Master in Management + Analytics

Reaction Prediction: The Case of Tweets From Luxury Fashion Brands

Autoría: Calviello Crusella, Chiara
Año académico: *2023*

UNIVERSIDAD TORCUATO DI TELLA

Master in Management + Analytics

Master's Thesis

Reaction Prediction: The Case of Tweets From Luxury Fashion Brands

---

by

Chiara Calviello Crusella

Advisor: **Santiago Cisco**

July 18, 2023

Reaction Prediction: The Case of Tweets From Luxury Fashion Brands

Chiara Calviello Crusella[1]

Abstract

Social media platforms represent an essential tool for both consumers and marketers. Meanwhile, luxury fashion brands play a key role in fashion, one of the most important industries of the world economy. Despite assumptions to the contrary, social media platforms and luxury fashion brands do mix, especially in the recent time. Consequently, it is worth asking whether it is possible to predict the reaction a post will generate in the audience of luxury fashion brands. This new question is the one this thesis intends to answer. To do so, the concept of reaction is defined through a novel composite index that is created and named Tweet reaction overall score (TROS), which is one of the solid and relevant contributions this thesis makes. Then, several predictive models are implemented, based on a wide range of different learning algorithms. The results show that it is indeed possible to predict the TROS that a post on Twitter will obtain in the audience of luxury fashion brands the day it is posted.

[1]ccalviellocrusella@mail.utdt.edu. Santiago Cisco is thanked for his advice.

Predicción de reacción: el caso de Tweets de marcas de moda de lujo

Chiara Calviello Crusella[2]

Resumen

Las redes sociales representan una herramienta esencial tanto para consumidores como para vendedores. Mientras tanto, las marcas de moda de lujo juegan un rol fundamental en la moda, una de las industrias más importantes de la economía mundial. A pesar de los supuestos contrarios, las redes sociales y las marcas de moda de lujo sí se mezclan, especialmente en los tiempos recientes. En consecuencia, vale preguntarse si es posible predecir la reacción que una publicación generará en la audiencia de las marcas de moda de lujo. Esta nueva pregunta es la que esta tesis se propone responder. Para ello, se define el concepto de reacción mediante un novedoso índice compuesto que se crea y nombra puntaje general de la reacción al Tweet (TROS, por sus siglas en inglés), que es una de las relevantes y sólidas contribuciones que realiza esta tesis. Luego, se implementan varios modelos predictivos, basados en un amplio rango de diferentes algoritmos de aprendizaje. Los resultados muestran que sí es posible predecir el TROS que una publicación en Twitter obtendrá en la audiencia de las marcas de moda de lujo el día en que se publica.

[2]ccalviellocrusella@mail.utdt.edu. Se agradece a Santiago Cisco por su asesoramiento.

Contents

List of Tables

List of Figures

1. Introduction

1.1. Context and Motivation

Social media platforms represent an essential resource for consumers in their decision-making process, influencing each step of their journey; as well as for marketers to develop and maintain a close brand-customer relationship (Vinerean and Opreana, 2019). As Chen (2021) states, when used well, social networks can be an effective tool to improve brand value and establish strong connections. Meanwhile, fashion is considered one of the most important industries, as it represents a significant part of the world economy (Vinerean and Opreana, 2019). In this industry, luxury fashion plays a key role (Vinerean and Opreana, 2019). Luxury fashion brands are distinct. They are consumed differently and require specific marketing strategies (Bazi et al., 2020).

A wide assumption is that social networks represent a threat to high-end brands that dare to try them because of the following. The main characteristic of luxury brands is exclusivity. In contrast, the access to social networks is easy and open, and they help develop a mass appeal (Vinerean and Opreana, 2019). In fact, during the mid 2000s, due to the penetration of social media, there was an increase in the awareness level of luxury fashion brands, which made these brands worry about their credibility, exclusivity, and distinctiveness in the potential consumers' mind (Hemantha, 2020).

Despite this logical inconsistency between the exclusivity of luxury and the accessibility of social networks, the findings of Tack et al. (2020) support the fact that social media marketing is an important and cost-effective tool for luxury brands in the digital age. Actually, existing studies show that social networks help develop trust with a brand of this kind, and that they can have an important impact on the brand's success. For instance, a study found that a well-established social media strategy can have positive results in word of mouth marketing and distribution of the message, leading to increased online visibility (Dhaoui, 2014, as cited in Vinerean and Opreana, 2019). Furthermore, the marketing of these brands on social networks has been found to have a positive impact on consumers' favorable perceptions of luxury, the desire for it, and the purchase intentions. As a matter of fact, the social media presence of Burberry (one of the first luxury fashion brands to embrace social media) was analyzed and the results show that due to this presence, the company's profits increased by 39.8% (Phan et al., 2011, as cited in Vinerean and Opreana, 2019).

Consequently, luxury fashion brands are all about the experience and the online one also counts. Eastman et al. (2018) state that since luxury marketers must focus on the entire luxury experience, especially when marketing to young adults, social networks are a necessity. Young customers are the fastest growing segment of buyers of luxury brands and are strong followers of luxury brands on social networks (Bazi et al., 2020). Also, building a brand takes a lot of time and money. So, once one becomes an internationally well-known luxury brand, one must take care of that reputation.

Currently, there is a huge transformation in the luxury landscape. Many luxury fashion brands have entered social networks and many consumers are engaging online (Hemantha, 2020). As increasingly more luxury fashion brands start using social media to connect with their consumers, it is important to comprehend how luxury fashion brands can generate the best possible reaction in their audience. In order to do so, understanding what customers need, what they like, and what helps them meet their needs is a crucial issue (Zohourian et al., 2018). In addition, for

luxury brands, there is no single selling proposition: Desire is the priority of luxe consumers, and thus the advertisement must be done according to the desires of consumers (Hemantha, 2020). Therefore, in the case of luxury fashion brands, it is even more important to understand what was said above (how luxury fashion brands can generate the best possible reaction in their audience). Moreover, like Graziani et al. (2019) state, social media marketing strategies can be changed according to the estimated emotions triggered when posting content. Likes, comments, and shares provide information on the willingness of the public to participate, and can help brands better understand their audience (Zohourian et al., 2018) and help explore themes aligned with consumer expectations (Chen, 2021). As Cuevas-Molano et al. (2021) assert, knowing which characteristics of branded content create value for consumers helps to foster their engagement levels, through their interactions with the brands' posts, and this represents a crucial part of a company's social media marketing strategy.

1.2. Problem and Objective

Given the previously described context, this thesis' research question is whether it is possible to predict the reaction a post will generate in the audience of luxury fashion brands. The hypothesis is that it is indeed possible.

In this way, a supervised learning problem is addressed. For each observation of the predictor measurements, there is an associated response measurement. A model that relates the response to the predictors is the one that wants to be fitted, to accurately predict the response for future observations, and to better understand the relationship between the response and the predictors. Regarding the success criteria, the baseline is considered to be predicting the historical median reaction associated with the brand author of the post.

Finally, it is worth anticipating that, as for the type of posts to analyze, the focus is on Tweets. This is due to the ease to access Twitter data (at the moment this project began), the continuous rising of Twitter user numbers, and the fact that this platform's largest audience share corresponds to a new core group of luxury consumers.[3] Therefore, this thesis tries to predict the reaction that a post *on Twitter* will generate in the audience of luxury fashion brands. Meanwhile, data from Google Trends is used to instead create some more attributes to better predict the aforementioned.[4]

1.3. Relevance for Business

By developing the proposed model, this thesis helps solve the business problem of how to design a post to maximize a desired reaction. In other words, the development of this model enables a luxury fashion brand, aiming to generate a specific reaction in its audience, to know better in advance what characteristics a post should have in order to boost that specific wanted reaction. In such manner, the extracted knowledge helps the decision-maker validate a post before publishing it and, thus, adapt and improve the post, as well as the marketing strategy in general.

Additionally, this contribution takes place in a context where research on social media strategy of luxury fashion brands is scarce (Oliveira and Fernandes, 2022; Vinerean and Opreana, 2019). Despite the increase in luxury brands on social media, research into social media luxury marketing is limited (Athwal et al., 2018). Moreover, in spite of its relevance, only a few studies have

[3]This is explained more in detail in the unit of analysis and selection of brands' section.
[4]This is explained more in detail in the data collection's and the feature engineering's sections.

implemented a predictive model to forecast the engagement of a post (Al Rawashdeh, 2017), and little is known about how different post-criteria influence different levels of engagement with social networks (Cuevas-Molano et al., 2021).

Furthermore, the dependent variable refers to the concept of reaction, and it is defined as a composite index named Tweet reaction overall score (TROS), made up of different indicators related to likes, Retweets, and sentiment and emotion of replies and Quote Tweets.[5] This definition is new and more comprehensive (relative to existing ones in both the academic and business fields), can also be used by a brand to calculate other descriptive statistics like its average TROS in its Twitter profile which could, in fact, be a new key performance indicator (KPI), and by slightly adjusting only a few of its components, it can be molded to measuring a specific desired kind of reaction, as well as be adapted to other types of social media posts. Consequently, the TROS, created in this thesis, represents a relevant contribution for both academia and business.

1.4. Structure of the Thesis

In the following sections, first, the state of the art is described, regarding both domain and methodology. Second, the data are presented: the unit of analysis, the selection of the brands, the construction of the dependent variable, the data collection, the exploratory data analysis, and the feature engineering. Third, the used methodology is explained; specifically, regarding the performance metrics, the test error estimation technique, and the predictive models. Fourth, the obtained results are revealed and analyzed. Finally, the conclusions are stated, which are followed by the references.

2. State of the Art

Little research has addressed some important issues for luxury brands in comparison to non-luxury ones (Becker et al., 2018). As mentioned above, luxury brands were initially reluctant to adopt an online presence due to the potential hazards to their core values of exclusivity, scarcity, and uniqueness. Thus, research on social media engagement with luxury brands is scarce (Oliveira and Fernandes, 2022). Although social media communication strongly influences the image of a brand, academic work dedicated to the social media strategy of luxury fashion brands is limited (Vinerean and Opreana, 2019). Previous studies have rarely explored how different advertising strategies can increase the effectiveness of luxury advertising on social networks (Y. K. Choi et al., 2020). In addition, despite its relevance, only a few studies have developed a predictive model to forecast engagement of a post (Al Rawashdeh, 2017), and little is known about how the characteristics of different posts influence engagement (Cuevas-Molano et al., 2021).

2.1. Current Brands' Practices

Hemantha (2020) asks how luxury and fashion brands formulate strategies to maintain their distinctiveness on social media consistent with their brand philosophy. To obtain an answer, Hemantha (2020) performs a qualitative study: a content analysis of the top five luxury fashion brands according to the Brandwatch Q4 2019 Fashion Index on social media platforms[6]. Namely,

[5] This is explained more in detail in the dependent variable's section.
[6] https://www.brandwatch.com/brandwatch-index/top-fashion.

Nike, Hermès, Gucci, adidas, and Louis Vuitton.

Before moving on, it is worth pointing out the following. Although Hemantha (2020) considers Nike, Hermès, Gucci, adidas, and Louis Vuitton as all luxury fashion brands, Nike and adidas differ from the other three in terms that the former are more related to the sports world than the latter. In effect, Hemantha (2020) presents Nike and adidas as brands specialized in sportswear while, for instance, Louis Vuitton as a brand specialized in clothes and leather goods. In fact, there are those who consider that the differences are so significant that conceive Nike and adidas not as luxury fashion brands, but as fast fashion ones. One of them is, indeed, Chen (2021). Nonetheless, the categorization of brands like Nike and adidas into sportswear versus general wear brands or luxury versus fast fashion brands is controversial. Actually, the issue has recently become more blurred given some collaborations like the one among adidas and Gucci[7] (which then appears in the exploratory data analysis).

Going back to Hemantha (2020), this study was carried out from December 2019 to June 2020 and finds that these luxury brands have showcased their brand philosophy and heritage through storytelling, creative advertising campaigns, and social media events, featuring celebrities and virtual collections.

More in detail, Hemantha (2020) finds that the role of advertising in luxury brands is to sell not just a product, but dreams, since the essence of luxury is to keep dreams alive by maintaining exclusivity and elite status. These desires are created through visual storytelling, while working with global celebrities in the advertisements. In addition, luxury consumers buy products irrespective of economic situations, and luxury communication is done through events and shows exclusive to a few selected customers. Furthermore, due to the increased usage of mobile devices by Generations Y and Z, luxury brands have ventured into social media space by incorporating, for instance, virtual fashion shows. Finally, Hemantha (2020) states eight factors that are key to the sustainability of luxury brands: high quality, distinctiveness, status, exclusivity, history, timelessness, feeling good factor, and experiential.

In contrast to Hemantha (2020), first, this thesis' research question is whether it is possible to predict the reaction that a post will generate in the audience of luxury fashion brands. Second, in this thesis, not only a descriptive, but also a predictive and a prescriptive analysis are carried out. Third, the number of brands studied is expanded. Fourth, as Hemantha (2020) suggests for future research, text mining and sentiment analysis are used for these types of brands, as well as suitable metrics to analyze the posts they make on social networks.

Meanwhile, Vinerean and Opreana (2019) present the concept of luxury brands and evaluate the marketing practices of luxury fashion brands on Instagram. More specifically, first of all, they state that a luxury brand represents a branded product or service that consumers perceive to be of high quality, that offers authentic value via desired benefits (whether functional or emotional), has a prestigious image within the market, is worthy of commanding a premium price, and is capable of inspiring a deep connection with the consumer.

Second, Vinerean and Opreana (2019) explain that, in an effort to bring the brand closer to its audience (made up of current, potential, and aspirational customers), luxury fashion brands develop different practices as part of their Instagram marketing strategy. Namely, stories to present ads of their products; live videos to present their fashion shows; Instagram TV to showcase backstage accesses; Instagram Shopping to provide luxury shoppers the opportunity to explore

[7] https://www.gucci.com/us/en/ca/whats-new/adidas-x-gucci-c-adidas-gucci-products.

and shop different products they encounter on Instagram; highlights (i.e., different old stories grouped in thematic sections, located below the biographical section of the account) to showcase past runway shows, celebrities wearing the brand, various products, different collaborations with other brands, or store experiences in distinct cities; special augmented reality (AR) filters to give Instagram users the opportunity to interact with the brand in a fun way, while building brand awareness; and custom hashtags to generate higher volume of engaging posts.

Finally, Vinerean and Opreana (2019) describe that luxury brands, especially in the fashion sector, partner with celebrities to gain access to new markets. These brands contract them to post about the brand in exchange for some sort of payment. This is done since celebrities appeal to a common reference group[8], and their profiles are at the top of the list of the most-followed ones. Additionally, luxury brands work with *influencers*[9], who advertise their products among their base of followers and potential customers (Vinerean and Opreana, 2019).

Before continuing, it is worth mentioning the following. It could be said that the celebrities and influencers that luxury fashion brands work with nowadays tend to be younger than before. This can be related to one of the facts pointed out in this thesis' introduction: Young customers are the fastest growing segment of buyers of luxury brands and are strong followers of luxury brands on social networks (Bazi et al., 2020). In fact, Hemantha (2020) presents Gucci as an Italian luxury brand that has become a point of reference among youngsters. Another change that has happened among these brands' ambassadors is the use of avatars and virtual influencers or models, that are generated by a computer and have realistic human features and even personalities. For instance, in 2016, as the face of one of its campaigns in Japan, Louis Vuitton chose Lightning, a character from the video game Final Fantasy XIII, very well known in that country[10]. Also, Prada has worked with a pre-existing virtual influencer called Miquela[11]; Dior, with imma[12]; and CHANEL, with Bermuda[13]. Meanwhile, Balmain works with not only pre-existing but also own virtual influencers and models: In 2018, Balmain launched a campaign starring two models exclusive to the brand who are part of its "virtual model army" and a third one who is Shudu Gram[14], a pre-existing virtual model hired by the brand[15].[16]

Going back to Vinerean and Opreana (2019), the authors conclude that Instagram allows luxury brands to have a visual storytelling approach. More and more, it is considered the new destination for inspiration and a new form of "window shopping", as consumers tend to consult it especially in the discovery and consideration phases; and it is adopted by luxury brands to reach potential audiences in a creative way.

Unlike Vinerean and Opreana (2019), in this thesis, the focus is on Twitter and, as they suggest for future research, a quantitative marketing research with primary data is conducted, as well as qualitative research based on text mining and sentiment analysis, to collect new information

[8] A reference group consists of a person or a group of people that serve as a reference to an individual in forming values and attitudes, and in doing so, provides consumers with a reference in their purchasing decisions (Vinerean and Opreana, 2019).

[9] Compared to celebrities, influencers seem more personable, credible, and easy to relate to, due to their shared experiences and snippets of their life on social media (Vinerean and Opreana, 2019).

[10] https://www.youtube.com/watch?v=HxCr4q1lUa0.

[11] https://www.instagram.com/p/Bfi3sd9l3yX/.

[12] https://www.instagram.com/p/BwBCds8jl9V/.

[13] https://www.instagram.com/p/B-XsXx1jBUK/.

[14] https://www.instagram.com/shudu.gram/.

[15] https://projects.balmain.com/gb/balmain/balmains-new-virtual-army.

[16] https://vs-lb.com/virtual-models-meet-luxury-brands/.

from user comments.

Tack et al. (2020), like Vinerean and Opreana (2019), also examine the marketing activities on Instagram, but of a single luxury brand: Delvaux. To do so, they use a conceptual framework made up of five dimensions: entertainment, interaction, trendiness, customization, and word of mouth. Then, based on a survey among 195 luxury consumers, they develop a structural equation model to explore how each of these five dimensions affects customer brand equity and purchase intentions. Contrary to Tack et al. (2020), in this thesis, the focus is on Twitter and a survey is not conducted.

Continuing with Instagram, L. Liu et al. (2018) introduce a "visual listening in" approach to measure how brands are portrayed on social networks, by mining visual content posted by users on Instagram. They study 56 brands in the apparel and beverage categories.

In the first stage, L. Liu et al. (2018) build and examine image classifiers to predict whether a particular brand attribute (glamorous, rugged, healthy, and fun)[17] is expressed in a given image. They use two supervised machine learning methods: support vector machine (SVM) classifiers and deep *convolutional* neural networks (NNs). The latter achieves better out-of-sample prediction accuracy, but the former provides easier-to-interpret insights. To train the classifiers, they gather an annotated training set from Flickr, an online photo-sharing website that provides a search engine that returns the most relevant photos for a keyword, based on text labels provided by users, image content, and clickstream data. All classifiers outperform the benchmark (i.e., 50% accuracy by randomly guessing, since the training set is balanced); the NNs fine-tuned from the Flickr style model perform the best across all perceptual attributes, as well as on average; and the accuracy of this classifier is high for this type of prediction task.

In the second stage, L. Liu et al. (2018) apply the image classifiers of the first stage to the images on Instagram created by consumers (and *hashtagged* with the name of the brand) and by the brands themselves. The authors create metrics, derived from those images, that allow brands to compare how they are portrayed on social media relative to competitors. They compute the ratio of the brands' images that express the perceptual attribute, to capture the brands' image portrayed by consumers on social networks. This is closely related to usage context and consumption experience of brands. Besides, they compute the proportion of brand-created images that are classified as positive on an attribute, which captures part of the firms' marketing efforts to create their brand identities. Finally, they also get a brand perception measure from a large national survey, to capture the perception of a nationally representative sample of consumers. To compare these brand attribute metrics, they conduct two empirical studies. First, they observe, given a pair of brands, which one is more associated with a certain attribute. They find that brand image portrayed on social media reflects consumers' brand perception. Additionally, in the apparel category, they see high consistency in the glamorous, rugged, and fun attributes, which are key factors that make the difference for apparel brands. Second, they create maps of the brands in each product category, which are useful for seeing where the brands fall in the competitive landscape and identifying gaps in position strategies.

In contrast to L. Liu et al. (2018), this thesis' objective has to do with predicting reaction to a post, and its focus is on text and luxury fashion brands.

[17]They focus on intangible brand attributes, which go beyond functional ones. In categories such as apparel, where many brands offer products with very similar functionality, what usually makes a bigger difference is the feeling consumers have about the brand. Thus, positioning brands along intangible attributes allows themselves to differentiate from one another (L. Liu et al., 2018).

2.2. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is an active area of study in the field of natural language processing (NLP) that computationally treats subjectivity in text (Hutto and Gilbert, 2014). It implies a common text categorization task that is based on the extraction of sentiment: the negative or positive orientation that a writer expresses toward some object. Extracting consumer or public sentiment is relevant in fields ranging from marketing to politics. Generally, there are three classes: negative, neutral, and positive. Nonetheless, more classes are common for tasks like emotion detection (Jurafsky and Martin, 2021). In fact, in the case of categorical emotion detection, sentences are usually classified into six universal emotion classes: anger, disgust, fear, happiness, sadness, and surprise (Graziani et al., 2019).

Pantano et al. (2019) aim to develop an understanding of content generated online by consumers in terms of negative or positive comments to increase marketing intelligence, through a sentiment analysis based on machine learning. To achieve that objective, they collect and evaluate 9,652 Tweets referring to three fast fashion retailers of different sizes operating in the United Kingdom (UK), which have been shared among consumers and between consumer and brand, and posted in February 2018. To perform sentiment analysis of the collected Tweets, they use an unsupervised machine learning already available in Wolfram Mathematica and choose the Classify previously-trained function. They find different amounts of negative, neutral, and positive Tweets for the three retailers.

In this thesis, when constructing the dependent variable, consumers online generated contents are also considered, by taking into account users' replies and Quote Tweets to brands' original Tweets[18]. However, in contrast to Pantano et al. (2019), this thesis' main unit of analysis is a brand's original Tweet, and its focus is on luxury fashion brands[19].

Y. Choi et al. (2021), like Pantano et al. (2019), also conduct a sentiment analysis. Y. Choi et al. (2021) investigate the perception and evaluation of consumers of the Big 4 Fashion Weeks (New York, London, Milan, and Paris), held in 2019. They study Tweets about these four fashion weeks and perform three steps. First, they identify the keywords that appeared in the Fall-Winter 2019 Fashion Week through a social network analysis. Second, they make a city-wise examination of the themes and topics associated with the collections through topic modeling. Third, they analyze the sentimental evaluation of the brands that participated in these four fashion weeks through a sentiment analysis.

Some of their main results are the following. First, each city's fashion week reflects the city's characteristics. Most of the keywords that appear at London Fashion Week are related to British fashion brands and designers; Tweets from Milan Fashion Week feature more references to the materials and patterns used in the produced clothes; the frequently mentioned keywords are in line with the idea that Milan Fashion Week tends to feature collections that combine both New York's practicality and Paris' creativity; external factors such as fashion bloggers and their social media activities are influential in Paris Fashion Week; and unlike the results of the other three, the top keywords mentioned in Paris Fashion Week are mostly related to fashion brands rather than influencers. Second, similar design inspirations, collection themes, and brands correspond to the same topic. Third, the most popular item calculated is the "Monogram handbag" from Michael Kors' New York Fashion Week collection. Four, fashion brands and designers can popularize

[18]This is explained more in detail in the dependent variable's section.
[19]This is explained more in detail in the unit of analysis and selection of brands' section.

their products by leveraging social media influencers (Y. Choi et al., 2021).

Compared to Y. Choi et al. (2021), this thesis has a different objective that is not exclusively focused on fashion weeks. Furthermore, in addition to analyzing negative, neutral, and positive sentiment reactions, responses such as anger and sadness are studied[20], just as Y. Choi et al. (2021) suggest for future research.

2.3. Evaluation of Sentiment Lexicons

A sentiment lexicon refers to a list of lexical features, like words, labeled according to their semantic orientation (Hutto and Gilbert, 2014). In other words, a dictionary of opinion words with a semantic score. It focuses only on certain words: those that carry particularly strong sentiment cues. It is used to identify and determine the sentiment orientation of a piece of text as negative, neutral, or positive (Dhaoui et al., 2017; Hasan et al., 2018; Jurafsky and Martin, 2021).

Hasan et al. (2018) calculate sentiments with three lexicon-based analyzers named SentiWordNet, TextBlob, and W-WSD (Word Sense Disambiguation), and test their results with two supervised machine learning classifiers: Naive Bayes and SVM. To do so, they gather 100,000 Tweets, selecting hashtags of a political nature. They keep Tweets in English and Urdu (the national language of Pakistan) and translate the latter into English. They find that TextBlob and W-WSD are much better than the SentiWordNet approach to analyze election sentiments and to make more accurate predictions. Unlike Hasan et al. (2018), testing sentiment lexicons is not this thesis' objective, and its focus is on luxury fashion brands instead of politics.

Meanwhile, Dhaoui et al. (2017) use a sample of 850 consumer comments on 83 pages of Facebook luxury fashion brands to compare the performance of lexicon-based and machine learning approaches to sentiment analysis, as well as their combination.

Both approaches typically classify any given text into negative, neutral, or positive according to the polarity of the content. However, the lexicon-based approach relies on a sentiment lexicon. Dhaoui et al. (2017) study the lexicon named Linguistic Inquiry and Word Count (LIWC). In contrast, the machine learning approach uses a fraction of the full data as a manually classified training data set and trains classifiers to learn by examples, thus supervising the classification and without relying on any prior lexicon. This approach requires manual labeling of training examples, whose size and quality affect the performance of the trained model. High-quality labeling of a large training data set can be time-consuming, while limiting the size of the training data set leads to poorer classification accuracy. Dhaoui et al. (2017) test different machine learning algorithms, and the top two performing ones are maximum entropy modeling (Maxent), which uses a *multinomial* logistic regression, for positive sentiment and bagging, in which each tree is constructed from a bootstrap sample drawn with replacement from the training data set, for negative sentiment classification. These two then constitute what the authors call the machine learning approach.

Dhaoui et al. (2017) find that both approaches are similar in accuracy, achieving a higher one when classifying positive sentiment, but that the combined approach (a version that still requires manual classification of data) significantly improves the performance of classifying positive sentiment without penalizing the performance of classifying the negative. The relatively lower

[20]This is explained more in detail in the dependent variable's section.

performance of classifying negative sentiment has to do with the difficulty of analyzing sarcasm, which is often a limitation for manual approaches too.

Dhaoui et al. (2017) suggest applying the combined approach to other types of customer-generated content on social networks, such as Tweets. In this thesis, Tweets are worked with, but this combined approach is not applied, since the time nor the human resources to carry out the manual classification are available. Instead, the lexicon-based approach is applied[21], which is widely used in the marketing research community, since it does not require any previous processing or training of the classifier, and achieves a similar accuracy relative to the machine learning approach (Dhaoui et al., 2017).

2.4. Reaction Prediction

Hogg et al. (2013) apply a stochastic modeling framework to predict how followers of an advocate for a topic respond to the advocate's posts. Stochastic modeling is an approach to modeling user behavior. It is a probabilistic framework that represents each user as a stochastic process that transitions between states with some probability. The probability representation captures the uncertainty about individual actions. On Twitter, the states include visiting the site, seeing a post, and responding to it; while transitions represent dependencies between states, like the fact that responding to a post is conditioned on seeing it and being interested in it.

The authors concretely analyze Twitter posts related to initiatives appearing on the California November 2012 ballot[22]. They estimate the parameters of the model and use the model to predict how users will respond to posts about a specific topic. They find that both response prediction and classification are better when they account for transitions between user states involved in social media than when using a statistical regression based on overall activity. In other words, they demonstrate that a model that accounts for the likelihood of seeing posts and user's interest better predicts response than just using the user's activity. In this way, the response is conditioned by both interest and visibility (i.e., how many newer posts are above it on the user's list and how likely the user is to scan through at least that many posts) of the item. Therefore, a lack of response should not necessarily indicate a lack of interest. Failing to account for the visibility of items can lead to erroneous estimates of interest and influence (Hogg et al., 2013). Comparatively, this thesis also studies reaction to a post but applied to fashion luxury brands.

In the meantime, Al Rawashdeh (2017) aims to predict the engagement of users with Facebook, Twitter, and Instagram posts during the period before, during, and after Ramadan. Concretely, the research questions are four: Which social network application is the best for advertising, what is the best type of media for the post to attract users, when is the best time to publish the post for maximum engagement, and how to predict engagement before publishing the post. The study approach implies both qualitative and quantitative techniques. Some of the findings include that users are more active on working days than on weekends and that while beauty pages had the maximum engagement before and after Ramadan, celebrities' pages had so before Ramadan and fashion pages, during Ramadan. Although this thesis' research question is related (principally to the fourth one and indirectly to the third one), luxury fashion brands are analyzed,

[21]This is explained more in detail in the dependent variable's section.

[22]A ballot initiative, or proposition, is a political process that enables citizens of some states, like California, to place new legislation on the ballot. If the proposition wins the popular vote, then it becomes law (Hogg et al., 2013).

instead of specifically the Arab world.

Graziani et al. (2019) also study Facebook. They focus on Facebook posts paired with "reactions" of multiple users. They propose a neural model that is able to jointly learn to detect emotions and predict Facebook reactions. They use First-Order Logic (FOL) formulas to express how reactions are connected to emotion classes and vice-versa. Each class is associated to a predicate, whose truth degree is computed using a function. Then, they convert these FOL formulas into polynomial constraints and softly enforce them into the learning problem, thus tolerating some violations. The model is trained using posts that include reactions from Facebook pages of newspapers and text labeled with emotions from popular data sets. The results show that the tasks of emotion classification and reaction prediction can both benefit from their interaction. In comparison, in this thesis, also emotion is classified and reaction is predicted, but focusing on luxury fashion brands instead of on newspapers and on Twitter instead of on Facebook.

Continuing with Facebook, Chen (2021) constructs a fashion brand image model to determine key image cues in Facebook posts made between January 1, 2011 and December 31, 2019, both by luxury and fast fashion brands. The selected luxury fashion brands are CHANEL, Hermès, and Louis Vuitton; while the selected fast fashion brands are adidas, Nike, and Zara. The results suggest, among other things, that luxury fashion brands use key image cues in their fan page posts and that these cues significantly affect participation and behavioral responses in the form of likes, comments, and shares.

From the findings, Chen (2021) divides the cues used by the selected brands into two modules: the image cue module and the image and theme cue module. The first focuses on the behavioral response of comments and image cues in the content. Research shows a high correlation between information familiarity and memory recall. Thus, brand pages that continuously provide key information along with brand familiarity are more likely to evoke stronger recall and more easily trigger participation in the form of comments. Meanwhile, the second cue module combines likes, comments, and shares; highlights image and theme cues that generate public focus on and interest in the brand; and primarily aims to transform image cues into a unique brand personality.

The research questions of Chen (2021) are specifically two. The first is whether luxury and fast fashion brands use image cues in their posts to position themselves. The second is whether data analysis and machine learning techniques can be applied to public data to identify preferences and predict participation characteristics. This second research question could be seen similar to this thesis one, but the latter is different in several aspects, like the ones mentioned next. First, the focus is only on luxury fashion brands and the analyzed quantity is higher than three. Second, Chen (2021) considers only likes, comments, and shares as the indicators of reaction; while in this thesis other indicators are added, related to sentiment and emotion[23]. Third, in this thesis, Facebook posts are not considered; the focus is on Tweets.

