U.Porto Journal of Engineering, 10:3 (2024) 47-58
ISSN 2183-6493
DOI: 10.24840/2183-6493_0010-03_002630

Received: 16 April, 2024
Accepted: 1 July, 2024
Published: 16 July, 2024

Textual Patterns and Virality in X: An Analysis of Engagement in Telenovela Posts

William Ferreira
Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 PORTO, Portugal (ferreira.williamsilva@gmail.com) ORCID 0000-0002-3146-1905

Jefferson Lima
Informatics Center, Federal University of Pernambuco, Av. Jorn. Aníbal Fernandes, s/n, RECIFE-PE, 50740-560, Brazil (jeffersonlima163@gmail.com) ORCID 0009-0009-4078-9804

Author Keywords

Virality; Text Analysis; Natural Language Processing; Engagement; Social Media.

Type: Research Article

Open Access
Peer Reviewed

Abstract

X, previously known as Twitter, boasts 556 million active users and is widely used by businesses to engage with their audiences. In our study, we focused on TV Globo's telenovela "*Terra e Paixão*" broadcast in 2023, to analyze the impact of textual patterns on post virality using natural language processing techniques. Techniques like sentiment analysis, Part-Of-Speech Tagging, reinforcement scoring, TF-IDF, semantic similarity, and cosine similarity were utilized to identify attributes that contribute to a post's success, aiming to enhance marketing strategies. We employed language models like BERT, RoBERTa, and e5 in our analysis. Our findings indicate that while various metrics affect post engagement, the challenge remains complex. Textual characteristics, although essential, do not fully explain a publication's popularity, underscoring the need for a multifaceted approach to understanding social media dynamics.

1. Introduction

Understanding the factors that drive user engagement on social media platforms is crucial for effective digital marketing. This study focuses on the social media platform Twitter, which was acquired by Elon Musk in 2022 and underwent significant changes to become a super app, now known as "X" (Hurst 2023). Our research aims to explore the impact of these changes on user engagement and identify key content elements that contribute to the virality of posts.

In the updated version of the platform, basic terms such as "tweet" or "retweet" were replaced by "post" and "repost," respectively. Despite these changes, the platform retained a base of 556 million active users, with Brazil ranking as the fourth country in terms of audience, accounting for 24 million users (Kemp 2023). The microblog, known for its concise content, established itself to disseminate news and trends and form communities. According to the We Are Social report (Kemp 2023), 47.1% of social media users use them to contact friends and family, while only 22.7% interact with their favorite brands. This mismatch between the interests of the public and the brand environment presents challenges to marketing managers, who must continuously understand their audience, behaviors, and expectations to promote attractive and relevant content that ensures brand involvement in online discussions.

The metrics for monitoring the engagement of a post on "X" include the total number of comments, likes, and reposts. These indicators demonstrate the post's importance and help the algorithm discern how relevant that content is to other users, extending its reach. Thus, ensuring attractive content has become a predominant challenge in digital marketing. Our study aims to verify whether the key to a viral post lies in the content of publications that have already achieved success in engagement. Our focus is on identifying, through data analysis and natural language processing techniques, which elements of a post's content significantly contribute to its success with the audience of a specific profile or theme. Aware that engagement is based on three key variables, we turn to the challenge of understanding the impact of content structure on the number of likes received. The methodology adopted is based on the structure of the information collected in our database, which will be detailed later.

From a social behavior perspective, Berger (2013, cited in Pressgrove 2017) proposes a marketing methodology that investigates factors capable of propelling content to the potential for virality. Named STEPPS, this methodology encompasses six variables identified in the analysis: (1) Social Currency, related to people's concern about appearing informed and intelligent, aiming to gain recognition; (2) Trigger involves exposure to messages stimulating action, directly impacting the receiver's behavior; (3) Emotions, where feelings of high arousal such as admiration, surprise, enthusiasm, humor, anger, or anxiety tend to provoke intense reactions, in contrast to more apathetic emotions, except in messages linked to charitable causes; (4) Public, where the visibility of a message fosters actions in a "herd mentality" model; (5) Practical Value, emphasizing the tendency to share content perceived as applicable; (6) Stories, highlighting the effectiveness of narratives in conveying meaningful messages, are more prone to virality than advertisements due to the perception of greater credibility. This social approach to engagement allows us to examine whether publications promoted by a specific brand on social media can guarantee the desired engagement.

