2023
ISSN 2177-3866

WHAT VIDEO ENGAGES THE MOST? AN ANALYSIS OF SOCIAL MEDIA INFLUENCERS' CONTENT ON YOUTUBE

ANA CRISTINA MUNARO
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ (PUCPR)

JOÃO PEDRO SANTOS RODRIGUES
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ (PUCPR)

ELIANE CRISTINE FRANCISCO-MAFFEZZOLLI
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ (PUCPR)

WHAT VIDEO ENGAGES THE MOST? AN ANALYSIS OF SOCIAL MEDIA INFLUENCERS' CONTENT ON YOUTUBE

1 INTRODUCTION

Visual information and video influencers have become 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; Chen, Yan, & Smith, 2023). 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 rebroadcasting of social media messages (Zhang, Moe, & Schweidel, 2017) and consumer engagement is the content that creators and brands disseminate through social media (Lee et al., 2018). With the predominance of content creators and influencer marketing in company strategies, research is needed to understand content consumption behavior on social media. Brand-generated communication on social networks generates traceable attention, affects consumers' attitudes toward branded content, engages consumers cognitively and emotionally, and can drive consumers to advocate for the brand (Gavilanes, Flatten, & Brettel, 2018), it rises the customers' trust beliefs in the integrity, ability, and benevolence of salespeople (Yaghtin, Safarzadeh, & Zand, 2020).

High-quality content is imperative for brand communication (Voorveld, 2019). The content is elastic, and it is much broader than advertising itself (van Noort, Himelboim, Martin, & Collinger, 2020), then understanding the strategy of content created on social networks is crucial. And we still do not understand precisely what kinds of content work better for which companies and in what ways (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). Further, there has been no standard set of measures of audiovisual information on YouTube. Thus far, academic research focuses mainly on online behavioral advertising and social media advertising (van Noort et al., 2020). And little is known about the factors that drive the success of online engagement with social media influencers (SMI) or influencers (Hughes, Swaminathan, & Brooks, 2019; Ladhari, Massa, & Skandrani, 2020), and regarding the choice of a given SMI as a brand endorser.

SMIs have gained increasing popularity across several domains, including marketing planning. For brands, working with an SMI who resonates with their target audience has become pivotal to the success of their campaigns, irrespective of the sector or niche (Hughes, 2020). According to the Digital Marketing Institute, businesses earn an average of $5.20 for every dollar spent on influencer marketing (Hughes, 2020). However, the partnership with SMIs can be a two-edged sword, given the risk of associating the brand with controversial eWOM, making selecting an influencer from numerous alternatives very challenging (Chung & Cho, 2017).

Besides, while text information has been widely studied and used, academic research 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). Today, unstructured data plays a key role in the consumer decision-making process (Li et al., 2019). Thereat, we adopted a method for knowledge extraction from videos through audio

transcriptions (Rodrigues & Paraiso, 2020). Text analytics methods such as Topic Modeling and Sentiment Analysis are among the most popular methods that researchers employ to study themes, sentiments, viewpoints, etc. which can be conducted by machine learning algorithms (Rouhani & Mozaffari, 2022).

Due to these gaps in the literature, the goal of the study is to investigate what the most popular content and valence-associated social media influencers discuss on YouTube. Then, we propose a unified framework for (1) extracting the latent content-related topics from social media influencers' channels on YouTube; (2) ascertaining the labels, valence, and heterogeneity of those dimensions; and (3) using those dimensions for strategy analysis considering digital engagement measures. This study attempts to this kind of effect, i.e., the influence of social media influencer content on user engagement, in the context of YouTube videos. We use SMI as content creators, digital engagement, and video features on YouTube as a theoretical framework.

Our study helps answer how can marketers use machine learning for extracting useful information from video data sets for optimizing brand communication. Sentiment analysis and topic modeling embrace a powerful tool for businesses and researchers to explore and study community attitudes, interpretations, and insightful consequences for decision support. Furthermore, several studies in the area focus on the consumer's point of view, user-generated content (UGC), considering reviews/tweets and comments (e.g., Tirunillai & Tellis, 2014; Büschken & Allenby, 2016; Guo, Barnes, & Jia, 2017; Liu et al., 2017), ignoring the content creator's side. Thus, this study fills a gap in the literature on the dynamics of video content on social media, from the point of view of SMIs as brand-generated content.

2 SOCIAL MEDIA INFLUENCERS (SMIS)

Social media influencers have become a global phenomenon. Brands are investing a huge portion of their budget in influencer marketing, while social media companies are investing in building the best digital platforms for influencers. Meanwhile, the number of people who wish to become an influencer has increased. An SMI is a person who consistently creates content for a given audience on a social media platform, establishes a relationship with their audience, stands out as an opinion leader in a community, and ultimately influences people to adopt certain behaviors (e.g., purchasing sponsored products) (Uzunoğlu & Misci Kip, 2014; Sette & Brito, 2020). An influencer is thus someone who has credibility in the group, persists in attempting to influence other individuals' attitudes or behavior, and introduces ideas that others pick up on or support (Zhao et al., 2018).

SMI nowadays has a more significant impact on brand attitudes and purchase behaviors than traditional celebrities (Schouten, Janssen, & Verspaget, 2020; Chen et al., 2023). Therefore, they may be more effective product endorsers and affect positively advertising effectiveness, which can be explained by processes of wishful proximity, trust, similarity, and identification (consumers feel more similar and identify more to influencers than celebrities) (Aleti et al., 2019; Schouten et al., 2020).

It is no surprise then that the literature on influencers has shown that several content characteristics play an important role in consumers' perception of information credibility and usefulness and consequently on their brand attitude and intention to buy (Hughes et al., 2019; Munaro et al., 2021). For instance, Casaló, Flavián, & Ibáñez-Sánchez (2020) suggest that the content's originality and uniqueness are crucial factors if a user is to be perceived as an influencer and an opinion leader. Although much of the literature identifies individual-related factors as antecedents of digital engagement, few studies have studied the impact of social media content on

engagement behavior. Thus, a more in-depth analysis of the impact of social media content is warranted (Shahbaznezhad, Dolan, & Rashidirad, 2020).

And, while there is certainly not a single, definite content attribute that explains social influence, the literature on influencer marketing has already provided clues to different elements driving their success. For instance, influencers' relationship with followers often relies on their communication strategies, content choices, and social presence (Jacobson, Hodson, & Mittelman, 2022). In this sense, not only the nature and content of posts but also how an influencer communicates with their followers should contribute to the maintenance of influencer-follower relationships and, therefore, their degree of influence (Jacobson et al., 2022).

Therefore, online content 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 (Jacobson et al., 2022), 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. Brands could identify the appropriate content to disseminate to their unique audience, the most appropriate valence, and develop a message by using words that have a high probability of being associated with that content topic (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), contributing to the maintenance of influencer-follower relationships and, therefore, their degree of influence (Jacobson et al., 2022).

2.1 Video content on YouTube

Video influencer marketing is increasingly popular amongst marketers (Chen et al., 2023). Overall, producing videos on YouTube with SMIs can be an effective way for companies to reach a wider audience, build brand awareness, and increase their chances of generating engagement and building a relationship with their target audience; normally it is also a more affordable option for companies with limited advertising budgets.