More recently, Vassio et al. (2022) provide an experimental analysis of the time evolution of interactions with posts and develop an analytical model that captures the main aspects of user interactions on social networks. To do so, they monitor the posts of Italian influencers (with at least 10,000 followers on June 1, 2021 and from different categories, such as politicians, musicians, and athletes) on Facebook and Instagram, from January 1, 2016 to June 1, 2021. Their experimental analysis shows, among other things, that followers tend to be more active later in the evening with respect to influencers. Meanwhile, their proposed model is able to predict a

[23]This is explained more in detail in the dependent variable's section.

post's popularity and shows considerable improvements over a simpler baseline. In contrast to Vassio et al. (2022), in this thesis, first, the focus is on the category of luxury fashion brands. Second, not influencer-generated but firm and user-generated posts are studied, and these posts can come from people all over the world, not necessarily Italy. Third, neither Facebook nor Instagram, but Twitter is analyzed.

2.5. Convenient Posts' Features

X. Liu et al. (2021) use big data to investigate the impact of social media marketing activities of a luxury brand on consumer engagement (CE). They analyze 3.78 million Tweets, retrieved from a 60-month period extending from July 2012 to June 2017, from the top 15 luxury brands with the highest number of Twitter followers. These brands are: CHANEL, Marc Jacobs, Burberry, Dior, Louis Vuitton, Dolce & Gabbana, Gucci, Saint Laurent, Versace, Michael Kors, Armani, Christian Louboutin, Ralph Lauren, Valentino, and Alexander McQueen. Their results show that focusing on the entertainment, interaction, and trendiness dimensions of a luxury brand's social media marketing efforts significantly increases CE; while focusing on the customization dimension does not. They suggest further studies to examine a more comprehensive and diverse sample of luxury brands. Keeping that in mind, in this thesis, a shorter period of time is analyzed, but for more brands, chosen not based on their number of followers on Twitter, but on several other sources[24]. Additionally, not only CE but also sentiment and emotion are measured[25].

Meanwhile, Ratnakumar (2021) explores what kind of content banks exactly need to post and at what hour and day. More specifically, the research question is what kind of post characteristics drive CE on the Sri Lankan bank's Facebook and Instagram profile pages. To answer this question, Ratnakumar (2021) constructs eight different types of regression models. Comparatively, this thesis' research question is similar but applied to luxury fashion brands. In addition, its focus is on Twitter instead of on Facebook and Instagram.

At the same time, Cuevas-Molano et al. (2021) seek to answer which characteristics of branded content create value for consumers to foster their engagement levels through their interactions with a brand's post. They perform a statistical content analysis on the social media engagement on the Instagram pages of 14 Spanish brands, belonging to 10 different sectors, from December 1, 2019 to January 31, 2020. The sectors are: Automotive, Cosmetics, Energy, Financial, Gambling, Government, Retail, Technological, Telecommunications, and Travel. The brands are: Caixabank, DGT, El Corte Inglés, Garnier, Ikea, Lidl, Mapfre, Nautalia Viajes, Seat, ONCE, Orange, Repsol, Samsung, and Vodafone. These 14 brands are extracted from a ranking of the brands with the greatest interaction on Spanish social networks and the highest advertising investment.

They develop ordinary least squares (OLS) linear regressions, based on the uses and gratifications (U&G) and personality systems interactions (PSI) theories. The former is a sociological and psychological approach for understanding people's motivations to actively seek and use specific media to satisfy specific needs. Its underlying assumption is that people are actively involved in media usage. Meanwhile, the latter explains how people interact with traditional media, such as television, where several resources are used to intensify perceived interactivity, such as subjective camera angle and fixation of visual and verbal directions toward viewers (Cuevas-Molano et al., 2021).

[24]This is explained more in detail in the unit of analysis and selection of brands' section.
[25]This is explained more in detail in the dependent variable's section.

The authors find that the guidelines for improving likes differ from those suggested to increase comments. The results reveal that videos with sound, carousel posts with multiple photos, and posts with hashtags achieve higher levels of engagement in terms of likes. In contrast, graphics and interactive content that involves voting, contests, and questions reach higher engagement with respect to comments. Additionally, they find a low influence of temporal factors that prevent them from making recommendations related to time (Cuevas-Molano et al., 2021).

Comparatively, the research question of Cuevas-Molano et al. (2021) is similar to the problem addressed in this thesis, but the latter is different in various aspects, like the following. First, instead of focusing on Spanish brands and covering kind of unrelated sectors, in this thesis the focus is on internationally renowned brands and only the luxury fashion sector is covered. Second, not only engagement, but also sentiment and emotion are measured[26]. Third, the selected social network is Twitter, instead of Instagram. Four, the number of units of analysis is expanded, exactly as Cuevas-Molano et al. (2021) suggest, since this is one of their main limitations ($n = 680$).

Continuing with Instagram, Romão et al. (2019) study how interactions on various social networks influence the number of likes on that platform. Using posts published by a small Portuguese shoe and bag luxury brand named Josefinas between September 1, 2015 and October 31, 2016 on Facebook, Instagram, Twitter, Google+, and Pinterest; they build an SVM model, fed with features related to the brand's social networks, product characteristics, and visibility in external media such as magazines.

In order to extract useful information from the SVM model and see the influence of each of the features on the number of Instagram's likes, Romão et al. (2019) conduct a sensitivity analysis. This is a technique that assesses how much is the output variation when the input features are assorted through their range of possible values. Concretely, Romão et al. (2019) perform a data-based sensitivity analysis (DSA), which extracts a randomly selected subset from the original training data set and changes simultaneously each of the input features through the possible values within the subset to assess output variation. DSA is less computationally demanding when compared to varying all features, while it also addresses input features' influence on each other (Romão et al., 2019).

The authors find that there is not a single feature that generates greater visibility on Instagram. However, they identify two features that stand out the most, which are the number of products and the Facebook video views. Regarding the number of products, the number of likes has the tendency to decrease as the number of products presented on the post increases above nine, whereas a number of products below three has the opposite effect. Meanwhile, in terms of the number of views from Facebook videos, the number of likes on Instagram tends to decrease as the number of visualizations increases (Romão et al., 2019).

Compared to Romão et al. (2019), this thesis' objective is different. Additionally, they suggest future research to study luxury brands with larger dimensions and to employ text mining and sentiment analysis to extract additional knowledge from users' comments published on social networks, which both are done by this thesis.

Lastly, Zohourian et al. (2018) ask themselves what makes a post popular, what features most affect the audience's sentiments and result in achieving a lot of attention and admiration, and what aspects should entities take into consideration in order to upload content that is more effective.

[26]This is explained more in detail in the dependent variable's section.

To answer these questions, they collect images and videos from three Iranian Instagram business accounts and apply different regression methods (namely, linear regression, local polynomial regression, SVM, and linear SVM) to predict the Popularity Score (i.e., the number of likes, divided by the number of followers). They categorize this score into three labels (namely, High, Medium, and Low) testing K-nearest neighbors, random forest, naive Bayes, and decision trees. They conclude that local polynomial regressions and decision trees are the algorithms that end up outperforming the others. Even though their questions are kind of similar to the problem this thesis tries to solve, the focus is on luxury fashion brands instead of on Iranian businesses, and reaction instead of popularity is predicted (the latter is based on likes so it is a subset of the former, which includes likes, comments, sentiment, and emotion[27]).

2.6. Recommendations for the General Strategy

de los Santos (2009) analyzes how the field of public relations has evolved and focuses on the area of luxury branding. She analyzes the successes and failures of communication methods in case studies from the high-end areas of the fashion, automotive, travel, and hospitality industries. She emphasizes the importance of customization and experience for the luxury audience and concludes her study with some more recommendations on how public relations practitioners can be more effective to their respective audiences.

The author states that luxury brands are inaccessible but not impossible. They are extremely high priced, but still available to those who can afford them. Luxury is about how consumers feel when using it and how it differentiates them from everyone else because the experience is unique as their own. According to de los Santos (2009), there are six features to luxury: heroic myth, exquisite product, iconic communication, carefully engineered celebrity, ultra-selective distribution, and the cool power.

Additionally, de los Santos (2009) considers that consumer loyalty, rather than awareness, becomes much more critical in the branding strategy. She suggests that the following three ideas should be kept in mind: Consumers must believe that the brand has an extraordinary history, that the product or service has genuine value, and that it reflects forward thinking. Moreover, the author states that brand managers must develop both a product reputation and secure consumer loyalty if their product or service is to succeed in the luxury realm. Managers must work to provide and maintain a relationship between the brand and its consumers, so that a heritage can be created and added to the brand's longevity. Furthermore, the brand needs to be innovated and refreshed at different intervals, so that it evolves with the consumer. Finally, the psychology of the consumer must be kept in mind. The brand must connect with the consumers' psychology, which, in turn, encourages them to purchase since they will be in-tune with the most important aspect of luxury: the emotional, experiential realm (de los Santos, 2009).

de los Santos (2009) concludes that communicators need to master social networks and their audiences to reach them the most effectively. This thesis aims to take a step in that direction.

Moving on, M. Park et al. (2020) ask whether a high level of brand-consumer engagement is always beneficial for luxury brands and how social networks can backfire in the context of these brands. They conduct three studies whose results imply that luxury fashion brands should maintain psychological distance on social media to protect the perceptions of brands' core values. Overly active and friendly brand-consumer engagement on social media may backfire them because

[27]This is explained more in detail in the dependent variable's section.

consumers may perceive them to be too accessible and approachable to everyday consumers. Consequently, these brands must sustain the myth and the dream of luxury, selectively engage with consumers, and only follow a certain group of them, like high-profile celebrities or artists. However, a high level of brand-consumer engagement can offer potential positive outcomes, such as word of mouth and increased brand awareness. So, these brands have to weigh the benefits of actively engaging with consumers against the cost of reducing core value perceptions of themselves.

Finally, Godey et al. (2016) investigate how social media marketing efforts influence brand equity and consumer behavior toward luxury brands. They select five of them (specifically, Burberry, Dior, Gucci, Hermès, and Louis Vuitton) and conduct a survey of 845 consumers of luxury brands from China, France, India, and Italy who follow these five brands on social media. Italy and France represent traditional luxury markets, while China and India have rapidly growing luxury consumer populations who only more recently have gained access to these kinds of goods. Based on this survey, they construct a structural equation model and state, according to their findings, as a general recommendation (although the study shows differences in the results between the four consumer cultures examined), that brands should seek to promote content that is entertaining, current, and likely to stimulate engagement and interaction on their social media sites. This thesis studies how to do so; that is, what specific characteristics the posts should have.

2.7. Reasons for Consumers Buying Luxury Goods

Wang (2022) defines "luxury" as expensive and exclusive products and brands that are differentiated from other offers based on their exquisite design and craftsmanship, sensory appeal, and distinct sociological and cultural narratives. From these unique features of luxury, Wang (2022) distinguishes three types of competencies: expertise in design and workmanship, aesthetic taste, and sensitivity to luxury's symbolism. Consumers employ them when they judge, purchase, and use luxury products and brands (Wang, 2022).

The author explains that although luxury products are expensive and exclusive and their possession signals wealth, achievement, and success[28]; this perspective does not sufficiently explain the realities of contemporary luxury consumption. First, luxury is moving from class to mass. Luxury is considered much less exclusive and elitist today and less associated with high status. In addition, consumers can also rent or lease products on online sites at a much lower cost. Thus, the wealth signal has been diluted. Second, luxury today comprises a wide range of goods and services, including "affordable luxuries", like perfumes and accessories. Third, the display of expensive possessions is no longer in fashion. Therefore, according to Wang (2022), the wealth-based perspective must be supplemented with the competency-based perspective. Luxury consumers not only use luxury to signal their accumulated wealth, they are also motivated to spend resources on learning and enjoying the unique features of luxury products. When consuming luxury, consumers develop and use the three aforementioned luxury competencies (i.e., expertise in design and workmanship, aesthetic taste, and sensitivity to luxury symbolism), which emphasize different consumption goals; considerations, choice, and usage; and consumption outcomes relative to wealth-based competencies. Namely, when consumers engage in wealth-based

[28]In fact, Vinerean and Opreana (2019) consider that the main reasons for purchasing luxury brands extend beyond functionality: Customers acquire luxury brands to gain exclusivity, status, and prestige. They add that luxury brands are status symbols that have a profound psychological value for consumers. Buying a luxury brand is a highly involved consumption experience that is strongly congruent with a person's self-concept (Vinerean and Opreana, 2019).

luxury consumption, they seek extrinsic status-related social rewards; they are mostly focused on conspicuous products and brands, their usage is ownership-focused; and the nature of the consumer-brand relationship is one-sided or hierarchical, resulting in reverential influence. On the contrary, when consumers engage in competency-based consumption, they seek intrinsic benefits by learning about and enjoying luxury features; their consumption is more inconspicuous, their usage is experience-focused; and the consumer-brand relationship is a mutual or equitable partnership, resulting in social influence by persuasion (Wang, 2022).

Wang (2022) adds that luxury consumption is usually a mix of both types along a continuum, that it is also shaped by situation and context, and that it depends on consumer learning and the level of knowledge about luxury. Furthermore, the author presents individual and societal moderating factors to establish which consumers pursue wealth-based and competencies-based consumption and under what conditions they follow more or less one or the other. For instance, achieved status is more likely to lead to wealth-based consumption, whereas high status or class endowed at birth facilitates competencies acquisition; and as power distance decreases in society, elites are seen as less legitimate, and society is structured more equally, consumers engage more in competencies-based rather than wealth-based consumption. Finally, Wang (2022) advises luxury managers to develop competency-based strategies to address contemporary luxury challenges.

This thesis considers the fact, stated by Wang (2022), that users cooperate and compete with companies to communicate and recommend products, and therefore the need for companies to manage the brand images and impressions created by these users. This is done by incorporating user-generated posts (namely, replies and Quote Tweets) and data of online searches into the analysis.

Meanwhile, de los Santos (2009) explains that people become luxury consumers for different reasons. There are those who see luxury as functional: They believe that it serves a purpose in their lives; it is a necessity. Others see it as a reward and purchase the product because they believe they deserve it. Also, there are those who simply give in to indulgence: They are luxury consumers because they want to be so and because it makes them feel good, and they usually make spur-of-the-moment purchases. Each of these reasons requires different types of messages, so knowing the target audience is of paramount importance (de los Santos, 2009).

The author adds that there are four different types of luxury clients. First, people who show: They are status driven and do whatever it takes to satisfy their perceived audience; purchasing luxury is an indicator of their personal success. Second, people who cannot be shown up: They are much more reserved and look to luxury not for exuberance but for confidence building. Third, people who show that they know: They are status driven like the first ones, but they differ in the fact that they take pride in having sufficient knowledge about what they are buying; they believe that their knowledge of the brand rationalizes its acquisition. Lastly, people who know: They are fascinated by luxury simply because it is luxury, they generally do not care about what others think, and they feel a genuine connection to brands (de los Santos, 2009).

Moving on, Dubois et al. (2021) synthesize the latest advances in the psychology behind luxury consumption. They review how biological, psychological, and structural factors drive the desire for luxury. Additionally, they propose that the psychology of luxury consumption is governed by a set of tensions between what luxury means to the self and the external forces that define luxury consumption. These tensions shape consumer behavior: from the level of desire for luxury, to the types of signals viewed as luxury and acquired and displayed as such, and to post-consumption consequences of consuming luxury. Knowing which are these factors and tensions and how they

operate improves marketing strategies.

Lastly, Becker et al. (2018) try to understand the meanings consumers bring to their lives when they are involved in social relationships with luxury brands. To do so, they aim to articulate a definition of luxury brands, to propose a framework for consumer luxury brand relationships, and to provide empirical evidence of the proposed model. They conduct two surveys, one in Lisbon and Porto, Portugal and another one in Boston, the United States of America (USA), in 2009 and 2013 respectively, for 13 well-known luxury brands. These brands are: Mercedes, BMW, Audi, CHANEL, Christian Dior, Gucci, Burberry, Calvin Klein, Hugo Boss, Armani, Ray Ban, Moët et Chandon, and Ralph Lauren.

They employ factor analysis and structural equation modeling techniques to test their hypotheses and build a luxury brand model (which they name the BECKER luxury brand model) that illustrates how the characteristics of the luxury product combine with the psychological characteristics of the consumer, to create the relationship between the luxury brand and the consumer. It has three stages, forming a pyramid with a foundation on which other characteristics of the luxury brand are based. In other words, they find a link between luxury products and consumers' psychological association in a hierarchical order. On a basic level, consumers develop a perception of quality that is combined with the aesthetics and the price of a luxury product, likely inducing satisfaction. At a secondary level, consumers use luxury products as an extrinsic signal of high social status and association with specific social groups, relying on the exclusivity and extraordinary physical characteristics to create self-connection. At the highest level, consumers who use luxury products become more intrinsic and spiritual; the product is likely associated with increasing symbolic features for consumers' self-image and self-identification. As luxury products represent higher degrees of desirable attributes, consumers express higher levels of psychological bonds, such as commitment, loyalty, and intimacy with the luxury product or brand (Becker et al., 2018). Knowing this luxury brand model can improve marketing strategies.

2.8. Reasons for Users Engaging With Luxury Brands on Social Media

Jahn et al. (2013) discuss the relevance of social networks for luxury brands and study how social media brand pages affect the relationship between the brand and the customer. Specifically, they present a general framework that describes how brand pages can contribute to customer brand loyalty and how participation on the brand page is influenced by various consumer values. They believe there are three main motivation areas for consumers' using social networks. First, a relationship area, where the focus of the individual is to stay connected and interact with others. Second, a content acquisition and distribution area, based on the individuals' interests; this content can be functional or hedonistic. Third, a self-presentation area, which is related to the social context, but also serves the purpose of self-assurance and personal identity.

To test their framework in a field environment, Jahn et al. (2013) conduct a Facebook survey to which members of different fan pages of luxury and non-luxury brands are invited, and test the proposed hypotheses using a structural equation model. All the coefficients, except two, of the proposed model are highly significant. They show as a central result of their study that social networks can be seen as a business opportunity. They conclude that fan pages are an excellent tool for brand management today and brand managers should embrace this channel and understand how to work with it in a contemporary fashion. The critical factor is not the number of fans, but the level of interaction. This thesis focuses on the level of interaction by studying how a brand's post affects users' reactions.

Meanwhile, Bazi et al. (2020) ask themselves why customers engage with luxury brands on social media platforms. To obtain an answer, they employ a qualitative approach. They carry out 25 semi-structured interviews with consumers, between 18 and 35 years old, of different nationalities and occupations, that follow luxury brands on social media, and study what motivates them to engage with these brands, using thematic analysis. The findings reveal 13 motivations that Bazi et al. (2020) group into six dimensions: perceived relevance of content (brand news, post quality, and celebrity endorsement), brand-customer relationship (brand love and brand *ethereality*), *hedonic* (entertainment), aesthetic (design appeal), psychology (actual self-congruence, status signaling, and enhance and maintain face), brand equity (perceived brand quality), and technology factors (ease of use and convenience).

In the meantime, Athwal et al. (2018) focus on Millennials (i.e., Generation Y), a new core group of luxury consumers. The authors examine what gratifications Millennials seek and obtain when following and connecting with luxury brands on social media. They use a two-stage data collection method: online observations and in-depth interviews. They collect data from the Facebook, Instagram, and Twitter accounts of the top five fashion brands according to brand value (namely, Louis Vuitton, Gucci, Hermès, Cartier, and Tiffany & Co.) and conduct 30 in-depth interviews with Millennials. To analyze the data and present the findings, they employ thematic analysis, like Bazi et al. (2020) do. The analysis reveals that two types of need are sought by social media users who follow and connect with luxury brands: emotional and cognitive. Concretely, the social media marketing activities of luxury brands satisfy two primary types of emotional needs: aesthetic appreciation and entertainment. Meanwhile, users are also able to satisfy their cognitive needs as they acquire, process, and share information. In addition to emotional and cognitive gratifications, users obtain interrelated gratifications, such as escapism and passing time. Athwal et al. (2018) conclude that it might be useful to conduct a content analysis of users' posts and comments on luxury brands' social media accounts. This thesis does so by incorporating to the analysis replies and Quote Tweets regarding the brands' original Tweets.

More recently, Oliveira and Fernandes (2022), based on data collected from a multi-national sample of 243 followers of luxury brands on Instagram, try to understand the drivers and results of CE on Instagram. They conclude that both consumer participation and brand self-expressiveness significantly impact social media engagements with luxury brands, which in turn predict outcomes such as brand image and loyalty.

2.9. Segmentation of Luxury Brands' Followers

Ramadan et al. (2018) conduct 24 in-depth interviews with Lebanese followers of an online luxury brand's social media pages to understand the different types of online luxury followers and the strategies needed to engage with them. They identify six main categories: pragmatists, bystanders, trend hunters, image seekers, passionate owners, and prime consumers. They assume these categories are both exhaustive and mutually exclusive.

Each group has specific engagement and propensity-to-buy levels. That is why the authors consider that one size does not fit all and state that a broad marketing strategy might fail to fulfill the needs and wishes of all the set of audiences within the luxury consumer base. Specifically, pragmatists are people who do not care about owning luxury brands. They have the means to buy luxury items but they are unlikely to buy them unless they believe these items will be highly practical to them. They focus on functional attributes. Consequently, this group can be targeted by highlighting the functional benefits of these luxury brands, which generates increased sales, as

this group has the means to purchase luxury products (Ramadan et al., 2018).

Bystanders are passive followers: They usually remain silent. They find satisfaction in observing these products from the sidelines, given their inability to buy them. This segment should mainly be used to spread positive word of mouth, which enhances the brand image and attracts more followers and potential clients. Contrarily, trend hunters are more active followers. They believe luxury brands are trendsetters and, since they cannot afford them, they hunt for similar but more affordable items. They show minimal regard for the actual brand quality, authenticity, and performance. This segment should be shifted toward buying the original luxury brands. One of the strategies to do so is to design online marketing campaigns in a way that leads users to the physical stores, where they can experience the brands' environment and product quality (Ramadan et al., 2018).

Image seekers love luxury brands, which they associate with success and status. They consider these brands to be part of their self-improvement process, they are the most concerned ones with displaying prestige and impressing their peers, and they treasure the intangible value as much as the functionality of the products. This segment should be targeted using emotional product appeals that communicate the brand's values and image, and by emphasizing the brand's role as a statement of success and social status. Meanwhile, passionate owners have a very strong passion for specific luxury brands. They enter into a relationship with the brand before purchasing it and, thus, purchase it to ensure self-satisfaction. They treat luxury products as sensory rewards; symbols to be attained at times of certain life events and transitions depicting the fulfillment of a certain goal or aspiration. Brands should work on shifting these customers from one-time buyers to regular ones, by establishing strong emotional ties and maintaining communications with them (Ramadan et al., 2018).

Lastly, prime consumers have the means to afford luxury products to the extent they do not categorize these brands as luxurious. They regularly buy and use luxury brands and they are satisfied with and loyal to them. They feel reassured that luxury brands offer superior levels of performance and quality. They feel a sense of security when purchasing a familiar high-functioning brand. Brand marketing managers must ensure that these clients are satisfied so that their buying patterns do not change and reaffirm the brand's value and quality in their communications with this segment. Additionally, as they are recurrent buyers, the brand has a lot of information regarding their buying behavior, allowing it to cater to the needs of each of them more effectively (Ramadan et al., 2018).

Moving on, J. Park et al. (2011) analyze which characteristics of social networks influence loyalty to luxury brands. They conduct field research aimed at subjects in their 20s and 30s who have experience with both luxury brands and social networks. They analyze the 331 responses using descriptive statistics, exploratory factor analysis, reliability analysis, cluster analysis, one-way analysis of variance (ANOVA), and multiple regression analysis.

They divide social networks' users into four lifestyle groups: active leisure-oriented, ego-expression, fashion leader, and early adopter. They find significant differences by user lifestyle in the relationship between social networks' characteristics and brand loyalty. Namely, the active leisure-oriented group is made up of young high-income earners, and its members use social networks to check others' information and socialize. They perceive all social networks' characteristics as highly important. Members of the ego-expression group enjoy updating their pictures and sending personal messages, and are relatively open to divulging private information. The fashion leader group members have self-oriented and appearance-oriented values and, contrary

to ego-expressionists, they stress privacy. They strongly relate reliability, which results from their tendency to protect their privacy. The early adopter group has no aversion to digital equipment and actively uses social networks. All groups, with the exception of the ego-expression one, show that social networks' characteristics have significant effects on brand loyalty (J. Park et al., 2011).

J. Park et al. (2011) conclude that their study is significant in analyzing the lifestyle groups of users of social networks to predict consumer behavior toward the social networks of luxury brands and in providing essential data to establish basic data target marketing strategies through consumer segmentation.

2.10. Variation Across Cultures

Different cultures perceive and use luxury brands differently (Bazi et al., 2020). Eastman et al. (2018) research how cultural variables influence the desire to purchase luxury fashion from young adults in the USA. To do so, they examine college-age consumers from the USA in two studies and young adults between the ages of 18 and 35 also from the USA in another one. The results of the three studies indicate that consumption of status has a positive impact on the intention to purchase luxury fashion, that cultural variables (specifically, collectivism, uncertainty avoidance, power distance, and masculinity) mediate this relationship, and that the bandwagon effect (i.e., the tendency for people to adopt certain behaviors just because others are also doing so) has a significant moderating impact in the same relationship for some cultural variables (namely, uncertainty avoidance, long-term orientation, and power distance).

The purchasing intentions of young adults from the USA for luxury fashion are more influenced by the need to fit in, the bandwagon effect, than by the need to stand out, the Veblen effect (i.e., when the demand of a good increases as the price increases, resulting in an upward-sloping demand curve, due to the good's exclusive nature and appeal as a status symbol) or the snob effect (i.e., when the demand for a certain good by individuals of a higher income level is inversely related to its demand by those of a lower one, because of the desire to own unusual and unique goods). The need to fit in influences luxury fashion purchase to reduce uncertainty and meet short-term needs for gratification; respondents want to positively compare with others while enhancing their view of self (Eastman et al., 2018).

Additionally, young adults from the USA buy luxury fashion items to reduce uncertainty, meet short-term needs for gratification, compare positively with others, and improve their view of themselves. Lastly, they prefer codes (for instance, the red soles on Christian Louboutin shoes) to logos, as the former are subtler ways of conveying the brand's identity. This has to do with the following. They tend to view themselves in a more positive light when it comes to discussing status symbols, but judge peers in a more negative light for owning the same status items. This favors the idea of private consumption versus public consumption and suggests less of a need for young adults to consume for status as a means of standing out. Therefore, they want to show status but in a discrete way, so that others do not think negatively of them in the way they would do so of others (Eastman et al., 2018).

Furthermore, the authors mention that other researchers suggest that East Asian young adult consumers are more driven than Western consumers to buy status products to conform, prefer fashion clothing brands from the West, and hold more positive attitudes and are willing to pay more for status fashion brands. Eastman et al. (2018) conclude that, even within one country, the cultural variables can play a role in mediating the relationship between motivation and purchase intention.

Finally, Y. K. Choi et al. (2020) ask how contemporary consumers evaluate luxury advertising on social networks and conduct three empirical studies in Korea, the USA, and Germany. They find that consumer perception of psychological distance associated with luxury brands influences whether benefit-based or attribute-based appeals are most effective, that this effect depends on the consumers' attitude functions toward luxury consumption, and that it varies across cultural contexts. More specifically, benefit-based rather than attribute-based message appeals are more effective for luxury brand advertising; consumers who have value-expressive, rather than social-adjustable, attitudes toward luxury consumption respond with greater purchase intentions when luxury brand messages use benefit-based, rather than attribute-based, appeals; and cultural context can moderate the positive congruence effect that occurs when benefit-based luxury appeals are matched with value-expressive attitudes.

It is worth adding that although consumers in different regions of the world purchase luxury products and engage with luxury brands on social media for a variety of reasons, giving rise to different segments of luxury brands' followers; the characteristics of luxury brand consumers appear to possess similar values regardless of their country of origin (Becker et al., 2018). In fact, in this thesis' train set[29], 33 of the brands sampled[30] have preferred to set their Twitter profile location to `Worldwide` or another phrase related to it like `Global` or `The World`, instead of setting it to a specific city or country. This shows that for many luxury fashion brands making their location explicit is not important, probably because they are international, just like their audience. Therefore, this thesis allows itself to consider the audience of luxury fashion brands as a whole and the predictive models chosen as able to identify the differences inside it if they are relevant to improve the prediction. In addition, at the time of analyzing the results, attention is paid to whether any of the previously described segments can be identified.

2.11. Brief State of the Art Summary

Several dimensions have been the ones analyzed in the literature review: beginning with the current brands' practices; following with techniques regarding sentiment analysis, lexicons, and reaction prediction; continuing with convenient posts' features; and finishing with works more of a psychological type, specifically about recommendations for the general strategy, reasons for consumers buying luxury goods and for users engaging with luxury brands on social media, segments of luxury brands' followers, and variation across cultures.

Hemantha (2020) analyzes the strategies on social media of luxury fashion brands. Vinerean and Opreana (2019) do so as well, but focusing on Instagram. Meanwhile, Tack et al. (2020), like Vinerean and Opreana (2019), also examine the marketing activities on Instagram, but of a single luxury brand. Continuing with Instagram, L. Liu et al. (2018) measure how brands are portrayed on social networks, by mining visual content posted by users on Instagram. Unlike these authors, not only a descriptive, but also a predictive and a prescriptive analysis are carried out in this thesis.

Moving on to sentiment analysis, Pantano et al. (2019) implement one for Tweets referring to three fast fashion retailers in the UK. Y. Choi et al. (2021) also conduct a sentiment analysis on Tweets, but about the Big 4 Fashion Weeks held in 2019. In contrast to these studies, the sentiment analysis that is carried out in this thesis is focused on luxury fashion brands, while not exclusively on the Big 4 Fashion Weeks.

[29]This is explained more in detail in the train, validation, and test sets' section.
[30]This is explained more in detail in the unit of analysis and selection of brands' section.

As to sentiment lexicons in particular, Hasan et al. (2018) evaluate three of the available ones on Tweets about politics. Dhaoui et al. (2017) also evaluate lexicons, but as an approach against the one based on machine learning; and considering Facebook pages, instead of Tweets. Contrary to these authors, testing sentiment lexicons is not this thesis' main objective.

Regarding reaction prediction, Hogg et al. (2013) aim to predict how followers of an advocate for a topic respond to the advocate's posts, applied to Tweets related to the California November 2012 ballot. In the meantime, Al Rawashdeh (2017) aims to predict the engagement of users with Facebook, Twitter, and Instagram posts during the period before, during, and after Ramadan. Graziani et al. (2019) also study Facebook: They aim to predict Facebook reactions. And, continuing with Facebook, Chen (2021) studies whether participation characteristics can be predicted by analyzing image cues in luxury and fast fashion brands' posts. More recently, Vassio et al. (2022) propose a model to predict a post's popularity, based on posts on Facebook and Instagram of Italian influencers. Comparatively, in this thesis, reaction to a post is also aimed to be predicted, but applied to fashion luxury brands and Twitter.

