



ARTICLE



<https://doi.org/10.1057/s41599-023-01544-x>

OPEN

Comparing content marketing strategies of digital brands using machine learning

Yulin Chen  ¹✉

This study identifies and recommends key cues in brand community and public behavioral data. It proposes a research framework to strengthen social monitoring and data analysis, as well as to review digital commercial brands and competition through continuous data capture and analysis. The proposed model integrates multiple technologies, analyzes unstructured data through ensemble learning, and combines social media and text exploration technologies to examine key cues in public behaviors and brand communities. The results reveal three main characteristics of the six major digital brands: notification and diversion module; interaction and diversion module; and notification, interaction, and diversion module. This study analyzes data to explore consumer focus on social media. Prompt insights on public behavior equip companies to respond quickly and improve their competitive advantage. In addition, the use of community content exploration technology combined with artificial intelligence data analysis helps grasp consumers' information demands and discover unstructured elements hidden in the information using available Facebook resources.

¹ Department of Marketing and Logistics Management of National Penghu University of Science and Technology, Penghu, Taiwan. ✉email: 143530@mail.tku.edu.tw

Introduction

Social media has drastically changed how we interact. Numerous companies have adopted social media to actively create business value in the community; increase popularity, loyalty, and return rate through content; improve public satisfaction; and strengthen brand awareness and reputation (Pitt et al., 2019). A growing number of individuals are also using social media to express their views about services and products. Exploring public behavioral reactions on social media can help businesses understand consumer demands, create meaningful brand content, improve the quality of service experience for various target audiences, and enhance competitiveness (Zhao et al., 2020).

Analyzing social media content, however, can be a challenging and time-consuming task (Semenov, 2013). The rapid growth of social media emphasizes the need for more automated exploration techniques. Companies constantly explore ways to adjust management risks and improve decision-making by performing competitive intelligence analyses. Companies monitor not only their own social media but also that of their competitors to gain reference information and insight into different perspectives such as public opinions, product promotions, and market trends. Social media generates expansive text data that provide additional cues. These cues can be converted into effective competitor insights (Gálvez Rodríguez et al., 2012). Thus, obtaining competitive intelligence enables organizations to recognize their own strengths and weaknesses and improve social media marketing efficiency and public gratification (Maier et al., 2015; Shrapnel, 2012).

The study develops a research framework that integrates multiple technologies, including unstructured data analysis, text mining, and ensemble learning, with social media and text exploration techniques to identify key cues in brand community and public behavioral data. The objective is to help companies perform more optimized information analyses by comparing various competitors' social media content and enhancing management strategies based on social media content. A total of six globally known digital brands are examined, of which three are digital product brands (Disney, Netflix, and Microsoft) and the remaining are digital platform brands (Google, YouTube, and LinkedIn). The research attempts to answer the following two questions by analyzing unstructured data on brands' fan pages:

- (A) Do fan pages of digital product brands and digital platform brands have key cues in their posts that improve brand image and marketing promotions?
- (B) Can actively identify public preferences for brand content using interactive data as well as analyze and predict competitor behaviors?

Fan pages provide a landing page for brands, businesses, organizations, and public figures and a point of contact with their customers and audiences. For example, information concerning promotional events or new product releases can be posted on fan pages to reach followers instantly. Businesses, brands, organizations, and public figures can also use fan pages to share updates or connect with their followers. Users who like and follow fan pages receive updates in their news feeds. Anyone can create a fan page, but only official representatives can create a fan page to be managed for organizations, business brands, and public figures. Therefore, in this study, the research only collected and analyzed the posts of fan pages linked to brands and excluded posts or content from personal/private fan pages.

Social media monitoring and analyses can be challenging (Weinberg and Pehlivan, 2011) given the vast amount of community data generated daily. This raises the question of whether more intuitive analysis techniques can help digital brands better

understand the public and behavior. Further, can data analyses identify more effective competitive strategies, information characteristics and corresponding methods, weaknesses, and unsatisfied public demands?

First, the advantage of this study is to combine three different ensemble learning methods to identify whether the suggested information provided by machine learning is significantly helpful to the interaction of each brand, with the main purpose of discovering the characteristics of different types of competitors. In terms of monitoring frequency and resource allocation, brands generally monitor direct competitors continuously while strategically monitoring indirect or potential competitors to assess their threat.

Secondly, this study examines content published on brand pages and information in posts using a data-mining program. This study also fully investigates the three sets of interaction modules (notification and diversion module; interaction and diversion module; and notification, interaction, and diversion module) to infer the key clues in each brand's postings. The evaluation model reveals positive image cues as indicators of public participation. Following an evaluation of the image cues, the study explores the behavioral impact of different types of information. Finally, the characteristics of posts on the digital brands' pages are integrated as a reference for content management and brand image construction.

The remainder of this paper is organized as follows. Section "Literature review" reviews the literature on community content exploration and text mining. Section "Research hypotheses" proposes the relevant hypotheses. Section "Research methodology" discusses the technical background and data analysis standards. Section "Data analyses and results" presents the verification results and explores ways to improve the quality of posts based on key cues. Section "Conclusions" highlights the contributions of this research and offers implementation suggestions.

Literature review

Text mining and competitor analysis. Researchers have shown that text information on social media platforms is more important than data information because it provides key insights through, for example, images, videos, and even emoticons, a characteristic that adds considerable information value to text mining (Bag et al., 2021). Data mining can be used to identify one or more specific problems in the database. However, text mining is a more complex process than data mining (Ahmed et al., 2016) because texts contain irregular and unstructured data patterns, while data mining largely focuses on processing structured data. Text mining extracts appropriate words and sentences from texts and then converts them into structured data using data restoration, machine learning, information mining, and computer linguistics (Sett et al., 2016).

Scholars have applied various techniques when conducting competitive analyses on social media content, such as statistical analysis, content analysis, text mining, and sentiment analysis. These methods involve data collection from competing social media content to obtain and compare public experiences and sentiments and the automated extraction of specific information useful to trend analyses (Zikopoulos et al., 2012). Text mining can be used to identify diverse trends ranging from those of online posts to influenza patients. In recent years, machine learning, cluster analysis, information analysis, and associative analysis have gained much popularity (Zikopoulos et al., 2012). Cluster analysis, in particular, identifies key information through applications, feature filtering, cluster calculations, and verification, the findings of which can serve as a reference for effective decision-making.

Thus, analyzing social media competition is data-intensive research that integrates computer science, statistical, data science, social science, and specialized methods with artificial intelligence techniques to derive comprehensive insights into text information and sentiment trends (Zikopoulos et al., 2012). However, few studies focus on digital brands in general and the continuous improvement of social media analysis and text mining technology in particular. Companies should attach greater importance to useful commercial content on social media (He et al., 2013) and explore text mining and the collection of competitive brand information to enhance their brand value and gain valuable insights.

Thus, many companies advocate the use of social media to further understand community benefits, grasp public response to new products, evaluate public perceptions and actions, and gradually cultivate loyal communities (Mostafa, 2013). In addition to brand reviews, it is necessary to continuously collect and analyze data on competitor characteristics. Given the wide range of companies on social media, it is possible to simultaneously monitor one's own and competitors' social media content to improve brand management. This also helps identify competitors' actions and trigger communities' self-protection mechanisms (Kaiser et al., 2011). Social media not only provides competitor information but also allows for a direct comparison of public participation among competing brands. To elaborate, users often share their opinions and compare competing brands. Understanding such differences contributes to a better understanding of public attitudes and related changes. Companies that consistently collect and analyze data on competing brands can better identify the advantages and disadvantages of their brands, and this information serves as a useful reference when formulating implementation strategies and guidelines. In sum, the proper use of social media to conduct competitor analysis enables companies to design strategies that can more effectively attract and retain customers.

Social media content and interaction. Several studies on the effects of concrete signs found that concrete cues evoke trust (Elliott and Wright, 2018) and enhance the willingness of users to participate in online events (Turel, 2015), share positive product information, and read marketing descriptions. Besides concrete signs, abstract community terminology and phrases, such as "team," "welcome," and "join us," can also incite a sense of belonging on social media. Concrete signs paint vivid imagery, while abstract signs solicit emotional responses. Psychologists have found that concrete and abstract concepts are organized differently in the mind, and different pathways are used to process concrete and abstract informational signs. Abstract concepts are organized through association, while concrete concepts are organized through categorization. Regardless of the cognitive processes, both types of signs help to strengthen users' sense of identity within their community.

Many early studies on social media advertising and user engagement mainly discuss the concepts of engagement (Dolan et al., 2019) or use likes, shares, or comments to verify engagement (Gavilanes et al., 2018). Methods included coding post content, analyzing user emotions, conducting regression analysis to determine the correlation between user engagement and behaviors, and cross-examining different social media content (rational and emotional), context (platform type and content format), and active/passive post engagement.

Drawing on previous studies, social media content has been conceptualized into three major categories: rational content (information, function, education, and current events), interactive content (experiences, personal, brand community, and user), and

marketing content (rewards, brand resonance, and promotions). Many scholars have attempted to investigate the properties of different content and their effects on user engagement. Many researchers examined the role of rational content in social media by adopting inconsistent methods to promote passive or active online engagement behaviors and find empirical support for both likes and comments (Coelho et al., 2016).

Behavioral verification of social media. Studies have shown that social media discussions affect corporate image (Kumar et al., 2016). Other studies found that the number of Facebook likes directly influences the business reputation. In terms of the factors that evoke users to like, comment, and share (Mochon et al., 2017), studies have found that brand and promotional content increases social media activity (Kacholia, 2013). Studies on user behaviors on Facebook can be broadly classified into two categories. The first is studies on user motivation and online behaviors through the theoretical lens of psychology and sociology, whereas the second study is on social media content followed by users.

Behavior is defined as activities promoted by interaction (Bowden et al., 2017), and engagement can be measured by analyzing visible behavior to determine the cognitive and affective components of social media engagement. User comments on Instagram like on Twitter, or shares on Facebook are why companies are active on social media (Beckers et al., 2017). Interactions on social media tend to trigger participatory experiences (Fradkin et al., 2018) used to build a solid and long-lasting company-user relationships. Additionally, user behaviors, such as like, comment, and share, can be used to measure content performance (Wallace et al., 2014). Attention (like), engagement (comment), and recognition (share) represent the steady progression of user participation. Therefore, Facebook likes, comments, and shares can be viewed as indicators of interactional progression (Nieto-Garcia et al., 2019; Viglia and Dolnicar, 2020).