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). While much of the literature identifies individual-related factors as antecedents of customer engagement, few studies have studied the impact of social media content strategy on engagement behavior (Shahbaznezhad et al., 2020).

YouTube offers an exported transcript of the closed captioning text. This data is potentially valuable to understand or discover terms when trying to explore an emerging technology that may have been discussed in YouTube video content. A named 'bag of words' can be assembled from this data which could perhaps establish a structure to be analyzed and tell the story of an emerging technology (Daniel & Dutta, 2018).

Especially since the quality of content may affect the influence of posts. Then, the task of identifying highly influential and quality user-generated content on social media sites is becoming increasingly important (Cheng et al., 2012). To Cheng et al. (2012) the effectiveness of a post and the likelihood to rebroadcast content depends not just on its content but also on how the post relates to its users' 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). To Shahbaznezhad et al. (2020) video format posts encourage users to actively engage by sharing their opinion and comments toward firms' posts, while photo formatted content stimulates passive users' engagement through liking behavior.

An aspect explored in the literature related to the content of videos refers to emotional valence. The effect of valence is complex and leads to different perceptions depending on the context. Emotional valence refers to the degree to which people express positive emotion or negative emotions (Chen, 2020). To Shahbaznezhad et al. (2020), emotional content in the video format stimulates an increase in active engagement (commenting). Thus, emotional content may be best suited to a higher media richness format (video), as it can convey greater levels of emotional stimuli compared to a photo post.

Contents that evoke positive emotional states (such as amusement, excitement, joy, warmth, inspiration, and pride) should make the receiver feel a positive attitude toward the sharer, enhancing the sharer's opportunities for self-enhancement in the present and reciprocity by the recipient in the future (Tellis et al., 2019). Content that evokes positive emotions is generally more effective than information-focused content in driving social sharing (Tellis et al., 2019). In the same way, people usually feel more inclined to socialize with those who make them feel good (Aleti et al., 2019; Tellis et al., 2019). Endorsing, to Yaghtin et al. (2020), task-oriented and emotion-oriented content classes are the most valuable ones from the audiences' viewpoints and the most efficient content classes in persuading audiences to participate in conversations. To Tellis et al. (2019) strong drama, surprise, and the use of celebrities, babies, and animals are effective in arousing emotions and creating social shares.

On the other hand, negative messages tend to include more diagnostic features associated with the product/service and thus tend to be more informative (Chen, 2020). Analyzing consumer sentiments toward brands from 1.7 million tweets, Liu et al. (2017) find that the percentage of negative tweets is much higher than that of positive tweets among all brands. These findings provide empirical support that unhappy customers are about three times more likely to engage in negative eWOM than happy customers are to engage in positive eWOM (Liu et al., 2017). Supporting Shahbaznezhad et al. (2020), who indicates that fans tend to express their opinion and write more comments with a negative sentiment, than positive.

Finally, 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 with their respective results of digital engagement (number of views, likes, comments) is a strategy to understand the idiosyncrasies of the public. Practitioners and academics alike have suggested the use of likes and comments as important metrics for consumer engagement behavior in social media (Oh, Roumani, Nwankpa, & Hu, 2017).

2.2 Digital Engagement with SMIs

Digital consumer engagement refers to consumers' interactions with brands or influencers in a digital environment. It strengthens consumers' investment in the sponsored brand at different levels and phases and produces traceable reactions such as impressions, clicks, likes, comments, and shares (Gavilanes, Flatten, & Brettel, 2018).

Although these digital actions can be conceptually regarded as representing "consumer engagement," they qualitatively reflect different types of digital engagement (Yoon et al., 2018). We may expect different reactions to different content when considering the effects of a video's

characteristics on the four engagement-related marketing outcomes; that is, views, likes, dislikes, and comments. Therefore, a complete understanding of digital engagement with SMIs depends on considering the effect of an SMI's post on different user behaviors, or, in the case of YouTube videos, the number of views, likes, dislikes, and comments (Munaro et al., 2021). These metrics show the online popularity of the SMIs and their videos on YouTube as well as the viewers' satisfaction with the SMIs' content. Hence, digital engagement may indicate whether the featured product/service will be successful in the market (Aggrawal, Arora, Anand, & Irshad, 2018).

The trend of the distribution of comments and likes shows that many posts receive few or no "comments/like" and only a few posts can gain very high influence. However, there is a long tail of the distribution, and this matches the famous "power law", which demonstrates that there are huge differences among the posts for gaining influence (Cheng et al., 2012).

To Lee et al. (2018), brand personality-related content, such as emotional and humorous content, is positively associated with higher engagement. This suggests that firms gain from sharing their brand personality and information about their social initiatives on social media. Further, directly informative content is associated with lower engagement on social media, but certain types of informative content can induce higher click-throughs. Thus, brand personality-related content is primarily associated with positive engagement and seems key for long-term brand building, while directly informative content is primarily associated with direct response and seems key to performance marketing (Lee et al., 2018).

Moreover, a complete understanding of digital engagement also depends on the factors that drive users to engage with influencers (Aggrawal et al., 2018). One of the most important factors affecting consumer engagement is the content that brands disseminate through social media (Lee et al., 2018; Hughes et al., 2019). Hence, the choice of specific content attributes, such as the word choice and the way messages are communicated to the audience, should significantly impact the different forms of consumer engagement.

3. METHOD

To answer the study objectives and the associated questions: What is the most popular content (in terms of the number of videos and user digital engagement) on YouTube? What are the most popular channels among Brazilian influencers? Which content categories generate the most engagement? What is the predominant valence in each category? The process used in our study comprises collecting audio transcriptions from videos, after executing a text preprocessing, performing a topic modeling stage, the Latent Dirichlet Allocation algorithm (LDA) (Blei, Ng, & Jordan, 2003), and text analysis. Figure 1 summarizes the methodological process: Data acquisition (collection of metadata and collection of video transcripts); Preprocessing (removing stopwords, text normalization, lemmatization, generating n-grams), and generating multiples LDA models (models with different k topics).

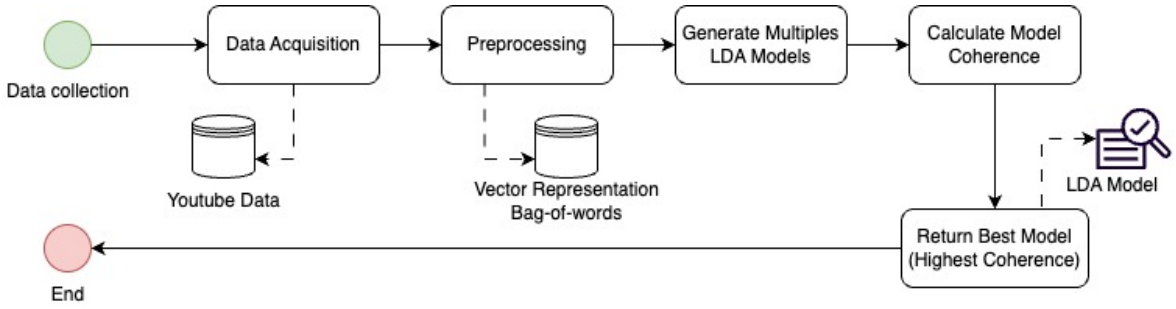


Figure 1. Adopted model.