Passing onto convenient posts' features, X. Liu et al. (2021) investigate the impact that the social media marketing activities of a luxury brand on Twitter have on CE. Meanwhile, Ratnakumar (2021) analyzes what kind of post characteristics drive CE on the Sri Lankan bank's Facebook and Instagram profile pages. At the same time, Cuevas-Molano et al. (2021) study which characteristics of branded content create value for consumers to foster their engagement levels, considering the Instagram pages of 14 Spanish brands belonging to different sectors. Continuing with Instagram, Romão et al. (2019) study how interactions on various social networks influence the number of likes on Instagram, but they focus on a single small luxury brand. Lastly, Zohourian et al. (2018) investigate what makes a post popular, but only consider three Iranian Instagram business accounts. As opposed to these studies, in this thesis, the analyzed sector is exclusively luxury fashion, more brands are considered, Twitter is the social network selected, and when measuring reaction not only CE but also sentiment and emotion are taken into account.

Moving on to the works more of a psychological type, de los Santos (2009) gives recommendations on how public relations practitioners on luxury branding can be more effective. Then, M. Park et al. (2020) recommend luxury brands to weigh the benefits of actively engaging with consumers against the cost of reducing core value perceptions of themselves. Finally, Godey et al. (2016) state that luxury brands should promote content that is entertaining, current, and likely to stimulate engagement. This thesis studies how to do so; that is, what specific characteristics the posts should have.

As to reasons for consumers buying luxury goods, Wang (2022) distinguishes three types of competencies consumers employ when purchasing these goods. Meanwhile, de los Santos (2009) gives different reasons for people becoming luxury consumers and establishes four different types of luxury clients. Also, Dubois et al. (2021) review how biological, psychological, and structural factors drive the desire for luxury. Lastly, Becker et al. (2018) build a luxury brand model that accounts for the relationship between a luxury brand and a consumer.

Now, as to reasons for users engaging with luxury brands on social media, Jahn et al. (2013) state that there are three main motivation areas for consumers' using social networks: relationship, content acquisition and distribution, and self-presentation. Then, Bazi et al. (2020) find 13 motivations behind users engaging with these brands. In the meantime, Athwal et al. (2018) consider that emotional, cognitive, and interrelated gratifications are obtained by users when following and connecting with luxury brands. More recently, Oliveira and Fernandes (2022) try

to understand the drivers of CE related to luxury brands, but particularly on Instagram.

Regarding segments of luxury brands' followers, Ramadan et al. (2018) identify six main categories: pragmatists, bystanders, trend hunters, image seekers, passionate owners, and prime consumers. In contrast, J. Park et al. (2011) divide social networks' users into four lifestyle groups: active leisure-oriented, ego-expression, fashion leader, and early adopter.

Finally, concerning variation across cultures, Eastman et al. (2018) research how cultural variables influence the desire to purchase luxury fashion and suggest differences between Western and East Asian young adult consumers. Similarly, Y. K. Choi et al. (2020) ask how contemporary consumers evaluate luxury advertising on social networks and conduct comparative studies in Korea, the USA, and Germany.

3. Data

3.1. Unit of Analysis and Selection of Brands

The focus is on Twitter posts, known as Tweets. This decision is based on the ease to access Twitter data relative to other social networks (at the moment this project began), as well as on the fact that Twitter user numbers continue rising. In the second quarter of 2022, the number of active *monetizable* daily users increased by more than 15% compared to the same quarter of the previous year. Also, its largest audience share corresponds to users between the ages of 25 and 34 (Dixon, 2022), who are among a new core group of luxury consumers (see Athwal et al., 2018).

It is worth adding that Tweets only in English are studied, since this is the main language used by luxury fashion brands for carrying out their marketing strategies (indeed, the fact that Italian or French brands generally post in English is an indicator of that) and since the machine learning techniques for NLP are nowadays better developed for the English language. Furthermore, like Pantano et al. (2019) point out, focusing only on English Tweets avoids possible issues emerging from the analysis of multilingual posts.

Regarding the selection of the brands' accounts, Vinerean and Opreana (2019) state that different studies distinguish between eight luxury product types. Namely, fashion, jewelry, cosmetics, wine, automobiles, hotels, tourism, and private banking. In this thesis, the focus is on the first one: fashion. Therefore, despite jewelry perhaps being considered as fashion by some people, the aforementioned classification is followed and, thus, luxury jewelry brands are excluded and, instead, the focus is on clothing, shoes, and bags luxury brands.

Furthermore, it is taken into consideration that the Fortune magazine considers an account active if at least one publication has been made in the last 30 days (Muñoz et al., 2022). At the time of the selection of the brands, this meant considering Twitter accounts that had posted at least one Tweet since August 2022.

Several sources were used to find luxury fashion brands that matched those criteria, including the luxury version of the Brandwatch Q4 2019 Fashion Index[31], three Harvard Dataverse data sets related to luxury fashion brands[32] [33] [34], the academic articles reviewed, and the list of fashion

[31] https://www.brandwatch.com/brandwatch-index/luxury-fashion.
[32] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BCOSKY.
[33] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/K7AW6F.
[34] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8BNXES.

brands that can be found on the Harrods website[35], the world's leading luxury department store.

The selected brands ended up being 100. Namely, Acne Studios, Alexander McQueen, Anya Hindmarch, Aquazzura, Armani, Axel Arigato, Balenciaga, Bally, Balmain, Barbour, Belstaff, Blumarine, Burberry, Canada Goose, Carolina Herrera, CELINE, CHANEL, Chloé, Christian Louboutin, Coach, Dior, Dolce & Gabbana, DSQUARED2, ELIE SAAB, ERDEM, Etro, Ettinger, Fendi, FERRAGAMO, Fusalp, Georges Hobeika, Gianvito Rossi, Givenchy, Globe-Trotter, Goyard, Gucci, HELMUT LANG, Hermès, Herno, J.Crew, Jean Paul Gaultier, Jenny Packham, Jimmy Choo, Johnstons of Elgin, Judith Leiber, KARL LAGERFELD, KENZO, Lafayette 148 New York, LANVIN, LOEWE, Longchamp, Louis Vuitton, Maison Margiela, Manolo Blahnik, Marchesa, Margaret Howell, MARNI, Marysia Swim, Max Mara, MCM, Michael Kors, Missoni, Miu Miu, Moncler, Monique Lhuillier, Moschino, Mugler, Mulberry, Needle & Thread, Oscar de la Renta, Paco Rabanne, PAIGE, Paul Smith, Prada, Proenza Schouler, PUCCI, Ralph Lauren, René Caovilla, RIMOWA, Roger Vivier, Saint Laurent, Smythson, Sportmax, Stella McCartney, Stuart Weitzman, Temperley London, Theory, Thom Browne, Tod's, TOM FORD, Tory Burch, TUMI, UGG, Valentino, Vera Wang, Versace, Victoria Beckham, Vivienne Westwood, Yohji Yamamoto, and ZIMMERMANN.

3.2. Dependent Variable

The dependent variable refers to the concept of reaction. Various are the ways in which users can react to a post (Hogg et al., 2013) and, thus, one can identify different indicators.

In fact, diverse measures have been used in the existing literature. Romão et al. (2019) consider the number of likes. Hogg et al. (2013) choose to focus on Retweets and admit that their particular definition of response is somewhat arbitrary. Vassio et al. (2022) measure the number of likes, but state that several studies on online social networks have analyzed the popularity of content as a function of the total number of interactions, measured at the time the data were collected, and the authors add that the number of likes metric could be complemented or substituted by the number of shares or comments. As a matter of fact, Chen (2021) analyzes Facebook and focuses on likes, shares, and comments. And Cuevas-Molano et al. (2021) measure the number of likes and the number of comments separately, and divide each of these numbers by the count of followers, to include the community size of each brand fan page and since it is the formula used to measure engagement on social networks in the professional practice. Zohourian et al. (2018) also consider likes and the size of the community, by measuring what they call the Popularity Score, which is just the normalized number of likes, by dividing it by the number of followers. As they want to perform classification methods on their data, they need to make the labels discrete, and they do this by categorizing the Popularity Score into three labels (namely, High, Medium, and Low), thus creating the Popularity Class.

Meanwhile, as a formula for the sentimental evaluation of a brand, Y. Choi et al. (2021) use the number of positive Tweets of the brand divided by the total number of Tweets of that same brand. Hasan et al. (2018) measure sentiment using three polarity classes: negative, neutral, and positive. They determine the polarity of each Tweet by assigning a score from -1 to 1 based on the words used, where a negative score means a negative sentiment, a positive score means a positive sentiment, and a value equal to zero means a neutral sentiment. Lastly, Graziani et al. (2019), for each short input text, output a probability distribution on the possible reactions and

[35] https://www.harrods.com/.

another on the possible classes of emotions, and they select the reaction-emotion pair associated with the highest probabilities.

Furthermore, the ratio is another metric that is often mentioned[36]. Instead of being a business or an academic metric, it is more informal and more linked to people's experience on social media. It indicates how many comments a post receives compared to likes and *reposts*. The mathematical formula is the number of comments divided by the sum of the number of likes and the number of reposts. By definition, likes indicate that other people view the Tweet in good light, and Retweets are generally endorsements, too, unless the user instead turns it into a Quote Tweet writing something sarcastically. On the contrary, comments, even though they might support a Tweet, usually debate or criticize the Tweet. Consequently, if comments exceed likes plus Retweets, that is a sign of getting a negative reaction. In contrast, a ratio of 0.5 means that there is a balance between comments, likes, and Retweets, showing that the Tweet is pretty solid and that the comments tend to be neutral or even positive (Salamander, 2018).

However, in this thesis, the ratio is not considered since, based on the way it is interpreted, it tries to approximately capture the sentiment of the responses, and this thesis already includes explicitly the sentiment, which is thought to be more precise, as it considers the actual content of the reply or Quote Tweet. Recall that, in line with this decision, Romão et al. (2019) suggest future research to analyze the sentiment of users' comments published on the social networks' profiles of luxury brands, and Athwal et al. (2018) conclude that it might be useful to analyze the content of users' comments on luxury brands' social media accounts. Additionally, as it is later explained in detail, what the number of replies represents in terms of sentiment is controversial and so, in line with Romão et al. (2019) and Athwal et al. (2018), Al Rawashdeh (2017) suggests future work to analyze the content of the replies.

As shown, several are the possible indicators of reaction and many of them are important in one way or another. In fact, Muñoz et al. (2022) state the need to use a holistic view when measuring the reaction to a Tweet. So, how can these indicators be combined into one single metric?

Al Rawashdeh (2017) considers the total number of engagements in a post by adding the number of likes, comments, shares, and Retweets, and categorizes this variable into five categories: very low, low, moderate, high, and very high. The problem with this kind of aggregation is that not always more is better. More specifically, the number of replies is controversial. The fact that comments are made indicates a Tweet's capacity to generate a reaction in others and, thus, should be kept in mind when measuring its influence. Nonetheless, although the number of replies could be considered an indicator of engagement in principle, some authors are not as convinced. That is because the content of the replies would have to be analyzed to confirm that the reply represents a favorable reaction, as established in the definition of engagement (Muñoz et al., 2022). Indeed, Al Rawashdeh (2017) suggests future work to analyze the content of comments to improve the engagement and the prediction model. Due to all this, instead of considering the number of replies or Quote Tweets, in this thesis, their sentiment and their emotion are considered.

Meanwhile, Muñoz et al. (2022) construct a composite index to measure user engagement on Twitter using technique for order of preference by similarity to ideal solution (TOPSIS). This is a multi-criteria method that is based on minimizing the distance to an ideal point and maximizing the distance to an anti-ideal one (Muñoz et al., 2022), and it is the preferred one among other

[36]Ramiro H. Gálvez is thanked for pointing this out.

approaches (Mohamaddoust et al., 2021). Like in Mohamaddoust et al. (2021) and Muñoz et al. (2022), this is the aggregation method used in this thesis.[37]

Following Muñoz et al. (2022), let m be the number of original Tweets (in other words, Tweets from the brands) being studied and n the number of indicators or criteria (in this thesis' case, $n = 16$[38]). Let $A = A_1, A_2, ..., A_m$ be the set of original Tweets, and let $C_1, C_2, ..., C_n$ be the indicators with which they are evaluated. x_{ij} is used to denote the value of the original Tweet A_i with respect to indicator C_j. Finally, suppose that all indicators are of the more-is-better type[39]. The steps for this method are as stated in the next paragraphs. It is worth clarifying that the following explanation of each of the six steps is based on Muñoz et al. (2022).

First, normalize the value of each original Tweet for each indicator. To do it, Muñoz et al. (2022) use the L2 norm, also known as the Euclidean norm:

$$y_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}, \quad i = 1, 2, ..., m; \quad j = 1, 2, ..., n. \tag{1}$$

To prevent data leakage[40], the denominator must be calculated using only the observations that belong to the train set[41]. At the beginning, in this thesis, the normalization was done using the Euclidean norm. However, observing the train set, it was noticed that the final index value did not satisfactorily illustrate the original Tweets' situation: The final index value tended to be extremely low but the corresponding reaction was not absolutely appalling. Consequently, another very common normalization technique was tried: the min-max normalization. Its formula is

$$y_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}, \quad i = 1, 2, ..., m; \quad j = 1, 2, ..., n. \tag{2}$$

To prevent data leakage, the minimums and maximums must also be calculated using only the observations that belong to the train set. With this second normalization technique, all indicators turn to be in a range of 0 to 1. Observing the train set, it was noticed that the final index values had become much more reasonable. Additionally, those values were now less sensitive to the extreme values in the first two indicators. Therefore, in this thesis, to normalize the value of each original Tweet for each indicator, the min-max normalization is applied.

Second, establish a weight for each indicator such that $w_j > 0$, $j = 1, 2, ..., n$, $\sum_{j=1}^{n} w_j = 1$ (how to establish these weights is later discussed in this same section). Third, calculate the elements of the normalized matrix:

$$v_{ij} = w_j \cdot y_{ij}, \quad i = 1, 2, ..., m; \quad j = 1, 2, ..., n. \tag{3}$$

[37] Nevertheless, in contrast to Mohamaddoust et al. (2021) who apply it to charisma, here it is done to reaction; and contrary to Muñoz et al. (2022) who apply it to a group of Spanish social media influencers, here it is done to a group of Tweets posted by global fashion luxury brands.

[38] More on this later in this same section.

[39] This is taken into account when later defining the indicators.

[40] Data leakage consists on evaluating and selecting models including in them information that is not going to be available at the moment of production. When this happens, usually the performance of the proposed model is being overestimated. The key to avoid it is to validate the model in situations comparable to those that it is going to have in production.

[41] This is explained more in detail in the train, validation, and test sets' section.

Fourth, define two artificial original Tweet profiles: the ideal original Tweet, $\mathbf{v}^+$, which is assigned the best value for each indicator, and the anti-ideal original Tweet, $\mathbf{v}^-$, which is assigned the worst value for each indicator. Artificial because, generally, these two points are virtual alternatives: Very rarely is there an observation that is the best or the worst on all indicators. Given that all indicators want to be maximized[42], the best value for each is the maximum value, $v_j^+ = v_{ij}$, and the worst is the minimum, $v_j^- = v_{ij}$:

$$\mathbf{v}^+ = (v_1^+, ..., v_n^+), \quad \mathbf{v}^- = (v_1^-, ..., v_n^-). \tag{4}$$

To prevent data leakage, these maximum and minimum values are also calculated using only the observations belonging to the train set[43].

Fifth, for each original Tweet, calculate the weighted distance from the ideal and anti-ideal points, using a measure of distance, like the Euclidean distance:

$$D_i^+ = \sqrt{\sum_{j=1}^{n}(v_j^+ - v_{ij})^2}, \quad D_i^- = \sqrt{\sum_{j=1}^{n}(v_j^- - v_{ij})^2}, \quad i = 1, 2, ..., m. \tag{5}$$

It is verified that $0 \leq D_i^+, D_i^- \leq 1$. Additionally, an original Tweet is better the closer it is to the ideal point, and the further it is from the anti-ideal one. Consequently, one can identify an original Tweet's strength as its closeness to the ideal (distance to minimize) and its distance from the anti-ideal (distance to maximize).

Sixth, for each original Tweet, calculate the relative closeness coefficient, which is the distance to the anti-ideal divided by the sum of the distance to the ideal and the distance to the anti-ideal:

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-}, \quad i = 1, 2, ..., m. \tag{6}$$

The relative closeness coefficient is a value between 0 and 1. If an observation is close to the ideal and far from the anti-ideal, its distance to the ideal equals a value close to 0, and its distance to the anti-ideal equals a value close to 1, so the quotient is close to the value 1. Therefore, high values of this coefficient are preferable (Muñoz et al., 2022). In this thesis, this relative closeness coefficient is the metric to predict.

Muñoz et al. (2022) suggest performing all these calculations using the `topsis` package[44] based on R. To better understand and check how this package implements the TOPSIS method, its definition[45] was looked at. It was verified that the function includes the normalization step and that the input weights are divided by their sum[46]. Since, in this thesis, a different normalization technique is chosen and the minimums and maximums of Equations 2 and 4 must be calculated using only the observations from the train set to avoid data leakage, some adjustments are made to the previous definition and this adapted version is used instead of the `topsis` package.

Regarding the indicators, this thesis does not use the same ones as Muñoz et al. (2022). Taking into consideration the ones used in the existing literature and their pros and cons, the following

[42] Just for the sake of recalling, this is taken into account when later defining the indicators.
[43] This is explained more in detail in the train, validation, and test sets' section.
[44] https://cran.r-project.org/web/packages/topsis/topsis.pdf.
[45] https://rdrr.io/cran/topsis/src/R/topsis.R.
[46] This is the reason why the example shows values like 1 and 2, instead of between 0 and 1. More about it can be read at https://or.stackexchange.com/questions/8061/conflicts-with-weights-of-the-topsis-method-in-r.

are included.

First, the number of likes of the original Tweet A_i (L_i) divided by the number of followers of the brand author of A_i (F_i). It goes from 0 to $+\infty$. The worst value is 0, while the best one is $+\infty$. It is worth adding that making this division instead of considering only its numerator has to do with the fact that, on average, there is a linear dependency between the total number of likes received by a post and the author's current number of followers (Vassio et al., 2022).

Second, the number of Retweets of the original Tweet A_i (RT_i) divided by the number of followers of the brand author of A_i (F_i). It also goes from 0 to $+\infty$, the former being the worst and the latter being the best. It is worth clarifying that this indicator is not condensed with the previous one into a single component, in case the decision-maker would later want to assign them different weights.

Third, the median sentiment in the replies to and Quote Tweets of the original Tweet A_i, respectively represented by SRP_i and SQT_i. It is worth clarifying that the median refers to the value lying at the midpoint of the frequency distribution of (sorted) observed sentiment values such that there is an equal probability of the observed values falling above or below it. Additionally, it is worth mentioning that the median instead of the mean is calculated since, exploring the train set, it was found that for various original Tweets the mean and the median differ, indicating the presence of outliers, and since the median is more robust to those extreme values than the mean.

Much of the applied research that takes advantage of sentiment analysis is heavily based on pre-existing lexicons (Hutto and Gilbert, 2014). This thesis also bases on them. To calculate sentiment, the previously trained sentiment classifier named Valence Aware Dictionary and sEntiment Reasoner (VADER) is used. One of its outputs is the compound score, which is computed by summing the valence (i.e., intensity) scores of each word in the lexicon, adjusted according to some rules[47], and then normalized to be between -1 (the most negative) and 1 (the most positive). As a single measure of sentiment for a given Tweet is looked for, this is VADER's most useful metric for this study.

As an alternative, transformers[48] were looked into, especially BERTweet[49]. However, VADER is chosen for the following reasons. It is a rule-based model for sentiment analysis tuned to the text of social media, specifically to *microblog*-like contexts, such as Twitter, and it has been validated by humans. Additionally, it performs as well as (and in most cases, better than) other highly regarded sentiment analysis tools, it outperforms individual human raters, and it performs exceptionally well in the social media domain (Hutto and Gilbert, 2014). Moreover, León-Sandoval et al. (2022) compare VADER with BERTweet and RoBERTa and conclude that all models show similar trends and react similarly to real-world events, making all three good options for large-scale sentiment analysis. Furthermore, when compared to sophisticated machine learning techniques, VADER has several advantages. For instance, it is both quick and computationally economical, without sacrificing accuracy; and the lexicon and rules it uses are directly accessible, not hidden within a machine-access-only black-box. In fact, VADER is a gold standard. Besides, it is freely available for download and use, and its methodology is described in the original article so that everyone can see how it works (Hutto and Gilbert, 2014). Finally,

[47] More on this later in this same section.
[48] https://huggingface.co/docs/transformers/index.
[49] https://huggingface.co/docs/transformers/model_doc/bertweet.

it is available for not only Python, but also R[50].

It is worth adding that VADER implies five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. These five rules are the following. First, the exclamation point (!) increases the magnitude of the sentiment intensity without modifying the semantic orientation. Following the example given by Hutto and Gilbert (2014), "The food here is good!!!" is more intense than "The food here is good." Second, using ALL-CAPS to emphasize a sentiment-relevant word, in the presence of other non-capitalized words, also increases the magnitude of the sentiment intensity without affecting the semantic orientation. Continuing with the previous example, "The food here is GREAT!" conveys more intensity than "The food here is great!" Third, degree modifiers impact sentiment intensity by either increasing or decreasing it. For instance, "The service here is extremely good" is more intense than "The service here is good", while "The service here is marginally good" reduces the intensity. Fourth, the contrasting conjunction "but" signals a shift in the sentiment polarity, dominating the sentiment of the text after the conjunction. As an example, "The food here is great, but the service is horrible" has mixed sentiments and, in VADER, the latter half is the one that determines the overall rating. Fifth, VADER catches nearly 90% of cases where negation flips the polarity of the text, by examining the *trigram* preceding a sentiment-laden lexical feature. To clarify, a negated sentence would be "The food here isn't really all that great." (Hutto and Gilbert, 2014).

Due to these five rules, before entering the text into VADER, punctuation (especially exclamation points) are not removed, upper case characters are not converted into lower case characters, neither *lemmatization*[51] nor stemming[52] are carried out to avoid losing the degree modifiers, the word `but` as well as negated sentences are not treated since VADER already does so internally, and to avoid damaging that treatment stop words are not removed.

Additionally, before entering the text into VADER, the text is normalized; i.e., converting it to a more convenient standard form (Jurafsky and Martin, 2021). Concretely, spaced versions of words are replaced with their non-spaced versions, word elongations (like `whyyyyy`) with their standard versions, mixed text-numeric represented ordinal numbers with words (e.g., `1st` with `first`), and ratings like `five stars` with more common adjectives. Also, mentioned accounts, the hashtag symbol (`#`), and Uniform Resource Locators (URLs) are removed. Furthermore, emojis (but not emoticons since, as Hutto and Gilbert (2014) explain, VADER includes a full list of Western-style emoticons, like `:-)` which denotes a `smiley face` and generally indicates positive sentiment) are replaced with their word equivalents, non-American Standard Code for Information Interchange (ASCII) characters are discarded, and symbols used as abbreviations (such as `%` for `percent`, `w/` for `with`, and `&` for `and`) are changed into their word equivalents, while one or more white space characters as well as line jumps such as `\n` into a single space.

It is worth clarifying that contractions are not replaced with their expanded versions nor slang with standard words, since by the examples given by Hutto and Gilbert (2014) it can be said that VADER deals with contractions, and since it incorporates commonly used slang with sentiment

[50] https://cran.r-project.org/web/packages/vader/vader.pdf.

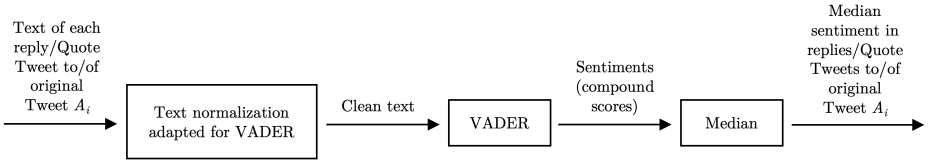
[51] Lemmatization refers to the task of determining that two words have the same root, despite their surface differences. For instance, the words `sings`, `sang`, and `sung` are forms of the verb `sing`. The word `sing` is the common lemma (i.e., a set of lexical forms having the same stem, major part of speech, and word sense) of these words, and a *lemmatizer* maps from all these to `sing` (Jurafsky and Martin, 2021).

[52] Stemming refers to a simpler version of lemmatization in which mainly suffixes are stripped from the end of the word (Jurafsky and Martin, 2021).

value (such as `nah` and `meh`).

Figure 1 summarizes the construction process of these two sentiment indicators.

Figure 1: Construction Process of Sentiment Indicators



Fourth and last, the median of each type of emotion in the replies to and Quote Tweets of the original Tweet A_i. To analyze emotion, this thesis uses the National Research Council of Canada Word-Emotion Association Lexicon (EmoLex)[53], which identifies eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.[54] Neither anticipation nor surprise are considered, given their ambiguous meaning.

So, for each of the emotions considered and for each reply to and Quote Tweet of the original Tweet A_i, the resulting number (i.e., the number of occurrences of words associated with that emotion) is divided by the total number of associations with any of the six analyzed emotions in that reply or Quote Tweet, thus getting a value between 0 and 1. In case the total number of associations with any of the six analyzed emotions in that reply or Quote Tweet is equal to 0, a 0 is directly assigned since it serves to indicate that the emotion in question is not present whatsoever, and since it is not possible to divide by 0.

Then, for each of the emotions considered, what is calculated is the median in the replies (thus obtaining $EARP_i$, $EDRP_i$, $EFRP_i$, $EJRP_i$, $ESRP_i$, and $ETRP_i$), as well as the median in Quote Tweets (thus obtaining $EAQT_i$, $EDQT_i$, $EFQT_i$, $EJQT_i$, $ESQT_i$, and $ETQT_i$). In this way, the median anger, disgust, fear, joy, sadness, and trust in both replies to and Quote Tweets of the original Tweet A_i are obtained. For these twelve indicators the median is calculated instead of the mean for the same reasons as for the two sentiment ones: Exploring the train set, it was found that for several original Tweets the mean and the median differ, which indicates the presence of outliers, and the median is more robust to those extreme values than the mean.

Depending on the emotion, that value wants to be maximized or minimized. Concretely, median joy and trust want to be maximized; while median anger, disgust, fear, and sadness want to be minimized. To maintain consistency with the previous assumption that all indicators are of the more-is-better type, the ratio for anger, disgust, fear, and sadness should be inverted, leaving a 1 in the numerator and the median in the denominator. However, the `topsis` function in the aforementioned `topsis` R package, as well as the adapted version used, include a parameter called `impacts`, to which one has to specify the way in which each criterion influences on the alternatives with the character or `"+"` or `"-"` (the former for maximizing and the latter for minimizing). Internally, these functions take this input into account when calculating Equation 4: For criteria to maximize, the best value is the maximum value and the worst is the minimum, while for criteria to minimize, the best value is the minimum value and the worst is the maximum.

[53] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

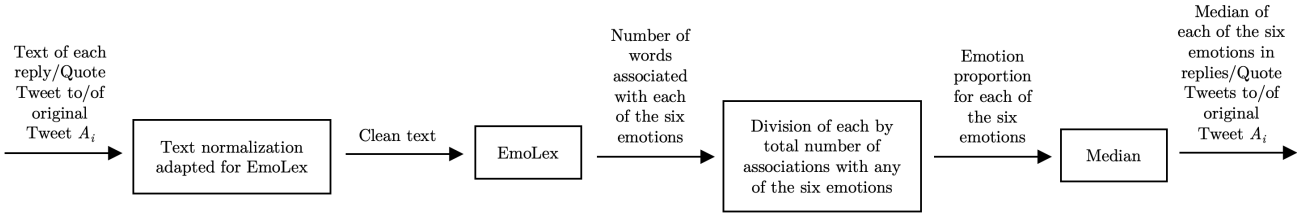
[54] An R implementation example is available at https://www.red-gate.com/simple-talk/databases/sql-server/bi-sql-server/text-mining-and-sentiment-analysis-with-r/.

Consequently, in practice, inverting the ratio is not necessary.[55]

Before inputting the text to EmoLex, the text is also normalized, but in a slightly different way than the one used for VADER. This is due to the fact that the text normalization process used for VADER implies some adaptations to VADER's own characteristics (like the ones regarding the emoticons, the contractions, and the slang), and some of these adaptations are no longer necessary when normalizing the text for EmoLex. So, for EmoLex, specifically, the spaced version of words is replaced with their non-spaced versions, word elongations with their standard versions, contractions with their expanded versions, slang with standard words, mixed text-numeric represented ordinal numbers with words, and ratings with more common adjectives. Also, mentioned accounts, the `#`, and URLs are removed. Furthermore, emoticons and emojis are replaced with their word equivalents, non-ASCII characters are discarded, and symbols used as abbreviations are changed into their word equivalents. Finally, upper case characters are converted into lower case ones, while one or more white space characters as well as line jumps into a single space; and lemmatization is carried out. Lemmatization is chosen over stemming because the former does the job more properly by using morphological vocabulary and analysis.

Figure 2 summarizes the construction process of these twelve emotion indicators.

Figure 2: Construction Process of Emotion Indicators



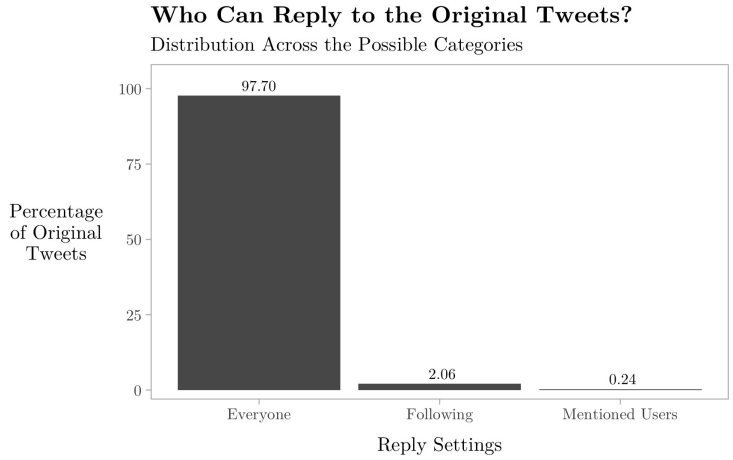
The relative closeness coefficient based on the 16 aforementioned indicators is what, in this thesis, is called the TROS (i.e., Tweet reaction overall score) and what this thesis tries to concretely predict. Mathematically, this composite index proposal can be expressed as follows:

$$\begin{aligned}\text{TROS} = C_i\Bigg(&\frac{L_i}{F_i}, \frac{RT_i}{F_i}, SRP_i, SQT_i,\\ &EJRP_i, ETRP_i, \frac{1}{EARP_i}, \frac{1}{EDRP_i}, \frac{1}{EFRP_i}, \frac{1}{ESRP_i},\\ &EJQT_i, ETQT_i, \frac{1}{EAQT_i}, \frac{1}{EDQT_i}, \frac{1}{EFQT_i}, \frac{1}{ESQT_i}\Bigg).\end{aligned} \tag{7}$$

It must be added that, observing the train set, it was discovered that, although most original Tweets can be replied to by everyone as Figure 3 illustrates, most of them have 0 replies and 0 Quote Tweets, like shown in Figure 4. Therefore, these Tweets would have `NA` (i.e., not available) as the value of the indicators related to sentiment and emotion, and this would impede the calculation of the TROS.