The rise of social media effectively transformed users into active participants through content (Lee et al., 2020). Consequently, social media platforms use co-creation to evoke user interactions (Dolan et al., 2019), including brand creation, contribution, and consumption (Malthouse et al., 2016). The popularity of social media also attracted the attention of scholars and managers (Breidbach and Brodie, 2017). The concept of engagement can be empirically validated in the fields of education (Baron et al., 2014), marketing (Bowden et al., 2017), information systems (Terkenli et al., 2020), and digital marketing (Breidbach and Brodie, 2017), and engagement can be defined as an essential interactive element in brand or company promotion (Hollebeek et al., 2014). Engaging content can be in the form of products or services, activities or events (Vivek et al., 2012), or PR campaigns or promotions (Malthouse et al., 2016), and the extent of engagement can be used to examine behavioral causality (Tian et al., 2021).

The study of engagement originated in the field of psychology. Nonetheless, engagement has prompted discussions in social science, political science, and organizational behavior and communication (Bag et al., 2021). The engagement has yet to be clearly defined in psychology and community science (Breidbach and Brodie, 2017). It is an adaptable structure that can be measured across different contexts and research topics (Table 1). Marketing researchers study engagement by referencing the synonyms of similar structures (Kannan and Kothamasu, 2020), such as word-of-mouth or brand community (Moran et al., 2020), commendations, and evaluations (García-Umaña and Tirado-Morueta, 2018), or the enthusiasm of following a topic of

Table 1 Summary of engagement sources.

Engagement element	Sources
Referencing the synonyms of similar structures	Kannan and Kothamasu (2020)
Word-of-mouth or brand community	Moran et al. (2020)
Commendations, and evaluations	García-Umaña and Tirado-Morueta (2018)
The enthusiasm for following a topic of interest	Yang et al. (2019)
Be examined on the basis of brand popularity	Villamediana-Pedrosa and Vila Lopez (2019)
Popularity, comment engagement, and virality.	Villamediana-Pedrosa and Vila Lopez (2019)
Popularity refers to likes, comment engagement refers to comments, and virality refers to shares	

interest (Yang et al., 2019). User responses in brand-focused social networks are seen as the manifestation of engagement (Ksiazek et al., 2016) and explain that responses to or opinions on a post are indicators of engagement and signs of intra-network user interactivity. However, engagement is largely studied in marketing (Dessart et al., 2015).

Engagement also refers to contextual interactions and can be validated by examining various experiences (Breidbach and Brodie, 2017). Studies on behavior have fairly consistent views as those on engagement (Dolan et al., 2019). Fanpage engagement can be broadly categorized into six types: creation, contribution, consumption, dormancy, active engagement, and passive/personalized engagement. The interactive characteristics of engagement may lead to different engagement intensities, such as passive (low) engagement or active (high) engagement, or optimistic and pessimistic engagement. Passive engagement refers to browsing without engaging in the community. By comparison, active engagement highlights the participatory interest of the user. Subsequently, users can validate whether the engagement of a specific social media platform is active or passive and engagement intensities by observing others' participation in activities and events, posting frequency, dissemination of information, and provision of emotional support (Peters, 2018). For example, likes can be viewed as emotional responses and comments as public deliberation, and both can be viewed as dynamic social media behavior. By contrast, clicking and reading content is viewed as passive engagement or psychological behaviors (Breidbach and Brodie, 2017).

Active participation is a multidimensional structure and is an indicator of a brand's positive value (Villamediana-Pedrosa and Vila Lopez, 2019; Azer and Alexander, 2018). Researchers have suggested that participatory public behaviors in social brand communities can be examined on the basis of brand popularity (Villamediana-Pedrosa and Vila Lopez, 2019). On Facebook, the degree of brand popularity can be verified by the rate of participation. Comments on posts can be measured to determine the positive and negative evaluation of brands (Villamediana-Pedrosa and Vila Lopez, 2019; Naumann et al., 2017). Participation is an adaptable structure and can be measured in different contexts and across fields (Breidbach and Brodie, 2017).

Research hypotheses

Digital brands attempt to enhance their brand recognition through online commercial content and services such as applications, gaming, shopping, product and service experiences, and knowledge. Digital brands have the advantage of setting up virtual stores, and this has contributed to the launch of numerous digital product and platform brands. Unlike traditional brands, digital brands must offer online experiences that satisfy consumers' sense of familiarity and convenience as well as trust in digital environments by improving the quality of information. In addition, they must create an environment that promotes identity expression with a focus on self-positioning, quality of life, and interactions with friends and family. Companies can publicize

and promote digital commodities through their fan page and enhance fan interaction by identifying keywords used at high frequencies. ML calculus, for example, helps identify high-frequency key cues in fan posts on, for example, Facebook. Brands can use this information to encourage active fan participation through likes, comments, and shares.

Digital brands are required to design information blueprints involving specific themes and product packaging when creating social media strategies, official website information, or online video content. All content is adjusted and arranged as per the blueprint and in line with the brand's positioning and promotion objectives. Brands use social media to develop an image and communicate with consumers (Lipsman et al., 2012), whereas the public does so to express their opinions using the like and comments features (Giannakos et al., 2014). Brands use such exchanges to establish relationships with their consumers (Brodie et al., 2013).

This research aims to identify key cues by examining public participation data. The model integrates the different dimensions of perception, cognition, and interaction on Facebook and conducts quantitative and artificial intelligence data analysis to understand the impact of community content on public participation. In addition, it analyzes the role of participatory posts in enhancing interactive public experiences and how content can be used to satisfy the public desire for brand information. The study is mainly based on artificial intelligence and ensemble analysis methods and focuses on digital product and platform brands. Finally, it analyzes the effects of key cues in fan page content on public perception and behavior. Accordingly, the following three hypotheses are proposed.

Digital platforms/products and post popularity. Events, locations, people, and brands are directly associated with the helpfulness of content and its effectiveness in soliciting likes and shares. On image/video-intensive platforms such as Instagram and Facebook, comments are less likely to impact engagement than likes (Dolan et al., 2019). For news, story, or sponsorship content on Twitter and Blogger.com (Moon et al., 2021), content with images or videos are more likely to gain user likes, comments, or share. By comparison, engagement on Facebook does not follow similar trends. On Facebook and Instagram, brand promotion posts that contain interesting content tend to attract a greater following and more likes (Nilashi et al., 2021). Past studies on company promotion on Facebook found that encouraging users to express their thoughts and feelings, regarding specific content had positive effects and that interesting content had a positive impact on user attitude. Therefore, it was concluded that humor was a significant factor influencing likes. In this study, we discussed the three dimensions proposed by Villamediana (2019): popularity, comment engagement, and virality. Popularity refers to likes, comment engagement refers to comments, and virality refers to shares (Villamediana-Pedrosa and Vila Lopez, 2019). We formulated the following two hypotheses on fan page posts concerning digital platforms/products:

H1a: Frequently using machine-recommended cues can improve the popularity of digital platform posts.

H1b: Frequently using machine-recommended cues can improve the popularity of digital product posts.

Digital platforms/products and comment engagement. A previous study on paid content found that discussion was dramatically lower for paid content than unpaid content (Pletikosa Cvijikj and Michahelles, 2013). Interestingly, the researchers found that paid content received more likes, and users were more willing to like and share their experiences. However, paid content did not affect comments (Dolan et al., 2016). Entertainment content refers to interesting media content. Interesting content significantly influences likes and shares. Therefore, researchers concluded that the purpose of Facebook content is to satisfy user needs and consolidate social interactions and the desire for community benefits (Foroudi et al., 2020). By comparison, brand-centered images and posts do not significantly affect comments. Therefore, we examined the comment engagement of social media posts and formulated the following hypotheses:

H2a: Frequently using machine-recommended cues can improve the comment engagement of digital platform posts.

H2b: Frequently using machine-recommended cues can improve the comment engagement of digital product posts.

Digital platforms/products and virality. Research shows that including rewards and discounts in marketing content diversifies appeal. For example, value perceptions are formed when content contains rewards. Such content takes advantage of low prices to influence user emotion and elicit positive reactions. However, users sometimes feel ashamed of taking advantage of discounts and may form negative perceptions (Dolan et al., 2019). The product specifications, product comments, and product recommendations on Facebook do not significantly affect likes and shares (Pletikosa Cvijikj and Michahelles, 2013). Facebook posts that contain specific product or brand information are more likely to elicit a positive response or comment and less likely to evoke negative sentiment. In this study, we focused on the associations between post cues and shares and formulated the following hypotheses:

H3a: Frequently using machine-recommended cues can improve the virality of digital platform posts.

H3b: Frequently using machine-recommended cues can improve the virality of digital product posts.

Research methodology

Social media data is generally directly extracted from the Web. For example, Twitter, Facebook, and YouTube offer APIs that allow users to define applications to track and collect data. The collected data can be stored in a backend database and filtered to prevent companies from deleting past information (Fig. 1). This study proposes a framework to strengthen social media monitoring and data exploration (Bruns and Stieglitz, 2014) through continuous data capture and analysis on social media goals and behaviors. The content reflects trends and problems, and accordingly, improvement measures are proposed. To ensure the validity of the measurement items and collected data, this study applies Facebook's functional classification to behavioral responses (i.e., like, comment, and share). This study references related research on visual perception to examine public cognitive

and behavioral gratification. In addition, it performs multiple verifications of the results by integrating research theory with computer science, data mining, big data, and ensemble analysis (e.g., AdaBoost, random decision forests, and XGboost).

Sample selection. This study examines the fan pages of Disney, Microsoft, and Netflix as digital commodities. Disney has developed into America's leading entertainment and media company with a wide range of services, including animated films, live film and television production, theme parks, drama, radio, music, publishing and online media, direct-to-consumer platforms, theaters, and production content for cable platforms. Netflix is an over-the-top (OTT) media platform that mainly provides online video streaming services in many countries. Subscribers can use various network devices to connect to Netflix's online content database and to access the platform's original content. Microsoft mainly develops, manufactures, licenses, and provides a wide range of computer software services. The most famous and best-selling products are the Microsoft Windows operating system and Microsoft Office software.

This study also examines the fan pages of two digital platforms, Google and LinkedIn. Google develops and provides a wide range of internet-based products and services such as internet advertising, internet searches, and cloud computing. Google Search is the most widely used internet search engine across many countries. LinkedIn is a social web service that has pioneered the concept of community-enabled online resumes. The platform automatically generates business cards upon user registration. It helps working professionals, entrepreneurs, and businesspeople expand their networks and find suitable or better job opportunities through these networks. It is important to proactively promote a platform through community platforms to understand its key features and introduce the latest features and services to enhance fan interest and use. YouTube an American video-sharing website owned by Google, is currently the world's largest video search and sharing platform, allowing users to upload, watch, share and comment on videos.

