REAd
Revista Eletrônica de Administração

# THE CONTRIBUTION OF UNSTRUCTURED DATA FROM SOCIAL MEDIA FOR PREDICTION IN MARKETING MANAGEMENT[1]

**Sylvio Ribeiro de Oliveira Santos**[2]

**Daniel Max de Sousa Oliveira**[3]

http://dx.doi.org/10.1590/1413-2311.392.117898

## ABSTRACT

The capacity to obtain market insights is a strategic need for companies to remain competitive. Despite this and the massive volume of data generated by consumers every second, companies rarely have the culture of making marketing decisions based on data and, when they do, rarely use consumer data widely available online, especially on social networks. One reason is that these data (e.g. texts) tend to be "dirty", disorganized and bulky, a so-called unstructured data. The purpose of this article is to discuss the benefits of new types of data that have become more abundant and accessible in Web 3.0 through popular social networks, as well as new methods of analysis, particularly learning methods for prediction. For this, an extensive literature review was carried out and a topic modeling was conducted to get an overview of the data and methods. At the end, the article suggests six main marketing challenges that unstructured data analytics can contribute to overcome, improving companies' competitiveness.
**Keywords:** Marketing Analytics. Unstructured Data. Predictive Models. Marketing Management. Social Media.

## LA CONTRIBUCIÓN DE DADOS NO ESTRUCTURADOS DE MIDIAS SOCIALES PARA LA PREDICCIÓN EN LA GESTIÓN DE MARKETING

La capacidad de obtener información sobre el mercado es una necesidad estratégica para que las empresas sigan siendo competitivas. A pesar de esto y del enorme volumen de datos que generan los consumidores cada segundo, las empresas rara vez tienen la cultura de tomar decisiones de marketing basadas en datos y, cuando lo hacen, rara vez utilizan datos de

[1] Recebido em 25/8/2021, aceito em 6/3/2023.
[2] Universidade Federal do Rio Grande do Sul – Programa de Pós-Graduação em Administração; Porto Alegre – RS (Brasil); https://orcid.org/0000-0002-2400-9187; Sylvio.ros@gmail.com.
[3] Universidade Federal do Rio Grande do Sul – Programa de Pós-Graduação em Administração; Porto Alegre – RS (Brasil); https://orcid.org/0000-0003-4661-4238; danielmax2011@gmail.com.

consumidores ampliamente disponibles en línea, especialmente en las redes sociales. Una razón es que estos datos (por ejemplo, textos) tienden a ser "sucios", desorganizados y voluminosos, los llamados datos no estructurados. A pesar de la complejidad que implica extraer valor informativo de estos datos, las empresas pueden obtener conocimientos que pueden mejorar la toma de decisiones y dar como resultado un mayor rendimiento competitivo. El propósito de este artículo es discutir los beneficios de los nuevos tipos de datos que se han vuelto más abundantes y accesibles en la Web 3.0, a través de redes sociales populares, así como los nuevos métodos de análisis, particularmente los métodos de aprendizaje. Para ello, se llevó a cabo una extensa revisión de la literatura y se realizó un modelado de temas para obtener una visión general de los datos y métodos. Al final, el artículo sugiere seis desafíos principales de marketing a los que puede contribuir el análisis de datos no estructurados, mejorando la competitividad de las empresas.
**Palabras clave**: Analítica de Marketing. Datos no Estructurados. Modelos Predictivos. Gestión de Marketing. Medios de Comunicación Social.



## A CONTRIBUIÇÃO DE DADOS NÃO ESTRUTURADOS DE MÍDIAS SOCIAIS PARA A PREDIÇÃO NA GESTÃO DE MARKETING

A capacidade de obter insights de mercado é uma necessidade estratégica para que as empresas se mantenham competitivas. Apesar disso e do enorme volume de dados gerados pelos consumidores a cada segundo, as empresas raramente têm a cultura de tomar decisões de marketing com base em dados e, quando o fazem, raramente usam os dados do consumidor amplamente disponíveis online, especialmente nas redes sociais. Uma razão é que esses dados (por exemplo, textos) tendem a ser “sujos”, desorganizados e volumosos, os chamados dados não estruturados. Apesar da complexidade envolvida na extração de valor informativo desses dados, as empresas podem obter insights que podem melhorar a tomada de decisões e resultar em maior desempenho competitivo. O objetivo deste artigo é discutir os benefícios de novos tipos de dados que se tornaram mais abundantes e acessíveis na Web 3.0, através das populares redes sociais, bem como novos métodos de análise, particularmente métodos de aprendizagem. Para isso, foi realizada uma extensa revisão da literatura e uma modelagem de tópicos para obter uma visão geral dos dados e métodos. No final, o artigo sugere seis desafios principais de marketing com os quais a análise de dados não estruturados pode contribuir, melhorando a competitividade das empresas.
**Palavras-chave:** Análise de Marketing. Dados não Estruturados. Modelos Preditivos. Gestão de Marketing. Mídia Social.

## INTRODUCTION

Obtaining consumer insights, a task conventionally under the responsibility of the marketing field, is a strategic need and, at the same time, a major challenge for companies (LEEFLANG et al., 2014). In recent years, the data-driven culture has gained popularity among companies by the promise to get to know consumers more deeply, make better decisions and increase earnings, which can be as high as 25% (BÖRINGER, 2022). It is a breakthrough for the marketing area, traditionally known for its creative culture and whose decisions are often

based on subjectivities (VERHOEF; KOOGE; WALK, 2016). An example of this is a survey carried out by IBM, at the beginning of the last decade, which found that 82% of marketers revealed that they make decisions based on experience and intuition (IBM, 2014). Fortunately, this seems to be changing. A 2022 survey showed that 73% of companies base their decisions on some data (BARC, 2022).



Marketing analytics is the key element behind a data-driven company, and can be defined as an interdisciplinary area capable of transforming insights into performance, increasing the return on investment through analytics that describe, diagnose, predict phenomena and prescribe actions (WEDEL; KANNAN, 2016). Through analytics, companies may use their own data and third parties' data to obtain different types of information. The results so far are very encouraging, companies that have adopted a data-based decision culture have seen their performance increase (SUNDSØY et al., 2014; WEDEL; KANNAN, 2016).

Wedel and Kannan (2016) called data "the oil of the digital economy". Companies have invested tens of billions of dollars in recent years in data collection and storage to secure their reserves (LARIVIERE et al., 2016), but little in transforming them into fuel for business, that is, investments in data analysis are too low (WEDEL; KANNAN, 2016). Through more advanced models and metrics, companies can explore the potential of the large volume of data made available in the last decade, mainly through social media, as it is estimated that between 80% and 95% of all business data are unstructured (BERGER et al., 2020). While not all social media data is unstructured, much of the unstructured data comes from social media (e.g., Twitter, Instagram, Reddit), especially in the business context. Unstructured data, such as text or large datasets from multiple sources, requires more complex analysis methods that seem to have not yet reached maturity (KALAMPOKIS; TAMBOURIS; TARABANIS, 2013; WEDEL; KANNAN, 2016), especially in marketing. The high accessibility and great informative potential of these data demonstrate the extreme relevance of the discussion about its use to guide strategic marketing decisions.

However, traditional statistical methods have difficulties in dealing with unstructured data (e.g. images, texts, spatial information) and massive datasets. One of the main problems is that many marketing phenomena are not linear. For example, estimating certain types of sales (SUN et al., 2008), demand for new products (HONG et al., 2010), seasonal demands or sales of "long tail" products (TEUNTER; SYNTETOS; BABAI, 2011). In these cases, the so-called learning methods, statistical learning and machine learning (e.g. neural networks, Bayesian methods, XGB, random forest, ensemble methods) usually present results superior to traditional statistical methods (HONG et al., 2010; HUANG; HO, 2012; KUO; XUE, 1999), being able to

handle nonlinear relationships with high precision (Guo et al., 2018), using more granular data (LI; KANNAN, 2014), more efficiently (TIMOSHENKO; HAUSER, 2019) and with fewer assumptions to be met (BAVARIA-PUIG; BUITRAGO-VERA; ESCRIBA-PEREZ, 2016). Thus, the combination of new data with new methods has helped companies in a series of activities such as forecasting demand, improving pricing through dynamic prices, predicting failures (LARIVIERE et al., 2016), increasing Customer Lifetime Value (CLV) (SIFA et al., 2015), optimize allocation of marketing mix resources and individualize offers (WEDEL; KANNAN, 2016), all with higher precision.