Numerous authors have dedicated efforts to develop models determining the probability of a post going viral on X. Among these, researchers focused on the processes and attributes of virality and prediction (Jenders 2013; Pressgrove 2017), the influence of users and content on a post's popularity (Maleewong 2016), and predictions related to message propagation (Petrović 2011; Kupavskii 2012) stand out. However, to facilitate comparison with this study, we suggest grouping them into two categories: those investigating the structural characteristics of posts and those examining the content of the published messages.

The findings of this study have the potential to guide content marketing actions more assertively, given the knowledge about the text structures that impact users on social media and how these structures can motivate them to interact with the content more frequently. This understanding has practical and managerial implications; it defines marketing strategies and guides the analysts responsible for content production so that content with a higher probability of engaging users is prioritized.

1.1. Structural analysis

In the field of structural analysis, scholars such as Bakshy et al. (2011) demonstrated that the prediction of a repost does not depend so much on its textual content but rather on the average retweets of the first user, thus estimating the potential for expansion of this cascade of shares. Based on Bakshy's findings, Kupavskii et al. (2012) introduced new predictive elements, including the user's PageRank in the repost network and the cascade dynamics. A user's PageRank in this context reflects their influence in the X sharing network, indicating that users with high PageRank frequently repost or are reposted by other influential users, occupying a central position in disseminating information on Twitter.

Jenders et al. (2013) investigated reposting trends using structural variables such as the number of followers. A surprising result indicated a higher propensity for reposting among users with less than 10,000 followers than those with over 300,000. Additionally, aspects such as the emotional valence of the post, its length, the number of hashtags, and URLs, among others, were examined. The analysis compared the Naive Bayes model with the Generalized Linear Model, revealing that the latter generally provides more accurate predictions as it does not assume conditional independence between variables. While Naive Bayes treats each feature in isolation, the generalized linear model considers their interactions.

On the same line, Maleewong (2016) applied a Multiple Linear Regression Model evaluating three variables: (1) characteristics of users who repost, including their popularity rate and activity; (2) content attributes, such as the presence of URL and the use of visual elements; (3) characteristics of the post author, such as the number of followers and post frequency.

This study demonstrated that these variables are more effective in predicting the popularity of posts. Maleewong's proposed model outperformed approaches that focused exclusively on content and author characteristics, highlighting the relevance of active and popular users in post propagation.

1.2. Content Analysis

In content analysis, researchers like Petrović et al. (2011) achieved significant results by studying the impact of reposts on X, comparing human predictability with technology. In their experiment, participants judged the viral potential of a post, and the results were then confronted with predictions of a machine learning model based on the Passive-Aggressive algorithm. This study highlighted that, besides quantitative factors such as the number of followers, elements like hashtag usage, post length, and popular keywords influence virality. The developed model achieved an accuracy of 82.7%, evidencing its ability to predict the reposting of content correctly.

Pressgrove (2017), although not using natural language processing, applied Berger's six factors (2013) to analyze the content of posts and their viral potential. This research established a strong correlation between STEPPS attributes and post engagement, with 16.8% reposted, 27.7% receiving likes, and 10.6% comments, especially those presenting “social currency” and “practical values.” This underscores the importance of textual analysis in understanding the virality of posts, which is the focus of this study.

Considering the extensive study of the structural elements of a post and its relation to popularity, this work is dedicated to exclusively analyzing posts' content, isolating aspects such as theme and textual elements without considering variables like followers, posting time, or the use of visual resources. We selected TV Globo and its publications about the novel “*Terra e Paixão*”, aired in 2023, to evaluate audience engagement based on the likes of each post.

We opted for a single brand and editorial line to minimize uncontrollable variables, focusing on the content of the postings, as TV Globo's publications are directed at the same audience and distributed at specific times and days, allowing for a more precise study of the impact of content on the popularity of posts.

2. Experimental Methodology

2.1. Overview

This study employs a social media monitoring approach to collect and analyze data from X. By focusing on the profile of TV Globo; we aim to understand engagement patterns in posts related to the telenovela "*Terra e Paixão*". The research design involves continuous data collection and filtering posts related to the telenovela. The study focuses exclusively on content elements to determine if specific textual practices enhance engagement on X. Structural variables such as the number of followers and posting time were excluded because all analyzed posts were made by the same profile (TV Globo) and during a particular time frame - the airing of the telenovela. This consistency in structural variables is advantageous for our analysis, as it allows us to isolate and examine the impact of textual content on engagement without the confounding influence of varying follower counts or posting times. By maintaining stable structural variables, we ensure that any observed differences in engagement can be attributed more confidently to the content itself, thus providing more precise insights into the textual strategies that drive audience interaction.