Text and data collection. The study aims to conduct an empirical analysis on the Facebook pages of digital product brands and digital platform brands and explore the impact of key information in the posts on public participation. Facebook's terms of service allow app designers to monitor interactions (Pletikosa Cvijikj and Michahelles, 2013). For the purpose of this research, app designers monitor interactions in line with Facebook's terms of service to obtain post data and content. Most information on Facebook fan pages is presented in four formats: text, photos, videos, and links. The behavioral benchmarks to measure online participation are likes, comments, and shares (Wallace et al., 2014). To ensure data accuracy, the study uses Facebook graph API to collect and extract information from daily posts, including post content, type, time, likes, shares, and comments. This information is then saved in the database which is subjected to a classification method. The research sample is composed of six global digital brands, of which three are digital product brands (Disney, Netflix, and Microsoft), and the remaining are digital platform brands (Google, YouTube, and LinkedIn). The fan pages must meet the criteria of continuous posting and must rank among the top ten in either digital industry. Post content, as well as the number and content of public replies to corresponding articles, are extracted for the period between January 1, 2011, and December 31, 2020. This study analyzes a total of 29,343 posts, including 14,028 posts on digital product brands and 15,315 posts on digital platform brands.

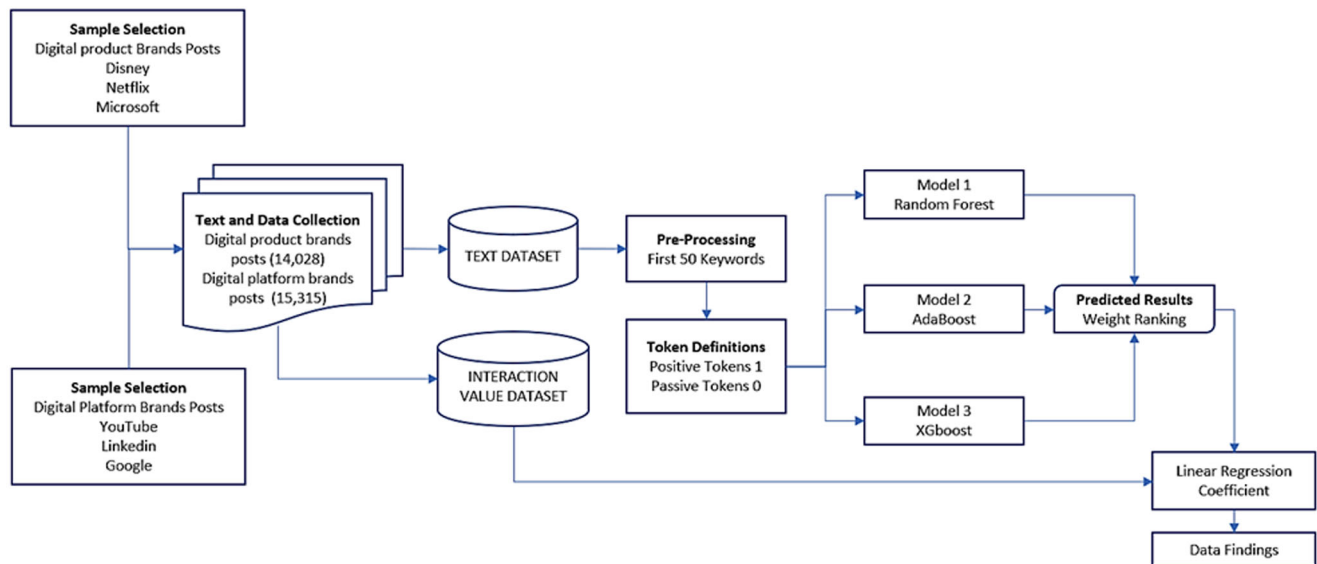


Fig. 1 Data analysis flow chart, this study was conducted in the process and proposed a framework to strengthen social media monitoring and data exploration.

Pre-processing. The user language on Facebook is largely casual, with punctuation errors, typographical errors, abbreviations, and emojis. This study pre-processed the data to reduce such errors and remove all website links, usernames, and garbled symbols. Words often provide information about a specific description focus and type. A semantic analysis employs a standard natural language parser to retrieve words used in certain texts. Semantic similarity measures the distance between words and categories and maps them into numerical vectors (Sánchez et al., 2012). The results are finally verified by researchers with expertise and who do not have a conflict of interest. The calculation of similarities between two items warrants a structured knowledge base of ontology. To this effect, a similarities analysis must account for the length of the path between terms and their prevalence in a hierarchy. Following the pre-processing, this study compared the topic categories and similar words in posts to identify the communication of certain key values and policies (Slimani, 2013). All unrelated words are discarded. The detailed results of the semantic analysis are referenced to discuss select images and key cues. Finally, the integrated model arranges the datasets and verifies the first 50 keywords from the fan page interactive data.

The data collected in this study contained a large amount of unrelated information. In the past, manually sorting through big data was extremely time intensive. Therefore, the research first calculated the compositions of the source topics and then adopted these compositions as the filtering condition to produce focused datasets, reduce data filtering and sorting times, and eliminate confounding or scattered datasets.

Imbalance depends on data size. Using the Weibo posts as an example, although data imbalance was present, the sheer number of samples in each category (each category contained at least 1000 samples) allowed us to overcome this issue easily. After data collection, it extracted the trainable data and then examined the data distribution. A sample size of over 5000 samples in each category denoted that the difference between positive and negative samples was within range, likely reducing data imbalance.

Token definitions. Data pre-processing is a critical step in ensuring classifier performance. Preprocessing enhances data quality and eliminates unnecessary information. For example,

tokenization, or text segmentation, or lexical analysis, is used to split long strings of text into smaller sentences, words, or characters. In this study, tokenization was applied to split text data into words. Sentences were refined, and stop words, special symbols, URLs, and irrelevant content were excluded. In the model, all three algorithms were used for text classification, and the number of interactions was used to represent active engagement (higher than average interaction) or passive engagement (lower than average interaction). Only official posts are discussed in this study. Official posts are reviewed carefully before posting. Therefore, these posts generally contain fewer grammatical, formatting, abbreviation, and symbol errors than casual posts.

We collected English posts associated with specific keywords. All posts retrieved from Facebook were pre-processed. During pre-processing, all stop words (this, the, at, on, he, etc.) and special symbols (% , @ , * , # , etc.) were eliminated. Word repetitions and quotes were also eliminated. The pre-processed data were then transferred to the classifiers. All three classifiers used trained datasets consisting of words and tokens. Within the datasets, positive tokens were set as 1, and passive tokens were set as 0. The three algorithms were Random Forest, XGboost, and AdaBoost. Three algorithms were used to avoid overfitting. The trained datasets were then split into smaller subsets and entered into the models to produce the aggregated results. The algorithms were specifically chosen based on their effects on the different datasets.

Ensemble learning Classification Calculus. The encoded text could be trained/tested through machine learning, and the model was developed based on a cross-validation/assessment framework. The experiment settings primarily included screened text from widely used machine learning algorithms. First, AdaBoost, Random Forest, and XGboost were selected. Then, the data were cross-validated to test the accuracy of the training outcomes. Third, we generated the ensemble algorithm and conducted a second cross-validation. All classifications and comparisons were carried out based on evaluation and validation metrics. True positives (TP) and true negatives (TN) corresponded to correct predictions, while false negatives (FN) and false positives (FP) corresponded to incorrect

predictions. Finally, the statistical validation indicator (K_p) was calculated. In this study, Random Forest, AdaBoost, and XGboost served as the classifiers in the ensemble learning system to classify interactions into active and passive behaviors.

In terms of accuracy, the ideal method is individually cross-validating pairs of algorithms, adjusting the parameters to ensure optimal solutions are returned, and then selecting the most favorable one. In machine learning, a fundamental principle is that no single algorithm can solve all problems perfectly (particularly supervised learning and predictive modeling). Therefore, different algorithms are selected depending on the specific problem when evaluating algorithmic performance with given test sets.

Random forest (RF) is similar to AdaBoost. RF is an alternative ensemble classifier that combines different decision tree models. RF contains a collection of unrelated individual decision trees, where each dataset transfer samples to all trees to predict class. The algorithm then selects the sample with the most votes as the prediction result. Random forest is a continuation algorithm for decision trees, in which multiple trees are constructed using the optimal segmentation of random subsets and the function of splitting nodes on each node and tree. This reduces the risk of overfitting. The final classification takes the form of various breakdowns produced by the tree. Random forest can achieve high accuracy in imbalanced data between two datasets; these imbalances may be observed in positive or negative data. Data imbalances affect the classification of results. Oversampling or under-sampling can be applied to balance the amount of positive and negative data (Cahyana et al., 2019). Randomness in redundant trees does not originate from bootstrap data but from the random separation of observations, which results in the multiple classifications of trees.

The research chose the random forest algorithm for its performance and advantages over other algorithms. Random forests can process high-order data (data containing many features) without running feature selection. After data training, random forests can identify important features, and they can also train data rapidly and produce independent trees. During training, random forests can also detect the influence between features. For imbalanced datasets, random forests can balance errors. It can provide an effective means to balance dataset errors when classification imbalance is detected. It can also maintain a high level of accuracy when features are missing. In addition, random forest is highly resistant to interference and can reduce the likeliness of overfitting when large portions of data are missing.

AdaBoost, or Adaptive Boost, is an ensemble classifier. It is an iterative ensemble method that consolidates several internal weak classifiers to enhance performance. The classifiers within the ensemble are added one at a time, and each classifier is trained using the data that the previous classifier could not correctly classify. In other words, AdaBoost incorporates a new training set of data for its learning model based on the previous training results. The prediction model of the gradient boosting classifier (GBC) is generally obtained by sequentially fitting the base learner to current pseudo-residuals. Therefore, the XGboost evaluates the model values of every training sample and gradually calculates the gradient of the minimized loss function in the current step.

The advantage of AdaBoost is the ability to cascade weak classifiers and set different classification algorithms as weak classifiers. AdaBoost is more accurate than bagging or random forest, and it sufficiently accounts for all classifier weights and utilizes binary classification and multi-classification scenarios. One of the reasons for choosing Adaboost as a classifier is its high

classification accuracy. Under the framework of Adaboost, various regression classification models can be adapted to flexibly construct weak learners without filtering features. As a binary classifier, Adaboost is simple to construct, and overfitting seldom occurs.