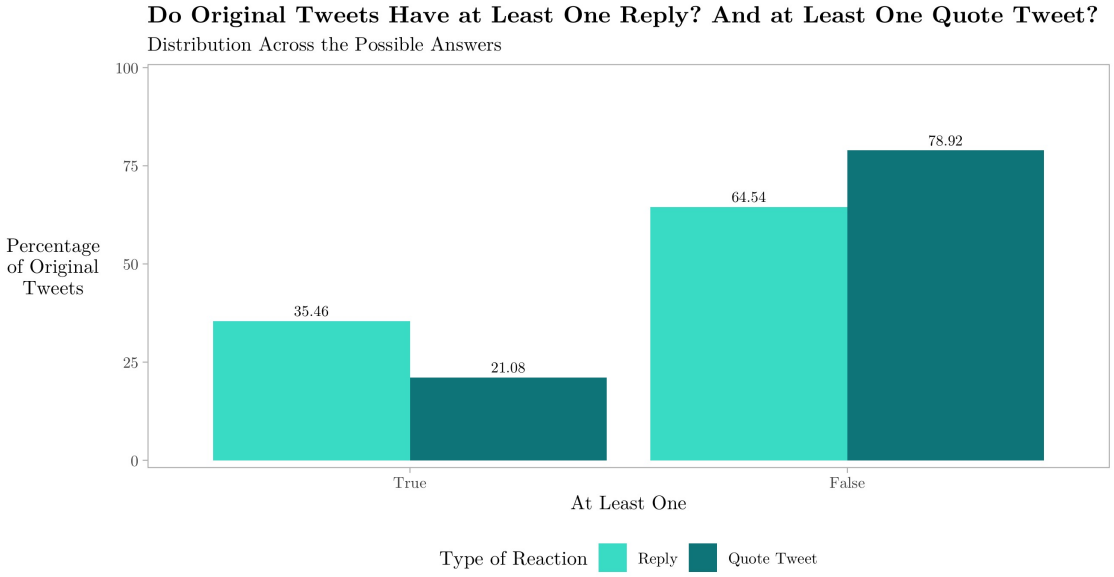
[55]Nonetheless, when next expressing mathematically the composite index, those ratios are inverted to visually state that those indicators should be minimized instead of maximized.

Figure 3: Distribution of Original Tweets Through Reply Settings



Note. This figure considers only the observations that belong to the train set.

Figure 4: Presence of Replies and Quote Tweets



Note. This figure considers only the observations that belong to the train set.

Regarding possible solutions, it is not correct to discard these observations because it can happen in real life that an original Tweet has 0 replies or 0 Quote Tweets, and because it would imply making an ex-post filter (i.e., no one knows with 100% certainty if an original Tweet is going to receive 0 replies or 0 Quote Tweets before it being published, but only after).

Why not consider only original Tweets to which everyone can reply to, as a possible solution, since this would represent an ex-ante, instead of an ex-post, filter? In Figure 3, it can be seen that 97.7% of the original Tweets (belonging to the train set) is set up so that everyone can reply to them. Meanwhile, in Figure 4, it can be seen that 35.46% and 21.08% of the original Tweets (belonging to the train set) have at least one reply or Quote Tweet, respectively. In consequence, to consider only Tweets that everyone can reply to, even though it is indeed an ex-ante filter, it would impact a very little number of original Tweets, together with the fact that everyone being able to respond does not necessarily imply that the original Tweet in question has obtained at least one reply. Therefore, to consider only original Tweets to which everyone can reply to does not address the problem in question.

In contrast, these indicators could be assigned for only this kind of observations one of the following options. First, a weight equal to 0. But, since this would imply that not all observations have their dependent variable assigned the same weights, this option is discarded. Second, a "neutral" score, like 0 for the two sentiment indicators, while 0.5 for the twelve emotion indicators. Third, the mean of those indicators (calculated using only the observations belonging to the train set[56], to avoid data leakage). Fourth, the median of those indicators (also calculated using only the observations belonging to the train set, to avoid data leakage).

Regarding the two sentiment indicators, looking at the train set, it was found that many of the rest of the observations have the neutral value (while none or very few have the mean or the median), as shown in Table 1, and thus the second option would make the differentiation between the two types of observation difficult: the ones that originally had `NA` and those that have always had the neutral score. Consequently, this second option is also discarded for these two indicators.

Table 1: Percentage of Original Tweets by Key Median Sentiment Values

	NA	Neutral	Mean	Median
Replies	64.55	7.95	0.00	0.49
Quote Tweets	78.93	5.98	0.00	0.32

Note. This table considers only the observations that belong to the train set.

Meanwhile, also looking at the train set, it was found that these indicators' distributions have some outliers. As shown in Table 2, for the median sentiment in both replies and Quote Tweets, the mean is lower than the median, indicating the presence of extreme values on the left side of the distribution. Thus, the third option is also discarded for these two indicators, since the mean is especially sensitive to those extreme values. Therefore, for the two sentiment indicators, the fourth option (i.e., assigning the median, calculated using only the observations belonging to the train set) is chosen.

Table 2: Presence of Outliers in Median Sentiment

	Mean	Median
Replies	0.290	0.361
Quote Tweets	0.290	0.318

Note. This table considers only the observations that belong to the train set.

In contrast, with regard to the twelve emotion indicators, looking at the train set, it was found that many of the rest of the observations have the median (while none or very few have the neutral score or the mean) as shown in Table 3, and thus the fourth option would make the differentiation between the two types of observation difficult: the ones that originally had `NA` and the ones that have always had the median. Consequently, this fourth option is also discarded for these twelve indicators.

[56]This is explained more in detail in the train, validation, and test sets' section.

Table 3: Percentage of Original Tweets by Key Median Emotion Values

	Emotion	NA	Neutral	Mean	Median
Replies	Joy	64.54	5.40	0.00	0.52
	Trust	64.54	3.71	0.00	17.98
	Anger	64.54	0.24	0.00	30.51
	Disgust	64.54	0.24	0.00	32.16
	Fear	64.54	0.72	0.00	29.05
	Sadness	64.54	0.58	0.00	29.66
Quote Tweets	Joy	79.92	3.42	0.00	10.59
	Trust	79.92	2.85	0.00	11.45
	Anger	79.92	0.15	0.00	19.02
	Disgust	79.92	0.14	0.00	19.54
	Fear	79.92	0.40	0.00	17.55
	Sadness	79.92	0.35	0.00	18.24

Note. This table considers only the observations that belong to the train set.

Meanwhile, also looking at the train set, it was found that these indicators' distributions also have outliers. As shown in Table 4, for all median emotions in both replies and Quote Tweets, the mean is higher than the median, indicating the presence of extreme values on the right-hand side of the distribution. Thus, the third option is also discarded for these two indicators, since the mean is especially sensitive to those extreme values. Therefore, for the twelve emotion indicators, the second option (i.e., assigning the neutral score of 0.5) is chosen.

Table 4: Presence of Outliers in Median Emotion

	Emotion	Mean	Median
Replies	Joy	0.256	0.200
	Trust	0.178	0.000
	Anger	0.030	0.000
	Disgust	0.024	0.000
	Fear	0.049	0.000
	Sadness	0.044	0.000
Quote Tweets	Joy	0.237	0.000
	Trust	0.187	0.000
	Anger	0.024	0.000
	Disgust	0.017	0.000
	Fear	0.058	0.000
	Sadness	0.043	0.000

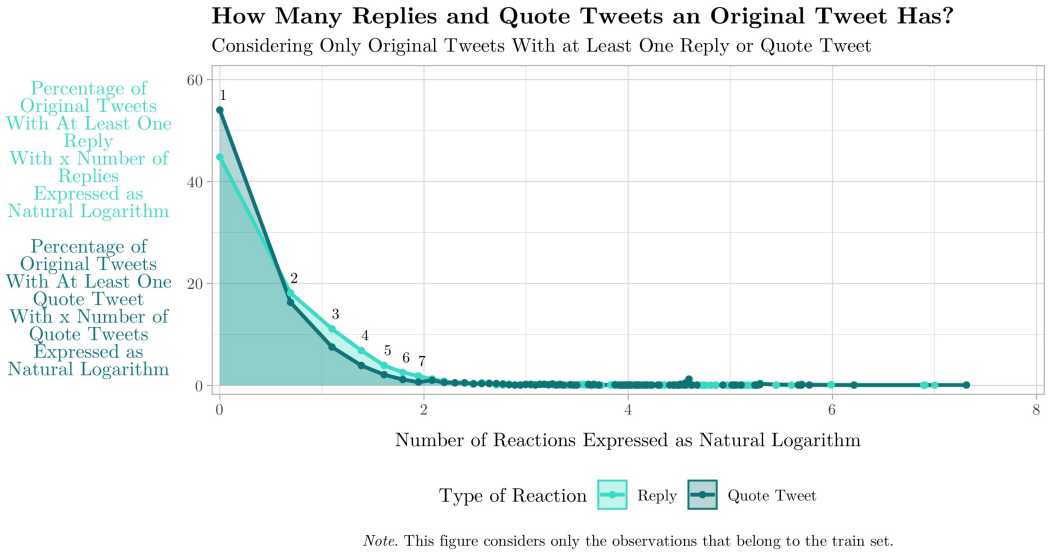
Note. This table considers only the observations that belong to the train set.

Before moving on, it is worth clarifying the following. Someone might point out that the carried-out "imputation" can reduce the variance in the dependent variable and modify the model's performance, thus representing a worry. However, in this case, the "imputation" is done not because the sentiment or the corresponding emotion value is unknown for the observation in question, but because that value directly does not exist since this observation has no replies or Quote Tweets at all. The difference can seem thin, but the key is that, in this thesis, the "imputation" does not consist on assigning, to an initially lost value, an estimation of the real

value that could not be captured; but instead, a way to equally represent those for which is known with certainty that their value does not exist and, at the same time, that this way enables to differentiate them as much as possible from the rest. Consequently, what here is done is not formally an imputation. Additionally, given the difference previously described, there is no dispersion whatsoever between the real value (non-existent) and the assigned value (the chosen to represent the non-existence) and, thus, when assigning the latter to the former, there is no considerable variance being reduced. In this way, the matter of the "imputation" and the variance does not actually represent a worry for this particular case.

Now, moving on, Figure 5 shows the distribution of the number of replies and Quote Tweets, for those original Tweets that do have at least one reply or Quote Tweet. To be able to better appreciate these distributions, the natural logarithm (i.e., logarithm with base e) is applied. The logarithm is undefined for negative and zero values, but the number of replies and the number of Quote Tweets are never negative, since they represent a count, and the counts equal to zero are not shown in this figure.

Figure 5: Distribution of Replies and Quote Tweets



Note. This figure considers only the observations that belong to the train set.

In Figure 5, it can be observed that both distributions are skewed to the right, which shows that some few original Tweets have an extremely large number of replies or Quote Tweets, while the majority of the original Tweets have a much smaller one. Also, it can be seen that the percentage of original Tweets with only one Quote Tweet is higher than the one with only one reply, and that there are more Original Tweets with higher numbers (between 2 and 7, both extremes included) of replies than of Quote Tweets.

It is worth adding that, observing the train set, regarding the replies, none of them have been edited, they have no URLs, 99.05% of them are not marked as containing material that may be sensitive, and their median length is equal to 75 characters; while, regarding the Quote Tweets, 99.97% of them have not been edited, they also have no URLs, 99.13% of them are not marked as containing material that may be sensitive, and their median length is also equal to 75 characters.

As mentioned in the second step, following Muñoz et al. (2022), this method requires a weight to be assigned to each indicator. This requirement is both an advantage and a disadvantage. The advantage is the freedom of the decision-maker to give greater importance to one criterion over another depending on the context in which the composite index is to be applied. Meanwhile,

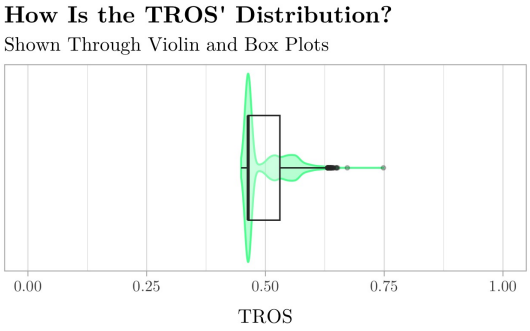
the disadvantage is the subjectivity involved in assigning one weight or another. However, this disadvantage can be mitigated by performing a weight sensitivity analysis; in other words, by determining how changes in the weights affect the value of the composite indicator. If the change in weights results in large shifts, then different solutions could be adopted, such as consulting various experts in the field in which the proposed index is to be applied. Muñoz et al. (2022) carry out this sensitivity analysis in the context of Twitter engagement and confirm that the variation in weights does not have a significant influence on the final ranking, especially at its extremes. Consequently, the aforementioned disadvantage is not severe.

To determine the weight of each indicator, Mohamaddoust et al. (2021) implement the entropy-Shannon, Spearman, and Tau-Kendall correlation methods and combine their resulting weights using a simple average. Meanwhile, Cuevas-Molano et al. (2021) say that commenting requires greater involvement than liking and that, thus, two different levels of social media engagement have been defined. On one hand, passive behavior, which refers to liking; and, on the other hand, active behavior, which refers to commenting. However, the more recent study of Muñoz et al. (2022) states that there is a lack of robust empirical evidence to suggest that some indicators are more important than others. Therefore, in this thesis, the same weight is assigned to all the chosen indicators, like Muñoz et al. (2022) do.

Finally, despite the possibility of transforming the TROS into a binary metric, this is not done because it implies establishing a threshold, which tends to be a subjective decision. Furthermore, in the end, one of the purposes of this study is to be able to compare different variations of one same Tweet: Design various options for the same Tweet and compare their predicted TROS in order to know which one to post. If the TROS is *binarized*, the following scenario would be possible. Given a threshold of, for instance, 0.5, two alternatives could be classified as being associated with a "good reaction" but when, in fact, one of them could have a score of 0.51 and the other, a score of 0.8, thus the latter probably being more recommendable. Leaving the TROS as a continuous variable allows to capture those differences, in contrast to the binary version.

Next, Figure 6 presents how the TROS is distributed. There, it can be observed that most original Tweets have a TROS slightly below the neutral score of 0.5, which shows that those Tweets are kind of fine: They are not terrible since they are quite far from 0, but they could be much better as they are still quite far from 1. In fact, there are many observations above the 0.5 threshold; not only outliers, but also several others that give rise to a second smaller bump. Thus, these latter observations show that, for the former ones, there is indeed room for improvement which can totally be achieved. Additionally, rounding every value to four decimals, from left to right, the minimum TROS is 0.4488; the first quartile is 0.4636, like the median; the mean is 0.4959; the third quartile is 0.5304; and the maximum is 0.7484.

Figure 6: Distribution of TROS

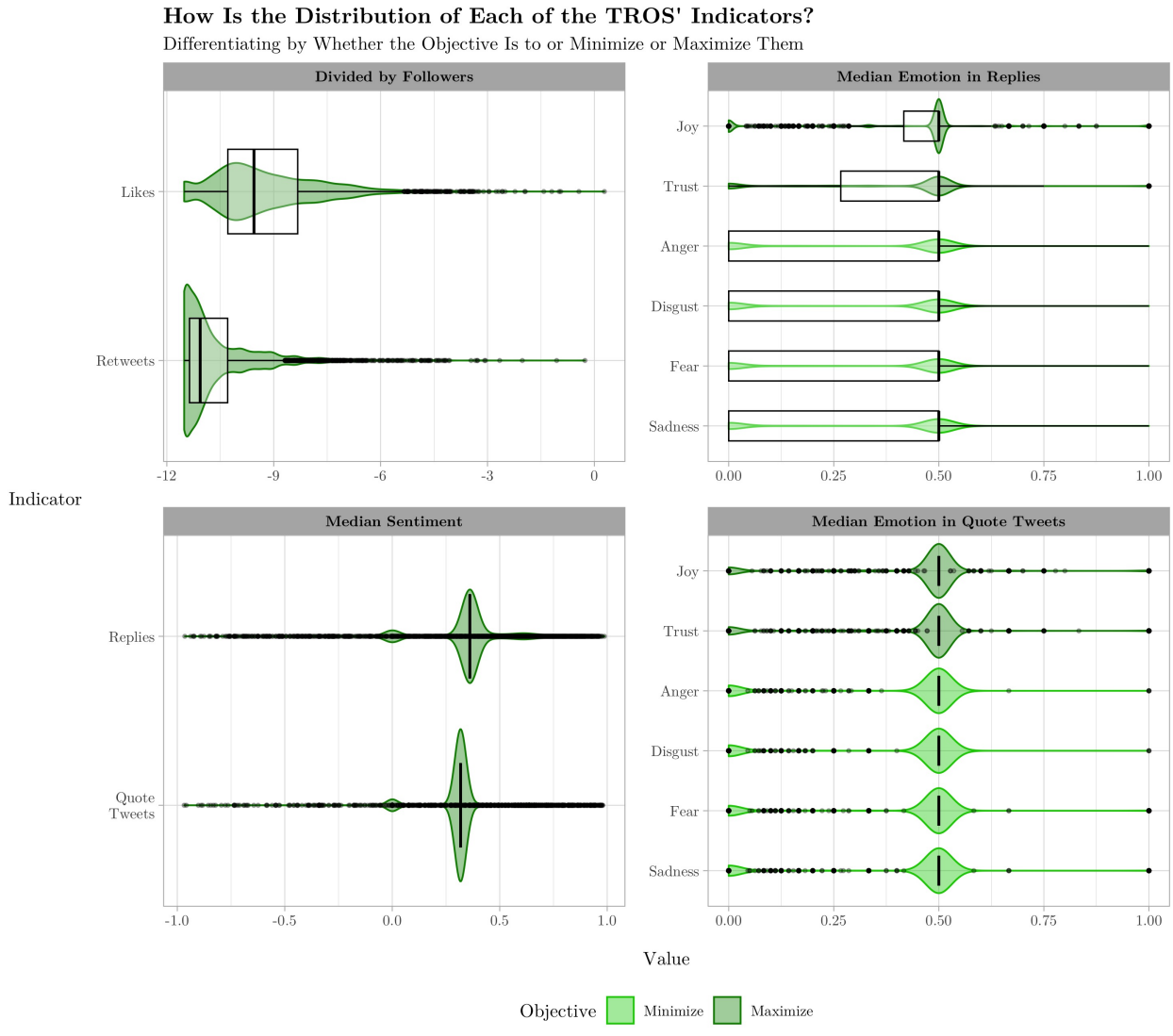


Note. This figure considers only the observations that belong to the train set.

It was just said that the minimum TROS is 0.4488. Why are there no original Tweets (at least in the train set) with a really low TROS, like 0 or 0.2? The intuition indicates that internationally renowned brands must have a communications or marketing department in charge of managing the posts' content and avoiding an absolutely appalling Tweet. As it was previously said in this thesis' introduction, building a brand takes a lot of time and money so, once one becomes an internationally well-known luxury brand, one must take care of that reputation. One way of doing so is being cautious regarding what to post. Additionally, as later shown in Figure 22, almost half of the original Tweets (that belong to the train set) are posted by accounts created on 2009. Therefore, these Tweets have behind them brands who have already acquired quite some experience on Twitter by 2022 and 2023, which are the posting years of these Tweets. In consequence, the lack of original Tweets with a really low TROS, like 0 or 0.2, seems to be completely reasonable.

Given that the TROS' distribution was just analyzed, the moment is seized to comment how the distribution of each of the TROS' indicators ended up being, after treating the cases in which there are no replies or Quote Tweets.

Figure 7: Distribution of TROS' Indicators



Note. This figure considers only the observations that belong to the train set.

In Figure 7, regarding the first two indicators, it can be seen that they have an asymmetric positive distribution, where most observations present very small values and only very few have bigger ones. The ratios are always, except once, lower than 1: The number of likes or Retweets obtained by the original Tweet is smaller than the number of followers that the account author of the original Tweet has. In addition, those ratios are not only lower than 1, but usually extremely near 0, which indicates that the number of likes or Retweets received by the original Tweet tends to be extremely lower than the number of followers that the account author of it has. Therefore, to be able to better appreciate their distribution, the natural logarithm is applied, plus a very small positive value (0.00001) to deal with those observations that have exactly 0 as their value, since the logarithm is not defined for them. The logarithm is also undefined for negative values, but these two indicators are never negative since their numerator or L_i or RT_i and their denominator F_i are always 0 or positive quantities.

Then, in Figure 7, it can also be seen that the ratio for the likes tends to be higher than the one for the Retweets. In relation to the rest of the indicators, they all have a bimodal distribution, where one of the modes corresponds to the most usual values in the sample, while the other, to the assigned value to the cases in which there are no replies or Quote Tweets. Particularly, as to the median sentiment indicators, it can be observed that the median sentiment tends to be positive in both replies and Quote Tweets but that the magnitude of positivism tends to be slightly higher in replies. Meanwhile, as to the median emotion indicators, they all have quite similar distributions. However, in replies, the negative ones (i.e., anger, disgust, fear, and sadness) present more observations on the lower side, and in Quote Tweets, the positive ones (i.e., joy and trust) have more outliers on the upper side. This is a good sign, since the former ones are to be minimized, while the latter ones are to be maximized.

Finally, to see whether the TROS is reasonably capturing the reactions to the original Tweets, the TROS' components are observed for those observations with the minimum and maximum TROS, as well as for three random ones.

Table 5: Sampled TROS

	TROS	Followers	Likes	Retweets
Minimum	0.4487858	22,073	18	2
Maximum	0.7483606	2,136	2,828	1,638
Random 1	0.5300074	1,936,028	221	18
Random 2	0.5210078	417,616	5	1
Random 3	0.4636043	396,431	74	6

Note. This table considers only the observations that belong to the train set.

Table 5 shows that the original Tweets associated with the minimum TROS or with any of the three random ones received extremely few likes and Retweets relative to the number of followers of their corresponding account. In contrast, the original Tweet associated with the maximum TROS received even more likes than its corresponding account's number of followers and a very good number of Retweets given those followers.

Additionally, the original Tweet associated with the minimum TROS received 0 replies and only 1 Quote Tweet, which has several loudly crying face emojis, during the day it was posted. Whereas, the one with the maximum TROS received 60 and 94, respectively. Some of those replies and Quote Tweets were randomly inspected and all the inspected are related to wishing a happy new Chinese year and flattering Xiao Zhan, the Chinese actor and singer mentioned and shown

in the original Tweet. Meanwhile, the original Tweet with the first random TROS received 0 replies and 2 Quote Tweets, with no negative feelings; the one with the second random TROS also received 0 replies but only 1 Quote Tweet, with a slightly positive sentiment; and the one with the third random TROS received 0 replies as well as Quote Tweets. Consequently, it seems that the TROS is reasonably capturing the reactions to the original Tweets.

3.3. Data Collection

To obtain the Tweets' data, the Twitter's application programming interface (API) was used, specifically its second version. An account on Twitter was created, a developer account was signed up for, and the key and tokens were saved[57]. Then, the recommended by Twitter R code samples for recent searches and user quests were used[58] [59], the query operators for the recent searches were customized using the guide provided by Twitter[60], and the requested Tweets' and Twitter accounts' fields were set up using the documentation also provided by Twitter[61] [62].

As to the data download, from November 6, 2022 Coordinated Universal Time (UTC) to April 6, 2023 UTC (both extremes fully included), every day at 00:00 UTC the accounts information was downloaded and this was then matched to the original Tweets posted later during that day. In this way, the account data were avoided from being biased by any original Tweet posted during that day.

Then, also from November 6, 2022 UTC to April 6, 2023 UTC (both extremes fully included), every day at 00:30 UTC the original Tweets posted during the entire previous day were downloaded, as well as the replies and Quote Tweets each of them had. It is worth clarifying that by original Tweets it is meant Tweets from the brands, and these do not include their Retweets. As Muñoz et al. (2022) state, a distinction must be made between the Tweets a user writes and the Tweets this user shares but that were written by another one. Although both of these kinds of Tweet appear on the user's profile and can be read by the user's followers, the true activity is the one generated by the former kind: the original messages (Muñoz et al., 2022).

By doing this second download every day instead of every 6 days as originally planned, the difference in the amount of time the original Tweets have been posted was reduced and this is important, since it is intuitive to think that the longer the time of exposure, the higher the chances of being seen and receiving more reactions. Specifically, from a possible maximum time difference of 143 hours with 59 minutes (an original Tweet being posted at 00:00 UTC of the first of the 6 days and another original Tweet being posted at 23:59 UTC of the last of the 6 days) it was passed onto a possible maximum time difference of only 23 hours 59 minutes. That is approximately an 83.34% reduction.

In this way, what is being tried to be predicted is the reaction that a post on Twitter will generate in the audience of luxury fashion brands *the day it is posted*. It is worth mentioning that Vassio et al. (2022) find that it is feasible to accurately predict the total number of interactions[63] after

[57] https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api.
[58] https://github.com/twitterdev/Twitter-API-v2-sample-code/blob/main/Recent-Search/recent-search.r.
[59] https://github.com/twitterdev/Twitter-API-v2-sample-code/blob/main/User-Lookup/get_users_with_bearer_token.r.
[60] https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query#list.
[61] https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet.
[62] https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/user.
[63] In this thesis, the type and the intensity of the reactions are instead tried to be predicted.

observing the very initial phase of a post's lifetime (i.e., the time at which the post has received 95% of its total interactions), in the context of Italian influencer-generated posts on Facebook and Instagram[64]. In fact, the authors state that the total number of interactions gathered by a post can be well predicted by measuring the interactions received within the first hour or even within the first few minutes. Additionally, there is no problem whatsoever with the accounts' data being downloaded every day instead of every smaller unit of time, such as hour or minute, because these data, especially the current number of followers, can be considered constant during the posts' lifetime (Vassio et al., 2022).

The download of both the accounts and the original Tweets' data were programmed to run automatically, in the author's personal computer, every day at 00:00 UTC and at 00:30 UTC, respectively, from November 6, 2022 UTC until manually stopped, which was carried out after the downloads of Day 152 were done.[65] To provide more clarity, Table 6 illustrates this organization.

Table 6: Data Download Organization

	Accounts (00:00)	Tweets (00:30)
November 6, 2022	Day 1	-
November 7, 2022	Day 2	Day 1
November 8, 2022	Day 3	Day 2
⋮	⋮	⋮
April 5, 2023	Day 151	Day 150
April 6, 2023	Day 152	Day 151
April 7, 2023	-	Day 152

Note. Time is in UTC.

Given the rhythm at which the collected volume of data increased and the possibility of having access to Twitter's complete archive, it is worth mentioning the following. The developer account in question is of the Essential kind, so what can be accessed is not the complete archive but the Tweets posted over the last week[66]. In order to be able to access the complete archive, one needs the Academic Research access[67], to which it can be applied[68]. However, those historical original Tweets retrieved have been exposed far longer than the most recent ones, thus extremely increasing the exposure time difference, which affects the indicators of the TROS. Furthermore, it would imply that this thesis would try to predict the reaction that a post on Twitter will generate in the audience of luxury fashion brands *from the time it was created to the moment of the data collection*, when instead this thesis wants to try to predict the reaction that a post on Twitter will generate in the audience of luxury fashion brands *the day it is posted*. Consequently, having access to the entire archive would not be useful for this study.

In addition, related to the volume of data, Cuevas-Molano et al. (2021) state that their Instagram

[64]In this thesis, brand and user-generated posts on Twitter are instead analyzed.

[65]On only a few days, these download times were slightly not respected (slightly is said since the time difference was not considerable). Concretely, regarding the accounts' data download, on 9 days while, regarding the Tweets' data download, on 10 days.

[66]This is why it had originally been planned to do the second download every 6 days (not every 7 days to avoid any completeness problems at the extremes).

[67]https://developer.twitter.com/en/products/twitter-api/academic-research.

[68]https://developer.twitter.com/en/products/twitter-api/academic-research/application-info.

posts' sample size is equal to 680, which represents one of their main limitations. Therefore, more than 680 observations need to be collected, in order to avoid that same limitation. This is something that has been achieved: Data has been collected for 152 days and more than 11,000 observations are available.[69]

The data from Twitter are complemented with data from searches on Google. These represent an indicator of what people are interested in at that moment. In addition, the top 10 luxury brands generate more than 30 million web searches per year, and potential customers discover content through search queries (Hemantha, 2020). Specifically, the global interest over time is downloaded for the following keywords. First, the name of each of the brands. Second, `luxury fashion` as a word related to luxury fashion brands. Third and last, `sustainable fashion` as a phrase related to ethical consumerism, due to its importance, which is on the rise (Fumagalli, 2021). In fact, ethical fashion is a hot topic among fashion brands and designers in Milan, and fashion trends in New York have shifted their focus from practical to ethical and sustainable designs (Y. K. Choi et al., 2020). Additionally, the new generation of consumers (which in 2018 alone drove 85% of the growth of luxury sales) consists mainly of Generations Y and Z and demands sustainable practices, messaging, and products. Accustomed to broadcast their lives on social media, these consumers perceive themselves as brands and thus choose labels that align with their personalized positioning, generally claiming to only associate with sustainable brands (Caïs, 2021). Therefore, it is extremely relevant for these brands, which are criticized for causing environmental pollution and animal abuse (Y. K. Choi et al., 2020), to be aware of the interest in these topics and to take it into account when planning the brands' actions (see Fumagalli, 2021). Luxury brands can no longer ignore the question of sustainability (Caïs, 2021).

To download the data related to searches on Google, the `gtrendsR` package[70] was used. This is an interface to retrieve the information returned online by Google Trends. Concretely, what was used was the `gtrends` function in that package.

It should be noted that Google Trends represents interest over time as the relative, not absolute, search volume. In fact, the retrieved number of hits reflects the interest in the search relative to the maximum number for that keyword in the same region and time period. Consequently, 100 indicates the highest popularity, while 50 indicates half of the highest popularity relative to the highest value. Since progress had to be made while the data were still being collected (given the thesis' deadline), it could not be waited till April 7, 2023 UTC to download the entire work period at once. Thus, the Google Trends data are downloaded per set as their corresponding time period[71] has passed. Given that the retrieved search volume is relative, to generate consistency between the retrieved values in each of the three downloads, when downloading the Google Trends' data corresponding to the validation set, in the query, the last day of the train set is included, a simple rule of three is calculated to know the "multiplier" between both sets, and this "multiplier" is applied to the validation's values to turn them consistent with the train ones.[72]

[69] By the way, some of the downloaded original Tweets were actually destined exclusively to individual users, and some others had the exact same text to others posted by the same account at the exact same time so it is intuitive to assume that they were also destined exclusively to individual users. Consequently, before computing any official calculus (i.e., all the ones included in this thesis), these original Tweets are discarded, together with their corresponding replies and Quote Tweets if they had any. In fact, the aforementioned 11,000 does not include these original Tweets.

[70] https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf

[71] This is explained more in detail in the train, validation, and test sets' section.