2.2. Data Collection

For data extraction, we utilized Buzzmonitor, a social media monitoring tool capable of collecting and storing information directly from X's API. This tool provides structured access to public posts, comments, likes, reposts, and user profiles. The platform's data collection method involves defining specific keywords or profiles for monitoring, and based on these definitions, data is collected continuously and in real-time, enabling updated analyses of trends and activities on social networks.

In this study, we were exclusively interested in the posts made by TV Globo's profile. Therefore, the tool was configured to collect all posts made by @tvglobo (https://x.com/tvglobo), regardless of the shared content. In other words, there was no need to define specific keywords. Using Buzzmonitor, we continuously indexed posts made by the broadcaster from May 1, 2023, to September 30, 2023 (from 00:00 to 23:59). During this period, TV Globo published 7,334 posts on X, covering various themes of its programming.

To analyze posts related to the telenovela "Terra e Paixão," we filtered all posts that included the hashtag #TerraePaixão. This hashtag was specifically chosen because TV Globo uses it to segment its editorial lines and categorize content related to the telenovela. By using this hashtag, TV Globo ensures that posts are easily identifiable as relevant to *"Terra e Paixão"*. Consequently, this filtering process guarantees that the posts included in our analysis are directly related to the telenovela, aligning with our research objectives of understanding engagement patterns specific to this content. This filtering resulted in a final set of 1,722 posts (23%) referring to the soap opera, as presented in Figure 1.

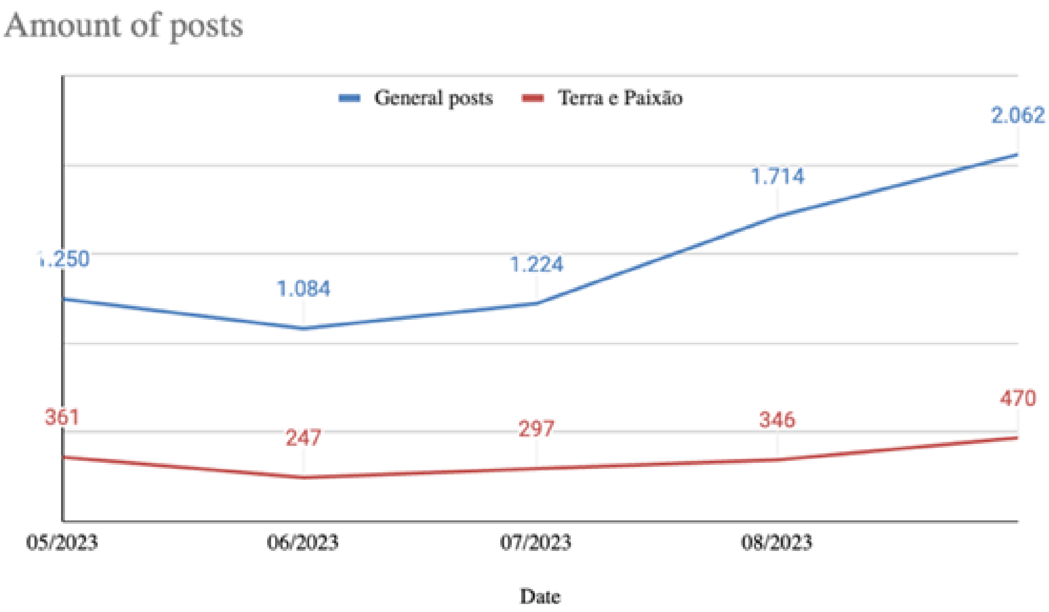


Figure 1: The image shows the number of posts on X by TV Globo in general and about “Terra e Paixão”.

Although the Buzzmonitor tool provides a comprehensive view of the posts made by TV Globo's profile on X, our selection criterion for the editorial line used a specific hashtag filter related to the telenovela for more accurate content control. However, this filter may inevitably result in an incomplete dataset if TV Globo's profile does not mention the hashtag in some posts. In any case, the complete dataset of TV Globo's posts during the period consists of 7,334 posts, all publicly available on their profile on X. This ensures the possibility of data extraction and reproducibility of the study at any time. To ensure the accuracy and reliability of the extracted data, we conducted a human verification process. Specifically, we manually reviewed all 1,722 posts to ensure they were related to the telenovela *"Terra e Paixão"*. This thorough review ensured that our dataset was accurate and relevant to the research objectives.

2.3 Data Analysis

Of the 1,722 analyzed publications about the telenovela, we divided them into two distinct sets using the Pareto rule (Investopedia 2022). This methodology aims to polarize the results by isolating posts with high engagement from the other posts in the dataset.