3.1 Data acquisition

The study collected data on the number of views, likes, dislikes, comments, topics content, and other video post characteristics, from 34,563 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 the "*Prêmio Influenciadores Digitais*" list of 2019-2020[i], which ranks influencers based on their relevance, popularity, and engagement to each influencer's area of expertise.

We relied on the application programming interfaces (APIs) provided by major social media platforms to extract the data. Data collection was conducted via Python programming language, and data persistence was performed using a MySQL relational database. We used the YouTube API (YouTube Data V3[ii]) for the extraction process (except for the transcriptions). There are different ways in which audio transcriptions of videos on YouTube can be generated. One way is using automated speech recognition (ASR) technology, which converts the spoken words in a video into text. Another way is through manual transcription, which involves a human transcribing the audio of the video. We used an open-source tool from Python API to extract all available auto-generated captions for the videos[iii]; it converts audio data into textual data (JSON file).

3.2 Data Preprocessing

In the sequence, we implemented text pre-processing by using modules of the Natural Language Toolkit (www.nltk.org) (Guo et al., 2017). And for the n-grams stage, we used the Gensim library (https://radimrehurek.com/gensim) (Řehůřek & Sojka, 2010). Text pre-processing used steps were 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, was removed stopwords, converting it to lowercase, lemmatization is also applied. Stopwords refer to frequently occurring words that are likely to diminish the interpretability of the results (Dantu, Dissanayake, & Nerur, 2020). Researchers often use stemming (i.e., reducing words to a root form) or lemmatization (i.e., reducing to a base dictionary form or a lemma, e.g., "reviews" and "reviewing" to "review") (Debortoli et al., 2016; Dantu et al., 2020). We use lemmatization to facilitate human interpretation because lemmatization is more advanced in the sense that it takes care of additional analysis that is not supported by stemming. For instance, lemmatization looks at the synonyms of a word unlike stemming, which results in more relevant documents (Balakrishnan & Lloyd-Yemoh, 2014).

We apply 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 of tokens that

frequently appear together in the input corpus were identified. Thus, only bi and trigrams that appeared more than 5 times in the documents and with a threshold greater than 10 were accepted[iv]. This task is important for obtaining the n-grams and also allows the dimensionality reduction of the terms while increasing their representativeness. Consequently, this tends to improve the model's outcome (Blei et al., 2003). Finally, the bag-of-words representation was generated from the count of the present terms.

3.3 Generate multiple LDA models

The Latent Dirichlet Allocation (LDA) algorithm is a Bayesian probabilistic model of text documents, well documented within the academic literature in terms of its application and automated topic generation from data sources (Daniel & Dutta, 2018; Feng, Mu, Wang, & Xu, 2021). The algorithm analyzes the number of words distribution found in a given document based on a known distribution of the word count by topic in advance. 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. Then, the document can be expressed using only the frequency of the words contained within it (Wallach, 2006; Feng, Mu, Wang, & Xu, 2021). Words with higher probabilities in a certain topic reflected the subject matter of that topic (Feng et al., 2021).

Using the LDA algorithm not only creates a distribution of topics populated with words but also includes the distribution of topics over documents (Daniel & Dutta, 2018). LDA requires us 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 we find a satisfactory result (de Souza & Souza, 2019), which is a key parameter of the analysis and often not easy to determine (Dantu et al., 2020). 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. In this way, 20 different models were trained, starting with 5 topics, and stopping at 100, incremented by 5 in 5. After creating 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 among the most frequent words of a single topic, it is necessary for topics to be interpretable and for finding meaningful topic labels (Büschken & Allenby, 2020).

The next step was to select the best model by choosing the model with the greatest coherence available, the model selected was the model with 50 topics and 44% coherence.

Sentiment Analysis: Sentiment analysis, a subfield in natural language processing, is a means to automatically classify texts by valence (Liu et al., 2017). Positive emotions can be captured by the frequency of words such as happy, excited, and thrilled, whereas negative emotions are related to 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 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. In the context of categorizing topic valence, 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 tallied the numbers of negative, positive, and neutral videos on each topic, considering values lower than -0.2 as negative valence, values greater than 0.2 as positive valence, and between -0.2 and 0.2 considered neutral valence.

And we determined the overall topic sentiment by calculating the proportion of negative, positive, and neutral videos relative to the total number of videos on each topic.

4. DATA ANALYSIS

Latent topics were extracted by their probability of occurrence. Based on each topic's terms, 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 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, considering their semantic qualities, it is important to consider topics meaningful, interpretable, coherent, and useful. Moreover, was analyzed channels and previous content categories related to each topic.

We name each latent topic based on the meaning conveyed by its top words as well as its weights (see Li & Ma, 2020). The labeling of dimensions was first conducted by one researcher and then confirmed by a second researcher. We named the 50 topics and then grouped some of them based on semantic similarities, video content, and channel participation among some topics. In this way, the 50 topics naming 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. Appendix A shows the 50 topics' names, groupings, and video proportions. After labeling the generated topics, a valence analysis, descriptive and statistical analyses were performed.

5. RESULTS

Each topic distribution includes all the words but assigns a different probability to each word. The words in a topic with high probabilities are the ones that tend to co-exist more often. Typically, the top 10 or 15 words are used to interpret topics and to label them semantically. Appendix B shows the top 10 most representative words by topic, each topic's terms are ordered by the probability that each term is assigned to a given topic. In other words, a weight is assigned to each term by each topic which indicates the probability of that term relates to the topic. In Table 1, we present the 15 topics with the highest percentage of videos to illustrate the topics, top words, and labeling process. The second column indicates the proportion of each topic, which represents the probability of the topic appearing in the dataset. Column 3 contains the top 10 most important words for each topic. In column 4, the descriptive labels, and in column 5 some examples of SMIs channels associate.

Table 1. The 15 most representative topics in the sample

Topic	Proportion (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, 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
3	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 (numbers 13, 26, 7, 0, 46) represent more than 26% of the total sample. *Topic 13* had the highest number of related videos and appeared on 95 different channels of influencers. The great dispersion for nearly all evaluated channels (n = 103) is a contributing factor to its popularity, associated with more generic keywords, such as power, fight, time, hero, and world. The major content categories for the topic on YouTube are Culture and Entertainment; Gaming; Humor; Economy, politics, and news. The most representative channels are Ei Nerd, Bibi, Meteoro Brasil, and Whindersson Nunes, with the topic representing 34%, 89%, 31%, and 39% respectively of the channels' content.

Topic 26 showed up in 43 different channels, its keywords represented cooking ingredients, preparation methods, and adjectives related to the gastronomy field. The major categories of the topic are Gastronomy, Fitness, Decoration, Organization and do-it-yourself (DIY). The representative channels are (topic proportion in parentheses): TPM by 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, the most likely words referring to politics and the Brazilian political and economic scenario. The major categories of the topic are Economy, Politics and current affairs,

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 refer to makeup and aesthetics products, body parts, and related adjectives. The YouTube video categories for that topic cover: beauty, fashion, behavior and 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 study's results, 'Education', 'Culture & Entertainment, and 'People, Behavior & Lifestyle' content categories are the most popular in quantity, i.e., have the bigger number of videos among the SMIs on YouTube (see Appendix A). Education is the most prevalent content among the sample analyzed. It covers 8 topics in 3,555 videos according to the grouping of related content, 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. As videos have gradually come to dominate people's online entertainment and information acquisition, many video influencers have emerged and become key opinion leaders in their respective domains of interest (Chen et al., 2023). Hence, considering this preference for information and troubleshooting videos on YouTube.