[72] In case this multiplier involves dividing by zero or dividing zero by something, 0.001 and 0.01 are respectively assigned to replace those zeros, to avoid making infinite or zero, respectively, the rest of the other set's values.

Then, the same is done for the test set, but including in the query the last day of the validation set and doing the simple rule of three with this day's already transformed value.

Furthermore, by doing the download by period of time, the time associated to each number of hits does not include the hour and there are 4 days of delay between the interest over time's data available for download and the present day. Therefore, every original Tweet is matched with the Google Trend's data corresponding to 4 days prior that Tweet's creation date. For example, original Tweets posted on November 6, 2022 are assigned the Google Trend's data corresponding to November 2, 2022.

The Google Trends' data are complemented with Twitter Trends' data. These latter trends, in comparison to the former, are used in this thesis to capture what people on Twitter are particularly interested in at the exact beginning (00:00 UTC) of the day in which the original Tweet is posted, so that then it can be calculated how related the original Tweet's text is to topics that are becoming distinctly visible on Twitter. Twitter Trends are determined by an algorithm that identifies topics that are popular at that moment, rather than topics that have been popular for a while or on a daily basis, to help discover the highly emerging topics of discussion on the platform. It is worth noting two things. First, the number of Tweets that are related to the Trends is only one of the factors the algorithm takes into account when ranking and determining the Trends. Second, if they are related to the same topic, Trends and hashtags are grouped together. For example, `#MondayMotivation` and `#MotivationMonday` might both be represented by `#MondayMotivation`.[73]

The second version of the Twitter API does not have a Trends' object to download, like it does with Tweets and users. However, then it was discovered that the standard first version of the Twitter API does have such a Trends' object. But, at the moment of the discovery, almost all the train set was already collected and the Trends' object does not have the option to specify from which date one wants the Trends. Instead, the downloaded Trends correspond to the time of the download. Thus, the standard first version of the Twitter API does not allow to recover those historical Trends. Luckily, a website named ExportData was found. This website has a section called Twitter Trends Worldwide, where one can access historical Twitter Trends[74]. Twitter recalculates trends on an hourly basis and this website has been collecting hourly Twitter Trends since August 2019. Additionally, the website explicitly states that one can use their page as a data source for one's research.

Therefore, the structure of the data base where the Twitter Trends' data are saved is created with R, it is exported as a comma-separated values (CSV) file, and it is imported into an Excel workbook. Then, the remaining columns (regarding the trends and their corresponding Tweet volume) are manually completed as the days pass by. To access the website information, URLs are used in the following format, where time is in UTC and date is expressed as yyyy/MM/dd: `https://www.exportdata.io/trends/worldwide/[date]/[hour]`. For instance, https://www.exportdata.io/trends/worldwide/2023-01-02/0 for January 2, 2023 at 00:00 UTC. The website recommends using that format and promises never to change that URL structure. As can be seen in the URL, the global results are downloaded, as done with Google Trends.

Finally, the list containing all the original Tweets' data is transformed into a data frame and merged with the accounts' data frame by brand and date. Regarding the replies', Quote Tweets'

[73] https://help.twitter.com/en/using-twitter/twitter-trending-faqs.
[74] https://www.exportdata.io/trends/worldwide.

and Google Trends' data, they are converted into a data frame. It is worth adding that the replies and the Quote Tweets' data frames can be merged with the main one by the original Tweet's ID, while the Google Trends and Twitter Trends ones, by date (and brand for the Google Trends' keywords related to the brands). However, the Twitter Trends' data are from every day in the sample at 00:00 UTC and so merged with original Tweets created on that same day. In contrast, the Google Trends' data are from one day in general; the hour is not specified. Consequently, as previously mentioned, to avoid data leakage, these data are merged with original Tweets created 4 days later.

3.4. Exploratory Data Analysis

In this section, the different data sets that contain the default attributes (or the default variables from which to construct attributes) to predict the TROS are explored. These data sets correspond to the original Tweets, the accounts author of the original Tweets, or the trends on or Google or Twitter. For each of them, their variables are stated and figures related to some of those variables are described. When doing so, to prevent data leakage, only the observations that belong to the train set are considered.

Table 7: Variables of Original Tweets

Name	Type	Function	NA
original_id	character	ID	0.00
original_created_at	character	Create attribute(s)	0.00
original_handle	character	Attribute	0.00
original_download_day	integer	ID	0.00
original_download_time	POSIXct	ID	0.00
original_brand	character	ID and attribute	0.00
original_attachments_media_keys	list	Create attribute(s)	3.98
original_author_id	character	None given	0.00
original_context_annotations	list	None given	14.14
original_conversation_id	character	None given	0.00
original_edit_controls_edits_remaining	integer	None given	0.00
original_edit_controls_is_edit_eligible	logical	Captured by other, so none	0.00
original_edit_controls_editable_until	character	None given	0.00
original_edit_history_tweet_ids	list	None given	0.00
original_entities_urls	logical	Captured by other, so none	100.00
original_entities_hashtags	list	Captured by other, so none	16.04
original_entities_mentions	list	Captured by other, so none	71.24
original_entities_annotations	list	None given	17.91
original_possibly_sensitive	logical	Attribute	0.00
original_reply_settings	character	Attribute	0.00
original_source	character	Attribute	52.89
original_text	character	Create attribute(s)	0.00
original_withheld	logical	None given	100.00
original_public_metrics_like_count	integer	Create response	0.00
original_public_metrics_retweet_count	integer	Create response	0.00
original_public_metrics_reply_count	integer	None given	0.00
original_public_metrics_quote_count	integer	None given	0.00

Note. This table considers only the observations that belong to the train set.

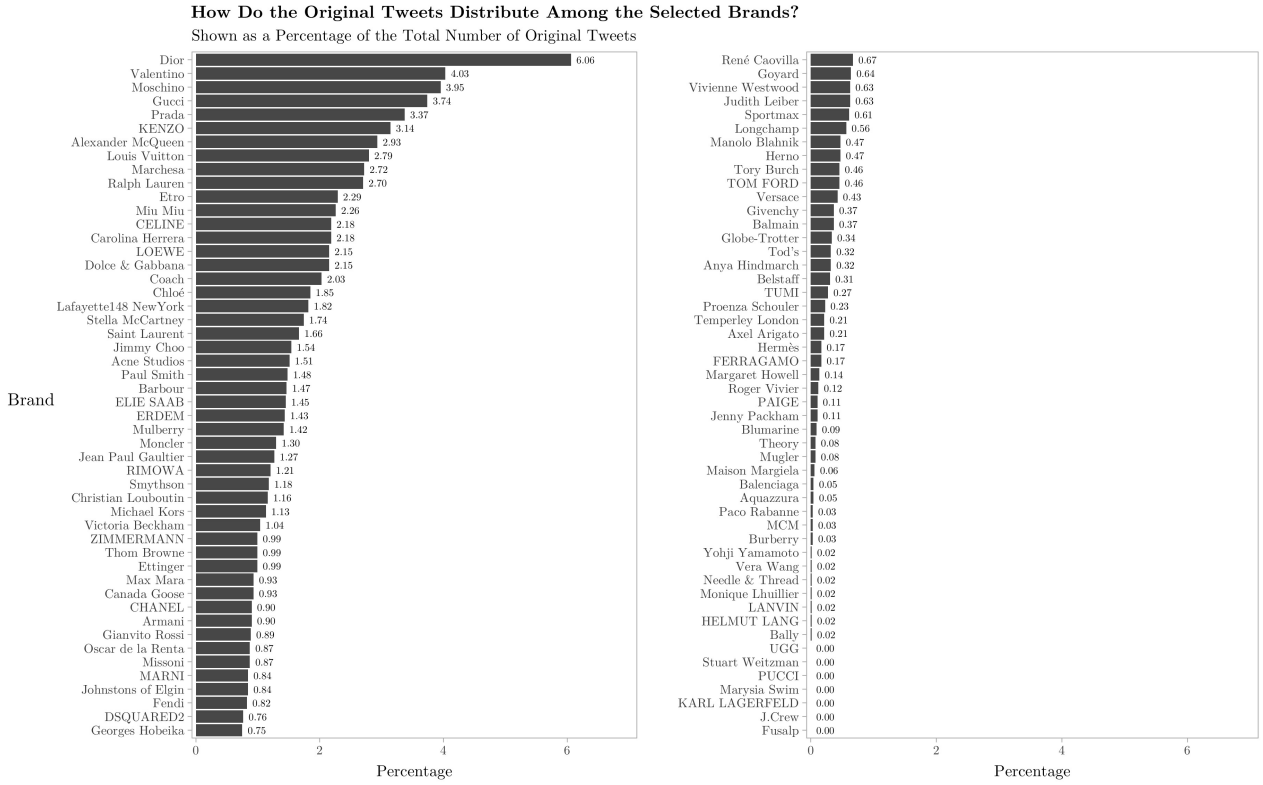
Beginning with the original Tweets' data set, Table 7 states its variables, showing their name, their type, their function, and their percentage of NA values.

It is worth mentioning that the variables in this table correspond to almost all attributes provided by Twitter for a particular Tweet. Almost is said because the table does not include the attributes `lang` since Tweets only in English are downloaded[75]; `non_public_metrics`, `organic_metrics`, and `promoted_metrics` because they require the context authentication of the user who posted the Tweet in question (which cannot be obtained by the author of this thesis due to not being the owner of any of the users under analysis); and `referenced_tweets` given that only original Tweets are downloaded[76] [77].

Additionally, it is worth clarifying that the NA value for the variable `original_withheld` indicates that there are no withholding details and so that the Tweet has not been marked as withheld. The fact that 100% of the (train set's) observations have NA for that variable is positive since it indicates that all users around the world can see the Tweet.

Next, some of these variables are described more in detail, and several figures related to these variables that help to have a better overview of them are presented. The first one is the following.

Figure 8: Distribution of Original Tweets Through Brands



Note. This figure considers only the observations that belong to the train set.

Figure 8 shows that 93 of the 100 selected brands end up appearing in the sample of original Tweets and that some brands have more original Tweets in the sample than others, being Dior the one with the highest number. It is worth noting that Balenciaga and BALLY stopped having

[75]The justification for this decision can be found in the second paragraph of the unit of analysis and selection of brands' section.

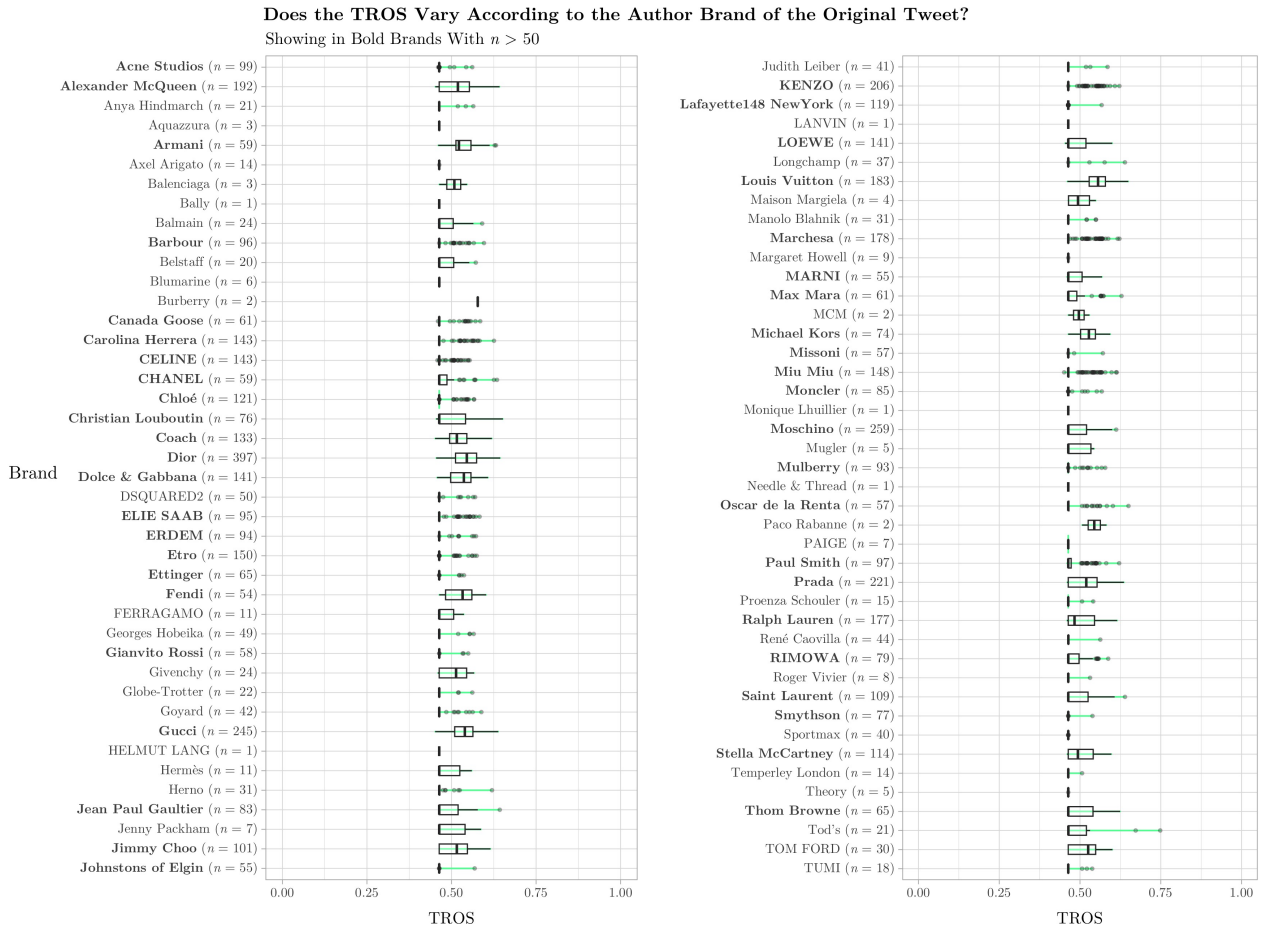
[76]The justification for this other decision can be found in the third paragraph of the data collection's section.

[77]https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet.

a Twitter profile during the data collection. The former had a Twitter profile only during the first 5 days of the data collection; while the latter, during the first 5, then stopped having it for 4 days, and finally had it for the following 7 days.

Meanwhile, Figure 9 illustrates that the median TROS for the observations corresponding to many of the brands is mainly slightly below 0.5, but the distribution of the TROS does vary between brands. In fact, there are brands with a relatively representative n (i.e., higher than 50) and associated with a relatively high median TROS (i.e., greater than 0.5), which are: Alexander McQueen, Armani, Coach, Dior, Dolce & Gabbana, Fendi, Gucci, Jimmy Choo, Louis Vuitton, Michael Kors, and Prada.

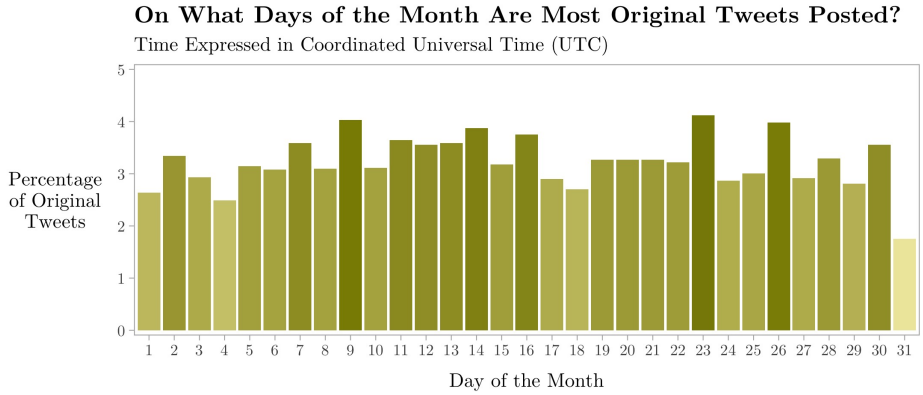
Figure 9: Distribution of TROS Through Brands



Note. This figure considers only the observations that belong to the train set and only the brands with at least one observation.

Moving onto the variable `original_created_at`, Figure 10 shows that more original Tweets are posted around the middle of the month (i.e., on the 9$^{\text{th}}$, 14$^{\text{th}}$, and 16$^{\text{th}}$) with some exceptions on the 23$^{\text{rd}}$ and 26$^{\text{th}}$.

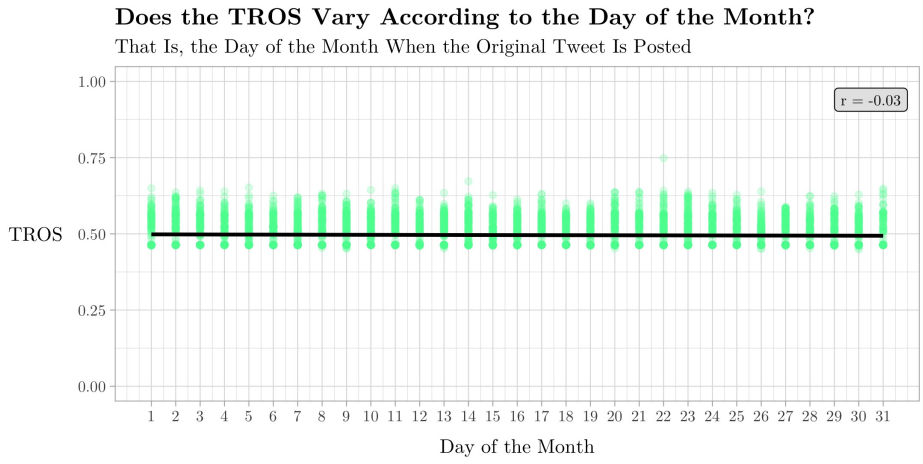
Figure 10: Distribution of Original Tweets Through Days of the Month



Note. This figure considers only the observations that belong to the train set.

Figure 11 indicates that the TROS almost does not vary through the month. In fact, the black smoothed line, whose aim is to aid the eye in seeing patterns, is a straight one at a TROS of 0.5. Meanwhile, the Pearson's correlation coefficient (r), which measures how linearly associated two variables are, is negative: The later during the month the original Tweet is posted, the lower the TROS. However, the magnitude of the negative linear association is extremely small.

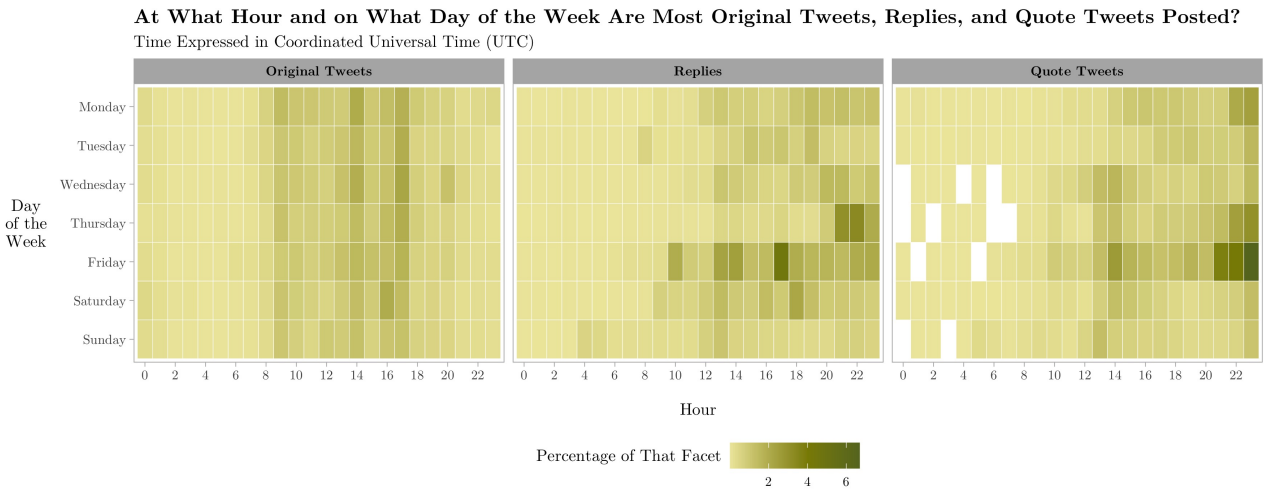
Figure 11: Distribution of TROS Through Days of the Month



Note. This figure considers only the observations that belong to the train set.

Figure 12 compares the distribution through hours and days of the week of original Tweets, replies, and Quote Tweets. In it, it can be seen that original Tweets are generally posted between 9:00 and 17:00 UTC, with Wednesdays extending to 20:00 UTC as an exception. In contrast, replies tend to be posted later during the day, mainly on Thursdays and Fridays. Finally, regarding Quote Tweets, they are usually posted at noon or late at night, especially on Mondays, Thursdays, and Fridays.

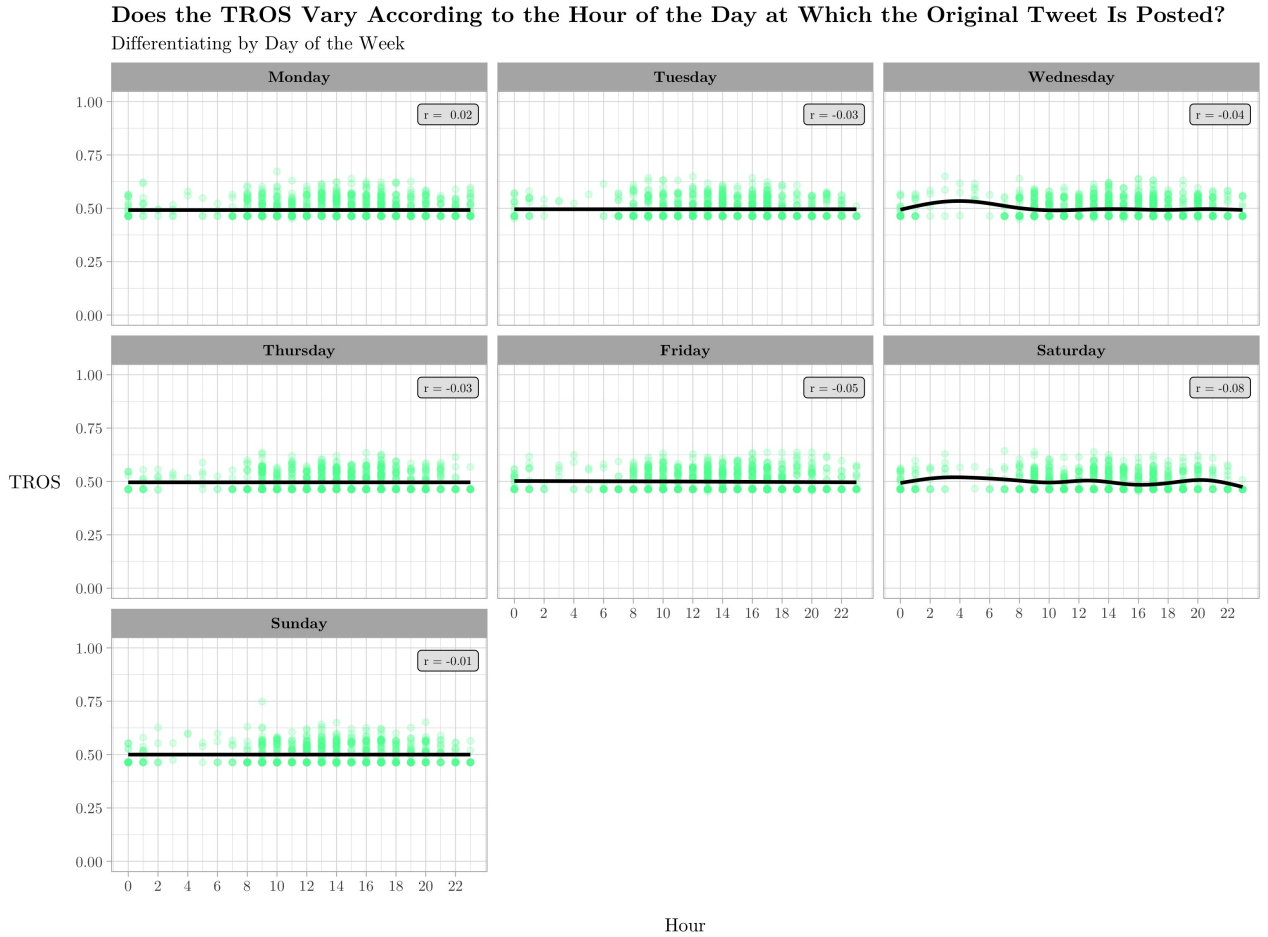
Figure 12: Distribution of Tweets Through Hours and Days of the Week



Note. This figure considers only the observations that belong to the train set.

Figure 13 shows that the TROS varies slightly throughout the hours of the day and the days of the week, but these variations are minimal, making the black smoothed line straight except on Wednesdays and Saturdays. Additionally, on every day, except Mondays, the r is negative (so the later the hour the original Tweet is posted, the lower the TROS), but minimal.

Figure 13: Distribution of TROS Through Hours and Days of the Week



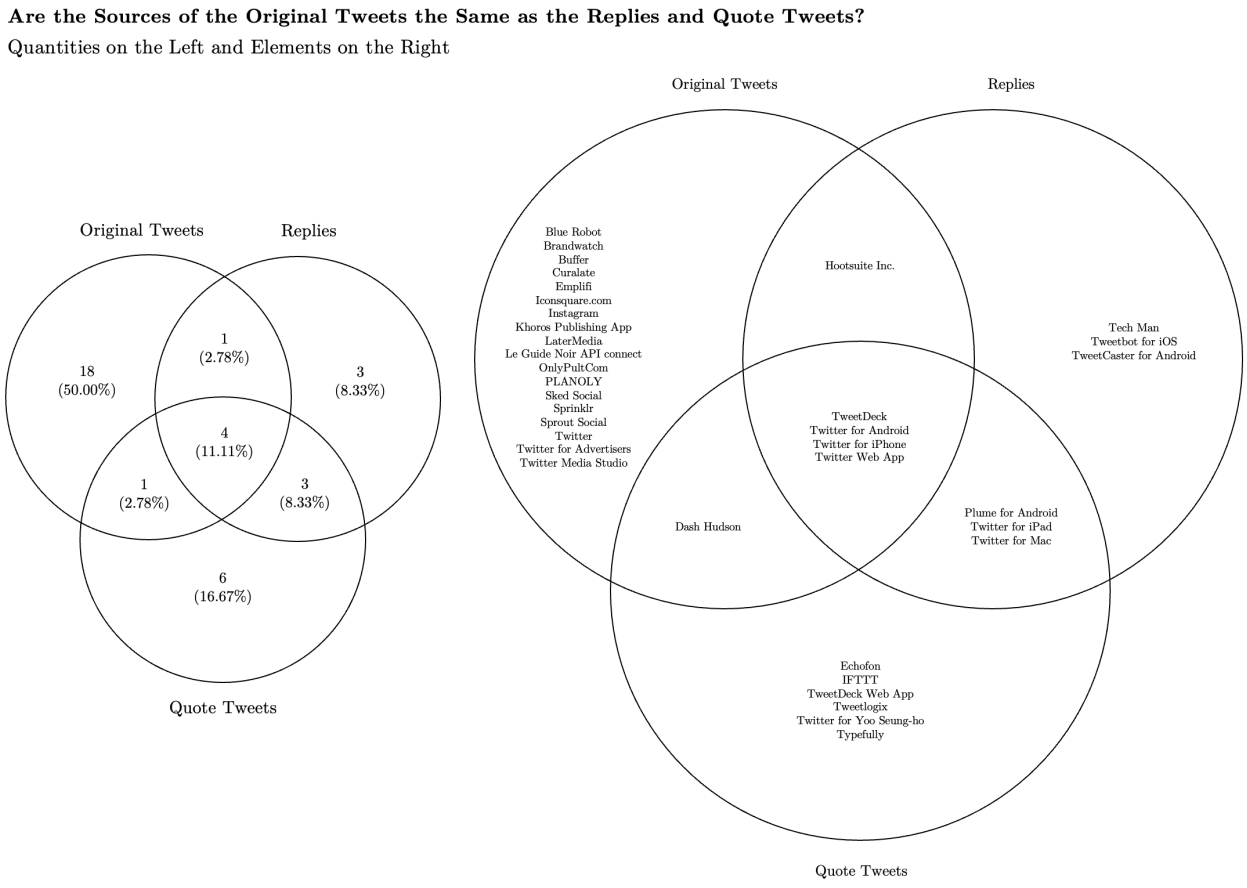
Note. This figure considers only the observations that belong to the train set.

Now, regarding the variable `original_edit_controls_edits_remaining`, this variable does not have any variation whatsoever. All the observations from the train set have the value `5`, which indicates that none of these original Tweets have been edited (during the day they were posted). This is the reason why no function is given to it.

Then, as to the variable `original_possibly_sensitive`, 99.65% of the original Tweets in the train set have the value `FALSE`. This is good since it indicates that the content of almost all original Tweets is shown directly to users, without them having to explicitly ask to view it.

Passing onto the variable `original_source`, Cuevas-Molano et al. (2021) advise studying whether the fact that brands' posts are organic or paid influences the consumers' interactions. The variable `original_source` contributes to measuring that. Figure 14 compares the sources of the original Tweets, replies, and Quote Tweets, using Venn diagrams.

Figure 14: Distribution of Tweets Through Sources



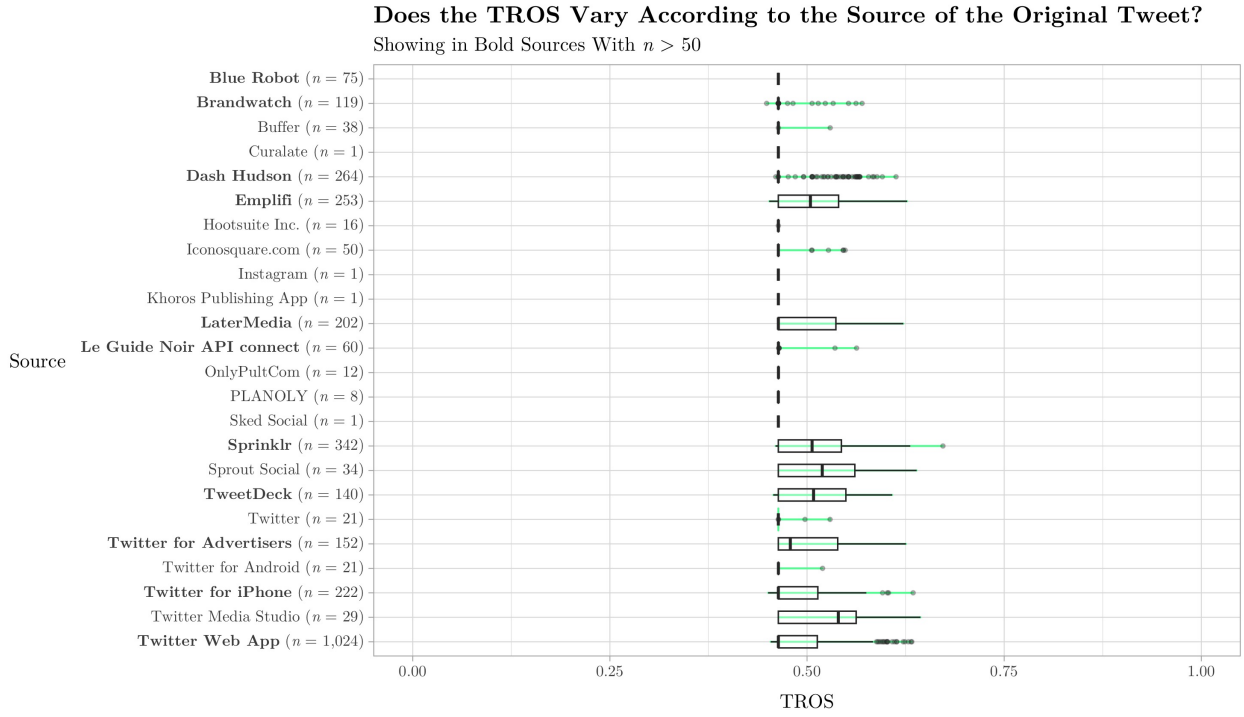
Note. This figure considers only the observations that belong to the train set.