- **Group 1:** Publications that generated engagement.
- **Group 2:** Publications that did not achieve significant engagement.

Group 1 includes 20% of the most popular posts, totaling 344, with likes ranging from 319 to 7,244. Group 2 comprises the remaining 80% (1,378 publications), each with likes between 0 and 318. This division allows for a detailed analysis of the texts in the first set to understand the characteristics that contributed to their popularity among the follower base. In this study, we consider a post to have achieved positive engagement if it received more than 319 likes.

Additionally, we apply natural language processing (NLP) techniques to analyze the textual content of posts using language models such as BERT, RoBERTa, and e5. This methodology allowed us to isolate the textual elements that most impacted engagement while keeping structural variables constant (number of followers, posting times, etc.) by focusing exclusively on publications from the same source and specific period.

3. Results and discussions

3.1. Sentiment of posts

We examined the emotions of posts, classifying them as positive, neutral, or negative, to check for a relationship between the emotional tone of the message and the engagement generated. The results indicate a predominance of posts with positive emotion, especially in Group 1, which represents 81.4% of the most liked posts, compared to Group 2, with 74.38%. These are followed by 10.76% and 14.01% of posts of neutral content and 7.85% and 11.61% of negative content, respectively. These findings point to a tendency of posts with a positive nature to promote greater involvement, highlighting the significant influence of emotional tone on user behavior on social media platforms.

The higher engagement with positive posts can also be attributed to the fictional nature of TV Globo's content. Telenovelas, by design, offer a form of escapism where viewers project their emotions and find solace in the narrative. Positive posts related to these stories have the potential to create deeper connections with users, who, in turn, react to express this synergy. This emotional investment is not just a reaction to the content but a way for viewers to extend their experience of the narrative beyond the screen. Additionally, it is natural for TV Globo to share parts of the most emotionally tense storylines in their posts to stimulate this connection, using humor, inspiration, or nostalgia as references.

3.2. Word class

Using Part-Of-Speech Tagging (POST), which involves identifying the grammatical class a word belongs to (Jurafsky e Martin 2000), and using a Hugging Face Transformers model, we performed individualized counting for Group 1 and Group 2. The goal was to identify the 20 most prevalent grammatical classes in each group, considering that each entry x1, x2, ..., xn corresponds to a specific class y1, y2, ..., yn.

The results showed similarity in data classification, with "N" (noun), "PU" (punctuation), "NPROP" (proper noun), and "V" (verb) being the most recurring categories. The central hypothesis of the analysis suggests that Group 1's posts would likely have a higher number of verbs, indicating action. At the same time, Group 2 would tend to contain more nouns, reflecting a more descriptive nature.

The chi-square test was applied to verify if there were significant differences between the two groups' counts of each grammatical class. The results showed that for most categories, such as "V" (verb), the differences were not statistically significant, with p-values greater than 0.05 (0.2985). However, the category "N" (noun) showed a significant difference, with a p-value of 0.00036. This indicates that although verbs, or actions, are not the decisive factor in the engagement of a post, the richness of the description contained in the content is.

This is evident as, upon calculating the harmonic mean, it was observed that Group 1 averages 15.06 nouns per post and 2.37 verbs, while Group 2 recorded 14.55 nouns per post and 2.24 verbs. The analysis is an important discovery for digital marketing practices, given the importance of the call to action in branded content.

3.3. Emphasis and punctuation

Question marks and exclamation points are often associated with curiosity or emphasis. In the context of soap opera narratives, they can be essential for reinforcing the emotional aspects of interaction with the public. Therefore, we investigated whether there is a correlation between these punctuation marks and the success of TV Globo’s posts. We conducted a quantitative analysis of the use of question marks "?" and exclamation points "!" to compare their usage between different groups. In Group 1, we observed the presence of 81 posts with "?" and 131 with "!". In Group 2, we recorded 422 "?" and 531 "!". Due to the discrepancy in quantity between the groups, we normalized the data in percentage terms. The results indicate that, in Group 1, there are 23.55% "?" and 38.08% "!", while in Group 2, we find 30.62% "?" and 38.53% "!". This implies that, despite using these punctuation marks in both cases, they are more frequent in posts that did not generate engagement. This finding suggests that, although these marks can express emotion or curiosity—characteristics closely associated with soap operas—they do not necessarily correlate with more effective results in social media postings.