That said, the aim of this article is to discuss the applications and implications of using unstructured data and learning methods for marketing management. For this, initially, a bibliometric analysis of studies that use learning models in the analysis of unstructured data was carried out in order to understand in which contexts this approach has been used more frequently. Interestingly, the marketing area is not among the five areas where this approach has been most applied, which highlights the need to demonstrate the potential of these new types of data and analysis methods for improving marketing management. Precisely for this reason, after the presentation of bibliometric findings, we discuss the potential applications and strategic and operational implications for the marketing management of unstructured data analysis through statistical learning models.

As a contribution, we present the benefits of data-driven management, where information guides decisions and reduces the complexity of various marketing challenges, such as managing the relationship with consumers, developing new products, customizing offers and communication, etc. The adoption of a more analytical management can solve a historical marketing challenge: measuring the contribution of marketing investments, changing the perception of a purely creative area, full of good ideas, but which contributes little to the company's financial performance (VERHOEF; LEEFLANG, 2011).

The article is structured as follows: the next section deals with the emergence and pervasiveness of unstructured data and learning models. Then, we analyze articles published on the topic to extract the main themes, methods and data sources to generate insights in different areas. At the end, we discuss implications and applications for strategic and operational marketing management.

## 1 THE BOOM OF UNSTRUCTURED DATA, PREDICTIVE METHODS AND STATISTICAL LEARNING MODELS



In recent years, investments in data storage in the order of US$40 billion (LARIVIERE et al., 2016) have allowed access to a greater amount of data, as well as to data of greater quality and richness of information. This improvement resulted in, for example, data with less bias (JANETZKO, 2017) and with different levels of aggregation (WEDEL; KANNAN, 2016).

Traditionally, data is usually organized before being analyzed, following a structure called “tidy data”, in which observations are in rows, variables in columns and rows share the same data types (WICKHAM, 2014). Although data extracted online, such as search volume, spatial data (georeferenced data), metadata, behavioral data (e.g. cookies, engagement), have a certain organization, most of the data available today are unstructured (BERGER et al., 2020), such as texts (tweets, reviews, social media posts), customer service chats, press-releases, as well as images, audio and videos. The last three can be placed in the last degree of unstructured data.

In general, unstructured data is all data not tabulated in the tidy pattern, it exists in abundance and has great informational potential. The variety of unstructured data makes it valuable for a large number of marketing applications. Wedel and Kannan (2016) argue that unstructured data can help to answer a much broader range of questions than conventional data, that is, structured data from surveys, interviews and other traditional collection techniques. While conventional data can, for example, measure the performance of advertising, promotional actions, changes in retail and marketing communication mix; unstructured data serves more varied and complex purposes such as trend analysis, sentiment analysis, buying journeys, customer segmentation, recommendation systems. These new possibilities are capable of offering greater competitive intelligence.

The greater availability of data, predominantly unstructured, was made possible by Web 3.0, and social media, today the most popular online activity (CHU; KIM, 2011; KEMP, 2020). Social media data are often good proxies for measuring performance in the real world (LASSEN; MADSEN; VATRAPU, 2014) or relevant consumer constructs such as: interest (KULKARNI et al., 2012; SAKAKI et al., 2016) and popularity (FAN; CHE; CHEN, 2017; LIU et al., 2016). For example, Kulkarni and colleagues (2012) used search volume to measure audience interest in movies before their release. Luo and Zhang (2013) used product review scores to measure brand equity and concluded that they are good predictors of variation.

According to Schoen (2013), for social media data to have predictive capacity, the prediction itself needs to be encoded in the data and must be preserved during collection, so the analysis method will work to unveil the prediction. The great value of this type of data is that it can often be linked to the user. This provides greater granularity in describing user behavior, with more disaggregated data generally increasing the accuracy of predictive models (BENNETT; STEWART; BEAL, 2013; LI; KANNAN, 2014; SOOD; JAMES; TELLIS, 2009; XIONG; BHARADWAJ, 2014).



With more data of different types at your disposal, you can increase the number of predictors in a model, measure them, and keep only those that provide satisfactory performance. For example, Sundsøy et al. (2014) used 350 variables to find the 20 most important variables (i.e., features) for their customer segmentation model. The result was a conversion of 6.29% new customers against 0.19% of the model with four variables based on the judgment of the researched company's marketing department.

A good example of how social media data can help improve predictions comes from the studies by Dellarocas, Zhang and Awad (2007) and Fan, Che and Chen (2017). The first extended the Bass Model to forecast sales by incorporating consumer review scores in addition to traditional sales data. The second extended the Bass Model even further, incorporating sentiment analysis of the content of reviews into notes and sales data. In both cases, the prediction models were more accurate.

Companies have always dreamed of being able to predict consumer behavior. Predictive models based on statistics and econometrics are, historically, the most used instruments to make this dream come true. There are a large number of prediction / forecasting techniques, classification and recommendation systems available today and which one to use depends on the type of data, strategy goals (GIRAUD-CARRIER; POVEL, 2003) and available resources. Perhaps for operational and/or financial reasons (BUCKLIN; GUPTA, 1999), most companies tend to opt for the use of simpler analysis techniques, such as linear regression and traditional time series methods. Sometimes these techniques are included in the management system used by the company. For example, the demand prediction tool included in SAP management software and Forecast Pro uses the Croston method, a method based on exponential smoothing. This method has been used to forecast demand for decades, but it can be quite imprecise, as in cases where demand varies in a non-systematic way, that is, without an apparent pattern (ALI et al., 2009; CANIATO et al., 2005). Despite the efforts of companies like SAP and SAS to stay relevant in the business analytics context, change is slow (FILDES, 2022), and cloud

computing platforms (e.g., Microsoft Azure, Google Cloud, Alteryx) especially when it comes to manage and manipulate unstructured data.

In addition to the methods, the metrics used by most companies are also often relatively simple. For example, two of the most used metrics in digital marketing to measure conversion (e.g. last-click, seven-day average) only consider the channels visited a few days before the conversion. These characteristics make them insufficient for evaluating decision hierarchies with multiple touchpoints or high-involvement products and services, in which case the consumer can take weeks to decide.



New types of data, especially in large volumes, demand new methods called learning methods, which include statistical learning and machine learning. Companies can use learning methods to increase their predictive capacity and obtain insights to formulate increasingly effective strategies in data-driven management.

Artificial neural networks (ANN) are a type of learning method widely used in predictive models. ANN[4] has proven to be superior to traditional forecasting methods (e.g. ARIMA, exponential smoothing, OLS) not only for its ability to capture nonlinear changes, but also for its ease of adapting to different contexts, with little or no assumptions to be met (HONG et al., 2010). ANN's advantages over traditional prediction methods have already been demonstrated in different contexts such as sales (HUANG; HO, 2012), electrical consumption (LIN; HOU; SU, 2012), demand fluctuations (CANIATO et al., 2005) and even demand for cash in ATMs (VENKATESH et al., 2014). There is a wide variety of learning models available and they can be combined with other learning methods or even traditional learning ones (BAVARIA-PUIG et al., 2016; CANIATO et al., 2005; HONG et al., 2010) to provide better insights. A review of data mining techniques applied to customer relationship management (NGAI; XIU; CHAU, 2009) listed 34 different techniques, with ANN being the most popular. While many of these techniques remain relevant today, new algorithms and more complex models emerge every year.

It is noteworthy that the implementation of more advanced models usually requires longer data preparation time, in addition to higher computational cost. Companies looking to embrace a data-driven culture will probably need to make big investments, as in big data environments capable of collecting, storing, retrieving, pre-processing and analyzing large volumes of data in real-time or periodically. Implementing these infrastructures involves different environments and languages such as NoSQL data structures, data management

[4] There are different types of neural networks and algorithms. Their functioning and the differences between them are beyond the scope of this paper.

environments such as Apache Hadoop, MongoDB, among others. Investment in medium and large data infrastructures is usually high, but the benefits should outweigh the price (ALI et al., 2009).

## 2 DATA COLLECTION AND ANALYSIS



Initially, we carried out a bibliometric search to better understand the use of unstructured data and new analysis techniques capable of producing better models, that is, with smaller errors and greater explanatory power.

The search for studies was performed on the Web of Science (WoS) using the keywords “predicting”, “forecasting”, “social media” and “social networks” in the title, abstract and keywords in the period between 2000 and 2022. It is worth mentioning that WoS has a huge collection of studies that meet the needs of this research, in addition to being frequently used in bibliometric studies in the business area. Regarding the search terms used, after several attempts, this was the combination of search terms that returned the set of documents with fewer false-positive results, that is, with fewer non-relevant documents. As for the time period, there are practically no studies on unstructured data and prediction or learning models before 2000, this explains the lower limit. On the other hand, 2022 is the last full year of data available, which explains the upper bound.