Gradient Boost is also an iterative forward distribution algorithm. However, its weak learner restricts the algorithm to CART regression tree models. Gradient Boost accumulates the results of all trees, which cannot be completed through classification. Therefore, GBDT trees are all CART regression trees rather than classification trees. Each calculation aims to reduce the residual of the previous calculation. To eliminate residuals, the model can be established in the gradient direction. Therefore, in Gradient Boost, the establishment of each new model is to reduce the previous residual in the gradient direction, which is different from other conventional boosting algorithms that largely focus on the correctness of sample weights. Hence, Gradient Boost can be used to process the negative gradient of loss functions to resolve both classification and regression problems.

XGboost has been optimized for model accuracy and computation speed. Therefore, deep learning typically always prevails in Kaggle competitions related to unstructured data, such as audio and video data. By comparison, XGboost is the obvious winner in competitions related to structured data, which highlights the power of XGboost. XGboost can include the complexity of tree models into regularization to prevent overfitting. Therefore, it outperforms GBDT in extrapolation. GBDT only supports CART as the base learner, while XGBoost supports CART and linear classifiers. XGBoost can also perform subsampling rather than training all data features to avoid overfitting, reduce computation time, and compensate for missing values in sparse datasets, much like random trees.

Data analyses and results

Reliability and validity. Principal component analysis is primarily based on exploratory factor analyses (EFA), which do not presume factor relationships and measures. Rather, EFA is predominantly an outcome-based method. Therefore, the first step was to evaluate the reliability and validity of the keywords. For reliability and validity analysis of the data, principal component factor analysis was performed to test the factor validity of the scale.

The KMO value measures the correctness of the features (keywords). A high validity denotes a stronger reflection of the features in the research results. The factor characteristic value of Digital product brands had a total variance of 80.292% and a KMO value of 0.849. The research adopted Cronbach's α coefficient as the measure of internal reliability to determine whether the items (keywords) exhibited a high level of internal consistency. The factor characteristic value of Digital platform brands had a total variance of 77.545% and a KMO value of 0.884. The expected load factor for all items is >0.5 , indicating good convergence and discriminant validity. In addition, the reliability test produced a Cronbach's alpha of 0.883 for Digital product brands and 0.881 for Digital platform brands. Each of these results shows good reliability.

Hypotheses testing. The results partially support H1, that is, digital product brands actively use image cues in the information packaging for their fan pages, which impacts fan behavior (i.e., likes, comments, and shares). The findings of the random decision forests and XGboost highlight that image cues are significant in predicting comments (Fig. 2).

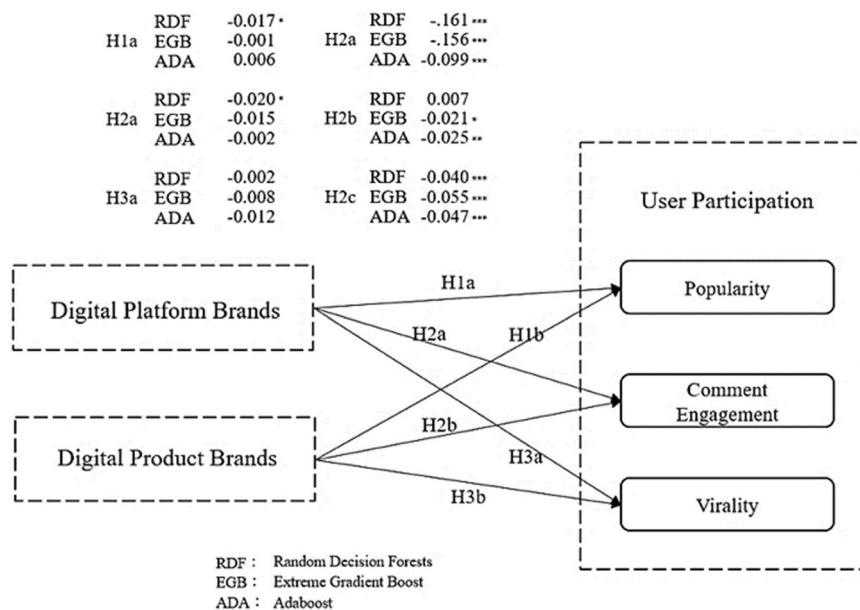


Fig. 2 Model results, in this study, all the hypotheses were validated except H3a.

Table 2 Summary of hypothesis.

ID	Hypothesis	Verdict
H1a	Frequently using machine-recommended cues can improve the popularity of digital platform posts.	Partially established
H2a	Frequently using machine-recommended cues can improve the popularity of digital product posts.	Established
H1b	Frequently using machine-recommended cues can improve the comment engagement of digital platform posts.	Partially established
H2b	Frequently using machine-recommended cues can improve the comment engagement of digital product posts.	Partially established
H3a	Frequently using machine-recommended cues can improve the virality of digital platform posts.	Not established
H3b	Frequently using machine-recommended cues can improve the virality of digital product posts.	Established

The results partially support H2, that is, digital platform brands actively use image cues in the information packaging for their fan pages, which impacts fan behavior (i.e., likes, comments, and shares). The findings of the random decision forests highlight that image cues are significant in predicting comments and shares, and those of AdaBoost and XGboost report significant results for interactions in the form of likes, comments, and shares. Accordingly, we propose the following hypothesis (Table 2).

Data verification. The analyses reveal that image cues in the social media content of digital brands significantly affect public behaviors and responses. In particular, image cues in content by digital product brands have a significant impact on public behavior. First, the hypothesis that information among digital product brands is supported, the behavior of users (Likes) was found to have a significant impact ($R = 0.017$, F Change = 4.423, $\beta = 0.017$, $t = -2.103$, $p = 0.035 < 0.05$) on Random Decision Forests made by users. Comments were found to have a significant impact ($R = 0.02$, F Change = 6.206, $\beta = 0.02$, $t = -2.491$, $p = 0.013 < 0.05$) on Random Decision Forests (Table 3).

Next, the hypothesis that information among digital platform brands is supported, the behavior of users (Likes) was found to have a significant impact ($R = 0.161$, F Change = 372.927, $\beta = 0.161$, $t = -19.311$, $p = 0.000 < 0.05$) on Random Decision Forests, ($R = 0.156$, F Change = 348.462, $\beta = 0.156$, $t = -18.667$, $p = 0.000 < 0.05$) on Extreme Gradient Boost and ($R = 0.099$, F Change = 138.411, $\beta = 0.099$, $t = -11.765$, $p = 0.000 < 0.05$) on Adaboost made by users. Comments was found to have a significant impact ($R = 0.021$, F Change = 60.038, $\beta = 0.021$, $t = -2.457$, $p = 0.014 < 0.05$) on Extreme Gradient Boost and

($R = 0.025$, F Change = 8.641, $\beta = 0.025$, $t = -2.94$, $p = 0.003 < 0.05$) on Adaboost made by users. Shares was found to have a significant impact ($R = 0.055$, F Change = 42.639, $\beta = 0.055$, $t = -6.53$, $p = 0.000 < 0.05$) on Random Decision Forests ($\beta = 0.055$, $p < 0.000$) on Extreme Gradient Boost and ($R = 0.047$, F Change = 31.517, $\beta = 0.047$, $t = -5.614$, $p = 0.000 < 0.05$) on Adaboost (Table 4).

Data findings. Drawing on the findings of the ensemble analyses, this study recommends image cues and integrates those with a significant impact.

Disney. The key cues that influence public behavior in the form of likes are little, Disney, http, theatre, and star according to the random decision forests; Disney, new, http, theatre, and now as per the XGboost; and day, Pixar, Disneyland, story, and history as highlighted by AdaBoost. In terms of shares, XGboost lists Disney, trailer, now, new, and http as key cues and AdaBoost reports Disneyland, Pixar, story, history, and day (Figs. 3–6, Tables A1–A6, <https://doi.org/10.7910/DVN/AU6TNV>).

Disney, trailer, now, and new are cues common to both random decision forests and XGboost. In other words, brands should incorporate information that captures public interest, such as updates on the latest movie, and that is easy to share with friends. The findings of AdaBoost (i.e., Disneyland, Pixar, story, and history) indicate public interest in amusement parks. Using these cues will help enhance the public's continued impression of and interest in the brand's physical products.

Table 3 Linear regression coefficient of determination and beta (digital product brands).

Digital product brands		R	R ²	ΔF	Durbin-Watson	B	Standard error	β	t	Sig.
Popularity (Likes)	Random Decision Forests	0.017	0.000	4.423	1.840	-113.768	540.096	0.017	-2.103	0.035
	Extreme Gradient Boost	0.001	0.000	0.032	1.839	-9.217	51.245	0.001	-0.180	0.857
	Adaboost	0.006	0.000	0.537	1.840	50.342	68.703	0.006	0.733	0.464
Comment Engagement (Comments)	Random Decision Forests	0.020	0.000	6.206	1.777	-14.423	5.790	0.020	-2.491	0.013
	Extreme Gradient Boost	0.015	0.000	3.595	1.777	-9.432	4.974	0.015	-1.896	0.058
	Adaboost	0.002	0.000	0.050	1.777	-1.408	6.282	0.002	-0.224	0.823
Virality (Shares)	Random Decision Forests	0.002	0.000	0.071	1.986	-8.424	31.545	0.002	-0.267	0.789
	Extreme Gradient Boost	0.008	0.000	10.093	1.986	-28.933	27.675	0.008	-10.045	0.296
	Adaboost	0.012	0.000	2.334	1.986	-55.134	360.089	0.012	-1.528	0.127

Table 4 Linear regression coefficient of determination and beta (digital platform brands).

Digital platform brands		R	R ²	ΔF	Durbin-Watson	B	Standard error	β	t	Sig.
Popularity (Likes)	Random Decision Forests	0.161	0.026	372.927	0.729	-3138.424	162.517	0.161	-19.311	0.000
	Extreme Gradient Boost	0.156	0.024	348.462	0.730	-2727.105	1460.091	0.156	-18.667	0.000
	Adaboost	0.099	0.010	138.411	0.714	-2076.970	176.541	0.099	-11.765	0.000
Comment Engagement (Comments)	Random Decision Forests	0.007	0.000	0.644	1.891	14.511	180.087	0.007	0.802	0.422
	Extreme Gradient Boost	0.021	0.000	60.038	1.891	-28.460	11.582	0.021	-2.457	0.014
	Adaboost	0.025	0.001	8.641	1.893	-40.995	13.946	0.025	-2.940	0.003
Virality (Shares)	Random Decision Forests	0.040	0.002	22.791	1.846	-306.851	64.276	0.040	-4.774	0.000
	Extreme Gradient Boost	0.055	0.003	42.639	1.846	-306.659	46.963	0.055	-6.530	0.000
	Adaboost	0.047	0.002	31.517	1.845	-3110.042	55.405	0.047	-5.614	0.000

Netflix. The key cues provided by Random Decision Forests for Likes included: Netflix has it, here, today, wants and met; by Extreme Gradient Boost for Likes included Netflix, Netflix has it, available, neither and cover; the cues for Comments included: Netflix, Netflix has it, watch, arrives and December; the cues for Shares included: Netflix has it, Netflix, cover, available and watch. The key cues provided by Adaboost for Likes included photo, cover, http, neither and March; the cues for Shares included: photo, http, October, neither and title (Fig. 7).