Moreover, as presented by Jung et al. (2022), some studies show punctuation as a form of clickbait, a strategy to catch the public's attention regardless of content quality. However, this increase does not necessarily translate into long-term engagement or user satisfaction. The negative perception associated with clickbait practices can eventually lead to declining user engagement and trust.

3.4. Frequent word

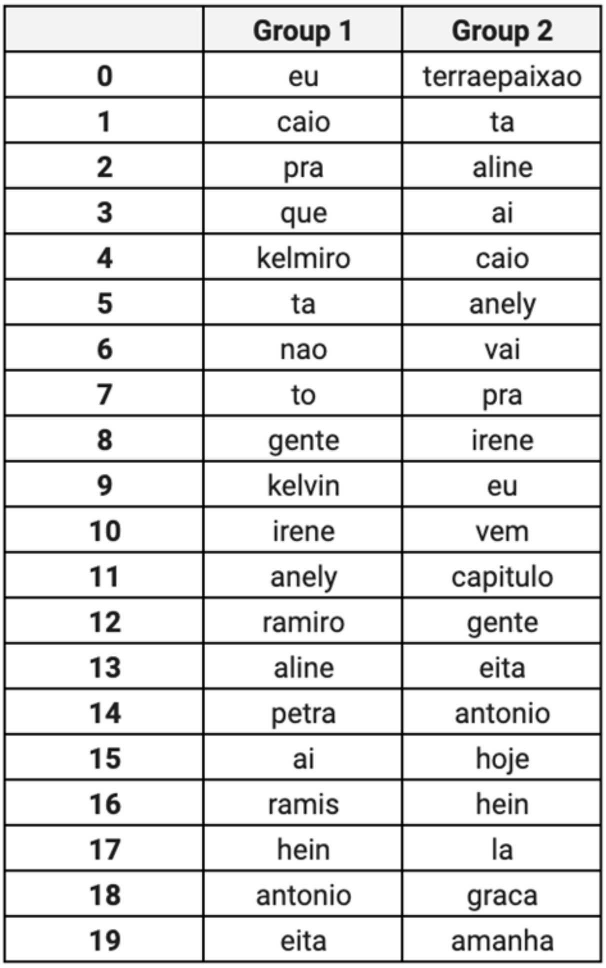
Some expressions can trigger reactions in the audience affected by the brand's post; therefore, we adopted the word counting technique to identify the 20 most frequent in each group. After the survey, we eliminated spaces and empty character strings, aiming for accuracy in counting. We observed that, in both contexts, the presence of URLs signals external content, such as images or videos, in addition to #TerraePaixao, reinforcing the brand and programming. Emojis (examples: "🗣", "🔥") also appear regularly, revealing a distinctive communication style, while "face with a speaking head" is often associated with a call to action, encouraging participation, open dialogue, or exchange of ideas. "fire" is generally used to convey enthusiasm, energy, and passion. In addition to being used to express the excellence or trend of something, it is also used to indicate that it is relevant to the moment.

However, Group 1's posts were distinguished by the frequency of emojis like "❤️", signaling the manifestation of affection or intense emotions, such as laughter ("😂") and crying ("😭"). Considering that TV Globo publications are accompanied by video content with scenes from the soap opera, the semantic implications of these emojis reinforce the publication's content by emphasizing loving, funny, or simply sad scenes. These indicators are crucial for guiding the message of the posts; that is, content expressing love, emotion, or comedy tends to increase engagement, indicating the public's apparent preference for interacting with emotional narratives.

3.5. Relevant word

A crucial factor for reader engagement in a novel is tied to the plot surrounding key characters. To analyze and identify relevant terms in a document that influence this plot, the TF-IDF (Term Frequency-Inverse Document Frequency) technique is employed by Jurafsky and Martin (2000) detail this methodology as a numerical statistic that reflects the importance of a word in a document collection or corpus. TF-IDF is a common technique in information retrieval, such as indexing and ranking documents in response to a query.

The analysis pointed out some words present only among the engaged posts in Table 1, such as "Kelmiro," a fusion created by internet users for the gay couple "Kelvin" and "Ramiro," in addition to "Kelvin," "Ramiro," "Petra," and "Ramis." Except for "Petra," the other variations refer to the couple, showing the plot's influence on the novel's audience. In this sense, it is essential to emphasize that the online audiences that interact with TV Globo's content demonstrated through the engagement identified in Group 1 of publications that themes linked to diversity and inclusion, such as the representation of an LGBTQIAPN+ couple, can be positive for the results of the brand, as it expands its reach through new audiences.