The search returned 6984 documents which comprise the full dataset. A new search was made on this dataset with the term "marketing", resulting in a second dataset with 469 articles. Both datasets were organized and analyzed using Metaknowledge, a bibliometric package for Python. The table 1 shows the research areas with most studies published and their impact through citations. Computer science is the most prolific area, although psychology studies are the most cited. Business & Economics field are close to CS in citations, even though B&E has less than half the number of articles published.

Table 1 - Proportion of studies and citations per research area

| Research areas | Articles published | Citations |
|---|---|---|
| Computer Science | 25,92% | 17,86% |
| Psychology | 16,70% | 22,19% |
| Business & Economics | 10,85% | 16,31% |
| Engineering | 9,25% | 5,71% |
| Communication | 6,52% | 4,97% |

| | | |
|---|---|---|
| Science & Technology - Others | 6,48% | 8,23% |
| Public Environmental Occupational Health | 6,17% | 5,93% |
| Information Science Library Science | 5,17% | 5,94% |
| Health Care Sciences & Medical Informatics | 5,06% | 5,91% |
| Telecommunications | 4,08% | 2,67% |

Source: Research data (2022).

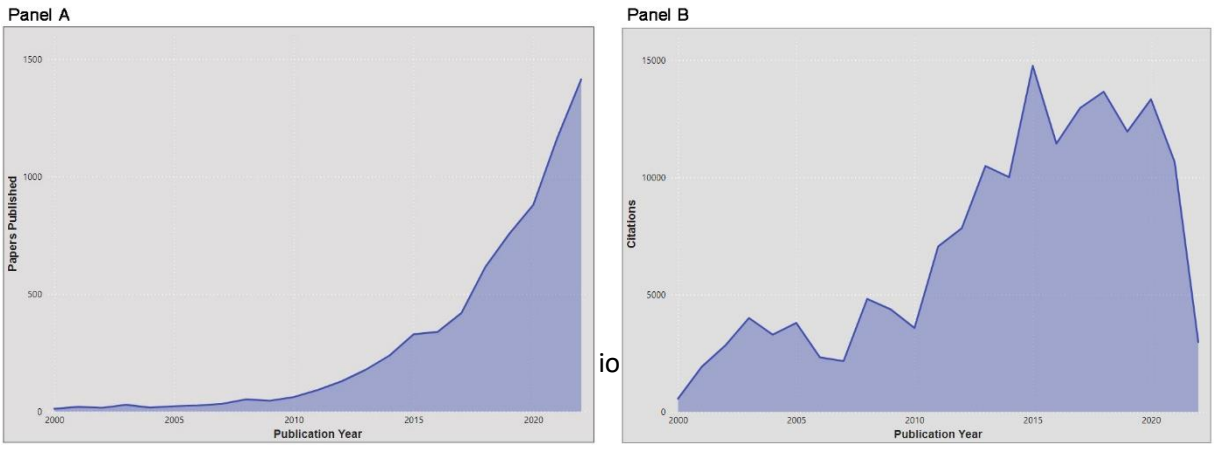


We performed topic modeling to get a picture of how the different areas of science are using unstructured data to predict and explain phenomena. The topic modeling used TF-IDF to obtain the most important words in each article in the full dataset, which afterward was used as an input for the LDA (latent dirichlet allocation). This unsupervised machine learning method calculates the probability of distribution of topics and articles and generates the most important topics. The process can be described in 3 steps. First, using the Metaknowledge package, we build the corpus comprising of each article's title and keywords. Second, we cleaned text data, removed common words and vectorized unique words using TF-IDF, a common NLP technique to value each word based on the specific document and the entire dataset. This step was performed using NLTK package. Lastly, the words obtained in the previous step were tokenized and used as input in LDA for topic modeling. This step was performed using Gensim package.

## 3 RESULTS

The first finding to be pointed out is the exponential growth of publications (see graph 1 - Panel A). From 2010 on, the number of published studies has increased considerably, and it is still not possible to verify a trend towards stability or decline. The number of citations has also grown steeply, specially at the beginning of the last decade.

Graph 1 - Panel A: Number of publications per year; Panel B: Total number of citations per year of publication



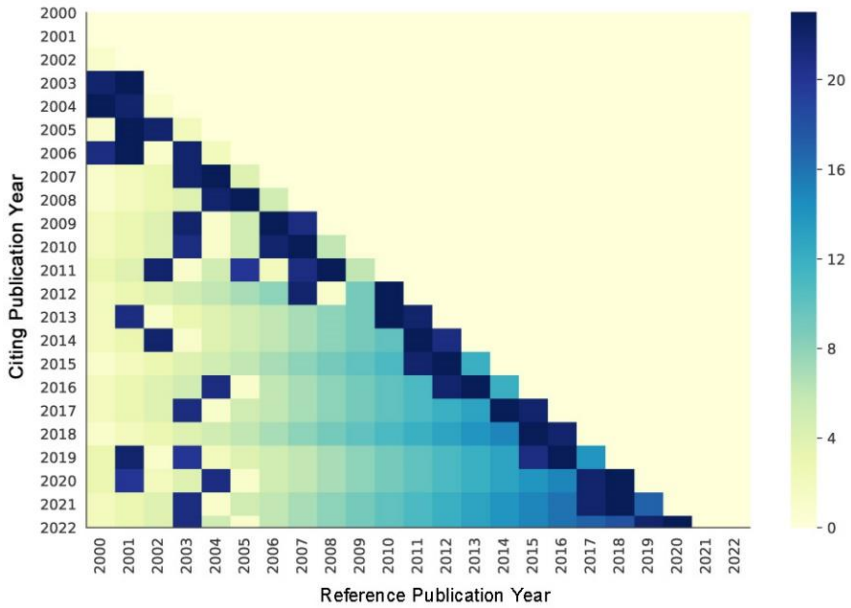

io

Source: Research data (2022).



Interestingly, articles published at early 2000s has similar impact as the late 2000s. A Multi RPYS analysis (COMINS; HUSSEY, 2015) show that papers published between 2001-2004 are the most impactful over time. Apart from them, the most influential studies are those published up to 3 years earlier, which indicates the field are changing fast.

Graph 2 - Multi Reference Publication Year Spectroscopy (RPYS) of studies published between 2000 and 2022




Source: Research data (2022).

Based on topic modeling, we have identified the main themes where unstructured data from social media has been used to predict phenomena. These studies help to visualize the potential of these new types of data and the different methods used and can contribute to advancing their applications in marketing management. All topics (in no particular order) are shown in Table 2.

Table 2 - Topic modeling results

| Topic # | Full dataset | Marketing dataset |
|---|---|---|
| 1 | Trust; Privacy; Election; Candidates; Selfies; Rumors | Sentiment; Stock; Market; Returns; Twitter; Investor; Price |
| 2 | Learning; Users; Networks; Methods; Accuracy; Time | Leadership; Team; Leaders; Group; Employees; Performance; Team |
| 3 | Health; Mental; Depression; Anxiety; Symptoms; Psychological | Personality; Travel; Search; Ties; Behavior; Digital |
| 4 | Risk; HIV; Climate; Weather; Disaster; Events; Drug; Community | Privacy; Employees; Value; Organizational; Factors; Work |
| 5 | Political; Participation; Groups; Individuals; Ties; Cultural | Brand; Consumers; Engagement; Facebook; Customers; Content; Advertising |
| 6 | Students; Adolescents; Sexual; Behavior; Alcohol; Body; Exposure | Consumers; Companies; Influence; Business; Adoption; Satisfaction |
| 7 | Public; Covid; Health; Content; Media; Pandemic | Intention; Knowledge; Engagement; Media; Behavior; Trust; Consumers |
| 8 | News; Sentiment; Stock; Market; Financial; Price; Sales | Innovation; Brand; Community; Performance; Sharing; Crowdfunding; Creativity |
| 9 | Online; Theory; Behavior; Factors; Influence; Knowledge | Business; Customers; Communication; Value; Influence |
| 10 | Disease; Population; Patterns; Size; Transmission; Contact; Cancer | Market; Twitter; Entrepreneurial; Bitcoin; Personal; Team; Share |

Source: Research data (2022).



In the following sections, we'll discuss the main applications of unstructured data in general and in the marketing field.