The results of the random decision forests and XGboost highlight Netflix has it, here, and today as key cues, which implies that most fans prefer real-time and new information. AdaBoost lists photo, http, neither, and October as image cues. In other words, to stimulate public willingness to share information, the brand should incorporate real-time features such as original works for the current week or month or a representative season of a series.

Microsoft. The key cues provided by Extreme Gradient Boost for Shares included http, Windows, Lumia, new and more; by Adaboost for Likes included party, RSVP, Store, zoom and music; the cues for Comments included: RSVP, party, Store, live and capture; the cues for Shares included: RSVP, party, live, Store and smartphone (Fig. 8).

According to XGboost, Windows, Lumia, new, and camera are key cues, indicating that fans are more willing to share exclusive brand information. The AdaBoost results (RSVP, store, party, live, and capture) suggest that proactively providing functional or practical information that meets consumer demands will increase public willingness to share.

YouTube. The key cues provided by Random Decision Forests for Comments included: year, MV, video, http and cover; the cues for Shares included: http, cover, MV, video and day; by Extreme Gradient Boost for Shares included http, YouTube, MV, video and day. The key cues provided by Adaboost for Comments included know, music, dance, best and look; the cues for Shares included: know, dance, look, music and only (Fig. 9).

Both the random decision forests and XGboost list MV, video, http, and YouTube as key cues. In particular, MV and video align with public demands and interest and stimulate consumer participation. The key cues, according to Adaboost, are dance, music, and best, indicating that keywords that match ethnic or cultural attributes of a region motivate sharing behaviors. In India, for example, dance, music, and Kapoor successfully stimulate public willingness to share.

LinkedIn. The key cues provided by Random Decision Forests for Comments included: headlines, most, week, http and

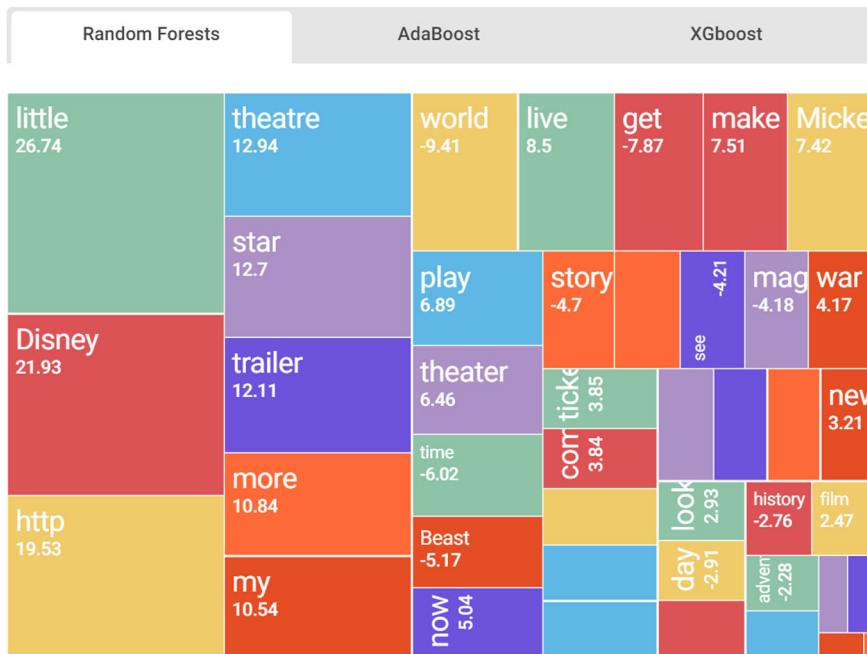


Fig. 3 Disney cues high weighting of random decision forests, the key cues that influence public behavior in the form of likes are little, Disney, http, theatre, and star according to the random decision forests.

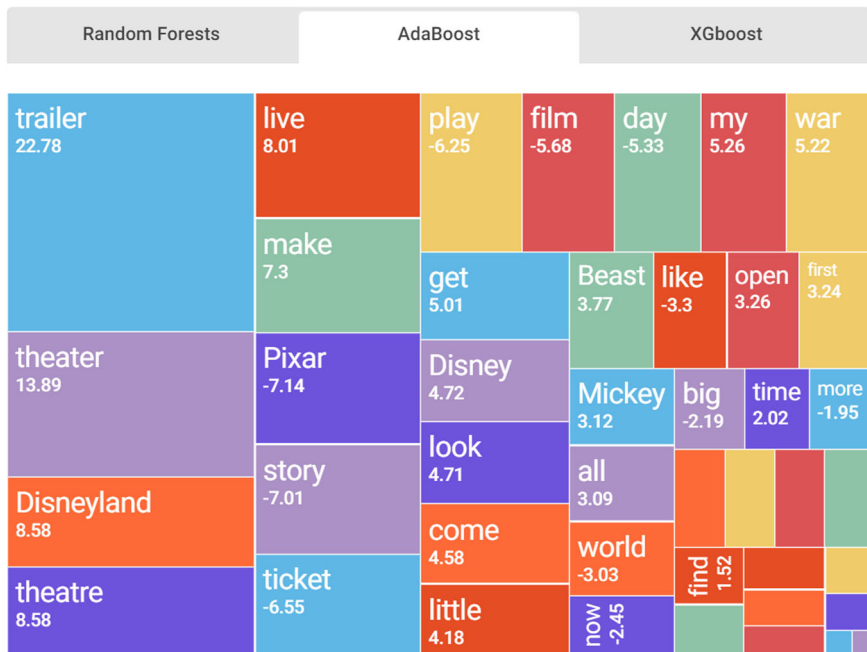


Fig. 4 Disney cues high weighting of Adaboost, in the form of likes are day, Pixar, Disneyland, story, and history as highlighted by AdaBoost.

opportunity; by Extreme Gradient Boost for Likes included headlines, http, LinkedIn, work and career; the cues for Comments included: headlines, http, LinkedIn, work and job; the cues for Shares included: headlines, LinkedIn, work, new and here (Fig. 10).

The random decision forests highlight that most, week, and opportunity stimulate interactions in the form of likes, comments, and shares. XGboost lists headlines, http, LinkedIn, and work. Thus, LinkedIn should consider information and keywords (e.g., opportunity and work) that directly satisfy

consumer interests and demands to stimulate participation and interaction behaviors in the form of comments and shares.

Google. Google is the only digital platform brand that did not report significant results (Fig. 11).

Conclusions

Research results. This research integrated the prediction results from the previous section and further discovered that the studied digital brands have three major characteristics: notification and

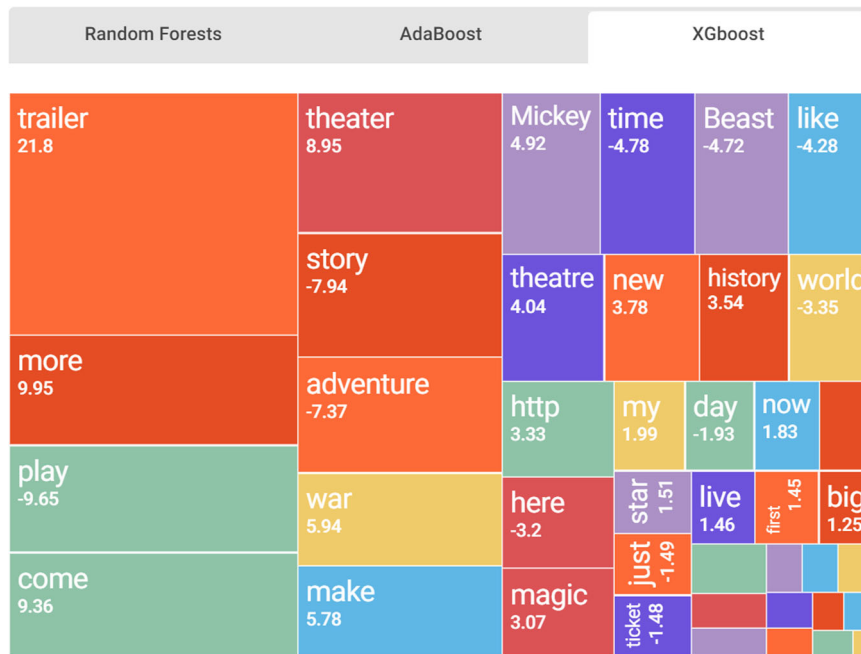


Fig. 5 Disney cues high weighting of XGboost, in the form of likes are Disney, new, http, theatre, and now as per the XGboost.

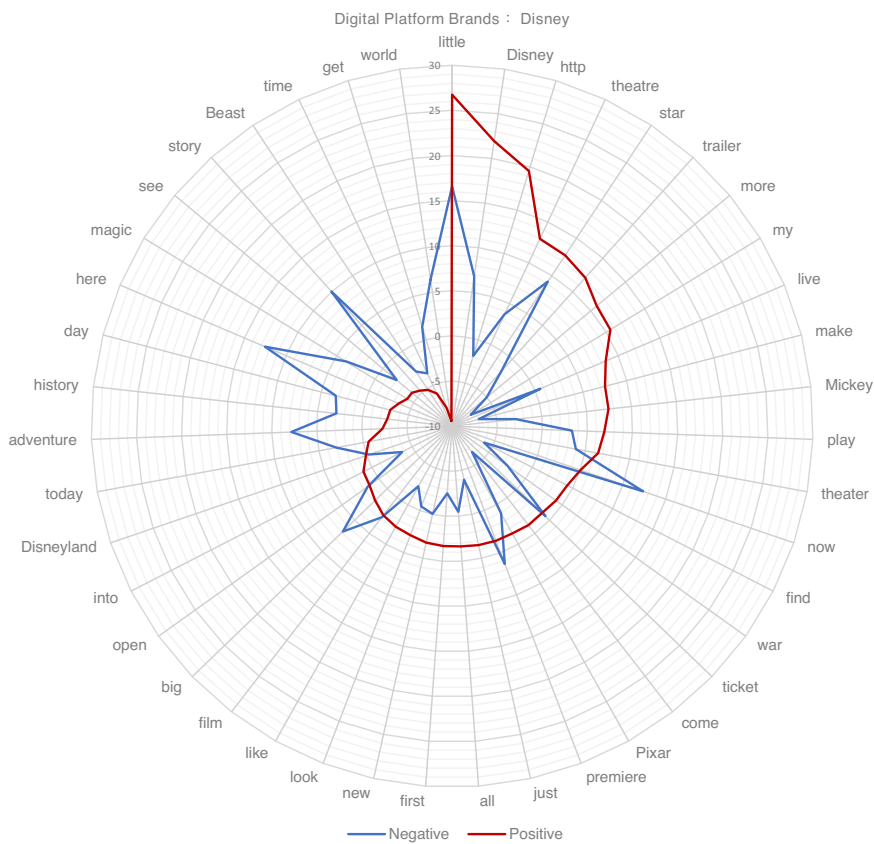


Fig. 6 Cues of Disney, using ensemble learning, this study obtained positive and negative clues from the posts of the “Disney” fan page.

diversion module; interaction and diversion module; and notification, interaction, and diversion module.