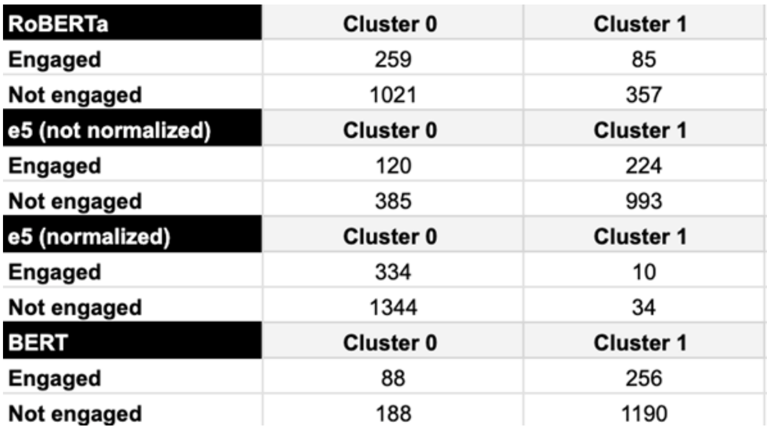
	Group 1	Group 2
0	eu	terraepaixao
1	caio	ta
2	pra	aline
3	que	ai
4	kelmiro	caio
5	ta	anely
6	nao	vai
7	to	pra
8	gente	irene
9	kelvin	eu
10	irene	vem
11	anely	capitulo
12	ramiro	gente
13	aline	eita
14	petra	antonio
15	ai	hoje
16	ramis	hein
17	hein	la
18	antonio	graca
19	eita	amanha

Table 1: The table shows a list of relevant word using TF-IDF, where. Group 1 are engaged posts, and Group 2, not engaged posts.

From a marketing management perspective, understanding the plot that enhances brand engagement is crucial for both strategic content cuts and investments in the continuity and prominence of characters in the novel. So, TF-IDF analysis also serves as valuable feedback to soap opera writers and producers. By identifying which characters and plots appeal most to the audience, the production can adapt the story to suit the audience's interests and preferences better, potentially adjusting the narrative focus to maintain or increase engagement.

3.6. Semantic grouping - Clustering

We employed unsupervised machine learning techniques and natural language processing to continue investigating the correlation between the semantics of posts and their engagement. For this purpose, we utilized the embedding-generating library algorithms BERT (Devlin et al. 2018), RoBERTA (Liu et al. 2019), and e5 (Wang et al. 2022). We used the KMeans (Jin and Han 2011) algorithm for semantic clustering. The objective was to group the posts into two clusters (Engaged and Not Engaged) based on their semantic similarity. The goal was to observe if the 2 clusters tend to have a predominance of one type of post (Engaged or Not Engaged). In this way, engagement would be present in the semantics.



RoBERTa	Cluster 0	Cluster 1
Engaged	259	85
Not engaged	1021	357
e5 (not normalized)	**Cluster 0**	**Cluster 1**
Engaged	120	224
Not engaged	385	993
e5 (normalized)	**Cluster 0**	**Cluster 1**
Engaged	334	10
Not engaged	1344	34
BERT	**Cluster 0**	**Cluster 1**
Engaged	88	256
Not engaged	188	1190

Table 2: The table shows clustering in comparison between models.

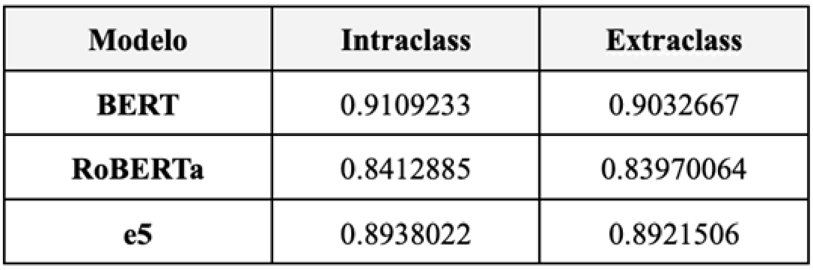
As can be observed in the table above, the groups generated needed to have a clear predominance of engaged or not engaged posts. This scenario reveals the complexity of using semantic similarity analyses to predict or understand engagement on social networks, as the interpretation of the content varied considerably among the employed models, even though there was more assertive clustering in some variables. The result indicates that, although the models can identify semantic nuances, this alone does not explain the engagement phenomenon, given the heterogeneity identified in each group.