First of all, it must be stated that, since they stand for the same concept, `Sprinklr Publishing` and `Sprinklr` are unified in the latter, as well as `Twitter Ads` and `Twitter for Advertisers` also in the latter. So, in Figure 14, it can be observed that original Tweets are the type of Tweets with most exclusive sources. These are mainly related to businesses that offer social media marketing services, while the sources that original Tweets, replies, and Quote Tweets have in common are more general, like Twitter for Android or iPhone. Original Tweets also have as one of their sources Twitter for Advertisers, which indicates that the Tweet is promoted

through Twitter Ads[78]. These Tweets are part of campaigns that can have one of several possible objectives related to awareness, consideration, or conversion[79]. It is worth noting that these Tweets are not necessarily shown to more people; it depends on the campaign's objective.

Additionally, Figure 15 shows that the original Tweets' sources with a representative n and a relatively high median TROS are not Twitter for Advertisers, but Emplifi, Sprinklr, and TweetDeck.[80]

Figure 15: Distribution of TROS Through Sources



Note. This figure considers only the observations that belong to the train set.

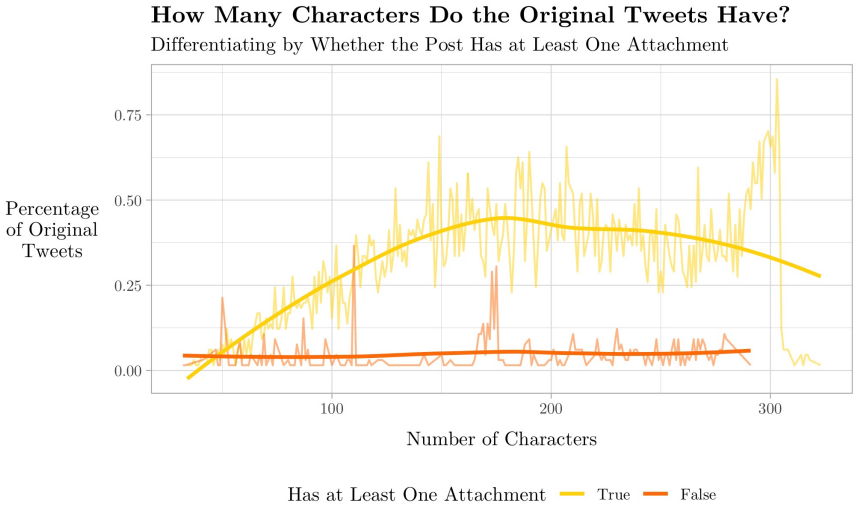
Next, regarding the variables `original_text` and `original_attachments_media_keys`, one might think that Tweets with attachments tend to be shorter in character's length since part of what they intend to transmit is being already expressed in the attachments. However, Figure 16 illustrates that, bearing in mind that there are more original Tweets with at least one attachment, these tend to be longer than those without attachments at all.

[78] https://business.twitter.com/en/advertising/get-started-with-twitter-ads.html.

[79] https://business.twitter.com/en/advertising/campaign-types.html.

[80] Original Tweets from Twitter Ads are not explored any further because, considering the train set, they represent less than 5% of the original Tweets with a source and only 2.3% of all original Tweets.

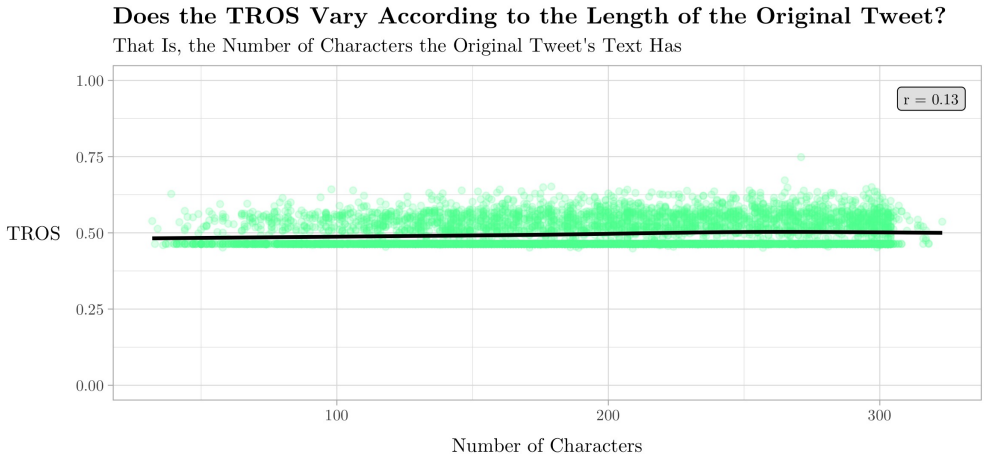
Figure 16: Distribution of Original Tweets Through Length and by Attachment



Note. This figure considers only the observations that belong to the train set.

Figure 17 shows that the TROS tends to be slightly higher the longer the original Tweet is. The r supports this statement, since it is a small positive value.

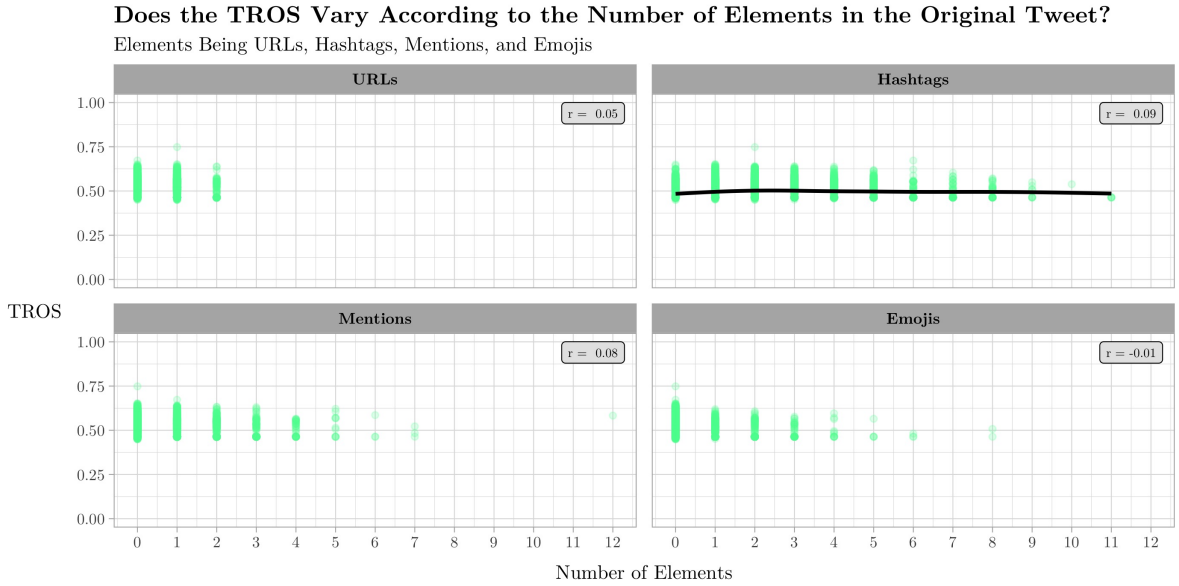
Figure 17: Distribution of TROS Through Length



Note. This figure considers only the observations that belong to the train set.

Figure 18 portrays how the TROS varies according to the number of URLs, hashtags, mentions, and emojis. First of all, it must be noted that the downloaded original Tweets tend to have a URL at the end of their text, which sends to the Tweet's web page. However, that URL is not visible in the posted Tweet's text, but only in the downloaded one. Consequently, when computing the number of URLs, 1 is subtracted to avoid considering that last "invisible" URL and, for the few cases in which that URL did not appear in the downloaded Tweet's text, the resulting -1 is replaced with 0 to illustrate that they did not have any visible URL in their posted text either. Having made that note, it can be seen that the number of URLs' range is lower than the one from the rest and that original Tweets tend to have few amounts of these types of elements, if any. Additionally, except for the number of emojis, the r is a very small positive value. Finally, it is worth clarifying that the black smoothed line appears only in the hashtags' facet, since the rest of the elements have insufficient unique values to be able to plot it.

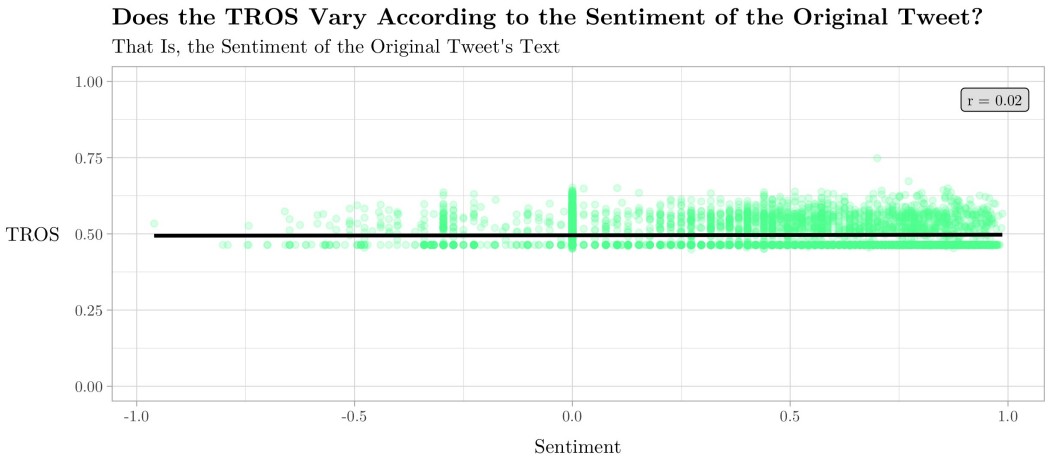
Figure 18: Distribution of TROS Through Number of Elements



Note. This figure considers only the observations that belong to the train set.

Figure 19 shows that the text of the original Tweets is more on the neutral and positive sentiment sides than on the negative one, and that the r is a very small positive value.

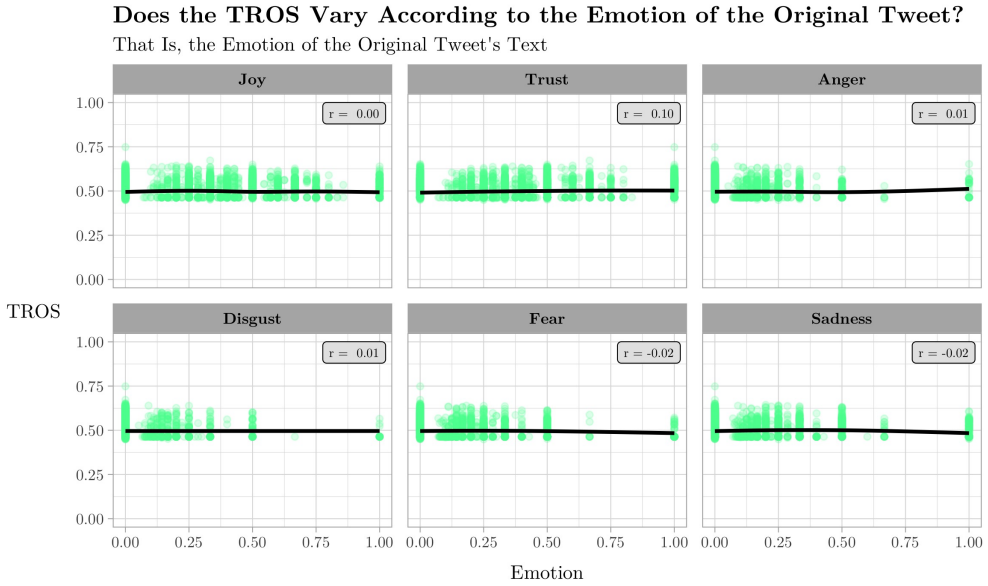
Figure 19: Distribution of TROS Through Sentiment of Original Tweets



Note. This figure considers only the observations that belong to the train set.

Meanwhile, Figure 20 illustrates that there are original Tweets all over the joy and trust continua, while they tend to concentrate on the lower side of the anger, disgust, fear, and sadness continua, which is good, as brands usually do not want to convey those kinds of emotion. Furthermore, regarding the r, this is neutral for joy, positive and small for trust, curiously positive but extremely small for anger and disgust, and coherently negative but kind of too small for fear and sadness. Additionally, recall that 12 of the 16 TROS' indicators are about the emotions either in the replies or the Quote Tweets. Then, Figure 20 also shows that the TROS and thus, indirectly, the emotions in the replies and the Quote Tweets are, interestingly, not very much related to the emotions in the text of the corresponding original Tweet.

Figure 20: Distribution of TROS Through Emotions of Original Tweets



Note. This figure considers only the observations that belong to the train set.

Finally, Figure 21 shows the most frequent words used in the original Tweets with lower versus higher TROS. To do this figure, first the text was cleaned following the same process as for EmoLex, but removing only the at the rate sign (`@`) instead of also the name of the account mentioned and also removing the stop words and all types of punctuation. In addition, only tokens of at least 4 characters and unique to each group are considered.

Figure 21: Frequency of Original Tweets' Words According to TROS

What Are the Most Common Words in the Original Tweet According to the TROS?

That Is, the Tokens That Appear at Least 10 Times in the Text of the Original Tweets With a TROS Lower or Equal (Left) Versus Higher (Right) Than the Median

Note. This figure considers only the observations that belong to the train set.

At the center with bigger font size, it can be seen that the most frequent words in original Tweets with a lower TROS are the brand Sportmax and the photographer Chloé Horseman; while in those with a higher TROS, the brand Gianvito Rossi and the artist Yayoi Kusama, who in

January 2023 collaborated in a campaign with Louis Vuitton. Furthermore, in the bag of words on the right, some celebrity's names can be seen. For instance, the full name of Dakota Johnson, who in January 2023 starred in the Gucci's Jackie 1961 campaign; as well as the surname of Robert Pattinson, who in February 2023 starred in the campaign for the perfume Dior Homme Sport.

The next data set explored is the accounts one. Like it was done with the originals data set, it is begun by presenting a general variables' description. This can be found in Table 8.

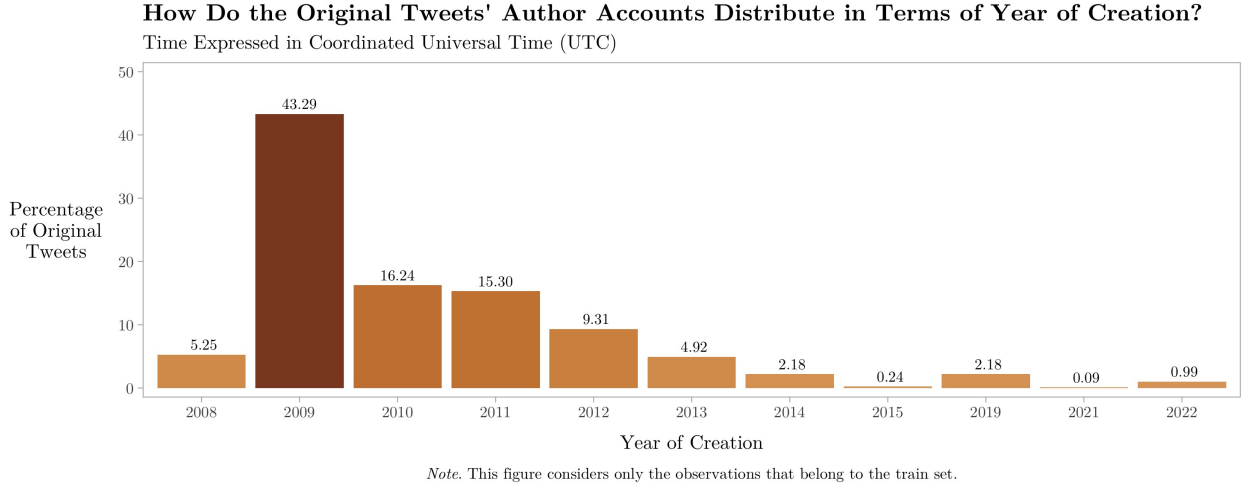
Table 8: Variables of Accounts

Name	Type	Function	NA
account_download_time	POSIXct	ID	0.00
account_id	character	ID	0.00
account_name	character	ID	0.00
account_created_at	character	Create attribute(s)	0.00
account_description	character	Create attribute(s)	0.00
account_location	character	Create attribute(s)	20.39
account_pinned_tweet_id	character	Create attribute(s)	69.00
account_profile_image_url	character	None given	0.00
account_protected	logical	None given	0.00
account_public_metrics_followers_count	integer	Attribute and create response	0.00
account_public_metrics_following_count	integer	Attribute	0.00
account_public_metrics_listed_count	integer	Attribute	0.00
account_public_metrics_tweet_count	integer	Attribute	0.00
account_url	character	Create attribute(s)	1.85
account_verified	logical	Attribute	0.00

Note. This table considers only the observations that belong to the train set.

A more in detail description of some of these variables follows, together with a presentation of several figures related to these variables that help to have a better overview of them. The first one is Figure 22, which shows that almost half of the original Tweets are posted by accounts created on 2009. It is worth noting that Twitter debuted in 2006, so it took around two years for luxury fashion brands to start using this social network.

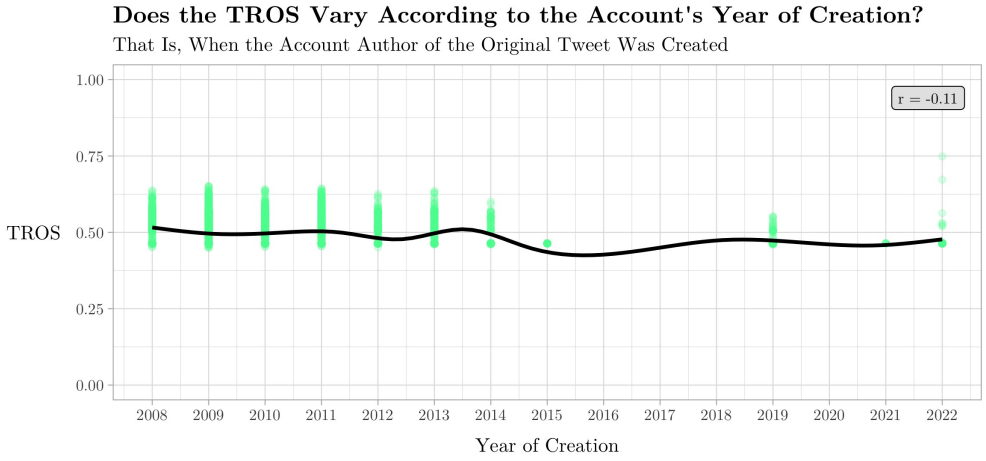
Figure 22: Distribution of Original Tweets Through Accounts' Year of Creation



Note. This figure considers only the observations that belong to the train set.

Meanwhile, Figure 23 illustrates that, except for two outliers, original Tweets posted by accounts created on the earlier years have a higher TROS. This is supported by the small but still negative r. It makes sense since it is intuitive to think that accounts created earlier have more experience on this social network and, thus, higher chances of generating a better reaction.

Figure 23: Distribution of TROS Through Accounts' Year of Creation

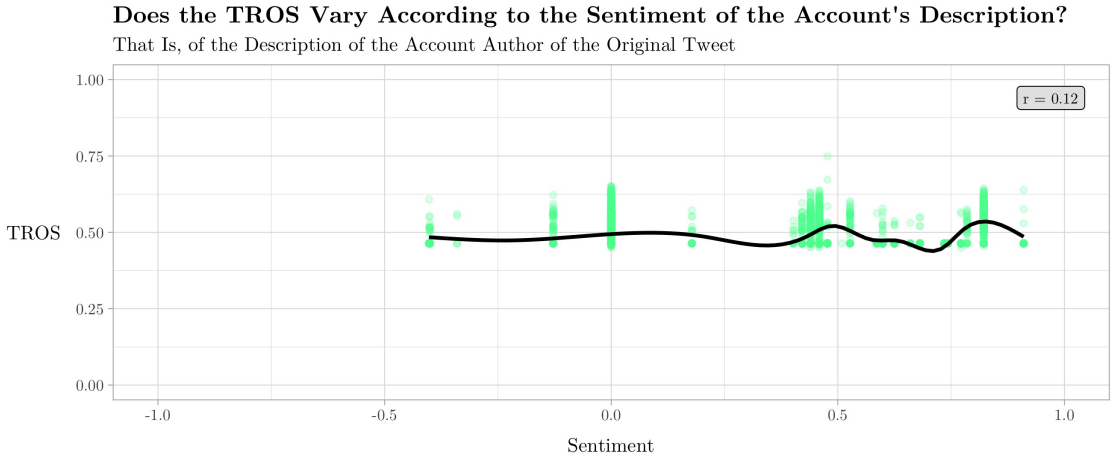


Note. This figure considers only the observations that belong to the train set.

Moving onto the variable `account_description`, the median number of characters had by the accounts' description is 48, and the r between the length and the TROS equals 0.06, which is a very small positive number, so the longer the account's description, slightly higher the TROS.

Then, Figure 24 shows that the accounts' description tends to be on the neutral or positive sides of the sentiment continuum. Also, it indicates that, in this case, the r is a small positive value, so the higher the sentiment, the slightly higher the TROS.

Figure 24: Distribution of TROS Through Sentiment of Accounts' Description

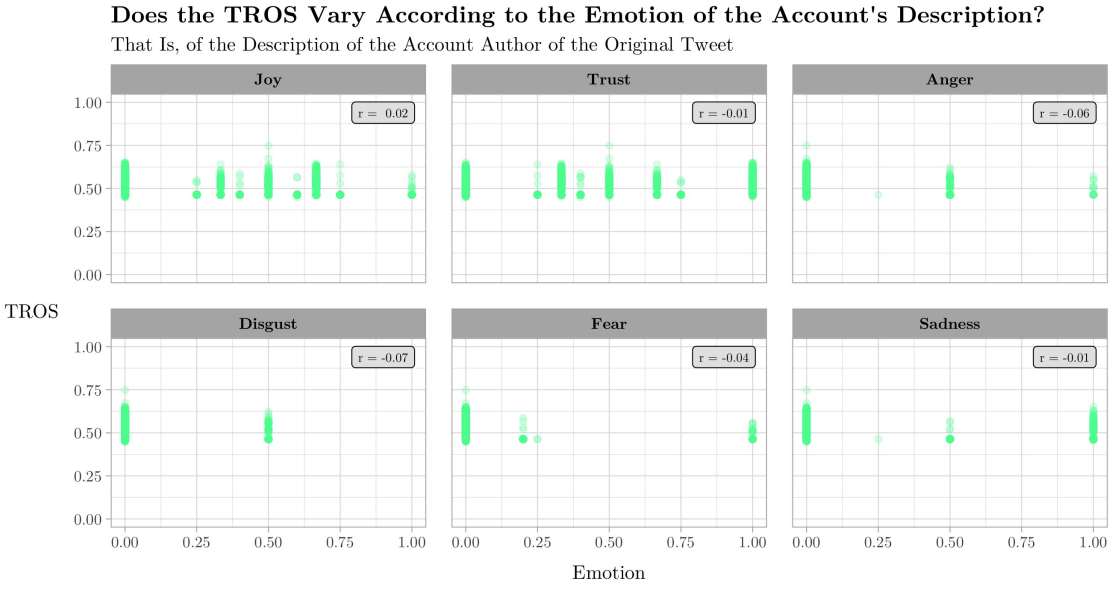


Note. This figure considers only the observations that belong to the train set.

Figure 25 illustrates that there are accounts' description with values quite all over the joy and trust continua, while mostly on 0, 0.5, and 1 for anger, disgust, fear, and sadness. Additionally, the sign of the r corresponding to each emotion is consistent with the intuitive direction of the linear association, except for trust: The higher the joy conveyed in the account's description, the higher the TROS; whereas, the higher the anger, disgust, fear or sadness, the lower the TROS.

However, the r of each emotion is extremely close to zero, which is the value that indicates a lack of linear association.

Figure 25: Distribution of TROS Through Emotions of Accounts' Description



Note. This figure considers only the observations that belong to the train set.

Figure 26 shows the most frequent words employed in the accounts' description whose Tweets have lower versus higher TROS. To do this figure, like with the previous word clouds, first the text was cleaned following the same process as for EmoLex, but removing only the @ instead of also the name of the account mentioned and also removing the stop words and all types of punctuation. In addition, only tokens of at least 4 characters and unique to each group are considered.

Figure 26: Frequency of Accounts' Description's Words According to TROS

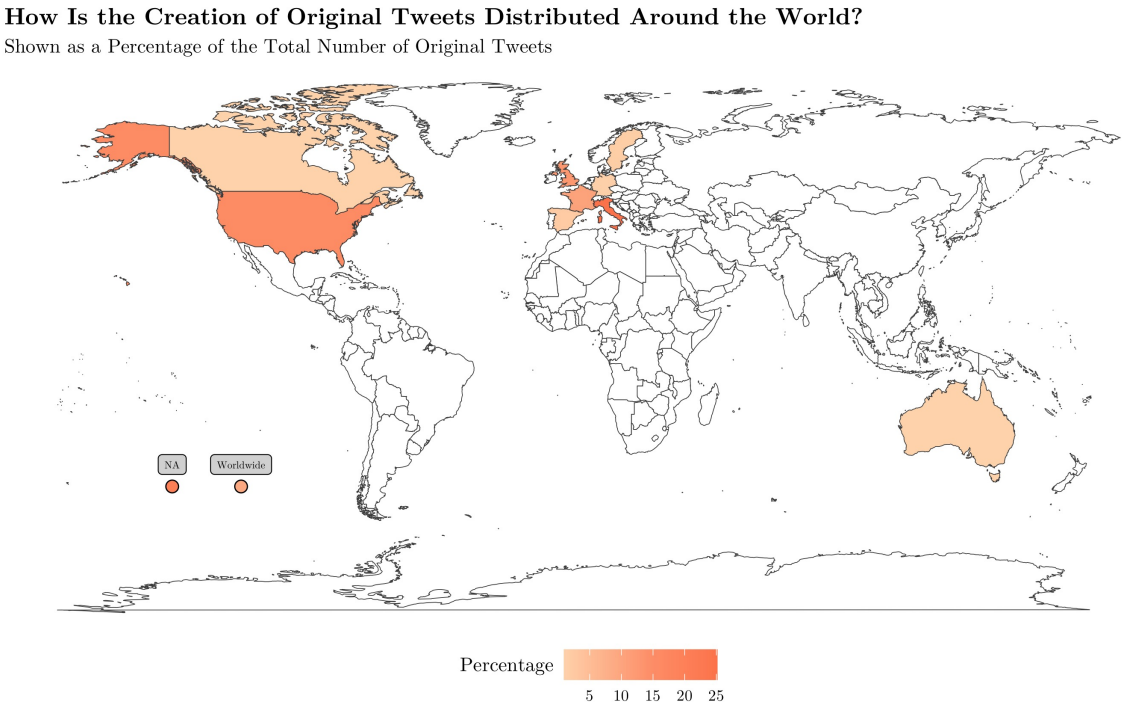
What Are the Most Common Words in the Accounts' Description According to the TROS?
That Is, the Tokens That Appear at Least 1 Time in the Description of the Accounts Author of the Original Tweets With a TROS Lower or Equal (Left) Versus Higher (Right) Than the Median

Note. This figure considers only the observations that belong to the train set.

In the descriptions associated with a lower TROS, it can be seen that some of the most frequent words are `strong`, `stand`, and `innovative`, as well as the names of the brands Monique Lhuillier and Needle & Thread. Meanwhile, in the descriptions associated with a higher TROS, it can be seen that Gianvito Rossi predominates, like it also does in the text of the original Tweets with a higher TROS. Also, `italy` can be seen, as well as some brand's names. For instance, `adidas`, which did a collaboration with Gucci; in fact, `adidasxgucci` appears as one of the most frequent tokens in the text of the original Tweets with a higher TROS. Other brand's names present are Aquazzura, Paco Rabanne, and RIMOWA.

Passing onto the variable `account_location`, Figure 27 indicates that the original Tweets' author accounts tend to have as their location a place referring to North America, Europe, or Oceania; while none of them have a place referring to Latin America nor Africa nor Asia. More specifically, they tend to have as their location a place referring to Italy, USA, or NA, followed by France, UK, or the world. In fact, as Figure 28 shows, the most frequent cities as the location of the accounts are Milan, New York, Paris, and London. These are precisely where the Big 4 Fashion Weeks take place (Y. Choi et al., 2021).

Figure 27: Distribution of Original Tweets Around the World



Note. This figure considers only the observations that belong to the train set.

Figure 28: Frequency of Accounts' City Location

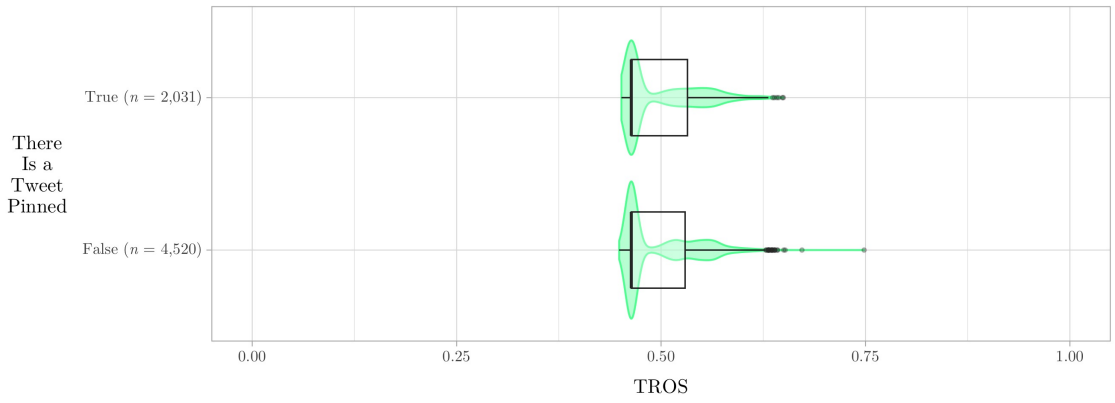
What Cities Are the Most Associated With the Original Tweets?
That Is, the Cities That Appear at Least Once as the Location of the Brands Author of the Original Tweets

Note. This figure considers only the observations that belong to the train set.