Notification and diversion module. The module uses the two features of notification and diversion and combines the

interactive behaviors of likes and shares. Disney’s brand page best exemplifies the notification and diversion module. The information motivation under this module stems from the demand for the latest products as well as discounts and offers. Such brands attract user attention by offering benefits. While this approach

Digital Platform Brands : Netflix

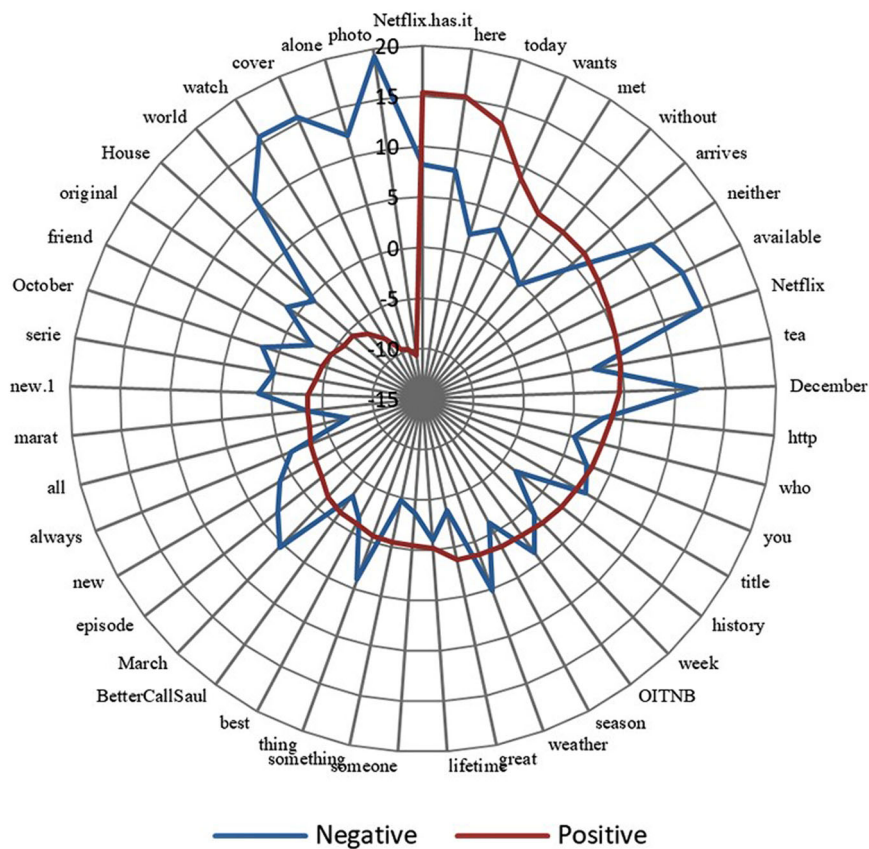


Fig. 7 Cues of Netflix, positive and negative clues in the postings on the “Netflix” fan page.

differs from maintaining consumer relationships, it generates stronger interactive relationships by enabling brands to satisfy information demands. Brands adopting the notification and diversion module can keep consumers engaged by offering information on new products, announcing offers and discounts, and using interactive content.

Interaction and diversion module. The module promotes the effects of interaction and diversion and combines the interaction behaviors of comments and shares. The brand page of the digital platform YouTube is representative of this operational feature. The module aims to encourage users to follow a brand on social networking sites (SNS) and engage in self-expression through interactive features. Brand interactions on SNS encourage the public to express themselves and share personal information with other potential consumers through comments and shares.

Notification, interaction, and diversion module. The module highlights notification, interaction, and diversion as the three main objectives of a brand page and combines the three interactive behaviors of likes, comments, and shares. The brand pages of digital product brands Netflix and Microsoft and the digital platform brand LinkedIn clearly reflect the operating characteristics of this module. The module highlights a clear difference in interactions between image and marketing information. Thus, it is necessary to promote consistent brand awareness using marketing approaches that reflect the brand.

This helps consumers identify brand information, and brands enhance participation frequency and level. Netflix, Microsoft, and LinkedIn adopt various marketing strategies such as combining brand prompts and converting them into interesting brand experiences. Using different types of images increases brand participation and strengthens brand reputation, thereby emphasizing the need to continuously maintain a positive dialog space and public trust in brand information.

Research discussion. First, this study uses a simple and easily comprehensible behavioral framework to explore the relationship between information on brands’ fan pages and public behavior. While existing research discusses public behavior, the present analysis identifies and compares key cues and interactive behaviors (i.e., likes, comments, and shares) using machine learning techniques. It further examines the key information benefits of competing for digital brands considering the dynamic nature of social media.

Recent studies found that rational content affects likes, but it does not affect active engagement (Dolan et al., 2019). Some argue that the appeal of reasonable content is less effective in prompting likes than the emotional appeal in user engagement. Subsequently, this argument contradicts mainstream assertions. In terms of the impact of brand content on social media users, a previous study found that entertainment content affected likes but failed to affect comments. For example, many brands on Facebook post humorous, interesting, and artistic content (entertainment) to solicit user likes. These posts consequently reduce the product and content value of serious brands (Tafesse, 2020). Previous studies

Digital Platform Brands : Microsoft

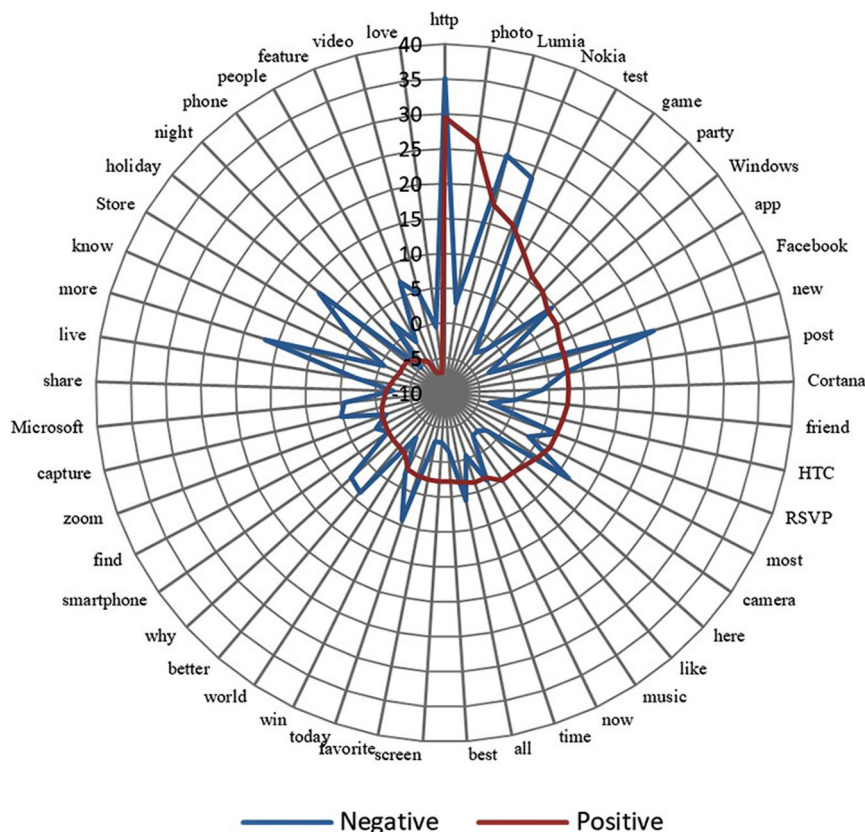


Fig. 8 Cues of Microsoft, positive and negative clues in the postings on the “Microsoft” fan page.

also found that emotional content effectively prompted user likes, suggesting that such content can generate more likes for services. These findings highlight the importance of moderating context to solicit engagement (Brown et al., 2019).

Marketers must take advantage of social media campaigns to promote digital products (Swani et al., 2017), such as promotional brand content, giveaways, lucky draws, or cash incentives. They can also provide bonuses related to work or daily life. Previous studies found that content with rewards or prizes significantly and positively impacted likes, shares, and comments on Facebook. Some found that posts containing quizzes promoted user engagement and positively impacted likes and comments (Nilashi et al., 2021). A previous study also found that brand or product advertisements on Facebook that contained persuasive or visually appealing information, such as new product/service announcements, online coupons, discounts, and contests or raffles, were more likely to receive likes, comments, and shares (Luo et al., 2021). By comparison, content that only contained product specifications, reviews, and suggestions could not achieve the expected likes and shares. Facebook content that simply encouraged consumption was unable to solicit likes. Moreover, although raffle posts gained negative likes, they gained positive comments (Foroudi et al., 2020).

Second, this study integrates social content exploration and artificial intelligence data analysis to re-examine public demands and interaction characteristics in the context of digital product and digital platform brands. In doing so, it highlights public motivation to engage with fan pages, information demands, and interactive characteristics. The content interaction model in this study can be

applied to various types of brand pages. The use of image cues to conduct information analyses can effectively enhance brand value and future content publishing (Gensler, 2013).

The relationship between people and brands has generated immense research interest in the field of marketing (Bowden et al., 2017), whether it is discussing different ways of defining and measuring engagement in marketing campaigns (Zheng et al., 2021) or examining the structures of positive/negative brand engagement (Villamediana-Pedrosa and Vila Lopez, 2019). Therefore, for digital goods or digital platforms, marketers must not only examine message content but also think of ways to adjust active/passive engagement structures. When measuring social network communities, active participation is considered a multidimensional construct that can uncover positive brand value. Researchers can determine the propaganda strength of different posts by examining likes, comments, and shares.

Third, the amount of data on the Internet will only continue to grow. Big data analysis is an inevitable trend. Therefore, effectively processing and analyzing useful data is a popular research topic. Manually reviewing data is extremely ineffective. Therefore, it is imperative that the research developed new tools to help researchers quickly and systematically sort through big data to identify themes for subsequent analysis. Broadly speaking, text mining is the unitization of text by keywords, key phrases, or concepts. The data can then be sorted, categorized, and used in predictive models to extract valuable information.