The variability in semantic clustering highlights that audience engagement is influenced by many factors beyond mere semantic content. This suggests that while semantic similarity can provide insights into the themes and tones that resonate with the audience, it does not capture the full spectrum of elements that drive engagement, among them, the different interpretations of the public, contextual influences, emotional and psychological triggers and how people project their emotions into the content they consume, especially in soap opera narratives.

3.7. Intra-class and Extra-class Similarity

To analyze the content and style of each group in our database, we used three language models, BERT, RoBERTa, and e5, with sentence embeddings where the mean of the intra-class and inter-class distances using cosine similarity (Jurafsky and Martin 2000) is calculated between two vectors A and B, where A . B is the dot product of the vectors, and ||A|| ||B|| are the norms of the vectors. Intra-class distance refers to the average of the distances between pairs of posts belonging to the same category (Group 1 or Group 2). In contrast, inter-class distance refers to the average of the distances between pairs of posts from different categories.

The choice of models significantly impacts how texts are grouped and interpreted due to their characteristics and training bases, leading to different data representations.



Modelo	Intraclass	Extraclass
BERT	0.9109233	0.9032667
RoBERTa	0.8412885	0.83970064
e5	0.8938022	0.8921506

Table 3: the table shows a comparative result of cosine similarity between RoBERTa, e5, and BERT focusing on intra-class and extra-class.

In the experiment, it was observed that the three embedding-generating models, BERT, RoBERTa, and e5, generated vectors whose extra-class distance was more significant than the intra-class distance. The expected result would be an intra-class distance more minor than the extra-class distance, indicating that the engaged posts are semantically similar, as are the non-engaged posts. When evaluating intra-class and extra-class similarity, BERT proved to be more effective in distinguishing characteristics of posts within the same class. However, the more minor extra-class similarity suggests that the semantic representation was insufficient to provide information that would guarantee a clear separation of the application of advanced machine learning techniques for semantic clustering offers profound insights into the intricate dynamics of post engagement. However, it is imperative to rigorously examine the implications of semantic variability on audience perception and behavior. Variability in semantic representations generated by different models can significantly influence how audiences interpret and interact with content. A comprehensive analysis of these semantic variations is essential for uncovering the subtleties of audience behavior, thereby enhancing the precision of content-targeting strategies for engaged and non-engaged posts.

Conclusions

Although some structural approaches are proposed to predict a post's engagement, in this study, natural language processing techniques were chosen to understand the impact of the textual content of posts about the soap opera "*Terra e Paixão*" on TV Globo's audience engagement. Our database was divided into two groups: Group 1, with publications that generated engagement, and Group 2, with those that did not.

It was observed that posts with positive sentiment tend to achieve better results, with a difference of 7 percentage points. Additionally, although the presence of verbs is not decisive for engagement, the descriptive richness of nouns (p-value of 0.00036) is crucial, highlighting the importance of well-structured content for digital marketing strategies.

On the other hand, reinforcement scores, typically used to express emotions or curiosities, like "?" and "!", did not prove to be crucial elements in the analysis, being present in both groups of publications and with greater representation among those that did not engage. The analysis suggests that the excessive use of these reinforcements may be perceived as less attractive by the audience.

Furthermore, the use of emojis proved to be an efficient attribute of the content, especially the exclusive presence of “❤️”, “😂”, and “😭” among the engaged publications. These words indicate extreme affections or emotions and correlate perfectly with the most frequent words involving the couple "Kelmiro", responsible for the romantic narrative and the script's comedy.

Regarding semantic similarity and its correlation with engagement, the three models used (RoBERTa, BERT, and e5) did not achieve satisfactory results in understanding the nuances of the text, indicating that they were not able to identify a consensus among the text segments, possibly due to the database's heterogeneity. When evaluating intra-class and extra-class similarity, BERT proved to be more effective in distinguishing characteristics of posts within the same class.

Finally, despite the approach focused on content rather than the structure of the post, it is concluded that analyzing engagement in social networks represents a multifaceted challenge and that, although necessary, it is insufficient to explain a publication's popularity.

Limitations and future research directions

Although the objective of this study was expressly the analysis of text as a potential for engagement in a publication, we recognize that the use of other characteristics, such as images and videos, is directly related to the result, forming a triad that deserves distinct and joint analyses. For this reason, although we know the choice to work exclusively with text, we acknowledge the limitation of this approach on the results. Additionally, although the data sample comprises all posts made by TV Globo's profile using the hashtag #TerraePaixao, we recognize the possibility that some posts about the telenovela may not have been considered if the hashtag was not used. However, we ensure that the database built from the hashtag exclusively includes posts about the theme.