Culture & Entertainment ranks second on the most created content in quantity on YouTube (10.13% of the total). It consists of two individual 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 the 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' more popular in quantity on YouTube among SMIs. 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%) that complete the list of 10 most created/disseminated content among SMIs on YouTube.

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

Table 2 shows the list of 20 individual channels of SMIs with the highest number of likes received in the sample, that is, the most popular channels among Brazilian influencers on YouTube. In column 2 there is the total number of topics identified in all videos collected from each channel, in column 3 we present the two most representative topics as a content proportion of the channels. Column 4 shows the total number of videos, and columns 5, 6, 7, and 8 the YouTube digital engagement metrics.

Table 2. Channels, representative topics, and digital engagement

Channel name	Total topics	Topics more representative (proportion)	Videos collected	Nº Views	Nº Likes	Nº Dislikes	Nº Comments
Você Sabia?	38	29 (19.3%), 36 (14.8%)	400	4,935,635,309	491,909,690	6,081,033	15,913,457
whinderssonnunes	30	13 (39.5%), 27 (9.3%)	248	4,260,945,954	476,620,000	4,103,510	10,435,220
Mari Maria	21	0 (72.5%), 27 (7.9%)	458	1,084,441,664	114,890,194	1,128,182	4,814,856
Canal Nostalgia	31	13 (42.9%), 38 (12.6%)	175	1,137,500,836	106,722,520	1,825,338	3,782,326
Ei Nerd	7	13 (89.4%), 38 (9.2%)	601	749,513,985	102,256,520	1,083,342	4,679,706
Desimpedidos	13	14 (66.9%), 48 (14.7%)	532	1,243,783,660	101,217,162	1,376,970	2,912,947
T3ddy	30	13 (17.8%), 6 (15.5%)	466	610,016,567	100,858,113	684,427	3,645,610
Mamaefalei	25	7 (49.6%), 11 (22.7%)	591	530,107,162	63,758,766	3,041,595	5,511,661
rezendeevil	25	47 (26.9%), 40 (19.9%)	513	685,158,204	56,538,920	1,204,574	1,702,760
Manual do Mundo	28	17 (30.3%), 16 (27.2%)	489	1,027,210,336	56,369,066	952,883	1,726,528
Fatos Desconhecidos	37	29 (36.0%), 36 (9.3%)	506	761,765,970	56,217,530	1,075,534	3,013,940
Pipocando	14	38 (85.1%), 13 (8.6%)	429	640,608,208	54,211,222	813,316	1,420,592
Receitas de Pai	8	26 (78.6%), 42 (18.1%)	420	727,090,748	52,918,171	759,013	1,398,566
Thiago Ventura	10	8 (67.5%), 13 (10.8%)	120	718,206,860	49,249,151	370,191	450,512
Me poupe!	32	34 (62.3%), 23 (6.3%)	559	499,746,734	48,408,148	617,059	1,538,671
Flavia Calina	26	40 (44.4%), 4 (18.8%)	522	1,727,891,009	48,300,582	1,144,768	1,636,034
JoutJout Prazer	40	13 (14.1%), 45 (11.5%)	426	435,302,856	42,278,430	435,555	1,692,430
Richard Rasmussen	32	44 (65.7%), 13 (16.6%)	470	549,408,472	39,825,492	450,693	965,401
malena010102	26	30 (72.6%), 13 (6.2%)	519	403,313,419	39,164,448	370,675	1,952,763
Danielle Noce	24	26 (58.0%), 42 (16.5%)	467	462,778,810	38,617,107	382,306	1,182,086

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

Observing the 5 channels with the best engagement numbers (ranked by the number of likes), the prevalence of latent content already highlighted is confirmed: Education (represented by topics 29, 36), Culture and Entertainment (represented by topics 13 and 38), People, Behavior and Lifestyle (topic 27) make up the largest proportion of the channels' content, adding the topic Beauty (Topic 0).

Analyzing the relationship between the content and the associated metrics of user digital engagement, Table 3 presents the top 20 topics that individually have the highest number of views, likes, comments, and dislikes. The data shows the digital engagement relationship of each topic. Topic 13 (Culture & Entertainment) is 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 40 (Family labeled) is the largest number of views, and the number of dislikes, the second topic is the number of likes, and the sixth is the 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.

Table 3. Topics associated with greater digital engagement on YouTube.

Topic	Label associated	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, learnings&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. The 20 most popular topics (column 1) are ranked by the number of likes. As a content category can be formed by several topics, the repetitions of some content labels can be noticed.

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

Observing the topics with the greatest negative valence among his videos, we can see topic 8 - Entertainment/general (90.6%), topics 11 (87.6%) and 7 (77.1%) of Politics, Economy and News, in the sequence, notes Education topics were 33 (75.6%), 32 (71.3%), 24 (71.2%) and 29 (65.7%), noting a theme with a more negative valence content. To complete the top 10, we also have topic 6 - Gaming (69.5%), topic 1 - Health and healthy lifestyle with 64%, and finally, topic 12 - Travel, learnings and curiosities with 59.2% of videos with negative valence. It is important to note that 6 of these topics are among those with the greatest digital engagement (topics 8, 11, 7, 6, 29, and 12 - see Table 3), which indicates the engagement potential of content that addresses more controversial thematic and that evoke negative emotions in the consumer.

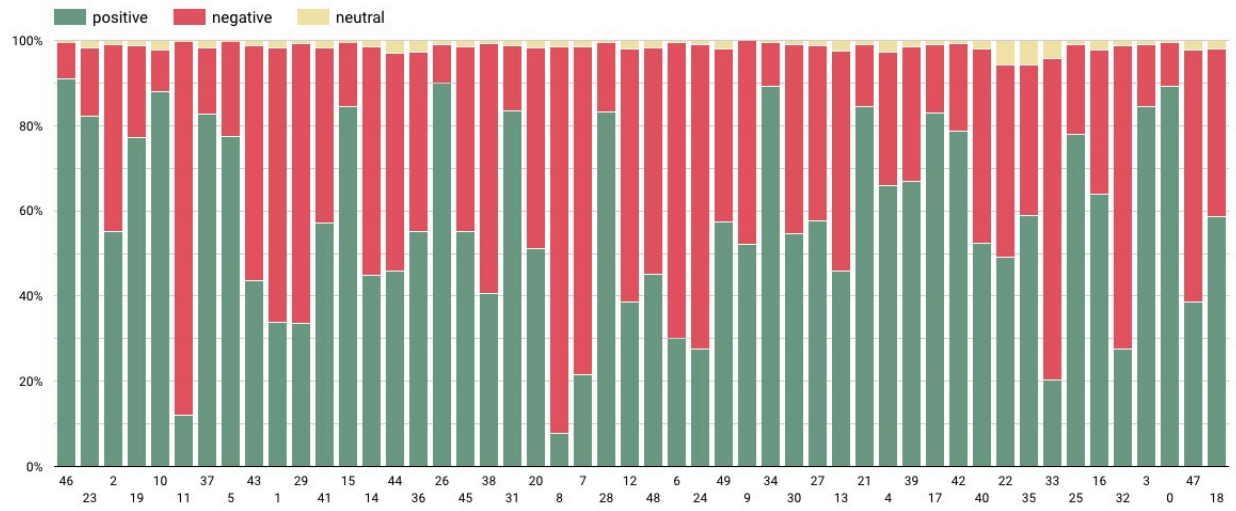


Figure 2. Sentiment analysis result: proportion of valence for each topic.