## 4 PRIMARY APPLICATIONS OF UNSTRUCTURED DATA AND LEARNING MODELS TO PREDICT PHENOMENONS

*Elections.* In the review by Kalampokis et al. (2013), 11 of the 54 prediction articles involved electoral themes. The study by Ceron, Curini, Iacus and Porro, (2014) is a good representation of the use of predictive models to predict election winners. The authors used Twitter data to measure user opinions based on expressed feelings and predict a candidate's chances of winning based on digital public opinion. Although a social media population cannot be considered a faithful representation of the actual population (RUTHS; PFEFFER, 2014), the authors obtained better results than classic sample-based surveys, possibly due to the more dynamic nature of the social media data that allows you to more quickly capture trends and changes in voters' opinions. Chauan et al. (2021) review shows that many studies use Twitter data and NLP techniques to predict election results with accuracy above 80%, similar to traditional election pools, sometimes even better. Undecided voters are a particular problem that researchers need to deal with. Silva (2020) tackled this issue using Bayesian method and the introduction of a variable called "voting decision."

*Stock market.* Few areas make as much use of predictive models as finance. Researchers in the field are often at the forefront of forecasting, developing advanced algorithms to obtain increasingly accurate estimates of financial market changes (Kimoto, Asakawa, Yoda, & Takeoka, 1990). Social media data has helped researchers create more accurate predictive models for the financial market. Bollen, Mao and Zeng (2011) are among the first researchers to use mood levels from Twitter messages to predict variations in the Dow Jones index, with satisfactory results in terms of accuracy and error reduction. Nofer and Hinz (2015) tried to replicate the results using a ten times larger database, 100 million tweets, but found significant results only when they added the contagion factor to the model, in the form of numbers of followers. Based on theory, the authors argue that the mood contained in the messages (e.g., anger, calm, positivity) is not enough to predict the behavior of the stock market, being necessary to take into account the capacity of these feelings to spread. This is one of the few studies that used a reasonable theoretical load to explain relationships between predictors and responses, whose importance was emphasized by Schoen et al. (2013). While the above studies used mood expressed in social networks as predictors, other authors (BING, CHAN & OO, 2014; NGUYEN; SHIRAI & VELCIN, 2015; RAO & SRIVASTAVA, 2013) used the valence of feeling contained in the messages, classifying each online post as positive, negative, or neutral, based on the words used. The model by Bing et al. (2014) managed to predict with 76% accuracy the share price of 30 companies listed on the NYSE and NASDAQ. Based on this, the authors defend that companies manage the feelings in the social media, since negative and positive feelings can impact the company's value through their actions. Recent studies have explored unstructured data through deep learning (MEHTA; PANDYA; KOTECHA, 2021) methods and by combining more advanced NLP techniques to preprocess textual data used as inputs in econometric models. As a result, researchers have been able to make more granular forecasting and real-time predictions. Backa, Boyd and Kannan (2021) analyzed tweets' content and found their impact on stock value over the next 30, 45, and 60 seconds and whether this impact would be temporary or permanent. Tweets with informative content tended to generate more lasting impact than tweets geared toward generating social media buzz.

*Sales.* Unlike the two previous themes, the use of unstructured data in sales prediction is more diversified. The topics analysis using LDA showed two distinct groups related to sales, the first and smallest is of a more attitudinal nature and is associated with terms such as "engagement", "advertising", "satisfaction" and generally use explanatory or conceptual models. The second group of studies is larger and more transactional in nature and is associated with terms such as "sales", "price", "adoption" and generally makes use of predictive models.

Examples of the latter group include predicting sales: in freemium games (SIFA et al., 2015); of electronics based on intentions posted on Twitter (KORPUSIK et al., 2016); using demographic and psychographic information extracted from Facebook (ZHANG; PENNACCHIOTTI, 2013); from the volume of searches performed (KULKARNI et al., 2012) or volume of blog posts (QIN, 2011). Movie box office predictions are considerably popular, possibly due to the ease of obtaining data. Kulkarni et al. (2012) was able to predict movie ticket sales in the opening week with greater effectiveness by adding into the model the search volume about the movie weeks before the movie release.



In their popular article, Ghose and Ipeirotis (2011) used robust methods to assess the impact of user reviews on the sales of 411 analyzed products. The authors found three review characteristics capable of affecting sales: author characteristics, subjectivity, and readability. For example, reviewing subjectivity (i.e., the author's opinion) is most important for experiencing products like movies. It can increase product sales and the usefulness of reviews as long as they do not contain severe errors in the text. The results highlight the importance of incorporating content data into the model for good predictive performance, as defended by Du et al. (2014). In addition to product reviews, tweets are widely used in sales forecasting. Korpusik et al. (2016) used neural networks to see if users who had mentioned purchase interest (e.g., "should I buy an iPhone?") on Twitter bought the product; the authors' model achieved 80% of accuracy. An exciting aspect of their method was using a sequence of posts and not just single posts. Using only Twitter data, Lassen et al. (2014) could predict iPhone sales with an accuracy equivalent to popular forecasts used by prominent consultants based on more complex models and different data types.

*Health.* Similar to sales, the analysis of topics showed that there were two distinct groups of studies. The first group has a psychological nature ("anxiety", "depression", "mental") while the second group has an infectious nature ("covid", "transmission", "patterns"), with only the second group usually using predictive models. All articles of the second group share the purpose of predicting demand for medical and hospital services or avoiding possible epidemic outbreaks. For example, Q. Zhang et al., 2017 used mechanistic models to develop real-time monitoring that captures the dynamic behavior of the Influenza virus. The authors used Twitter georeferencing, meteorological and socioeconomic data and were able to predict with four weeks in advance, and 90% accuracy, flu outbreaks in Spain. Similarly, predictive models used social media data to predict physician visits due to asthma (RAM et al., 2015), number of infected, exposed, and healthy people (CHEN et al., 2014), and

smog peaks in China. A feature of these studies is the use of predictors from a large number of sources, such as physical sensors, in addition to social media.

The covid pandemic contributed to many studies, including many studies from outside the health field. However, the large majority of these studies is survey-based and use common statistical methods. A small part of Covid studies addressed the contagious aspect of the disease, similar to those cited above, and used a variety of predictive methods with data types. For example, prediction of new cases and mortality (PINTER et al, 2020; AMINU et al, 2021); disease severity (ASHRAFI et al, 2021) and diagnostics using respiratory sound (LELLA; PJA, 2021) or imaging exams (SEDIK et al. 2021).



*Events.* The most recent among the themes seems to have emerged with the evolution of machine learning techniques capable of processing and extracting meaning from a large volume of data, especially georeferencing and language. Similar to health prediction studies, event prediction has a strong social presence and several studies seek to predict events of great impact on society such as protests, political crises, cyberactivism (BOECKING; HALL; SCHNEIDER, 2015; KALLUS, 2014; ZHAO et al., 2015). However, unlike predicting new cases of diseases, predicting events does not seem to depend on past events to predict the future, which practically makes the use of traditional econometric methods and historical data fruitless. In fact, the unique and rare nature of many events makes this type of prediction different from others (CEDERMAN; WEIDMANN, 2017). For example, Nguyen et al. (2020) predicted in real-time the population needs in the face of hurricanes using tweets and weather data. In events, detection is not enough (i.e., the number of people infected), it is necessary to predict in advance or almost in real time. Therefore, predicting events, whether it's a large-scale protest (KALLUS, 2014) or the popularity of the new iPhone launch (ZHANG et al., 2015) is often based on what people are talking about, how much the subject is being spoken and where the authors of the messages are at the time of posting. Advances in deep learning techniques should continue to contribute to improving the predictive capacity of models.

## 5 THE VALUE OF UNSTRUCTURED DATA AND LEARNING MODELS FOR MARKETING MANAGEMENT

The area of business, economics and management represented less than 10,85% of the articles found in the research carried out in the previous section. Marketing articles represent 6,87% of full dataset. Only a small portion of the marketing dataset is based on predictive

models[5], although many claims to be. This academic reality also has its correspondence in the business world, where professionals traditionally make marketing decisions based little on data and more on intuition and experience (IBM, 2014; VERHOEF et al., 2016), a culture that is difficult to change, even with studies showing that decisions based on data analysis are more efficient and increase company performance (HUANG; HO, 2012; SUNDSØY et al., 2014). According to Blattberg and Hoch (1990), it is the combination of intuition, experience, data and mathematical models that result in the best decisions. Academics and professionals agree that the biggest marketing challenges are in the digital universe (LEEFLANG et al., 2014). Fast digitization has brought not just an explosion of data, but a proliferation of channels and changes in behavior that have profoundly changed the dynamics of consumption. Thus, it remains for companies to adapt, using data to leverage new competitive advantages. The new types of data and the large volume generated in the digital age offer two paths: i) improving current models by generating more accurate estimates; ii) predict and understand complex phenomena with the potential to create something totally new. However, this innovative potential is still in its infancy and the vast majority of models still focus on improving processes (LARIVIERE et al., 2016).