Text mining has expanded beyond sorting and classifying documents as information overload continues to gain attention, and more researchers are developing advanced text mining

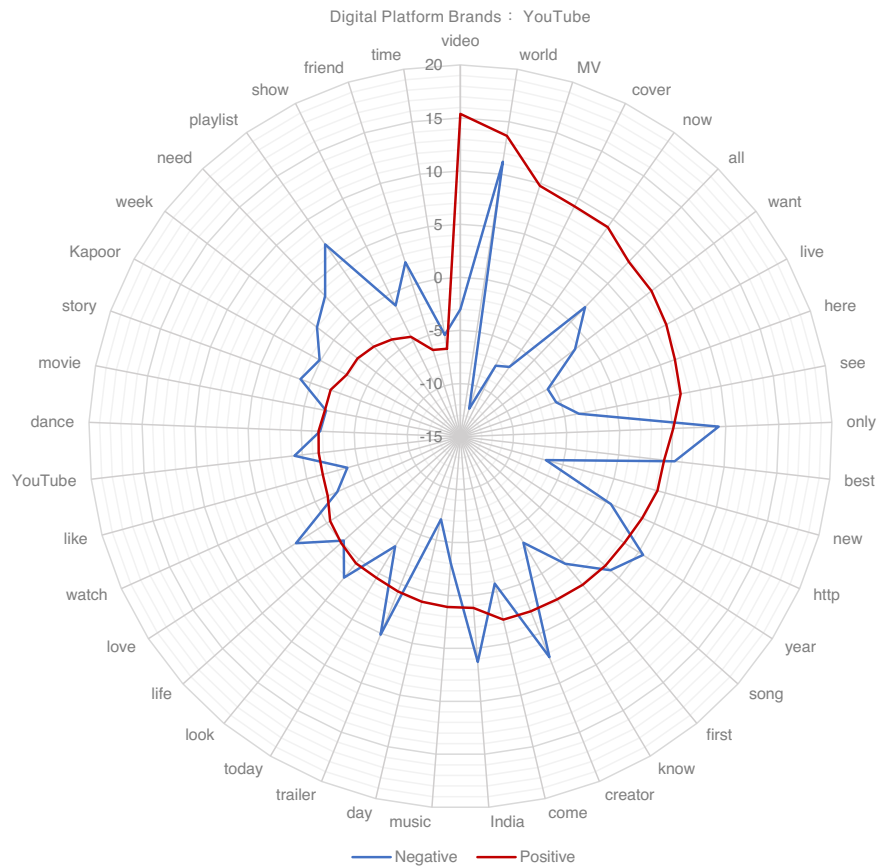


Fig. 9 Cues of YouTube, positive and negative clues in the postings on the “ YouTube” fan page.

algorithms and computational models for data extraction. The keywords lifted from the text examined in this study included nouns, verbs, adjectives, and adverbs. This research presented these keywords visually to enhance the presentation of the model results and facilitate interpretation.

In addition, the researchers observed the source composition of the keywords to determine whether the keywords centered on specific topics. It also cross-referenced each brand with its most relevant keywords and calculated the number of times each keyword appeared to elucidate the proportion and tendency of the keywords from each source. The research then manually assessed the contribution that the high-weighted keywords had to the brand posts. For example, the keyword “arrives” was prevalent in posts that announced the launch of popular or new products or services. It also linked special months, such as “December,” with anniversary sales. These keywords can be combined with intricate designs to create attractive event posts. Although only keywords were examined in this study, manually cross-referencing keywords with post content allowed us to assess the attractiveness of the keywords.

Research implications. First, this study performs data analysis to understand characteristics attracting public attention on social media. Information that is easily understood helps companies take immediate action and improve their competitive advantage. Page admins can examine social media data and integrate the results to improve the quality of products and service planning (Xu, 2015). Intensifying market competition has accelerated the demand for social media, and most companies use related platforms to regularly broadcast information and engage users

through various types of content. In a homogenous business environment, it is critical for managers to pay increasing attention to information obtained from social media content to enhance precise marketing in the community (Dey et al., 2011). The constant monitoring will provide brands with the insights necessary to improve public participation and marketing strategies that align with social trends and public demand. Unfortunately, numerous companies are unfamiliar with social media analysis and, particularly, social content analyses related to competing brands (Dai et al., 2011). The proposed research framework, including purposeful behavior and topic analyses, can be used to accurately convert social media data into a decision-making reference for brands.

Second, this study integrates community content exploration techniques with artificial intelligence data analysis using Facebook’s resources to understand public information demands and identify unstructured elements hidden in the information. The rapid growth of social networks has led to an exponential rise in unstructured data and the involvement of various organizations (i.e., government, enterprise, and non-government). Therefore, scholars must cautiously select representative variables to examine different issues (Injadat et al., 2016). Researchers can easily obtain detailed statistical data since Facebook’s community content and archive data is publicly available (Back et al., 2010). The original data can be transformed into knowledge that is useful for data management and market analysis. However, social media content mining is a time-consuming and labor-intensive task; therefore, it is imperative to fully utilize data mining and text mining technologies to meet enterprise demands and avoid resource wastage (Barbier and Liu, 2011). Artificial intelligence data analysis, a critical technical support tool especially for

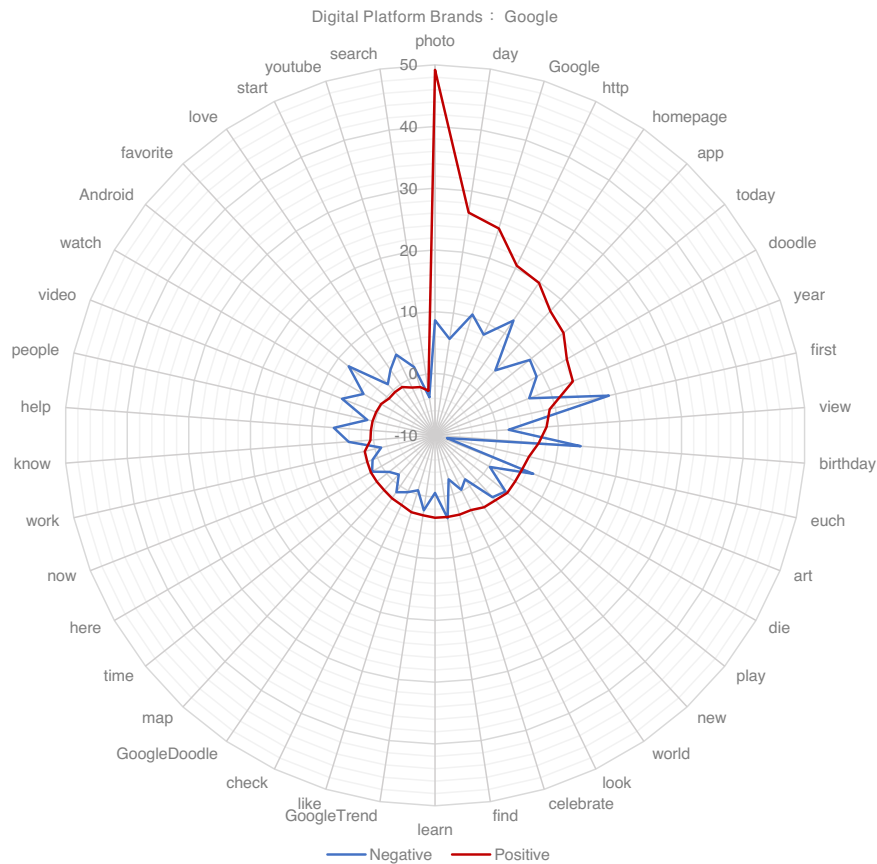


Fig. 11 Cues of Google, positive and negative clues in the postings on the “Google” fan page.

and interests. This can contribute to enhancing the diversity of community content.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Received: 31 July 2022; Accepted: 30 January 2023;

Published online: 14 February 2023

References

- Ahmed A-K, Weatherburn P, Reid D, Hickson F, Torres-Rueda S, Steinberg P, Bourne A (2016) Social norms related to combining drugs and sex (“chemsex”) among gay men in South London. *Int J Drug Policy* 38:29–35
- Azer J, Alexander MJ (2018) Conceptualizing negatively valenced influencing behavior: forms and triggers. *J Serv Manag* 29:468–490
- Back MD, Stopfer JM, Vazire S, Gaddis S, Schmukle SC, Egloff B, Gosling SD (2010) Facebook profiles reflect actual personality, not self-idealization. *Psychol Sci* 21:372–374
- Bag S, Pretorius JHC, Gupta S, Dwivedi YK (2021) Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities. *Technol Forecast Soc Change* 163:120420
- Barbier G, Liu H (2011) Data mining in social media. *Social network data analytics* 327–352
- Baron SD, Brouwer C, Garbayo A (2014) A model for delivering branding value through high-impact digital advertising: how high-impact digital media created a stronger connection to Kellogg’s special K. *J Advert Res* 54:286–291
- Beckers SFM, Doorn JV, Verhoef PC (2017) Good, better, engaged? The effect of company-initiated customer engagement behavior on shareholder value. *J Acad Mark Sci* 46:366–383
- Bowden JLH, Conduit J, Hollebeek LD, Luoma-Aho V, Solem BA (2017) Engagement valence duality and spillover effects in online brand communities. *J Serv Theory Pract* 27:877–897
- Breidbach CF, Brodie RJ (2017) Engagement platforms in the sharing economy: conceptual foundations and research directions. *J Serv Theory Pract* 27:761–777
- Brodie RJ, Ilic A, Juric B, Hollebeek L (2013) Consumer engagement in a virtual brand community: an exploratory analysis. *Journal of Business Research* 66:105–114
- Brown B, Gude WT, Blakeman T, van der Veer SN, Ivers N, Francis JJ, ..., Daker-White G (2019) Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. *Implement Sci* 14(1):1–25
- Bruns A, Steglitz S (2014) Metrics for understanding communication on Twitter. *Twitter Soc* 69–82
- Cahyana N, Khomsah S, Aribowo AS (2019) Improving imbalanced dataset classification using oversampling and gradient boosting. In *2019 5th International Conference on Science in Information Technology (ICSITech)* 217–222. IEEE
- Chakraborty G, Krishna M (2014) Analysis of unstructured data: Applications of text analytics and sentiment mining. In *SAS global forum 1288–2014*
- Coelho RLF, de Oliveira DS, de Almeida MIS (2016) Does social media matter for post typology? Impact of post content on Facebook and Instagram metrics. *Online Inf Rev* 40:458–471
- Dai Y, Kakkonen T, Sutinen E (2011) MinEDec: a decision-support model that combines text-mining technologies with two competitive intelligence analysis methods. *Int J Comput Inf Syst Ind Manag Appl* 3:165–173
- Dessart L, Veloutsou C, Morgan-Thomas A (2015) Consumer engagement in online brand communities: a social media perspective. *J Product Brand Manag* 24:28–42
- Dey L, Haque SM, Khurdiya A, Shroff G (2011) Acquiring competitive intelligence from social media. In *Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data* 1–9