Moving onto the variable `account_pinned_tweet_id`, Figure 29 illustrates that the number of original Tweets whose author account has a Tweet pinned is less than double the number of those whose author account does not have it. Also, this figure shows that the distribution of the TROS is very similar between those two groups of original Tweets. However, the group without a pinned Tweet presents some outliers with a higher TROS than the other. A possible explanation for this can be that, as there is no pinned Tweet, when looking at the account's Twitter profile there is less scrolling needed to get to the original Tweet in question, increasing in this way the chances of it being seen and, thus, of generating more engagement.

Figure 29: Distribution of TROS by Whether There Is a Tweet Pinned

Does the TROS Vary According to Whether There Is a Tweet Pinned?
To the Top of the Brand's Profile



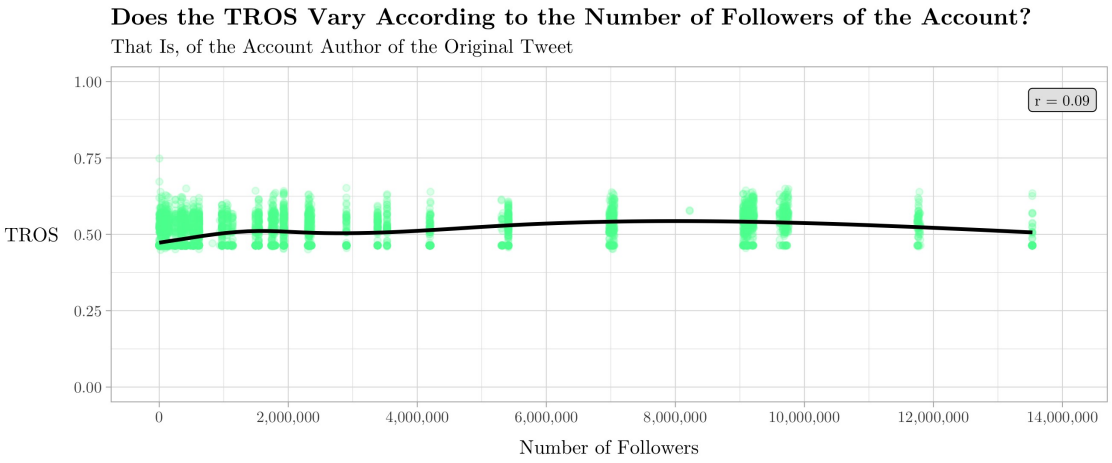
Note. This figure considers only the observations that belong to the train set.

Then, as to the variables `account_profile_image_url` and `account_protected`, all accounts associated with the original Tweets have a profile image and none of them are protected. This makes sense as it enables the accounts to have more visibility, which is something especially desirable for this kind of business accounts. It is worth adding that, initially, the variable `account_profile_image_url` was going to be given a function, which was to create an attribute

that indicates whether the account has a profile image. However, when observing the train set, it was found out that all observations have their corresponding URL. Therefore, they would all have had the value `TRUE` for this new attribute and so it lost its meaning. Meanwhile, the variable `account_protected` was going to be an attribute until it was found out that it has no variation whatsoever in the train set, thus also losing its meaning.

Now, with respect to the variable `account_public_metrics_followers_count`, Zohourian et al. (2018) also consider as an attribute the number of followers at the time the post is uploaded. Figure 30 shows that most original Tweets are associated with author accounts with less than 10,000,000 followers. Additionally, by looking at the black smoothed line, the TROS seems to have an upward trend up to that 10,000,000 threshold where it seems to start having a downward trend. The fact that the r is positive but very small kind of reflects the aforementioned.

Figure 30: Distribution of TROS Through Number of Followers

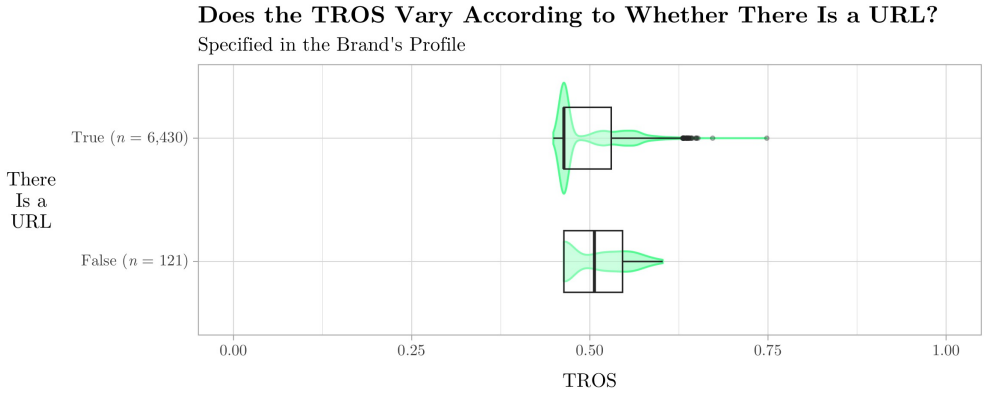


Note. This figure considers only the observations that belong to the train set.

As to the variable `account_public_metrics_following_count`, the r between it and the TROS equals -0.05, which is a very small negative value. Therefore, the higher the number of accounts followed by the author account, the slightly lower the TROS. Meanwhile, regarding the variable `account_public_metrics_listed_count`, it indicates the number of public lists of which the account is a member. In this case, the r equals 0.48, which is a relatively high positive value. Thus, the higher the number of public lists of which the author account is a member, the higher the TROS. This can be related to the fact that being a member of public lists probably increases one's Tweets' visibility and, thus, their chances of getting more engagement. Then, with respect to the variable `account_public_metrics_tweet_count`, the r in this case is 0.19, which is also a positive value but a little lower than the previous one. So, the higher the number of Tweets posted by the account author of the original Tweet in question, a little bit higher the TROS of the original Tweet.

Passing onto the variable `account_url`, Figure 31 illustrates that most original Tweets' author accounts have a URL specified in their profiles. Also, it shows that the median TROS for those original Tweets whose author accounts do not have a URL is higher than for those whose accounts do have one. Nevertheless, the latter present several outliers with a TROS higher than the maximum obtained by the former.

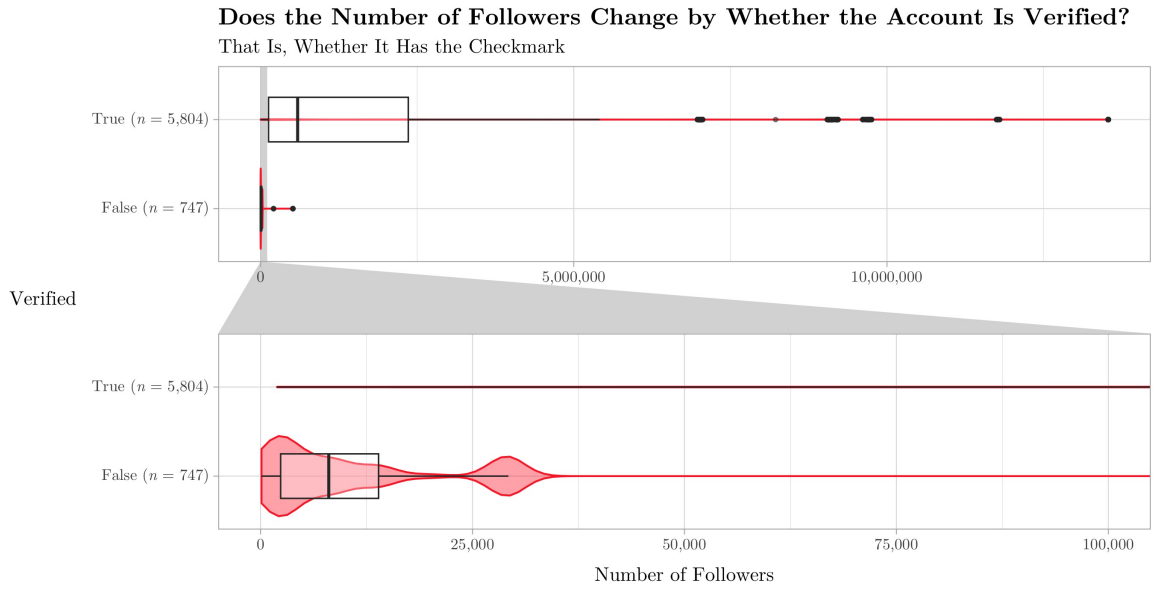
Figure 31: Distribution of TROS by Whether There Is a URL in the Accounts' Profile



Note. This figure considers only the observations that belong to the train set.

Finally, in terms of the variable `account_verified`, this variable indicates whether the account is verified. By manually inspecting the train set, it was noticed that the verified stands for both the blue and the gold check marks.[81] The blue one can mean two different things: Either the account has an active subscription to the Twitter Blue subscription service and has met the eligibility criteria (i.e., complete, active, secure, and non-deceptive), or the account was previously verified under the legacy verification criteria (i.e., active, notable, and authentic)[82]. Meanwhile, the gold check mark replaces the `Official` label on the businesses' accounts[83].

Figure 32: Distribution of Followers by Whether the Account Is Verified



Note. This figure considers only the observations that belong to the train set.

Having clarified what the `account_verified` variable stands for, Figure 32 demonstrates that most of the original Tweets are posted by verified accounts and that these accounts tend to have a higher number of followers than unverified ones. Meanwhile, Figure 33 shows that, although

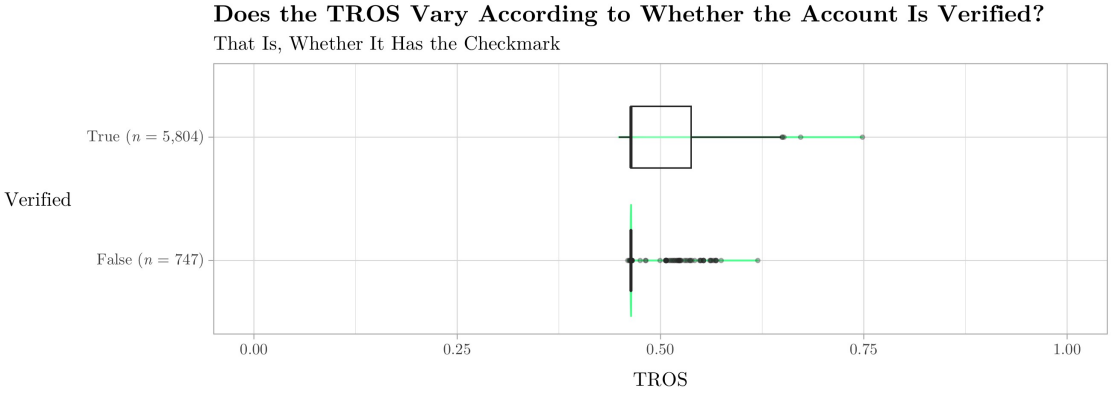
[81] It is not known if it also does so for the gray check mark, since none of the accounts in the sample correspond to a government institution or official, or a multilateral organization.

[82] https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.

[83] https://blog.twitter.com/en_us/topics/product/2022/twitter-blue-update.

the median is very similar, more original Tweets with verified author accounts have a higher TROS than the ones with unverified author accounts.

Figure 33: Distribution of TROS by Whether the Account Is Verified



Note. This figure considers only the observations that belong to the train set.

The following data set explored is the Google Trends one, which is made of separate smaller ones. Like with the previous data sets, it is begun by presenting a general description of the variables. This is shown in Tables 9, 10, and 11.

Table 9: Variables of Brands' Google Trends

Name	Type	Function	NA
`gtrend_brand_keyword`	character	ID	0.00
`gtrend_brand_date`	POSIXct	ID	0.00
`gtrend_brand_hits`	integer	Create attribute(s)	0.00

Note. This table considers only the observations that belong to the train set.

Table 10: Variables of Luxury's Google Trends

Name	Type	Function	NA
`gtrend_luxury_keyword`	character	ID	0.00
`gtrend_luxury_date`	POSIXct	ID	0.00
`gtrend_luxury_hits`	integer	Create attribute(s)	0.00

Note. This table considers only the observations that belong to the train set.

Table 11: Variables of Ethical Consumerism's Google Trends

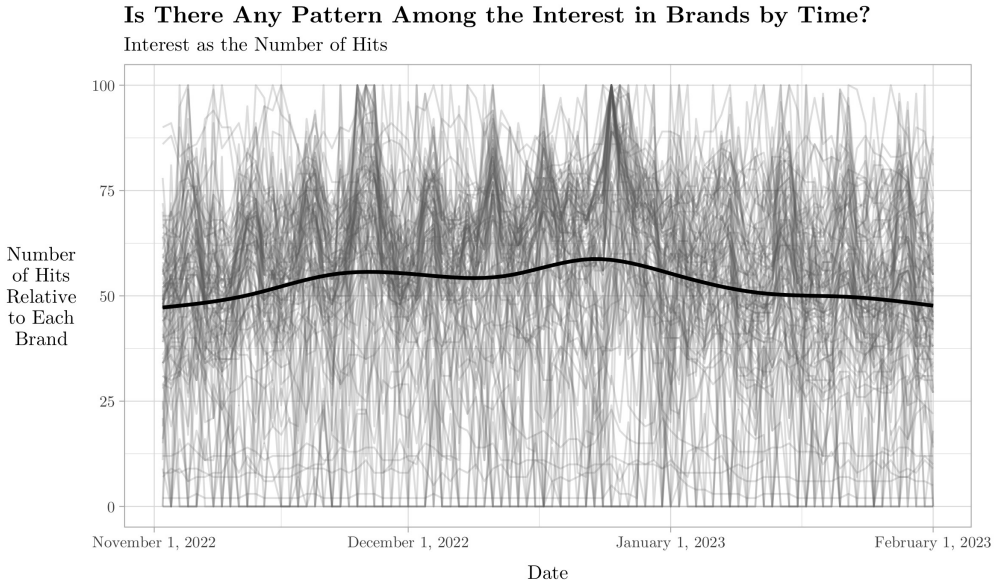
Name	Type	Function	NA
`gtrend_ethics_keyword`	character	ID	0.00
`gtrend_ethics_date`	POSIXct	ID	0.00
`gtrend_ethics_hits`	integer	Create attribute(s)	0.00

Note. This table considers only the observations that belong to the train set.

It is proceeded describing some of these variables more in detail and presenting several figures related to these variables that help to have a better overview of them. The first one is Figure 34, which represents with every thin line a different brand. In it, it can be seen that there is a

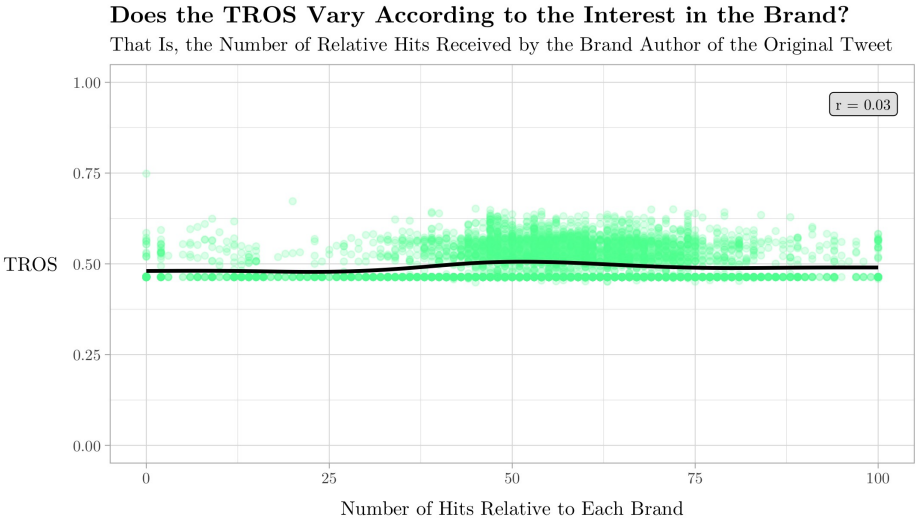
kind of darker shade formed by the agglomeration of several of these thin lines, which indicates that the interest in these brands tends to fluctuate similarly through time. Additionally, the black smoothed line shows that, generally, the interest in brands increased on the second half of November, as well as on the one of December, before the end of year festivities, which were followed by a decrease of the interest. Meanwhile, Figure 35 illustrates that there is only a very small positive r between the interest in the brand author of the original Tweet and the TROS of this Tweet.

Figure 34: Interest in Brands Through Time



Note. This figure considers only the observations that belong to the train set.

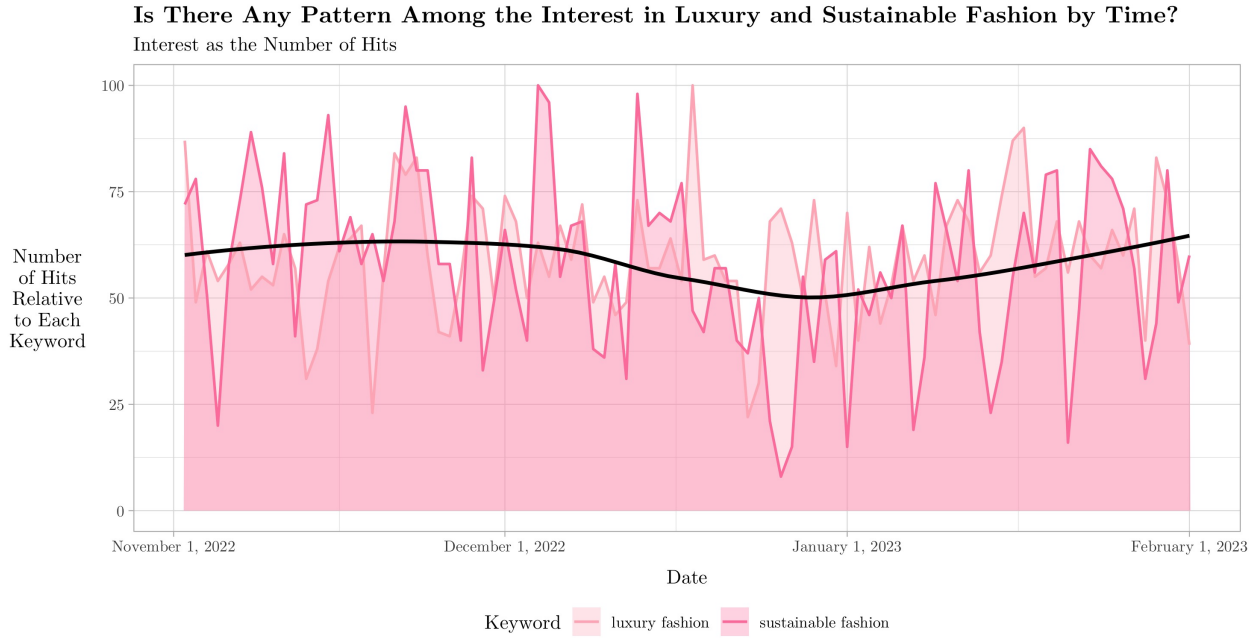
Figure 35: Distribution of TROS Through Interest in Brands



Note. This figure considers only the observations that belong to the train set.

Figure 36 shows that the interest in luxury and sustainable fashion tended to move quite similarly during November and December 2022 and then started doing so more differently. However, the black smoothed line shows that both interests increased during January, 2023. It is worth adding that it can also be seen that the minimum interest in sustainable fashion was lower than the one in luxury fashion.

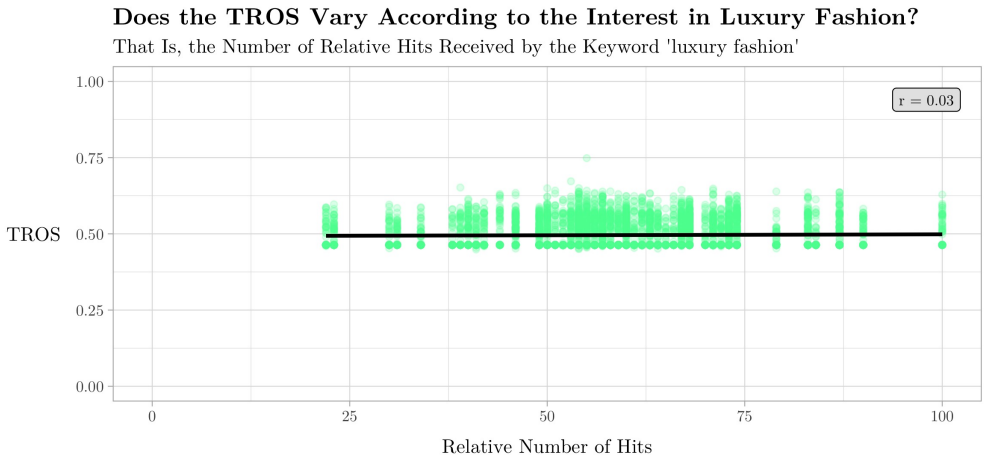
Figure 36: Interest in Luxury and Ethical Consumerism Through Time



Note. This figure considers only the observations that belong to the train set.

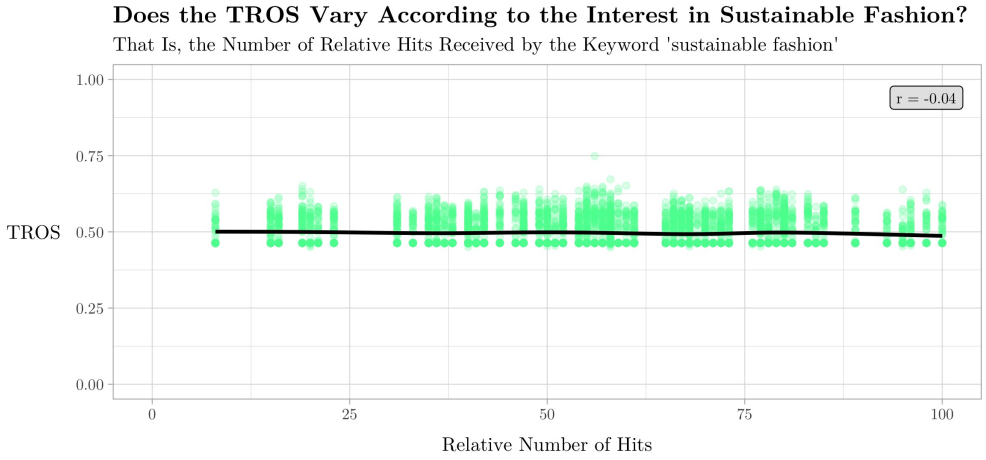
Figure 37 shows that there is a slightly positive linear association between the interest in luxury and the original Tweets' TROS. In contrast, in Figure 38, it can be seen that, interestingly, there is a slightly *negative* linear association between the interest in sustainable fashion and the original Tweets' TROS.

Figure 37: Distribution of TROS Through Interest in Luxury



Note. This figure considers only the observations that belong to the train set.

Figure 38: Distribution of TROS Through Interest in Ethical Consumerism



Note. This figure considers only the observations that belong to the train set.

The final data set explored is the Twitter Trends one. Like with the previous ones, it is started by presenting a general description of the variables. This is shown in Table 12.

Table 12: Variables of Twitter Trends

Name	Type	Function	NA
`ttrend_download_day`	numeric	ID	0.00
`ttrend_year`	numeric	ID	0.00
`ttrend_month`	numeric	ID	0.00
`ttrend_day`	numeric	ID	0.00
`ttrend_url`	character	ID	0.00
`ttrend_rank`	numeric	ID	0.00
`ttrend_trending_topic`	character	Create attribute(s)	0.00
`ttrend_tweet_volume`	character	None given	0.00
`ttrend_tweet_volume_in_k`	character	None given	0.00

Note. This table considers only the observations that belong to the train set.

It is proceeded describing some of these variables more in detail and presenting several figures related to these variables that help to have a better overview of them. First of all, there are usually 50 Twitter Trends per day at 00:00 UTC. It is said usually because there are some few cases in which there are less. For instance, for December 10, 2022 there are 49 instead of 50[84].

Then, Figure 39 shows the most frequent words in the Twitter Trends.[85] To do this figure, like with the previous word clouds, first the text was cleaned following the same process as for EmoLex, but removing only the `@` instead of also the name of the account mentioned and also removing the stop words and all types of punctuation. In addition, only tokens of at least 4 characters are considered.

[84] https://www.exportdata.io/trends/worldwide/2022-12-10/0.

[85] This analysis was done according to the TROS; like with the previous word clouds, but using the median TROS per day, since the Twitter Trends are per day and there can be more than one original Tweet per day. However, it was found out that the two groups of tokens are perfectly intersected. So, it was decided to do the analysis in general, to at least have an idea of what the Twitter Trends are usually about.

Figure 39: Frequency of Twitter Trends' Words

What Are the Most Common Words in the Twitter Trends?

That Is, the Tokens That Appear at Least 20,000 Times in the Text of the Twitter Trends

Note. This figure considers only the observations that belong to the train set.

In this figure, it can be seen that `copyright`, `register`, `trademark`, and `cent` are the most common words in the Twitter Trends. Furthermore, there are many words related to the Fédération Internationale de Football Association (FIFA) World Cup, last played at the end of 2022. Examples include names of football players representing their corresponding nations, like `neymar`, `mbappe`, `modric`, `ronaldo`, and `messi`; as well as of countries, such as `portugal`, `brazil`, `argentina`, and `francia`.

Figure 40 illustrates the results of calculating the cosine and Jaccard similarities between the original Tweets' text and the Twitter Trends. The cosine similarity is the most common similarity metric. Between two vectors $\mathbf{v}$ and $\mathbf{w}$, it can be computed as

$$\text{cosine}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| * \|\mathbf{w}\|} = \frac{\sum_{i=1}^{N}(v_i * w_i)}{\sqrt{\sum_{i=1}^{N} v_i^2} * \sqrt{\sum_{i=1}^{N} w_i^2}}. \tag{8}$$

The cosine value ranges from -1 for vectors pointing in opposite directions, 0 for orthogonal vectors, and 1 for vectors pointing in the same direction. Since frequency values are non-negative, the cosine for the vectors used here ranges from 0 to 1 (Jurafsky and Martin, 2021). Meanwhile,

the Jaccard similarity is also a metric of similarity. It is defined as

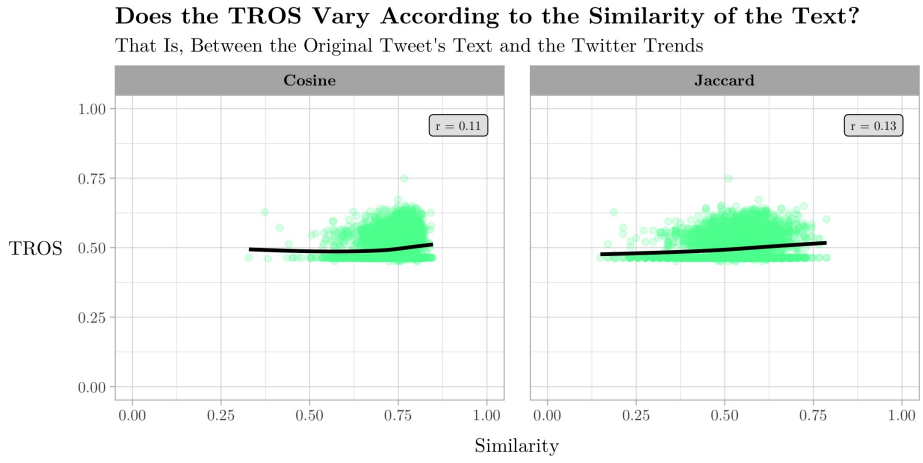
$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{9}$$

where $|C|$ refers to the number of elements in set C. The Jaccard similarity moves between 0, when A and B are completely different; and 1, when they are the same.

To calculate these two types of similarities, the R package `stringdist`[86] is used, specifically its function `stringsim`. This function considers similarity as a value between 0 and 1, where 0 corresponds to complete dissimilarity while 1, to perfect similarity. Before inputting the objects to this function, the text of both the original Tweets and the Twitter Trends is cleaned following the same process as for EmoLex, but removing only the `@` instead of also the name of the account mentioned and also removing the stop words and all types of punctuation.

It is worth adding that, for the text of each original Tweet, the similarity is computed trend by trend. In other words, the approximately 50 Twitter Trends of that day are not treated as one single text, but on their own. This is done because one day's Twitter Trends tend to be about quite different themes and, thus, it is very difficult for a single Tweet to be related to all of them. To get a single value for each original Tweet, the maximum among the obtained similarities is calculated. Therefore, the variables corresponding to the cosine and Jaccard similarities end up capturing the maximum similarity that the original Tweet's text has in regard to the Twitter Trends.

Figure 40: Distribution of TROS Through Texts' Similarity



Note. This figure considers only the observations that belong to the train set.

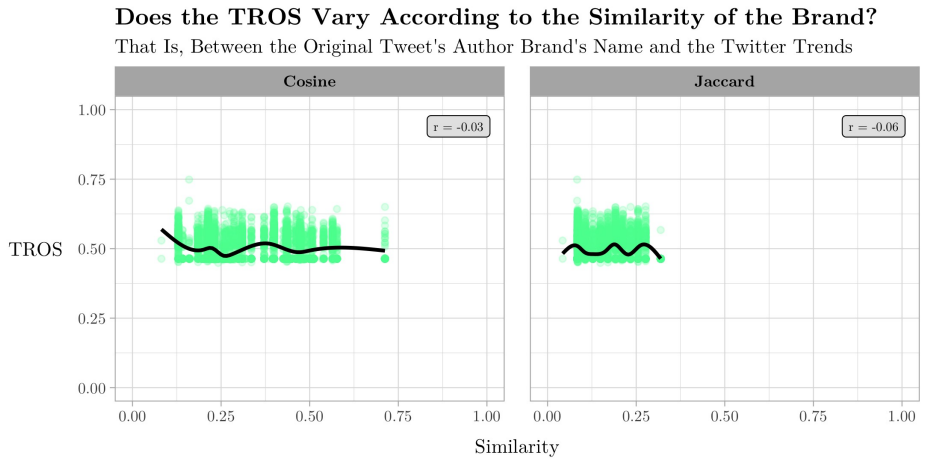
Having explained all the aforementioned, Figure 40 depicts that the higher any of the two types of similarities, the higher the TROS. This is supported not only by the dots and the black smoothed line, but also by the fact that both facets present a positive r.

Lastly, Figure 41 shows the same data as Figure 40, but between the name of the brands author of the original Tweets and the Twitter Trends. It is added because there are cases in which the Twitter Trend is precisely the name of a brand, which does not necessarily appear explicitly in the original Tweet's text. For instance, the word `Balenciaga` was the 43$^{\text{rd}}$ Twitter Trend on

[86] https://cran.r-project.org/web/packages/stringdist/stringdist.pdf.

November 29, 2022[87], while the word `Gucci` the 30th on January 10, 2023[88]. It is worth noting that, in this case, the cleaning process applied to the brands' name consists of only case folding, to be able to be matched to the potential brands in the Twitter Trends (that still pass through the same cleaning process as for EmoLex, but removing only the `@` instead of also the name of the account mentioned and also removing the stop words and all types of punctuation). So, in Figure 41, the black smoothed lines indicate that the relations in this case are not linear. This is supported by the fact that both facets contain an r extremely near zero.