The purpose of this section is to present strategic and operational implications of using unstructured data (usually scrapped from social media) and predictive learning models for marketing management. In other words, how marketing can benefit from the vast amount of data available, especially from digital platforms, and from the new quantitative methods developed to extract insights from this often noisy data. If advanced areas such as engineering and computer science have not yet reached maturity in the use of unstructured data and learning models, as we were able to see from the data collected (the majority of studies are survey-based) and it has been pointed out by Kalampokis et al. (2013) and Wedel and Kannan (2016), marketing analytics (a marketing area that involves collection, management and analysis to generate insights that increase company performance) is still far from maturity.

Wedel and Kannan (2016) point out four main applications of the use of marketing analytics: i) Customer management (CRM): in the development of models that aim to improve customer acquisition, retention and satisfaction, increasing lifetime value; ii) Marketing mix:

[5] Explanatory models are based on measures of association (e.g. p-value, $R^2$) that provide explanatory power, but not predictive power that comes from the difference between predicted and actual values. Predictive power can only be obtained from predictive analytics (SHMUELI; KOPPIUS, 2010), when using statistical or machine learning methods to predict future or unknown outcomes (ABBASI; LOU; BROWN, 2015). Explanatory measures may have low replicability and generalizability. In other words, they may not reflect the real world. Interestingly, Kalampokis et al., (2013) found in their review that about half of the articles committed to saying that their models predicted, when in fact they used explanatory methods.

supporting the best allocation of resources for greater investment efficiency; iii) Customization of the marketing mix: capturing the heterogeneity of consumers and customizing actions for different segments; and iv) Privacy and security: in the ethical use of customer data, ensuring their privacy.



There are several practical implications of predictive studies using unstructured data and social media. One of the most popular is the sales prediction with greater accuracy. In fact, a survey of nearly 800 marketing executives revealed that selling more is the main reason companies invest in data analysis (LEEFLANG et al., 2014). It can be said that predicting sales has both a strategic and an operational impact (ALI et al., 2009; CANIATO et al., 2005; KULKARNI et al., 2012; MIAH et al., 2017). On the strategic side, sales forecasts allow the adjustment or maintenance of plans, reallocating resources according to the forecast. On the operational side, it guides production by improving inventory management, avoiding unavailability and excess of products, and optimizing costs. Another use of sales prediction is to feed product recommendation systems (KORPUSIK et al., 2016). Predictions can have different time horizons. The shorter the horizon, the more the prediction is geared towards operational management. Over the past years, there has been a growth in both short-term and very short-term predictive studies, the latter often being called “nowcasting” (see KIM; CHA; LEE, 2017; RAM et al., 2015; ZHANG et al., 2017). In business, a company can then predict sales of a new product to be released at the end of the year, and after the launching it can implement models to calculate sales in the next 24 hours or weeks ahead; in the absence of real-time monitoring, these models can be used to simulate real-time sales.

Sales predictions are usually based on demand for the product. A very popular way to forecast demand with social media data is using buzz (DELLAROCAS et al., 2007; DHAR; CHANG, 2009; FAN et al., 2017; BOGAERT et al., 2021). Online buzz is typically measured primarily through two types of data: volume and content.

Volume is considered a proxy of consumer interest (KULKARNI et al., 2012; XIONG; BHARADWAJ, 2014), helping to measure the popularity of a given subject. Dhar and Chang (2009) demonstrated that the volume of blogs ‘talking’ about a new music album is a good predictor of sales before its release. Qin (2011) expanded to the film market and showed that the greater the word-of-mouth of a movie on blogs, the higher the movie’s box office. They also identified the inverse effect, the higher the box office, the greater the number of online publications about the film. Kulkarni et al. (2012) cemented these findings by using 4.3 billion searches to demonstrate that search engine volume can be used to measure consumer interest in a movie and help predict its sales a month before its release. Using more granular data, such

as daily buzz, can also help companies closely monitor a product launch. Xiong and Bharadwaj (2014) performed a functional data analysis of buzz about video games before launch and were able to visualize changes in online word-of-mouth over time and demonstrated that high buzz is a sign of good results in video games sales. The authors concluded with a warning: if buzz declines near launch, sales will be negatively affected. The authors emphasize that such results were only possible due to the use of disaggregated data extracted from social media.



While volume helps produce more accurate demand estimates, it is not revolutionary. The content, in turn, has greater informational load and can be used to supplement volume, increasing model accuracy (GEVA et al., 2017) and estimating purchase intent (KORPUSIK et al., 2016; LIU et al., 2016) rather than mere interest. Using only Twitter data, Korpusik et al. (2016) classified tweets as relevant or not relevant to predicting a purchase and predicted with more than 80% accuracy whether the author of a tweet who revealed an interest (e.g. "should buy", "need one") in an electronic device in fact purchased the product afterwards. The results revealed that only 16% of those interested actually compared the product. In addition to revealing attitudes, content analytics also provide insight into consumer sentiments in brand interactions. By monitoring these feelings, companies have valuable information to guide marketing decisions, both in the short and long term. An example of the practical application of feeling analysis is the study Kao and Huang (2017). The authors demonstrated that highly positive posts on a brand's Facebook page over a three-day period are associated with a jump in sales a few days later.

Bogaert et al. (2021) used 64 predictors based on volume and content to predict box office sales using social media and movie data. The researchers found that predictors based on social media data have higher predictive value than movie-based predictors (e.g., genre, top actor, rating). Through machine learning techniques, Facebook data showed higher accuracy than Twitter data across all models. Moreover, user generated content (UGC) is substantially more predictive than firm generated content (FGC).

A third type of data also used in predictions is metadata (e.g. number of followers, number of likes, ratings in ratings, number of words in a rating). Similar to volume, metadata also helps to measure popularity. Apala and colleagues (2013) found that the popularity of a movie's actor or lead actress (measured by the number of Twitter followers) is related to the success of his movie at the box office. Therefore, using information about the author or post can be a valuable predictor of attitudes and behaviors (APALA et al., 2013; BOECKING et al., 2015; MIAH et al., 2017; NOFER; HINZ, 2015), and can be used to supplement volume and content data to improve model's accuracy.

Other types of unstructured data extracted from social media are also used to predict demand or behavior and can be used in conjunction with the three types of data (volume, content, metadata) described above. Chen (2021) used chat conversations from Twitch.tv to predict the number of views of gaming live streaming, a highly popular online activity nowadays. Lassen et al. (2014) used a combination of text content and social graph[6] to predict iPhone sales using Twitter data only, and the result was an accuracy equivalent to that used by the Morgan Stanley bank. In the context of mobile applications, Sifa et al. (2015) used only app data to predict the likelihood that a user will purchase premium content from a free app and when. The authors also emphasize that the method used, Random Forest, is practical, quick to implement and scalable, in addition to having good explanatory power. In other words, it can be implemented by companies without major difficulties. Much like sales prediction, social media data is also used to predict the firm's stock market value, such as mood and feelings effects (BING et al., 2014; BOLLEN et al., 2011; NOFER; HINZ, 2015) and opinions posted online (BARTOV; FAUREL; MOHANRAM, 2018). Furthermore



In addition to predicting demand and company value with greater precision, there is a reasonable variety of applications for the use of social media data for marketing, mostly unstructured, but the themes are diffuse, with few studies published on each theme. Below are some relevant implications for marketing, both strategic and operational, published in recent years.

Discovery of new consumer needs: Reviews can be mined to reveal consumer needs and provide information for developing new products or improving current versions. Using a hybrid machine learning technique, Timoshenko and Hauser (2019) analyzed 12.000 reviews from the dental sector and concluded that this data can provide more insights than traditional research methods and with a 40% higher efficiency.

Designing the purchase journey and creating more accurate metrics: cookies record all online steps taken by consumers and can be used to analyze the purchase journey of online consumers. Specifically in the case of virtual stores, the day and time of visits are recorded, where the user came from and a purchase was made. Using real data from a company, Li and Kannan (2014) used cookies to measure the contribution that different communication channels made to purchase or not purchase. The authors criticize the metrics used in the market today, as they take into account only a few days before purchase and only the importance of the last channel, ignoring spillover and carryover effects from other channels. In this way, companies

[6] Data structure used to establish relationships between people and/or places and their interactions.

can analyze their customers' purchase journey to discover the most efficient channels (e.g. the one with the highest ROI), the average journey time to purchase, and develop more reliable metrics for their business. In the context of the offline shopping journey, location data recorded by social networks is promising and can help map the physical path consumers take during their purchases and predict where they will go next (HUANG, 2017).