Overall, several factors can contribute to why negative content may be more likely to be shared on YouTube. For example, we can presume that negative content can be attention-grabbing and can evoke a strong emotional response in viewers, which can make them more likely to engage with and share the content. Negative content that is shocking, controversial, or unexpected may be more likely to go viral on social media, such as content on *Politics, Economy and News*, or content relates to *Education* aspects, such as historical facts. In the same sense, as social identity, for example, if someone strongly identifies with a particular political or social movement, they may engage (like, share) negative content that supports their beliefs and values.

Finally, to present the results in an aggregated form in the 19 identified and labeled content categories and to answer, in general, which content categories generate the most engagement and which is the predominant valence in each category, we compiled in Table 4. Table 4 brings the digital engagement and associated valence for each content category identified on YouTube. It can be assumed that depending on the strategic marketing goal, the choice for predominant content changes. Overall, Family, Entertainment/general and Culture and Entertainment were the contents with the best engagement indicators (number of views, likes, comments). Regarding the level of comments, it is noted that the content on Culture and Entertainment is highlighted with the highest average among all (3,299.16), followed by Entertainment/general (2,957.33), Politics, Economy and News (2,603.09) and Gaming (2,582.22).

Table 4. Summary of 19 content categories with digital engagement and valence data

Label	Number videos (%)	Nº Views	Nº likes	Nº dislikes	Nº comments	Valence positive	Valence negative
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 and 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 and 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, learnings & curiosities	1198 (3.4)	612,750.54	50,825.90	719.61	1,358.29	69.45%	29.47%

Notes. The digital engagement values correspond to the average of the sum of videos in each content category. For valency, the average proportion (%) of each content category is presented.

The most popular content on YouTube varies over time and across different regions, but some general categories tend to have consistently high numbers of videos and digital engagement. The study shows that Entertainment content, such as comedy skits, TV show clips, movie trailers reviews, and reactions to memes, challenges, and situations involving humor are also very popular on YouTube. Many YouTubers have built large audiences by creating original content in this category. Also, YouTube is a platform where parents can find a wealth of information on a wide range of parenting topics. From advice on pregnancy and childbirth to tips on raising children of all ages, there is a vast amount of parenting content available on the platform. Family content can offer a sense of community and connection with other parents who may be going through similar experiences. Also, we can assume that YouTube allows creators to create highly specific content for niche audiences. This means that parents can find content that is tailored to their particular situation, such as single parenting, special needs parenting, or parenting multiples.

Sentiment analysis shows us that content about Beauty, Gastronomy and Economics, Entrepreneurship and Business are those with the highest proportional positive valence, signaling that their videos bring words that refer to positive emotions. On the other hand, Politics, Economy and News, Entertainment/general, and Gaming are contents with high percentages of negative valence and involve themes that are more contradictory, controversial or that arouse negative emotions in the consumer.

6. DISCUSSION AND CONCLUSIONS

With the fast-growing use of social media platforms, and a large volume of user-generated content, mostly unstructured texts, sentiment analysis has become useful in various applications and domains. Mining and analysis of this large volume of unstructured data necessitates applying text mining and Natural Language Processing techniques and encompass many challenges and at the same time different applications that have led to much work in this field and formed many literature studies. This framework effectively employs both LDA and sentiment analysis, which are two highly acclaimed machine-learning methods that are specifically designed to tackle the challenge of big data in textual formats (Liu et al., 2017)

We identify 19 key dimensions of video content on YouTube using a data mining approach in a data set that includes 34,563 videos transcript for 103 SMIs' channels. The study highlights the top 3 content categories with greater user digital engagement among influencers: 'Family', 'Entertainment/general' and 'Culture and Entertainment'. These classifications are different of the content categories that are the most popular in quantity: 'Education', 'Culture and Entertainment, and 'People, Behavior and Lifestyle'. Furthermore, the sentiment analysis shows that content about 'Beauty', 'Gastronomy' and 'Economics, Entrepreneurship and Business' are those with the highest proportional positive valence. And, 'Politics, Economy and News', 'Entertainment/general', and 'Gaming' contents with high percentages of negative valence.

Firstly, the study identified the most prevalent topics that the influencer is discussing or promoting, which can be useful in understanding their brand and messaging. This information can be used by businesses or marketers who are considering partnering with SMIs to ensure that their product or service aligns with the influencer's messaging and values.

Also, we provide detailed steps for uncovering and interpreting latent topics gleaned by LDA from brand-related content, and to analyze with digital engagement metrics. Furthermore, the sentiment analysis of video content topics is presented. For dimension extraction, the LDA analysis of content topics reveals meaningful dimensions that are not found via traditional means.

Secondly, since had organized large amounts of content, the study results can help identify the influencer's audience's interests and preferences. This information can be used to tailor marketing strategies or create products that appeal to the influencer's audience. The study demonstrates that we have 19 categories of content being produced by the top influencers in Brazil. The findings help to identify trends and patterns in social media content, including the most discussed topics, frequently used words, and sentiment analysis. This information can be used to inform marketing strategies, identify gaps in content, and improve social media engagement.

Thirdly, the study findings can help identify potential content gaps in the SMI's channel, such as environmental issues, gardening, and the animal world, which can be filled to attract a wider audience or increase engagement with existing followers.

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 incorporate 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). For managers and executives, the key to creating and sustaining consumer engagement may come from the ability to adequately use cohesive social media strategies across different channels and different online communities (Oh et al., 2017). Finally, 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 brand-generated content (BGC). SMIs have a more significant impact on brand attitudes and purchase behaviors than traditional celebrities, they may affect positively advertising effectiveness (Schouten et al., 2020).

LIMITATIONS AND FURTHER RESEARCH

This study has some important limitations. Firstly, our computerized technique extracted all available autogenerated captions (ASR) from the videos in our sample as textual data. Relying on these captions is not always ideal, as they are the products of the automatic conversion of audio into text, and the quality of this process is dependent on several factors, such as semantic errors,

accents, dialects, background noise, and technical jargon, which can result in errors or omissions in the captions.

Secondly, the study did not consider the personality traits of social media influencers and did not even consider the number of subscribers of each channel. Future research can analyze the identified topics and their relationship with the personal aspects of influencers at the channel level. Third, the study does not consider audience data from influencers, such as demographic and/or psychographic data, as the fit between the message content and the audience's interest is a significant driver of rebroadcasting behavior (Zhang et al., 2017), analyzing the target profile is an important factor to better understand the dynamics of content in digital media. Also, we do not analyze rare or infrequent words in the long tail of the distribution; these words could reflect emerging content topics that could be very helpful to understand the latent content on YouTube. Each of the above limitations could be a rich avenue for further research.

Finally, our LDA model is exploratory. The LDA model is sensitive to the model's tuning parameters, which leads to a considerable difference in the results. These parameters include the number of topics based on word frequency cut-off for common and rare words.

REFERENCES

Aggrawal, N., Arora, A., Anand, A., & Irshad, M. S. (2018). View-count based modeling for YouTube videos and weighted criteria-based ranking. In: Advanced Mathematical Techniques in Engineering Sciences. CRC Press, 149-160.