For future research, an integrative approach, considering the impact of images and videos together with the text, can yield better results, as well as the analysis of other programs or telenovelas that serve as comparative parameters. Considering the narratives of the telenovelas, a longitudinal approach could also be considered, crossing the script with the content of the posts, not only the posts on social media. Additionally, expanding the engagement metrics, such as shares, comments, and click-through rates, can provide a more comprehensive view of what constitutes a viral post as well as the use of other language models in addition to BERT, RoBERTa and e5, analyzing their efficiency in observing the semantic parameters of engagement.

References

Bakshy, Eytan, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. “Identifying Influencers on Twitter.” In Fourth ACM International Conference on Web Seach and Data Mining (WSDM). Vol. 2. https://www.academia.edu/download/46922984/Identifying_Influencers_on_Twitter20160630-14564-iqdg4g.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arXiv. http://arxiv.org/abs/1810.04805.

Hasan, Mahbub, Abdullah Mueen, Vassilis Tsotras, and Eamonn Keogh. 2012. “Diversifying Query Results on Semi-Structured Data.” In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2099–2103. Maui Hawaii USA: ACM. https://doi.org/10.1145/2396761.2398581

Hurst, Luke. 2023. “A Transformação do Twitter Para X | Euronews.” Accessed July 31, 2024. https://pt.euronews.com/next/2023/10/27/a-transformacao-do-twitter-para-x.

Investopedia. 2022. "What Is the Pareto Principle—aka the Pareto Rule or 80/20 Rule?" Accessed December 31, 2023. Investopedia. https://www.investopedia.com/terms/p/paretoprinciple.asp.

Jenders, Maximilian, Gjergji Kasneci, and Felix Naumann. 2013. “Analyzing and Predicting Viral Tweets.” In Proceedings of the 22nd International Conference on World Wide Web, 657–64. Rio de Janeiro Brazil: ACM. https://doi.org/10.1145/2487788.2488017.

Jin, Xin, and Jiawei Han. 2011. “K-Means Clustering.” In Encyclopedia of Machine Learning, edited by Claude Sammut and Geoffrey I. Webb, 563–64. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_425.

Jung, Anna-Katharina, Stefan Stieglitz, Tobias Kissmer, Milad Mirbabaie, and Tobias Kroll. 2022. “Click Me…! The Influence of Clickbait on User Engagement in Social Media and the

Role of Digital Nudging." Edited by Jarosław Jankowski. PLOS ONE 17 (6): e0266743. https://doi.org/10.1371/journal.pone.0266743.

Jurafsky, Daniel, and James Martin. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Vol. 2.

Kemp, Simon. 2023. "Global Overview Report." Accessed December 31, 2023.

https://datareportal.com/reports/digital-2023-global-overview-report.

Kupavskii, Andrey, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. 2012. "Prediction of Retweet Cascade Size over Time." In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2335–38. Maui Hawaii USA: ACM. https://doi.org/10.1145/2396761.2398634.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv. http://arxiv.org/abs/1907.11692.

Maleewong, Krissada. 2016. "An Analysis of Influential Users for Predicting the Popularity of News Tweets." In PRICAI 2016: Trends in Artificial Intelligence, edited by Richard Booth and Min-Ling Zhang, 9810:306–18. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-42911-3_26.

McKelvey, Karissa, and Filippo Menczer. 2013. "Design and Prototyping of a Social Media Observatory." In Proceedings of the 22nd International Conference on World Wide Web, 1351–58. Rio de Janeiro Brazil: ACM. https://doi.org/10.1145/2487788.2488174.

Petrovic, Sasa, Miles Osborne, and Victor Lavrenko. 2021. "RT to Win! Predicting Message Propagation in Twitter." Proceedings of the International AAAI Conference on Web and Social Media 5 (1): 586–89. https://doi.org/10.1609/icwsm.v5i1.14149.

Pressgrove, Geah, Brooke Mckeever, and Mo Jones - Jang. 2017. "What Is Contagious? Exploring Why Content Goes Viral on Twitter: A Case Study of the ALS Ice Bucket Challenge." International Journal of Nonprofit and Voluntary Sector Marketing 23 (August):e1586. https://doi.org/10.1002/nvsm.1586.

Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. "Text Embeddings by Weakly-Supervised Contrastive Pre-Training." arXiv. http://arxiv.org/abs/2212.03533.

Wu, Shaomei, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. "Who Says What to Whom on Twitter." In Proceedings of the 20th International Conference on World Wide Web, 705–14. Hyderabad India: ACM. https://doi.org/10.1145/1963405.1963504.