Impact of interactions on satisfaction: studies with data from social media and service learning models are scarce, although equally relevant. Using natural language processing (NLP) techniques, Packard and Berger (2020) showed that it is possible to measure the impact of interactions with an agent on customer satisfaction. More specifically, the authors demonstrated that the language (concrete vs. abstract) used by the employee can affect current satisfaction and even future purchases. Text analytics hold promise for digging deeper into consumer experiences (Van Laer, Edson Escalas, Ludwig, & Van Den Hende, 2019) and have spectacular informational potential, yet they are among the least structured types of existing data, demanding elaborated pre-processing steps and more complex methods to extract insights which use to involve deep learning techniques and knowledge in linguistics (NLU), in addition to high computational cost. The popularity of language models like ChatGPT are chaning human-machines interactions. It is not very complex for companies to create their own language model from a pre-trained model like BeRT or GPT-2 or even with your own data using Artificial Neural Networks and use it to feed a Chatbot, for example (KUSHWAHA; KAR, 2021). People like to interact with robots in the service context and they will become more present with the recent huge advance in the language field.

Greater efficiency in marketing communication: companies can use social media data to improve the result of their digital campaigns, both before (MORO; RITA; VALA, 2016) and before launch (VILLARROEL ORDENES et al., 2018) and analyze the dynamics of content dissemination in the social media used (BOURIGAULT et al., 2014; LI et al., 2014), such as what kind of word-of-mouth (explicit *vs.* implicit) your customers use the most and which are the most persuasive (PACKARD; BERGER, 2017). Ordenes et al. (2019) measured the engagement of posts on Facebook, both in text and image formats, based on the types of speeches. One of the results found is that images should complement the text and vice versa, that is, if the text is more directive ("start saving now"), the image should not direct the action, but facilitate (illustrate why, by example, a happy customer could have saved). AI can be used to valuate images used in ads or FGC. Philp et al. (2022) used Google Vision AI, an easy-to-use image classification tool to analyze posts' images of restaurants and discovered that the easier the image is to understand, the more engagement it will receive. The results were

replicated in a follow-up experimental study. Lastly, generative tools powered by AI like GPT-4, MidJourney, Dall-E can help marketers to create marketing messages in seconds with good quality, something wholly unprecedented.

Measuring social value: studies have used social media data to measure a network's social connections (e.g. company's followers on Instagram) based on its ability to spread content and influence purchases (DOMINGOS; RICHARDSON, 2001; SKEELS et al. al., 2009; ZHOU; ROGER; LEI, 2015). Companies can use these models, improving them with the use of new data types, new analytics techniques and investments in big data environments, through platforms like Hadoop, to monitor the value of a company based on their followers and fans.



Obtain information about consumers: digital data and machine learning techniques are powerful to extract, analyze and infer information, with lower cost and greater efficiency (RASMUSSEN, 2017; TIMOSHENKO; HAUSER, 2019). Texts can be used to obtain demographic information (TAUSCZIK; PENNEBAKER, 2010), emotional states (Chapman, 2020), brand personality (PAMUKSUZ; YUN; HUMPHREYS, 2021) and individual personality traits (LIU et al., 2016; NETZER; LEMAIRE; HERZENSTEIN, 2019; WINKLER; RIEGER; ENGELEN, 2020) with an accuracy similar to that of surveys (OBSCHONKA; FISCH; BOYD, 2017). Pamuksuz et al. (2021) developed a classification model to assess brand identities on social media, a valuable way to track the brand image and its competitors over time, as well as the consumer's opinion about them. The valuable information obtained by analyzing social media data is not always obtainable by traditional means and, when it is, can be costly. Z. Liu et al. (2016) showed that it is possible to use information inferred from text to predict sales in 100 product categories. Such information can be used to answer different problems. One of these problems typical of online stores is called 'cold start', which is the low performance of recommendation algorithms with new users due to the lack of information from new users as to their interests and preferences.

Finally, new data and techniques can help companies develop a skill that Abbasi, Lou, and Brown (2015) called the "holy grail" of predictive analytics and is at the heart of any marketing strategy (WEDEL; KANNAN, 2016): the ability to capture the heterogeneity of consumers and offer individual actions for different types and at different times of day, at the lowest possible level of aggregation. This is slowly becoming a reality with more granular data usage and big data environments capable of handling large volumes of data, but this is just the beginning.

## 6 FINAL CONSIDERATIONS



This article aims to discuss the benefits of unstructured data types and analysis methods for marketing management, in particular learning methods such as machine learning and deep learning. In line with several authors (BLATTBERG; HOCH, 1990; LEEFLANG et al., 2014; SUNDSØY et al., 2014; VERHOEF et al., 2016; WEDEL; KANNAN, 2016), this article argues that a data-driven marketing management is necessary to develop and sustain competitive advantages, yet there is a lot for improvement both for marketing researchers and practitioners (BARC, 2022). High costs, lack of skills and the need for a change in organizational culture are some of the obstacles that need to be overcome to adopt intelligent marketing management. This culture must be based on and able to guide decisions and gain valuable insights into consumers and the market quickly that reflect better measurable performance for brands, products and organizations.

## REFERENCES

ABBASI, Ahmed; LAU, Raymond YK; BROWN, Donald E. Predicting behavior. **IEEE Intelligent Systems**, v. 30, n. 3, p. 35-43, 2015.

ALI, Özden Gür et al. SKU demand forecasting in the presence of promotions. **Expert Systems with Applications**, v. 36, n. 10, p. 12340-12348, 2009.

AMINU, Muhammad; AHMAD, Noor Atinah; NOOR, Mohd Halim Mohd. Covid-19 detection via deep neural network and occlusion sensitivity maps. **Alexandria Engineering Journal**, v. 60, n. 5, p. 4829-4855, 2021.

APALA, Krushikanth R. et al. Prediction of movies box office performance using social media. In: **2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)**. IEEE, 2013. p. 1209-1214.

ASHRAFI.; ALNUMAY, W. S.; ALI, R., HUR, S.; BASHIR, A. K.; ZIKRIA, Y. B.. Prediction models for covid-19 integrating age groups, gender, and underlying conditions. **Computers, Materials and Continua**, 67(3), 3009-3044, 2021

BARC. BARC Data Culture Survey 23: How to Liberalize Data Access to Empower Data Users, 2022.

BARTOV, Eli; FAUREL, Lucile; MOHANRAM, Partha S. Can Twitter help predict firm-level earnings and stock returns?. **The Accounting Review**, v. 93, n. 3, p. 25-57, 2018.

BAVIERA-PUIG, Amparo; BUITRAGO-VERA, Juan; ESCRIBA-PEREZ, Carmen. Geomarketing models in supermarket location strategies. **Journal of Business Economics and Management**, v. 17, n. 6, p. 1205-1221, 2016.

BENNETT, Christopher; STEWART, Rodney A.; BEAL, Cara D. ANN-based residential water end-use demand forecasting model. **Expert systems with applications**, v. 40, n. 4, p. 1014-1023, 2013.



BERGER, Jonah et al. Uniting the tribes: Using text for marketing insight. **Journal of Marketing**, v. 84, n. 1, p. 1-25, 2020.

BING, Li; CHAN, Keith CC; OU, Carol. Public sentiment analysis in Twitter data for prediction of a company's stock price movements. In: **2014 IEEE 11th International Conference on e-Business Engineering**. IEEE, 2014. p. 232-239.

BLATTBERG, Robert C.; HOCH, Stephen J. Database models and managerial intuition: 50% model+ 50% manager. In: **Perspectives On Promotion And Database Marketing: The Collected Works of Robert C Blattberg**. 2010. p. 215-227.

BOECKING, Benedikt; HALL, Margeret; SCHNEIDER, Jeff. Event prediction with learning algorithms—A study of events surrounding the egyptian revolution of 2011 on the basis of micro blog data. **Policy & Internet**, v. 7, n. 2, p. 159-184, 2015.

BOGAERT, M.; BALLINGS, M.; VAN DEN POEL, D.; OZTEKIN, A. Box office sales and social media: a cross-platform comparison of predictive ability and mechanisms. **Decision Support Systems**, 147, 113517, 2021.