- Dolan R, Conduit J, Fahy J, Goodman S (2016) Social media engagement behaviour: a uses and gratifications perspective. *J Strateg Mark* 24:261–277
- Dolan R, Conduit J, Frethey-Bentham C, Fahy J, Goodman S (2019) Social media engagement behavior: a framework for engaging customers through social media content. *Eur J Mark* 53:2213–2243
- Elliott H, Wright T (2018) Canadian student leaders' conceptualizations of sustainability and sustainable universities. *J Educ Sustain Dev* 12:103–119
- Foroudi P, Nazarian A, Ziyadin S, Kitchen P, Hafeez K, Priporas C, Pantano E (2020) Co-creating brand image and reputation through stakeholder's social network. *J Bus Res* 114:42–59
- Fradkin A, Grewal E, Holtz D (2018) The determinants of online review informativeness: Evidence from field experiments on Airbnb. *SSRN Electron J* 41:1–12
- García-Umaña A, Tirado-Morueta R (2018) Comportamiento mediático digital de estudiantes escolares: uso abusivo de Internet. *J New Approaches Educ Res* 7(2):152–159
- Gavilanes JM, Flatten TC, Brettel M (2018) Content strategies for digital consumer engagement in social networks: why advertising is an antecedent of engagement. *J Advert* 47:4–23
- Gensler SS, Rosenthal LH (2013) Measuring the Quality of Judging: It All Adds Up to One. *New Eng L Rev* 48:475
- Giannakos MN, Pappas IO, Mikalef P (2014) Absolute price as a determinant of perceived service quality in hotels: a qualitative analysis of online customer reviews. *Int J Hosp Event Manag* 1:62–80
- Gálvez Rodríguez MM, Caba PeRez MC, Godoy ML (2012) Determining factors in online transparency of NGOs: a Spanish case study. *Voluntas* 23:661–683
- He W, Zha S, Li L (2013) Social media competitive analysis and text mining: a case study in the pizza industry. *Int J Inf Manag* 33:464–472
- Hollebeek LD, Glynn MS, Brodie RJ (2014) Consumer brand engagement in social media: conceptualization, scale development and validation. *J Interact Mark* 28:149–165
- Injadat M, Salo F, Nassif AB (2016) Data mining techniques in social media: a survey. *Neurocomputing* 214:654–670
- Kacholia V (2013) News feed FYI: showing more high quality content. Facebook.com
- Kaiser C, Schlick S, Bodendorf F (2011) Warning system for online market research—identifying critical situations in online opinion formation. *Knowl-Based Syst* 24:824–836
- Kannan E, Kothamasu LA (2020) A pattern based approach for sentiment analysis using ternary classification on Twitter data. *Int J Emerg Technol* 11:811–816
- Ksiazek TB, Peer L, Lessard K (2016) User engagement with online news: conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media Soc* 18:502–520
- Kumar A, Bezawada R, Rishika R, Janakiraman R, Kannan PK (2016) From social to sale: the effects of firm-generated content in social media on customer behavior. *J Mark* 80:7–25
- Lee D, Ng PML, Bogomolova S (2020) The impact of university brand identification and eWOM behaviour on students' psychological well-being: a multi-group analysis among active and passive social media users. *J Mark Manag* 36:384–403
- Lipsman A, Mudd G, Rich M, Bruich S (2012) The power of “like”: how brands reach (and influence) fans through social-media marketing. *J Advert Res* 52:40–52
- Luo L, Duan S, Shang S, Pan Y (2021) What makes a helpful online review? Empirical evidence on the effects of review and reviewer characteristics. *Online Inf Rev* 45:614–632
- Maier C, Laumer S, Eckhardt A, Weitzel T (2015) Giving too much social support: Social overload on social networking sites. *Eur J Inf Syst* 24:447–464
- Malthouse EC, Calder BJ, Kim SJ, Vandenbosch M (2016) Evidence that user-generated content that produces engagement increases purchase behaviours. *J Mark Manag* 32:427–444
- Mochon D, Johnson K, Schwartz J, Ariely D (2017) What are likes worth? A Facebook page field experiment. *J Mark Res* 54:306–317
- Moon S, Kim MY, Iacobucci D (2021) Content analysis of fake consumer reviews by survey-based text categorization. *Int J Res Mark* 38:343–364
- Moran G, Muzellec L, Johnson D (2020) Message content features and social media engagement: evidence from the media industry. *J Product Brand Manag* 29:533–545
- Mostafa MM (2013) More than words: social networks' text mining for consumer brand sentiments. *Expert Syst Appl* 40:4241–4251
- Naumann K, Lay-Hwa Bowden J, Gabbott M (2017) Exploring customer engagement valences in the social services. *Asia Pacif J Mark Logist* 29:890–912
- Nieto-García M, Resce G, Ishizaka A, Occhiocupo N, Viglia G (2019) The dimensions of hotel customer ratings that boost RevPAR. *Int J Hospitality Manag* 77:583–592
- Nilashi M, Asadi S, Minaei-Bidgoli B, Abumalloh RA, Samad S, Ghabban F, Ahani A (2021) Recommendation agents and information sharing through social media for coronavirus outbreak. *Telemat Inform* 61:101597
- Okoli C, Mehdi M, Mesgari M, Nielsen FÅ, Lanamöki A (2014) Wikipedia in the eyes of its beholders: a systematic review of scholarly research on Wikipedia readers and readership. *J Assoc Inf Sci Technol* 65:2381–2403
- Peters B (2018) The new Facebook algorithm: secrets behind how it works and what you can do to succeed. Buffer Social Blog
- Pitt CS, Plangger KA, Botha E, Kietzmann J, Pitt L (2019) How employees engage with B2B brands on social media: word choice and verbal tone. *Indu Mark Manag* 81:130–137
- Pletikosa Cvijikj I, Michahelles F (2013) Online engagement factors on Facebook brand pages. *Soc Netw Anal Min* 3:843–861
- Schoder D, Gloor P, Metaxas PT (2013) Special issue on social media (editorial). *Künstl Intell* 27:5–8
- Semenov A (2013) Principles of social media monitoring and analysis software. *Jyväskylä studies in computing*, (168)
- Sett N, Ranbir Singh S, Nandi S (2016) Influence of edge weight on node proximity based link prediction methods: an empirical analysis. *Neurocomputing* 172:71–83
- Shrapnel E (2012) Engaging young adults in museums: An audience research study. *Master of Museum Studies*
- Slimani T (2013) Description and evaluation of semantic similarity measures approaches. *Int J Comput Appl* 80:25–33
- Sánchez D, Batet M, Isern D, Valls A (2012) Ontology-based semantic similarity: a new feature-based approach. *Expert Syst Appl* 39:7718–7728
- Steinberger R (2012) A survey of methods to ease the development of highly multilingual text mining applications. *Language Resour Eval* 46:155–176
- Stieglitz S, Dang-Xuan L (2013) Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *J Manag Inf Syst* 29:217–247
- Swani K, Milne GR, Brown BP, Assaf AG, Donthu N (2017) What messages to post? Evaluating the popularity of social media communications in business versus consumer markets. *Ind Mark Manag* 62:77–87
- Tafesse W (2020) YouTube marketing: how marketers' video optimization practices influence video views. *Internet Res* 30:1689–1707
- Terkenli TS, Bell S, Tošković O, Dubljević-Tomičević J, Panagopoulos T, Straupe I, ..., Živojinović I (2020) Tourist perceptions and uses of urban green infrastructure: An exploratory cross-cultural investigation. *Urban For Urban Green* 49:126624
- Tian G, Lu L, McIntosh C (2021) What factors affect consumers' dining sentiments and their ratings: Evidence from restaurant online review data. *Food Qual Prefer* 88:104060
- Turel O (2015) Quitting the use of a habituated hedonic information system: a theoretical model and empirical examination of Facebook users. *Eur J Inf Syst* 24:431–446
- Viglia G, Dolnicar S (2020) A review of experiments in tourism and hospitality. *Ann Tour Res* 80:102858
- Villamediana-Pedrosa JD, Vila-Lopez N, Küster-Boluda I (2019) Secrets to design an effective message on Facebook: an application to a touristic destination based on big data analysis. *Curr Issues Tour* 22(15):1841–1861
- Villamediana-Pedrosa JD, Vila-Lopez N, Küster-Boluda I (2020) Predictors of tourist engagement: Travel motives and tourism destination profiles. *J Dest Mark Manag* 16:100412
- Vivek SD, Beatty SE, Morgan RM (2012) Customer engagement: exploring customer relationships beyond purchase. *J Mark Theory Pract* 20:122–146
- Wallace E, Buil I, de Chernatony L, Hogan M (2014) Who “likes” you...and why? A typology of facebook fans: from “fan”-atics and self-expressives to utilitarians and authentic. *J Advert Res* 54:92–109
- Weinberg BD, Pehlivan E (2011) Social spending: managing the social media mix. *Bus Horizons* 54:275–282
- Witten IH (2005) Text mining. *Practical handbook of Internet computing*: 14–1
- Xu M (2015) Graphic relation of multimodal discourse analysis: analysis based on discourse in Weibo, QQ, and Renren. *Chin Foreign Entrep* 29:253–255
- Yang Z, Zheng Y, Zhang Y, Jiang Y, Chao HT, Doong SC (2019) Bipolar influence of firm-generated content on customers' offline purchasing behavior: A field experiment in China. *Electron Commer Res Appl* 35:100844
- Zhao Y, Wen L, Feng X, Li R, Lin X (2020) How managerial responses to online reviews affect customer satisfaction: An empirical study based on additional reviews. *J Retail Consum Serv* 57:102205
- Zheng T, Wu F, Law R, Qiu Q, Wu R (2021) Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach. *Int J Hosp Manag* 92:102658
- Zikopoulos P, Deroos D, Parasuraman K, Deutsch T, Giles J, Corrigan D (2012) Harness the power of big data The IBM big data platform. McGraw Hill Professional

Acknowledgements

This study was funded by the Ministry of Science and Technology, Digital Humanities Program (MOST 110-2410-H-032-051).

Competing interests

The author declares no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Yulin Chen.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023