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.

Almeida, Rafael J. A. (2018). Leia-léxico para inferência adaptada. Retrieved from: https://github.com/rafjaa/LeIA. Accessed May 2023.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. Lecture Notes on Software Engineering, 2(3), 174-179.

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.

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.

Chen, M. J. (2020). Examining the influence of emotional expressions in online consumer reviews on perceived helpfulness. *Information Processing & Management*, 57(6), 1–15.

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.

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*, Beijing, China.

Chung, S., & Cho, H. (2017). Fostering Parasocial Relationships with Celebrities on social media: Implications for Celebrity Endorsement. *Psychology & Marketing, 34(4), 481–495.*

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*, p. 1762-1770.

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.

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.

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: Apr. 20, 2023).

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.

Hughes, D. (2020). Digital Marketing Institute. #TeamTrees, vol. 2019's Biggest Influencer-Driven Viral Success. Retrieved from: https://digitalmarketinginstitute.com/blog/teamtrees-2019s-biggest-influencer-driven-viral-success. Accessed May 2023.

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.

Jacobson, J., Hodson, J., & Mittelman, R. (2022). Pup-ularity contest: The advertising practices of popular animal influencers on Instagram. *Technological Forecasting and Social Change*, 174, 121226.

Ladhari, R., Massa, E., & Skandrani, H. (2020). YouTube vloggers' popularity and influence: The roles of homophily, emotional attachment, and expertise. *Journal of Retailing and Consumer Services*, *54*, 102027.

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*.

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.

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

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.

Rouhani, S., & Mozaffari, F. (2022). Sentiment analysis researches story narrated by topic modeling approach. *Social Sciences & Humanities Open*, 6(1), 100309.
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.
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*, 47-65.
Sette, G., & Brito, P. Q. (2020). To what extent are digital influencers creative? *Creativity and Innovation Management, 29*(S1), 90-102.
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.
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.
Uzunoğlu, E., & Misci Kip, S. (2014). Brand communication through digital influencers: Leveraging blogger engagement. *International Journal of Information Management*, *34*(5), 592–602.
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.
Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine Learning*, p. 977-984.
Yaghtin, S., Safarzadeh, H., & Zand, M. K. (2020). Planning a goal-oriented B2B content marketing strategy. *Marketing Intelligence & Planning*, 38(7), 1007-1020.
Yoon, G., Li, C., Ji, Y. G., North, M., Hong, C., & Liu, J. (2018). Attracting comments: Digital engagement metrics on Facebook and financial performance. *Journal of Advertising, 3367*, 1-14.
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.
Zhao, Y., Kou, G., Peng, Y., & Chen, Y. (2018). Understanding influence power of opinion leaders in e-commerce networks: An opinion dynamics theory perspective. *Information Sciences*, *426*, 131-147.

Appendix A

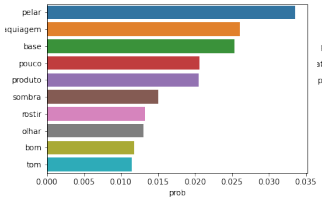
Topics names presentation, label grouping, and proportion

Topics	Label	% representativeness
Tópico 10 (N = 564): Questões matemáticas e lógicas simples e maneiras de resolvê-las	Educação (Education)	10.3
Tópico 16 (N = 607): Química e Limpeza - água, produto, plástico, temperatura.		
Tópico 24 (N = 362): História e Política - Brasil, mundo, governar, guerra, sociedade.		
Tópico 25 (N = 469): Educação - aula, aluno, professor, redação, exercício.		
Tópico 29 (N = 517): História antiga - épocas, nomes, fatos, lugares, reis		
Tópico 32 (N = 171): História do Brasil - regiões, épocas, fatos, personagens		
Tópico 33 (N = 336): Pandemia - vírus, vacinas, máscaras, números, população		
Tópico 36 (N = 529): Astronomia - planetas, estrelas, universo, espaço, ciência		
Tópico 13 (N = 2937): Super-heróis e vilões poderosos em lutas épicas	Cultura e entretenimento (Culture and Entertainment)	10.1
Tópico 38 (N = 564): Cinema e séries - história, personagens, atores, cenas		
Tópico 27 (N = 1199): Espiritualidade, comportamento, estilo de vida.	Pessoas, comportamento e estilo de vida (People, Behavior and Lifestyle)	8.4
Tópico 39 (N = 211): Rotina diária - trabalho, tempo, dicas, hábitos, manhã		
Tópico 41 (N = 457): Comunicação em mídias sociais, vida cotidiana		
Tópico 43 (N = 529): Festas, estilo de vida e eventos sociais		
Tópico 45 (N = 533): Relacionamentos amorosos		
Tópico 26 (N = 1861): Receitas, bolos, chocolate, massa, leite, açúcar.	Gastronomia (Gastronomy)	8.1
Tópico 42 (N = 941): Culinária e gastronomia		
Tópico 5 (N = 292): Conteúdo digital inovador	Tecnologia (Tech)	7.9
Tópico 9 (N = 452): Configuração e desempenho de computadores para jogos		
Tópico 21 (N = 669): Tecnologia - aplicativo, celular, preço, tela, produto.		
Tópico 46 (N = 1323): Fotografia, Aparelho, Celular, Câmera, Qualidade.		
Tópico 7 (N = 1589): Política e governo no Brasil e suas controvérsias	Política, Economia e notícias (Politics, Economy and News)	6.4
Tópico 11 (N = 629): Atualidades, Direitos na sociedade atual		
Tópico 1 (N = 445): Saúde e tratamentos médicos	Saúde e estilo de vida saudável (Health and healthy lifestyle)	5.9
Tópico 2 (N = 529): Alimentação saudável e equilibrada		
Tópico 18 (N = 495): Situações Interpessoais - relação, problema, possibilidade.		
Tópico 49 (N = 567): Exercícios, Academia, Corpo, Movimento, Força.		
Tópico 17 (N = 1088): Artesanato e Papelaria - papel, pedaço, formato, cor.	Decoração, Organização e Faça Você Mesmo (Decoration, Organization and DIY)	5.7
Tópico 31 (N = 894): Decoração de casa - móveis, ambientes, projetos, cores		
Tópico 23 (N = 374): Realização Pessoal - pessoa, mundo, dinheiro, trabalhar, sonhar.	Economia, Empreendedorismo e Negócios (Economics, Entrepreneurship and Business)	5.5
Tópico 28 (N = 694): Investimentos - ações, bolsas, valores, bancos, setores		
Tópico 34 (N = 837): Finanças pessoais - dinheiro, investimentos, dívidas, bancos		
Tópico 4 (N = 553): Laços familiares e afetivos	Família (Family)	5.4
Tópico 40 (N = 1303): Cotidiano em família		
Tópico 6 (N = 663): Jogo de ação emocionante	Jogos (Gaming)	4
Tópico 30 (N = 455): Jogos e entretenimento		
Tópico 47 (N = 280): Atualidades, Negócio, jogos, tecnologia.		
Tópico 0 (N = 1361): Maquiagem natural e sofisticada	Beleza (Beauty)	3.9
Tópico 3 (N = 763): Roupas e acessórios estilosos	Moda/estilo de vida (Fashion/Lifestyle)	3.7
Tópico 19 (N = 529): Moda e Estilo - roupa, calçados, moda, visual, loja.		
Tópico 12 (N = 393): Turismo, curiosidades, aventuras emocionantes	Viagens, aprendizados e curiosidades (Travel, learnings and curiosities)	3.4
Tópico 15 (N = 805): Viagens incríveis e bonitas pelo mundo		
Tópico 14 (N = 772): Jogos de futebol e desafios emocionantes	Esportes (Sports)	3.4
Tópico 48 (N = 421): Futebol, Time, Jogador, Gol, Vitória.		
Tópico 44 (N = 812): Animais de estimação	Animais de estimação (Pets & Animals)	2.3
Tópico 8 (N = 372): Gírias e expressões informais em conversas cotidianas	Entretenimento/geral (Entertainment/general)	2.2
Tópico 22 (N = 171): Entretenimento - show, noite, gol, linha, bandar.		
Tópico 35 (N = 207): Música ao vivo - shows, bandas, letras, rock, funk		
Tópico 20 (N = 648): Guerra e Militarismo - Estados Unidos, Rússia, soldados.	Militar (Military)	1.9
Tópico 37 (N = 391): Jardinagem - plantas, flores, solo, vaso, árvores	Jardinagem (Gardening)	1.1