BOLLEN, Johan; MAO, Huina; ZENG, Xiaojun. Twitter mood predicts the stock market. **Journal of computational science**, v. 2, n. 1, p. 1-8, 2011.

BÖRINGER, J.; DIERKS, A.; HUBER, I; SPILLECKE, D. Insights to impact: Creating and sustaining data-driven commercial growth. **McKinsey & Company**, 2022. **Available** in: https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/insights-to-impact-creating-and-sustaining-data-driven-commercial-growth. Accessed in: April 22th, 2023.

BOURIGAULT, Simon et al. Learning social network embeddings for predicting information diffusion. In: **Proceedings of the 7th ACM international conference on Web search and data mining**. 2014. p. 393-402.

BUCKLIN, Randolph E.; GUPTA, Sunil. Commercial use of UPC scanner data: Industry and academic perspectives. **Marketing Science**, v. 18, n. 3, p. 247-273, 1999.

CANIATO, Federico et al. Clustering customers to forecast demand. **Production Planning & Control**, v. 16, n. 1, p. 32-43, 2005.

CEDERMAN, Lars-Erik; WEIDMANN, Nils B. Predicting armed conflict: Time to adjust our expectations?. **Science**, v. 355, n. 6324, p. 474-476, 2017.

CERON, Andrea et al. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. **New media & society**, v. 16, n. 2, p. 340-358, 2014.

CHAUHAN, Priyavrat; SHARMA, Nonita; SIKKA, Geeta. The emergence of social media data and sentiment analysis in election prediction. Journal of Ambient Intelligence and **Humanized Computing**, v. 12, p. 2601-2627, 2021.



CHAPMAN, Chris. Commentary: Mind your text in marketing practice. **Journal of Marketing**, v. 84, n. 1, p. 26-31, 2020.

CHEN, Liangzhe et al. Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. In: **2014 IEEE international conference on data mining**. IEEE, 2014. p. 755-760.

CHEN, Wen-Kuo; CHEN, Long-Sheng; PAN, Yi-Ting. A text mining-based framework to discover the important factors in text reviews for predicting the views of live streaming. **Applied Soft Computing**, v. 111, p. 107704, 2021.

CHU, Shu-Chuan; KIM, Yoojung. Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. **International journal of Advertising**, v. 30, n. 1, p. 47-75, 2011.

COMINS, Jordan A.; HUSSEY, Thomas W. Compressing multiple scales of impact detection by Reference Publication Year Spectroscopy. **Journal of Informetrics**, v. 9, n. 3, p. 449-454, 2015.

DELLAROCAS, Chrysanthos; ZHANG, Xiaoquan Michael; AWAD, Neveen F. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. **Journal of Interactive marketing**, v. 21, n. 4, p. 23-45, 2007.

DHAR, Vasant; CHANG, Elaine A. Does chatter matter? The impact of user-generated content on music sales. **Journal of Interactive Marketing**, v. 23, n. 4, p. 300-307, 2009.

DOMINGOS, Pedro; RICHARDSON, Matt. Mining the network value of customers. In: **Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining**. 2001. p. 57-66.

FAN, Zhi-Ping; CHE, Yu-Jie; CHEN, Zhen-Yu. Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. **Journal of Business Research**, v. 74, p. 90-100, 2017.

FILDES, Robert; MA, Shaohui; KOLASSA, Stephan. Retail forecasting: Research and practice. **International Journal of Forecasting**, v. 38, n. 4, p. 1283-1318, 2022.

FILIERI, R.; LIN, Z.; LI, Y.; LU, X.; & YANG, X. Customer emotions in service robot encounters: A hybrid machine-human intelligence approach. **Journal of Service Research**, 25(4), 614-629, 2022.

GEVA, Tomer et al. Using forum and search data for sales prediction of high-involvement products. In: **Tomer Geva, Gal Oestreicher-Singer, Niv Efron, Yair Shimshoni." Using Forum and Search Data for Sales Prediction of High-Involvement Products"-MIS Quarterly, Forthcoming**. 2015.

GHOSE, Anindya; IPEIROTIS, Panagiotis G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. **IEEE transactions on knowledge and data engineering**, v. 23, n. 10, p. 1498-1512, 2010.

GIRAUD-CARRIER, C.; POVEL, Olivier. Characterising data mining software. **Intelligent Data Analysis**, v. 7, n. 3, p. 181-192, 2003.

GRIMALDI, Didier; CELY, Javier Diaz; ARBOLEDA, Hugo. Inferring the votes in a new political landscape: the case of the 2019 Spanish presidential elections. Journal of Big Data, v. 7, n. 1, p. 1-19, 2020.

GUO, Guancheng et al. Short-term water demand forecast based on deep learning method. **Journal of Water Resources Planning and Management**, v. 144, n. 12, p. 04018076, 2018.

HONG, Wei-Chiang et al. Taiwanese 3G mobile phone demand forecasting by SVR with hybrid evolutionary algorithms. **Expert Systems with Applications**, v. 37, n. 6, p. 4452-4462, 2010.

HUANG, Han-Chen et al. Back-propagation neural network combined with a particle swarm optimization algorithm for travel package demand forecasting. **International Journal of Digital Content Technology and its Applications**, v. 6, n. 17, p. 194-203, 2012.

HUANG, Qunying. Mining online footprints to predict user's next location. **International Journal of Geographical Information Science**, v. 31, n. 3, p. 523-541, 2017.

IBM. The 123s of CDOs: Transforming culture to be analytically driven. IBM, 2014. **Available** in: https://ibmcai.wordpress.com/2014/10/29/the-123s-of-cdos-transforming-culture-to-be analytically-driven/. Accessed in: June 4th, 2021.

JANETZKO, Dietmar. Nonreactive data collection online. **The SAGE handbook of online research methods**, p. 76-91, 2017.

KALAMPOKIS, Evangelos; TAMBOURIS, Efthimios; TARABANIS, Konstantinos. Understanding the predictive power of social media. **Internet Research**, 2013.

KALLUS, Nathan. Predicting crowd behavior with big public data. In: **Proceedings of the 23rd International Conference on World Wide Web**. 2014. p. 625-630.

KAO, Li-Jen; HUANG, Yo-Ping. Predicting purchase intention according to fan page users' sentiment. In: **2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)**. IEEE, 2017. p. 831-835.

KEMP, S. (2020). More than a half of the people on earth now use social media. DataReportal, 2020. Available in: https://datareportal.com/reports/more-than%20half-the-

world-now-uses-social-media. Accessed in: December 8th, 2020.

KIMOTO, Takashi et al. Stock market prediction system with modular neural networks. In: **1990 IJCNN international joint conference on neural networks**. IEEE, 1990. p. 1-6.

KORPUSIK, Mandy et al. Recurrent Neural Networks for Customer Purchase Prediction on Twitter. **CBREcsys@ recsys**, v. 1673, p. 47-50, 2016.



KULKARNI, Gauri; KANNAN, P. K.; MOE, Wendy. Using online search data to forecast new product sales. **Decision support systems**, v. 52, n. 3, p. 604-611, 2012.

KUO, R. J.; XUE, K. C. Fuzzy neural networks with application to sales forecasting. **Fuzzy Sets and Systems**, v. 108, n. 2, p. 123-143, 1999.

KUSHWAHA, Amit Kumar; KAR, Arpan Kumar. MarkBot–a language model-driven chatbot for interactive marketing in post-modern world. **Information systems frontiers**, p. 1-18, 2021.

LARIVIERE, Jacob et al. Where predictive analytics is having the biggest impact. **Harvard business review**, v. 25, n. May, 2016.

LASSEN, Niels Buus; MADSEN, Rene; VATRAPU, Ravi. Predicting iphone sales from iphone tweets. In: **2014 IEEE 18th International Enterprise Distributed Object Computing Conference**. IEEE, 2014. p. 81-90.

LEEFLANG, Peter SH et al. Challenges and solutions for marketing in a digital era. **European management journal**, v. 32, n. 1, p. 1-12, 2014.

LELLA, Kranthi Kumar; PJA, Alphonse. Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: cough, breath, and voice. **AIMS public health**, v. 8, n. 2, p. 240, 2021.

LI, Hongshuang; KANNAN, P. K. Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. **Journal of Marketing Research**, v. 51, n. 1, p. 40-56, 2014.

LI, Jingxuan et al. Social network user influence sense-making and dynamics prediction. **Expert Systems with Applications**, v. 41, n. 11, p. 5115-5124, 2014.