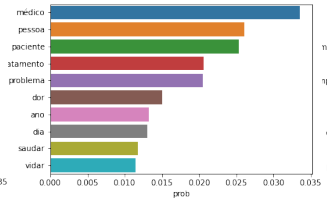
Appendix B

Probability distribution of top 10 most representative words by topic

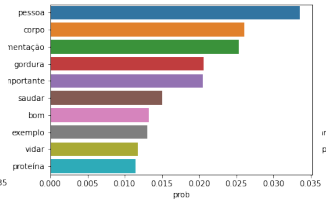
Topic 0



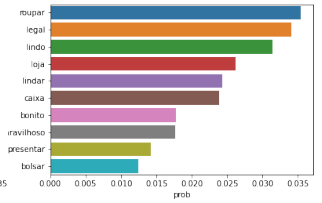
Topic 1



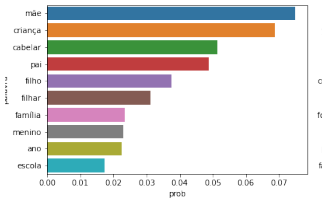
Topic 2



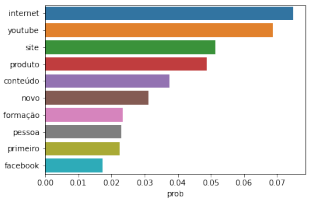
Topic 3



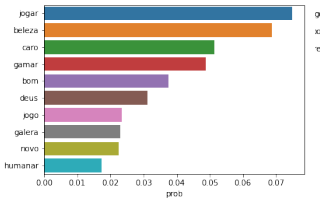
Topic 4



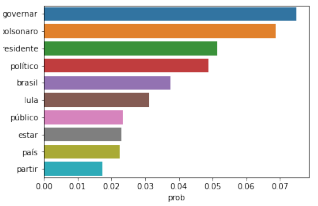
Topic 5



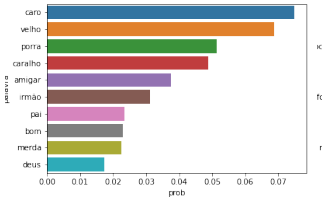
Topic 6



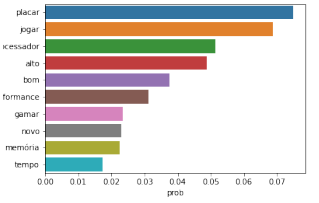
Topic 7



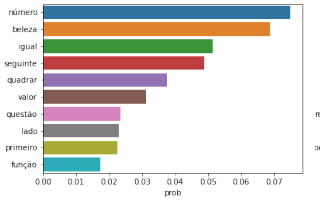
Topic 8



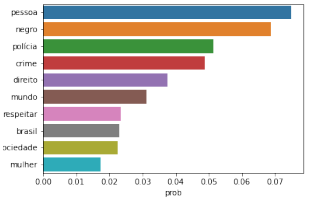
Topic 9



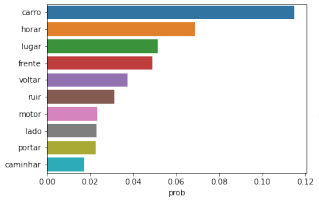
Topic 10



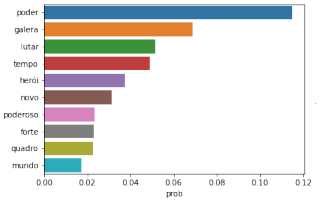
Topic 11



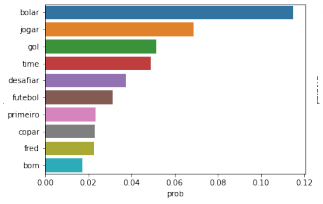
Topic 12



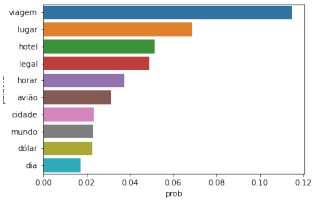
Topic 13



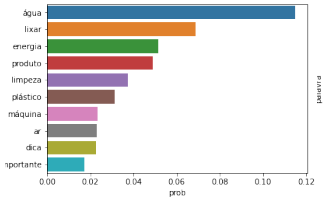
Topic 14



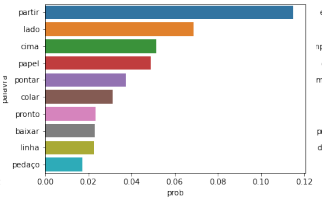
Topic 15



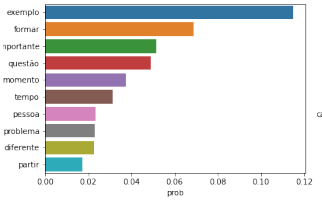
Topic 16



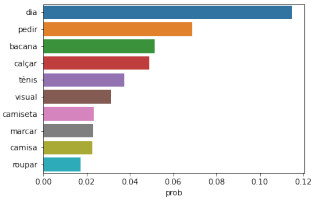
Topic 17



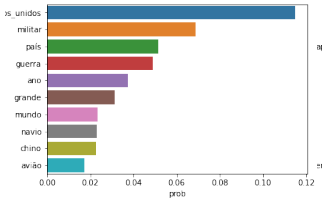
Topic 18



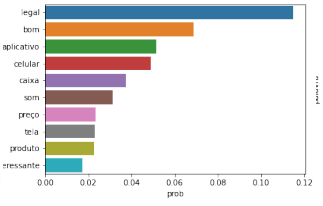
Topic 19



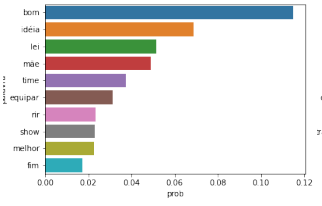
Topic 20



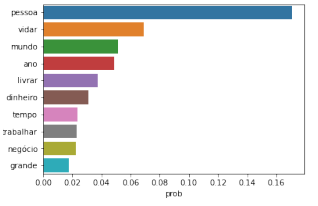
Topic 21



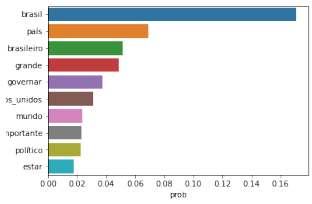
Topic 22



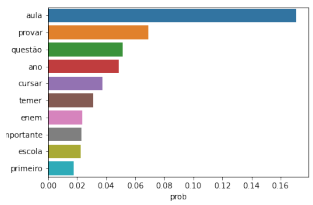
Topic 23



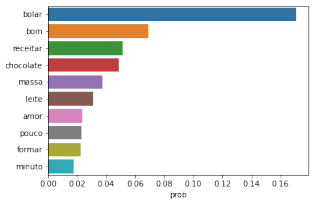
Topic 24



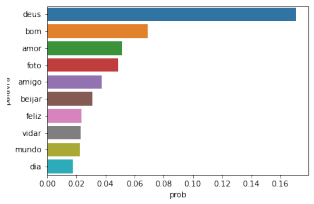
Topic 25



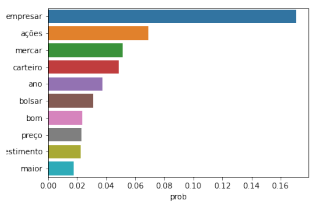
Topic 26



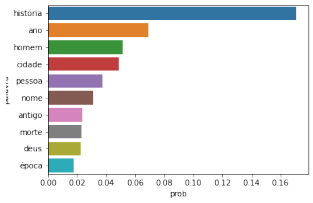
Topic 27



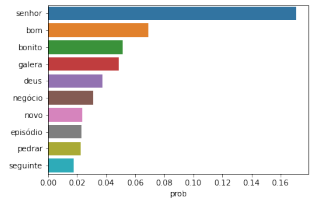
Topic 28



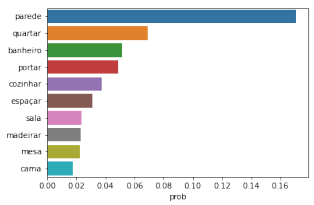
Topic 29



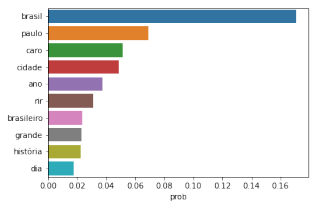
Topic 30



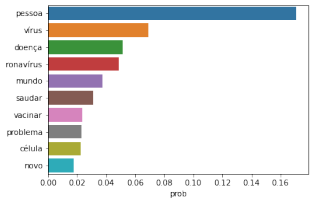
Topic 31



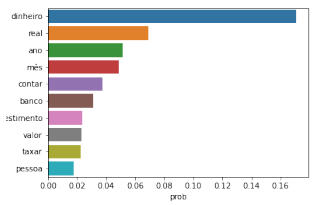
Topic 32



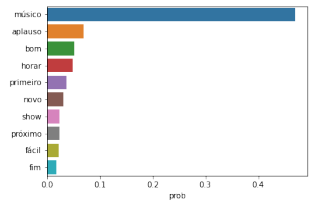
Topic 33



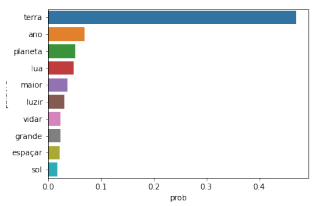
Topic 34



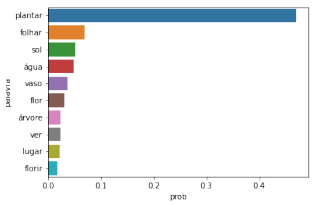
Topic 35



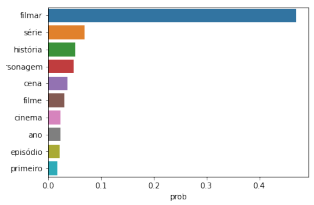
Topic 36



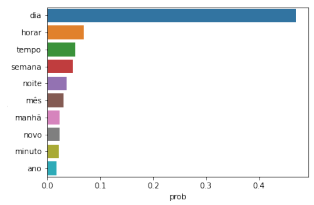
Topic 37



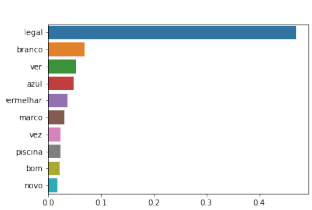
Topic 38



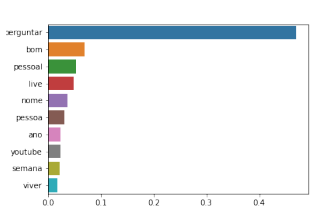
Topic 39



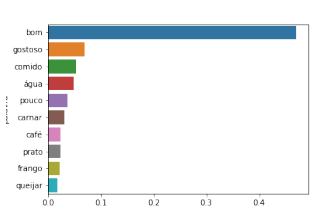
Topic 40



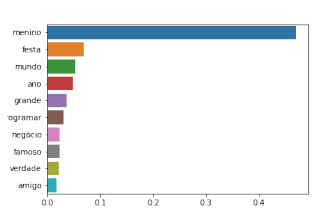
Topic 41



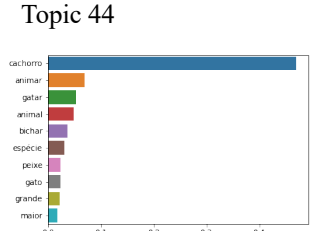
Topic 42



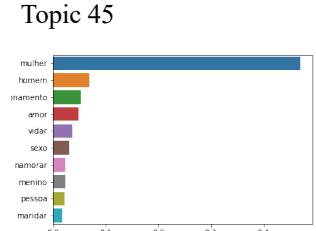
Topic 43



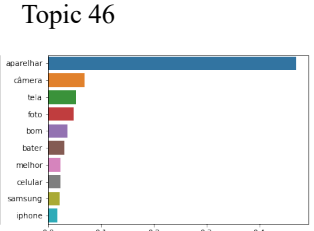
Topic 44



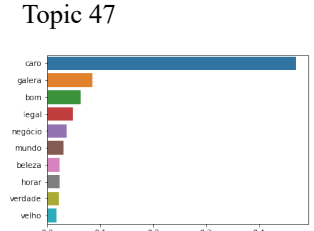
Topic 45



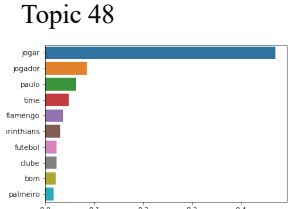
Topic 46



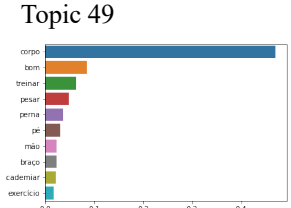
Topic 47



Topic 48



Topic 49



[i] The authors chose to select the sample before the Coronavirus pandemic period to avoid bias in the video content.
[ii] YouTube Data API: https://developers.google.com/youtube/v3.
[iii] YouTube Transcript/Subtitle API: https://github.com/jdepoix/youtube-transcript-api.
[iv] API Gensim - Phrase (collocation) detection (https://radimrehurek.com/gensim/models/phrases.html)