LIN, Chi-Chung; HOU, Tung-Hsu; SU, Chwen-Tzeng. Application of Data-mining Models to the Forecasting of Electricity Demand. **Advances in Information Sciences and Service Sciences**, v. 4, n. 21, 2012.

LIU, Ting et al. Predicting movie box-office revenues by exploiting large-scale social media content. **Multimedia Tools and Applications**, v. 75, n. 3, p. 1509-1528, 2016.

LIU, Zhe et al. To buy or not to buy? Understanding the role of personality traits in predicting consumer behaviors. In: **International Conference on Social Informatics**. Springer, Cham, 2016. p. 337-346.

LUO, Xueming; ZHANG, Jie. How do consumer buzz and traffic in social media marketing predict the value of the firm?. **Journal of Management Information Systems**, v. 30, n. 2, p. 213-238, 2013.

MENTZER, John T.; BIENSTOCK, Carol C. **Sales forecasting management: understanding the techniques, systems and management of the sales forecasting process**. SAGE Publications, Incorporated, 1998.



MIAH, Shah Jahan et al. A big data analytics method for tourist behaviour analysis. **Information & Management**, v. 54, n. 6, p. 771-785, 2017.

MORO, Sérgio; RITA, Paulo; VALA, Bernardo. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. **Journal of Business Research**, v. 69, n. 9, p. 3341-3351, 2016.

NETZER, Oded; LEMAIRE, Alain; HERZENSTEIN, Michal. When words sweat: Identifying signals for loan default in the text of loan applications. **Journal of Marketing Research**, v. 56, n. 6, p. 960-980, 2019.

NEWMAN, Mark EJ. The structure and function of complex networks. **SIAM review**, v. 45, n. 2, p. 167-256, 2003.

NGAI, Eric WT; XIU, Li; CHAU, Dorothy CK. Application of data mining techniques in customer relationship management: A literature review and classification. **Expert systems with applications**, v. 36, n. 2, p. 2592-2602, 2009.

NGUYEN, Thien Hai; SHIRAI, Kiyoaki; VELCIN, Julien. Sentiment analysis on social media for stock movement prediction. **Expert Systems with Applications**, v. 42, n. 24, p. 9603-9611, 2015.

NOFER, Michael; HINZ, Oliver. Using twitter to predict the stock market. **Business & Information Systems Engineering**, v. 57, n. 4, p. 229-242, 2015.

OBSCHONKA, Martin; FISCH, Christian; BOYD, Ryan. Using digital footprints in entrepreneurship research: A Twitter-based personality analysis of superstar entrepreneurs and managers. **Journal of Business Venturing Insights**, v. 8, p. 13-23, 2017.

PACKARD, Grant; BERGER, Jonah. How language shapes word of mouth's impact. **Journal of Marketing Research**, v. 54, n. 4, p. 572-588, 2017.

PACKARD, Grant; BERGER, Jonah. How Concrete Language Shapes Customer Satisfaction. **Journal of Consumer Research**, v. 47, n. 5, p. 787-806, 2021.

PAMUKSUZ, Utku; YUN, Joseph T.; HUMPHREYS, Ashlee. A brand-new look at you: predicting brand personality in social media networks with machine learning**. Journal of Interactive Marketing**, v. 56, n. 1, p. 1-15, 2021.

PHILP, Matthew; JACOBSON, Jenna; PANCER, Ethan. Predicting social media engagement with computer vision: An examination of food marketing on Instagram. **Journal of Business Research**, v. 149, p. 736-747, 2022.

QIN, Li. Word-of-blog for movies: A predictor and an outcome of box office revenue?. **Journal of Electronic Commerce Research**, v. 12, n. 3, p. 187, 2011.

RAM, Sudha et al. Predicting asthma-related emergency department visits using big data. **IEEE journal of biomedical and health informatics**, v. 19, n. 4, p. 1216-1223, 2015.



RAO, Tushar; SRIVASTAVA, Saket. Modeling movements in oil, gold, forex and market indices using search volume index and twitter sentiments. In: **Proceedings of the 5th annual ACM Web science conference**. 2013. p. 336-345.

RASMUSSEN, Karsten Boye. Data quality in online environments. **The SAGE handbook of online research methods**, p. 37-53, 2017.

RUTHS, Derek; PFEFFER, Jürgen. Social media for large studies of behavior. **Science**, v. 346, n. 6213, p. 1063-1064, 2014.

SAKAKI, Shigeyuki et al. Corpus for customer purchase behavior prediction in social media. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. 2016. p. 2976-2980.

SCHOEN, Harald et al. The power of prediction with social media. **Internet Research**, 2013.
SIFA, Rafet et al. Predicting purchase decisions in mobile free-to-play games. In: **Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment**. 2015.

SUN, Eric et al. Gesundheit! Modeling contagion through facebook news feed. In: **Third international AAAI conference on weblogs and social media**. 2009.

SOOD, Ashish; JAMES, Gareth M.; TELLIS, Gerard J. Functional regression: A new model for predicting market penetration of new products. **Marketing Science**, v. 28, n. 1, p. 36-51, 2009.

SUN, Zhan-Li et al. Sales forecasting using extreme learning machine with applications in fashion retailing. **Decision Support Systems**, v. 46, n. 1, p. 411-419, 2008.

SUNDSØY, Pål et al. Big data-driven marketing: how machine learning outperforms marketers' gut-feeling. In: **International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction**. Springer, Cham, 2014. p. 367-374.

TAUSCZIK, Yla R.; PENNEBAKER, James W. The psychological meaning of words: LIWC and computerized text analysis methods. **Journal of language and social psychology**, v. 29, n. 1, p. 24-54, 2010.

TEUNTER, Ruud H.; SYNTETOS, Aris A.; BABAI, M. Zied. Intermittent demand: Linking forecasting to inventory obsolescence. **European Journal of Operational Research**, v. 214, n. 3, p. 606-615, 2011.

TIMOSHENKO, Artem; HAUSER, John R. Identifying customer needs from user-generated content. **Marketing Science**, v. 38, n. 1, p. 1-20, 2019.

VAN LAER, Tom et al. What happens in Vegas stays on TripAdvisor? A theory and technique to understand narrativity in consumer reviews. **Journal of Consumer Research**, v. 46, n. 2, p. 267-285, 2019.

VENKATESH, Kamini et al. Cash demand forecasting in ATMs by clustering and neural networks. **European Journal of Operational Research**, v. 232, n. 2, p. 383-392, 2014.



VERHOEF, Peter; KOOGE, Edwin; WALK, Natasha. **Creating value with big data analytics: Making smarter marketing decisions**. Routledge, 2016.

VERHOEF, Peter C.; LEEFLANG, Peter SH. Accountability as a main ingredient of getting marketing back in the board room. **Marketing Review St. Gallen**, v. 28, n. 3, p. 26-32, 2011.

VILLARROEL ORDENES, Francisco et al. Cutting through content clutter: How speech and image acts drive consumer sharing of social media brand messages. **Journal of Consumer Research**, v. 45, n. 5, p. 988-1012, 2019.

WEDEL, Michel; KANNAN, P. K. Marketing analytics for data-rich environments. **Journal of Marketing**, v. 80, n. 6, p. 97-121, 2016.

WICKHAM, Hadley. Tidy data. **Journal of statistical software**, v. 59, n. 1, p. 1-23, 2014.

WINKLER, Hans-Jörg; RIEGER, Verena; ENGELEN, Andreas. Does the CMO's personality matter for web traffic? Evidence from technology-based new ventures. **Journal of the Academy of Marketing Science**, v. 48, n. 2, p. 308-330, 2020.

XIONG, Guiyang; BHARADWAJ, Sundar. Prerelease buzz evolution patterns and new product performance. **Marketing Science**, v. 33, n. 3, p. 401-421, 2014.

ZHANG, Qian et al. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In: **Proceedings of the 26th international conference on world wide web**. 2017. p. 311-319.

ZHANG, Xiaoming et al. Event detection and popularity prediction in microblogging. **Neurocomputing**, v. 149, p. 1469-1480, 2015.

ZHANG, Yongzheng; PENNACCHIOTTI, Marco. Predicting purchase behaviors from social media. In: **Proceedings of the 22nd international conference on World Wide Web**. 2013. p. 1521-1532.

ZHAO, Liang et al. Multi-task learning for spatio-temporal event forecasting. In: **Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining**. 2015. p. 1503-1512.

ZHOU, Feng; JIAO, Jianxin Roger; LEI, Baiying. A linear threshold-hurdle model for product adoption prediction incorporating social network effects. **Information Sciences**, v. 307, p. 95-109, 2015.