Figure 41: Distribution of TROS Through Brands' Similarity



Note. This figure considers only the observations that belong to the train set.

3.5. Feature Engineering

The process of feature engineering consists on creating new variables from the data that allow the model to predict in a better way. This step is one of the most important ones since what the model learns depends to a good extent on the attributes it has available. Garbage in, garbage out: If the quality of what is inputted to the model is not good, then the result is usually not good either. Therefore, in this section, the performed feature engineering is described. It is worth mentioning that most of the variables stated next have appeared in at least one of the figures shown in the exploratory data analysis. They were first created provisionally and now they are definitely incorporated into the set of attributes.

Table 13 states which are the attributes created and, for each of them, from which default variable this is done. Several comments regarding this table are worth making.

[87] https://www.exportdata.io/trends/worldwide/2022-11-29/0.
[88] https://www.exportdata.io/trends/worldwide/2023-01-10/0.

Table 13: Feature Engineering

Default	Created
original_attachments_media_keys	original_attachments_media_keys_present
original_created_at	original_created_at_year original_created_at_month original_created_at_day_week original_created_at_weekday original_created_at_day_number original_created_at_hour original_created_at_minute original_created_at_second
original_text	original_text_length original_text_n_urls original_text_n_hashtags original_text_n_mentions original_text_n_emojis original_text_sentiment original_text_emotion_n_words original_text_emotion_joy original_text_emotion_trust original_text_emotion_anger original_text_emotion_disgust original_text_emotion_fear original_text_emotion_sadness Bag of words
account_created_at	account_created_at_year
account_description	account_description_length account_description_sentiment account_description_emotion_n_words account_description_emotion_joy account_description_emotion_trust account_description_emotion_anger account_description_emotion_disgust account_description_emotion_fear account_description_emotion_sadness
account_location	account_location_country account_location_city
account_pinned_tweet_id	account_pinned_tweet_present
account_url	account_url_present
gtrend_brand_hits	gtrend_brand_hits_adjusted
gtrend_luxury_hits	gtrend_luxury_hits_adjusted
gtrend_ethics_hits	gtrend_ethics_hits_adjusted
ttrend_trending_topic	ttrend_trending_topic_text_cosine ttrend_trending_topic_text_jaccard ttrend_trending_topic_brand_cosine ttrend_trending_topic_brand_jaccard cluster

First, with respect to the attributes created from `original_created_at`, previous research on digital marketing found that temporary programming of brand post might increase a company's revenue (Cuevas-Molano et al., 2021). However, the results of Cuevas-Molano et al. (2021) also show that time and day have no influence on engagement. Additionally, in relation concretely with the created attribute `original_created_at_weekday`, Cuevas-Molano et al. (2021) state that various studies on digital marketing have offered contradictory results on the impact of being or not a weekend on CE, and so to investigate whether brand posting time may influence engagement behavior, they distinguish between posts published on weekdays and weekends, as done in this thesis. Furthermore, Zohourian et al. (2018) consider the season, the month, the day of the week, the time of the day, and whether it was a holiday the day the post was done. Those are also considered here, except for season and holiday, since they highly depend on geography, while an international audience is here being considered.

Second, concerning the created attribute `original_text_length`, Cuevas-Molano et al. (2021) confirm that previous works have reported mixed results for the length of a brand's post on engagement, and these authors find that the number of characters has a positive effect on comments. In other words, they find that longer posts capture greater interest and participation from the audience.

Third, Zohourian et al. (2018) consider in their study the number of hashtags used in the post, like it is done in this thesis with the created attribute `original_text_n_hashtags`.

Fourth, in contrast to when the sentiment and emotion indicators of the dependent variable are built, when the analog attributes from `original_text` and `account_description` are created, a value is not automatically assigned to the cases in which there are no replies or Quote Tweets. This difference is due to the fact that there could not be NA values in the dependent variable's indicators since they would impede the calculation of the TROS for the corresponding observations, while there is no problem if there are some NA values in some of the attributes. In fact, some of the predicted models here implemented can automatically deal with NA values[89].

Fifth, regarding the bag of words created from `original_text`, this consists of representing the original Tweets' text as if they were an unordered set of words, with their position ignored and keeping only their frequency in each text (Jurafsky and Martin, 2021). Before creating this bag of words, the original Tweets' text is applied the same cleaning process as for EmoLex, but removing only the `@` instead of also the name of the account mentioned and also removing the stop words and all types of punctuation. Also, only tokens of at least 4 characters are considered. Then, the bag of words is initially constructed using only the observations that belong to the train set[90], to prevent data leakage, and the already created columns (i.e., one for each found token) are later completed for the observations belonging to the validation and the test sets. Additionally, the process of constructing the bag of words is done twice, according to the TROS. First, it is done considering only the original Tweets with a TROS lower than the median and, then, it is done considering only the rest of the original Tweets. Furthermore, only the columns (i.e., the tokens) that appear in only one of the two bags of words are kept, so that these columns can better help to differentiate the two groups of original Tweets and, thus, to predict the TROS. Lastly, to be kept, besides of being unique to one of the groups, the tokens must have a frequency in the bag of words of at least 10. In this way, a total of 53 tokens is left. It is worth adding that a bag of words neither from the accounts' description nor from the Twitter Trends are created

[89]This is explained more in detail in the predictive models' section.
[90]This is explained more in detail in the train, validation, and test sets' section.

because doing that would extremely increase the number of attributes for the available number of observations, when the information they would communicate is already captured in one way or another by some of the already present features.

Sixth, recall that, like it is explained in the section of the data collection, the retrieved search volume from Google Trends is relative instead of absolute. Thus, to generate consistency between the retrieved values in each of the three downloads, the previously mentioned "multipliers" are calculated and applied to the corresponding values. The created attributes `gtrend_brand_hits_adjusted`, `gtrend_luxury_hits_adjusted`, and `gtrend_ethics_hits_adjusted` are the result of this process.

Seventh, the variable `cluster` is the result of a K-means clustering. Clustering refers to a set of techniques for finding subgroups in a data set, in a way in which the observations within each group are quite similar to each other, while the observations in different groups are quite different from each other. It is an unsupervised problem, since there is no associated response measurement associated to each observation; there is no single right answer (James et al., 2021). For instance, J. Park et al. (2011) implement clustering to distinguish groups among social networks' users, using the results of their field research regarding which characteristics of social networks influence loyalty to luxury brands.

Together with hierarchical clustering, K-means clustering is one of the best-known clustering approaches.[91] To perform K-means clustering, following James et al. (2021), first the desired number of clusters K must be specified and, then, the K-means algorithm assigns each observation to exactly one of the K distinct, non-overlapping clusters. The algorithm does this assignment in a way such that the total within-cluster variation, summed over all K clusters, is as small as possible. The within-cluster variation for cluster C_k is a measure $W(C_k)$ of how much the observations within a cluster differ from each other. Therefore, the problem to solve is

$$\underset{C_1,\ldots,C_k}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}. \tag{10}$$

The most common choice to define the within-cluster variation involves the squared Euclidean distance:

$$W(C_k) = \frac{\sum_{i,i' \in C_k} \sum_{j=1}^{p} (X_{ij} - x_{i'j})^2}{|C_k|}, \tag{11}$$

where $|C_k|$ denotes the number of observations in the k^{th} cluster. Consequently, combining

[91] Despite the fact hierarchical clustering, in contrast to K-means clustering, does not require to pre-specify the number of clusters K and does result in an attractive tree-based representation of the observations, known as *dendrogram*; K-means clustering is chosen instead, because hierarchical clustering implies making additional decisions that can have a strong impact on the results obtained, like what dissimilarity measure should be used, what type of linkage should be used, and where the dendrogram should be cut to obtain the clusters (James et al., 2021).

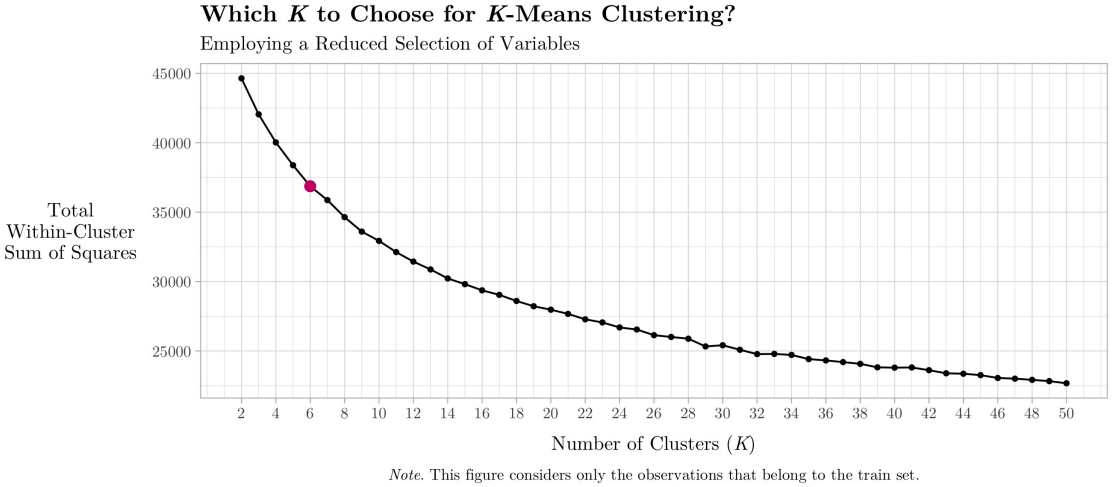
Equations 10 and 11, the optimization problem that defines K-means clustering is

$$\underset{C_1,\ldots,C_k}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{\sum_{i,i' \in C_k} \sum_{j=1}^{p} (X_{ij} - x_{i'j})^2}{|C_k|} \right\}. \tag{12}$$

This is a difficult problem to solve, since there are almost K^n ways to partition n observations into K clusters. This is a huge number unless K and n are tiny. So, an algorithm that provides a local optimum (which is a pretty good solution) to the K-means optimization problem is the following. Start by randomly assigning a number, from 1 to K, to each of the observations; these serve as initial cluster assignments for the observations. Then, iterate the following two steps until the cluster assignments stop changing: For each of the K clusters, compute the cluster *centroid* (i.e., the vector of the p feature means for the observations in the k^{th} cluster; this is from where K-means clustering derives its name) and assign each observation to the cluster whose centroid is closest (according to the Euclidean distance). This algorithm is guaranteed to decrease the value of the objective 12 at each step, until the result no longer changes and, thus, a local optimum has been reached (James et al., 2021).

Since the K-means algorithm finds a local rather than a global optimum, the results obtained depend on the initial random cluster assignment of each observation, in the first step of the algorithm. Therefore, it is important to run the algorithm multiple times with different random initial configurations. Then, the best solution is selected; in other words, the selected solution is the one for which the objective 12 is smallest. Additionally, to select K, different values are tried and the decision of which of them to select is usually based on a scree plot (which shows the total within-cluster sum of squares corresponding to each of the tried Ks) and following the (subjective) elbow criteria (i.e., the number where there is kind of an inflection point) (James et al., 2021).

Figure 42: Search for the Best Number of Clusters



Note. This figure considers only the observations that belong to the train set.

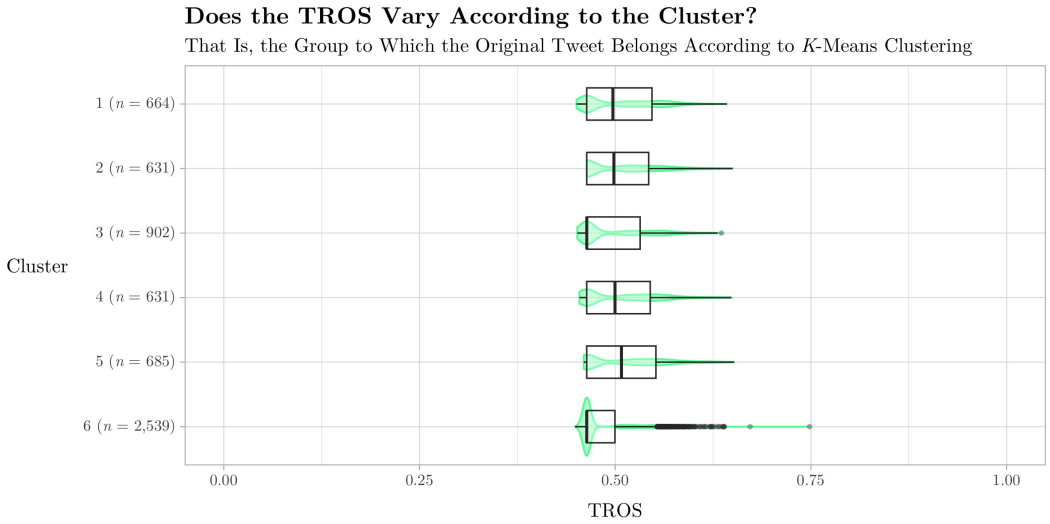
Figure 42 is the scree plot corresponding to the own K-means clustering analysis and it has $K = 6$ marked as the chosen number of clusters. This decision was based on the elbow criteria, as well as on the recommendation given by the `NbClust` function of the `NbClust` R package[92],

[92] https://cran.r-project.org/web/packages/NbClust/NbClust.pdf.

whose one of its indices also proposed 6 as the best number of clusters.

Meanwhile, Figure 43 illustrates that the dependent variable does vary according to the cluster. For instance, it can be seen that clusters 3 and 6 have the lowest median TROS, while clusters 1, 2, and 4 one almost equal to 0.5; and that cluster 5 has one slightly higher than 0.5.

Figure 43: Distribution of TROS Through Clusters



Note. This figure considers only the observations that belong to the train set.

The variables inputted to the K-means clustering algorithm were not all the available ones, but a selection of them, and they were previously scaled and had their extreme values removed. Specifically, the selected variables were

- `original_created_at_month`,
- `original_created_at_day_week`,
- `original_created_at_day_number`,
- `original_created_at_hour`,
- `original_created_at_minute`,
- `original_text_length`,
- `original_text_n_urls`,
- `original_text_n_hashtags`,
- `original_text_n_mentions`,
- `original_text_n_emojis`,
- `original_text_sentiment`,
- `original_text_emotion_n_words`,
- `original_text_emotion_joy`,
- `original_text_emotion_trust`,
- `original_text_emotion_anger`,
- `original_text_emotion_disgust`,
- `original_text_emotion_fear`,
- `original_text_emotion_sadness`,
- `account_description_length`,
- `account_description_sentiment`,

- `account_description_emotion_n_words`,
- `account_description_emotion_joy`,
- `account_description_emotion_trust`,
- `account_description_emotion_anger`,
- `account_description_emotion_disgust`,
- `account_description_emotion_fear`,
- `account_description_emotion_sadness`,
- `account_public_metrics_followers_count`,
- `account_public_metrics_following_count`,
- `account_public_metrics_listed_count`,
- `account_public_metrics_tweet_count`,
- `gtrend_brand_hits_adjusted`,
- `gtrend_luxury_hits_adjusted`,
- `gtrend_ethics_hits_adjusted`,
- `ttrend_trending_topic_text_jaccard`, and
- `ttrend_trending_topic_brand_jaccard`.

The Jaccard similarity variables were chosen instead of the cosine ones because, in the exploratory data analysis, it had been found that the former are slightly more correlated to the TROS than the latter.

Finally, for some particular predictive models, some more feature engineering is then carried out, since some models have specific requirements. For example, for regressions the log transformation matters and one-hot encoding does so for the learning algorithms that do not handle categorical attributes.[93]

3.6. Summary of Variables

Table 14 sums up the variables used to build the TROS, as well as the ones generally used as or main or control attributes to feed the models. It is said generally since, as previously mentioned, for some particular predictive models, some more feature engineering is carried out[94].

[93]This is explained more in detail in the predictive models' section.
[94]This is explained more in detail in the predictive models' section.

Table 14: Summary of Variables

Group	Members
Response construction	original_public_metrics_like_count original_public_metrics_retweet_count account_public_metrics_followers_count Text from replies Text from Quote Tweets
Main attributes	original_attachments_media_keys_present original_brand original_created_at_year original_created_at_month original_created_at_day_week original_created_at_weekday original_created_at_day_number original_created_at_hour original_created_at_minute original_created_at_second original_handle original_possibly_sensitive original_reply_settings original_source original_text_length original_text_n_urls original_text_n_hashtags original_text_n_mentions original_text_n_emojis original_text_sentiment original_text_emotion_n_words original_text_emotion_joy original_text_emotion_trust original_text_emotion_anger original_text_emotion_disgust original_text_emotion_fear original_text_emotion_sadness Bag of words from original_text
Control attributes	account_created_at_year account_description_length account_description_sentiment account_description_emotion_n_words account_description_emotion_joy account_description_emotion_trust account_description_emotion_anger account_description_emotion_disgust account_description_emotion_fear account_description_emotion_sadness account_location_country account_location_city account_pinned_tweet_id_present account_public_metrics_followers_count account_public_metrics_following_count account_public_metrics_listed_count account_public_metrics_tweet_count account_url_present account_verified gtrend_brand_hits_adjusted gtrend_luxury_hits_adjusted gtrend_ethics_hits_adjusted ttrend_trending_topic_text_cosine ttrend_trending_topic_text_jaccard ttrend_trending_topic_brand_cosine ttrend_trending_topic_brand_jaccard cluster

4. Methodology

4.1. Performance Metrics

To evaluate the performance of a learning method on the data set in question, a way to measure how well its predictions actually match the observed data is needed. In other words, one needs to quantify how close the predicted response value is against the real one for that observation. This is also needed to be able to compare the implemented models against the baseline (i.e., the historical median TROS associated with the brand author of the original Tweet), as well as among themselves.

Recall that the dependent variable is continuous. Problems with a quantitative response tend to be referred as regression problems, while those with a qualitative response are usually referred to as classification problems (James et al., 2021). Thus, this thesis' problem is of the regression type. In this context, various performance metrics are available.

Most of these metrics are based on the residual, which is the difference between the ith observed response value and the ith response value that is predicted by the model. Its formula is

$$e_i = y_i - \hat{y}_i, \tag{13}$$

where y_i is the observed value for the ith observation, and $\hat{y}_i$ is the predicted value for the ith observation using $\hat{f}$ which is the estimate for the unknown function f (James et al., 2021).

Having established their usual base, the most known performance metrics for regression problems are the following. To begin with, the mean absolute error (MAE) implies calculating the absolute difference between real and predicted values. The objective is to minimize it, as it is a loss. Its formula is

$$\text{MAE} = \frac{\sum_{i=1}^{n} |e_i|}{n}. \tag{14}$$

The MAE is not differentiable[95], so it cannot easily be used as a loss function. In addition, the MAE treats all errors the same, which can be beneficial in some business contexts but not in others.

Furthermore, the MAE is expressed in the same measure units as the variable to predict. Thus, it cannot be used to compare errors if their corresponding dependent variable is in a different measure unit. The mean absolute percentage error (MAPE) represents a solution to this. It is the MAE's percentage version, and its formula is

$$\text{MAPE} = \frac{\sum_{i=1}^{n} |p_i|}{n}, \tag{15}$$

where $p_i = e_i / y_i * 100$ is the residual expressed as a percentage.

The measure unit drawback is not relevant in this thesis' context, since all models to compare predict the same dependent variable defined as a unit-free measure. Nevertheless, it is useful to

[95] A differentiable function is characterized by its derivative existing at each point in its domain. Graphically, a differentiable function does not have a vertical tangent line at any of its interior points in its domain; it is smooth and does not contain any angle or break.

have the performance metric expressed as a percentage, since it clarifies the interpretation of the error and facilitates the determination of whether its magnitude is acceptable.

Then, one of the most commonly used metrics is the mean squared error (MSE), which can be seen as a sum of squared residuals. Its formula is

$$\text{MSE} = \frac{\sum_{i=1}^{n} e_i^2}{n}. \tag{16}$$

The MSE is small if the predicted responses are very close to the real responses, while large if for some of the observations the predicted and real responses differ substantially (James et al., 2021). By squaring the errors, it penalizes even the small ones, which leads to an overestimation of how bad the model is. Additionally, since it weighs all differences equally, large residuals have a high impact on the MSE. Consequently, this metric is sensitive to outliers.

Given the MSE's sensitivity to outliers, sometimes the median absolute deviation (MAD) is considered instead. Its formula is

$$\text{MAD} = \text{median}(|e_1|, ..., |e_n|). \tag{17}$$

The MAD is more robust to outliers than the MSE, but its mathematical properties are less favorable.

Furthermore, the MSE is on a different scale from the dependent variable. A variant that is easier to interpret for this metric is the root mean squared error (RMSE), which is another of the most commonly used metrics. It takes values in the range $[0, +\infty)$ and its formula is

$$\text{RMSE} = \sqrt{\text{MSE}}. \tag{18}$$

Therefore, the RMSE is in the same scale as the dependent variable, facilitating the interpretation. Besides, by square rooting the MSE, it handles the penalization of smaller errors and is less prone to struggle in the case of outliers.

Like the MAE, the RMSE also has a percentage version, named root mean squared percentage error (RMSPE) and whose formula is

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^{n} p_i^2}{n}}. \tag{19}$$

However, the RMSE and so the RMSPE too still penalize more the bigger errors, thus not being completely robust to outliers. The root mean squared logarithmic error (RMSLE) implies taking the logarithm of the RMSE, which slows down the scale of the error and thus drastically scales down the outliers, nullifying their effect. Its formula is

$$\text{RMSLE} = \sqrt{\frac{\sum_{i=1}^{n} (\log(y_i + 1) - \log(\hat{y}_i + 1))^2}{n}}. \tag{20}$$

The RMSLE considers the relative error between the predicted and real values. In addition, the scale of the error is not significant; what matters is only the percentage difference. Furthermore, the RMSLE penalizes the underestimation of the real value more severely than it does for the overestimation. More penalty is incurred when the predicted value is less than the real value,

while less penalty is incurred when the predicted value is more than the actual value. This is useful in business cases where the underestimation of the dependent variable is unacceptable but the overestimation can be tolerated.

Moving on, the residual standard error (RSE) is an estimation of the standard deviation; i.e., the average amount that the predicted response deviates from the real one. Its formula is

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}, \tag{21}$$

where $\text{RSS} = \sum_{i=1}^{n} e_i^2$ is the residual sum of squares[96] (James et al., 2021).

The RSE is considered a measure of the lack of fit of the model to the data. If the predictions obtained using the model are very close to the real outcome values, then the RSE is small, and it can be concluded that the model fits the data very well. In contrast, if $\hat{y}_i$ is very far from y_i for one or more observations, then the RSE can be quite large, indicating that the model does not fit the data well (James et al., 2021).

Finally, the R^2 represents the proportion of variance explained, so it takes on a value between 0 and 1 and is independent of the scale of Y. A perfectly fitting model leads to $R^2 = 1$, while $R^2 = 0$ means that the model in question does not do better than the baseline model[97]. Like James et al. (2021) state, its formula is

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}, \tag{22}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares, and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ is the y sample mean.

The R^2 has an advantage in terms of interpretation over the RSE since, unlike the RSE, it always lies between 0 and 1. However, it can still be challenging to determine what is a good R^2 value and, generally, this depends on the application. In addition, the R^2 always increases as more features are added (James et al., 2021), which is not a desired characteristic due to the following. If an irrelevant feature is added, the R^2 can stay constant or worse: start increasing, which is incorrect.

To control this aspect, there is the adjusted R^2. It is another common approach for selecting among a set of models that contain different numbers of variables. It is calculated as

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}. \tag{23}$$

A large value of adjusted R^2 indicates a model with a small test error. The model with the largest adjusted R^2 has only correct and no noise variables. Unlike the R^2, the adjusted R^2 charges a price for the inclusion of unnecessary variables in the model, through d (James et al., 2021).

For this study, the RMSPE is selected as the main performance metric because of the following reasons. First, the MAE and the MAPE are discarded to be considered as the main one since they

[96] Note that the RSS also appears in the numerator of the fraction in Equation 16.

[97] In the case of the classical linear regression, the baseline model includes only the intercept, which implies using the mean value of Y as the prediction for all observations (James et al., 2021).

treat all errors the same, which is not beneficial in this thesis' context, in which overestimating the TROS can be potentially more damaging than underestimating it; in other words, posting a Tweet that ends up being not so good as predicted is potentially more damaging than posting a Tweet that ends up being better than predicted. Second, the MSE is also discarded since it is on a different scale from the dependent variable, which hinders the interpretation. Third, the MAD too because the RMSPE also handles the MSE's sensitivity to outliers, while additionally being more commonly used. Fourth, the RMSE is also discarded because the RMSPE clarifies the interpretation of the error and facilitates the determination of whether its magnitude is acceptable. Fifth, the RMSLE too since it penalizes the underestimation of the real value more severely than it does for the overestimation when, in this thesis' context, it would be beneficial for it to be the other way around since, as mentioned before, overestimating the TROS can be potentially more damaging than underestimating it. Sixth, the RSE, the R^2, and the adjusted R^2 are also discarded because they are generally more secondary metrics that complement the main one. Seventh, the RMSPE shares the following RMSE's advantage: Through the square root, the penalization of smaller errors is handled, and the probability of struggling due to outliers is reduced, while large errors are still penalized a little bit more, which is desired for this context (to avoid posting a potentially bad Tweet as well as to avoid discarding a potentially good one, and to avoid wrongly getting the ranking among different Tweet options).

Therefore, the RMSPE is considered as the main metric and, thus, as the one to determine which model is best (among the tried ones). However, given their previously described characteristics, the RSE and the adjusted R^2 are reported as complementary performance metrics.

4.2. Train, Validation, and Test Sets

In this thesis, as it is generally the case, what is interesting is not how well the model performs on the data with which it was trained, but instead on previously unseen test data. Following the example given by James et al. (2021), suppose that the interest is in making an algorithm to forecast a stock's price based on the stock's previous returns. The model can be trained using stock returns from the last six months. However, what is really cared about is not how well the model predicts last week's stock price, but how well it predicts tomorrow's or next month's price. Mathematically, what is wanted to be known is whether $f(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a previously unseen test observation that has not been used to train the model (James et al., 2021).

In the absence of a very large designated test set that can be used to directly estimate the test error, there are a number of techniques that can be used to estimate this test error using the available training data. These techniques estimate the test error by holding out a subset of the training observations from the fitting process and then applying the learning method to those held out observations (James et al., 2021).

A first technique is the validation set approach. It is a very simple strategy which involves randomly dividing the available set of observations into two parts, a train set and a validation set. The model is fitted on the train set, and its performance is evaluated on the validation set. The resulting validation set error provides an estimate of the test error (James et al., 2021).

A second technique is the leave-one-out cross-validation (LOOCV). Like the previous approach, it involves splitting the set of observations into two parts. But, instead of creating two subsets of comparable size, a single observation is used for the validation set and the remaining ones for the train set. The model is fitted on the $n - 1$ training observations, and a prediction is made for the

excluded observation. This is a poor estimate of the test error because it is highly variable, as it is based on a single observation. Therefore, this procedure is repeated n times. In this way, the LOOCV estimate for the test error is the average of the n error estimates (James et al., 2021).

A third technique is the k-fold CV. It involves randomly dividing the set of observations into k groups, or folds, of approximately the same size. The first fold is treated as a validation set, and the model is fitted on the remaining $k-1$ folds. The error is then computed on the observations in the held-out fold. This procedure is repeated k times. Each time a different fold is treated as the validation set. Thus, this process results in k estimates of the test error. The k-fold CV estimate is computed by averaging those k estimates. It is worth noting that LOOCV is a special case of k-fold CV in which k is set to be equal to n (James et al., 2021).

Comparatively, first, in contrast to the validation approach which yields different results when applied repeatedly due to the randomness in the split, the LOOCV always yields the same results, as there is no such randomness. Second, the LOOCV has less bias than the validation set approach because the train sets of the former are larger than the ones of the latter. Therefore, the LOOCV tends not to overestimate the test error as much as the validation set approach does. However, the LOOCV has the potential to be expensive to implement, since the model has to be fitted n times. This can be very time consuming if n is large and if each individual model is slow to fit (James et al., 2021).

Given that in this thesis the n is large, the LOOCV is discarded, as well as the k-fold CV which is mainly for a medium n. The chosen approach is the validation set approach, which is simple and easy to implement (James et al., 2021). However, instead of being random, the split is done by date, since in practice, when having to train the model, one only has available Tweets from the past, not from the future. By splitting by date instead of randomly, this approach's drawback regarding the high variability of results if repeated is highly reduced. In addition, at least in this thesis' context, it is far worse for the approach to underestimate the test error than to overestimate it. Consequently, the drawback of the validation set approach regarding the overestimation of the test error is not major whatsoever. Instead, it should be seen as making this approach more conservative in terms of risk taking, which can be considered beneficial.

Moving on, as the flexibility of the model increases, there is a decrease in the training error, while there is a U-shaped movement in the test performance. This is a fundamental property of statistical learning that holds regardless of the particular data set and method used. As the model flexibility increases, the training error decreases, but the test error may not. In the extreme, *overfitting* can occur: When a given model yields a small training error but a large test one. This happens because the learning procedure is working too hard to find patterns in the training data and might be picking up some patterns that are caused just by random noise, instead of by true properties of the unknown function f. Thus, when the training data are *overfitted*, the test error is very large because the supposed patterns that the method found in the training data do not exist in the test data. It is worth clarifying that, regardless of whether overfitting has occurred, the training error is almost always expected to be smaller than the test one, because most learning methods directly or indirectly seek to minimize the training error. Overfitting refers concretely to the case in which a less flexible model would have generated a smaller test error (James et al., 2021).

If different tests are performed, it is possible to make overfitting on the validation set. In other words, if one tries many different options, it can happen that the model performs well on the validation set not because it captures true patterns, but because it is capturing noise from the

validation set. Consequently, many times three, instead of two, sets are used: the training, the validation, and the test sets. These work as follows. One trains many models on the train set and sees how they perform on the validation set. The model chosen is the one with the best performance on the validation set. Once that model is identified, it is trained again, but now with both the train and the validation sets' data. Then, one sees how this final model predicts on the test set, as an estimation of how this final model would perform in production. It is worth making explicit that the model is not chosen based on the performance on the test set because then this set would turn into a validation set.

In this thesis, these three sets are used: the train, the validation, and the test sets. Generally, the data are divided between those three, approximately, in the following percentages: 60% for the train set; while 20% for the validation and the test sets, each. 11,008 observations were collected, approximately 2,201 per month. Thus, the observations from November 6, 2022 to February 5, 2023 (both extremes fully included) correspond to the train set; while those from February 6, 2023 to March 7, 2023 (both extremes also fully included) correspond to the validation set and those from March 8, 2023 to April 6, 2023 (both extremes also fully included) to the test set. This resulted in 6,551 observations for the train set, 2,476 for the validation set, and 1,981 for the test set; 59.51, 22.49, and 18%, respectively.

4.3. Predictive Models

As explained in the section of problem and objective, a supervised learning problem is addressed. Many classical statistical learning methods, like linear regression, as well as more modern approaches, such as boosting and SVM, operate in the supervised learning domain. There are many approaches, rather than just a single best method, because there is no free lunch in statistics: No one method dominates all others over all possible data sets. On a particular data set, one specific method might work best, but another method may work better on a similar but different data set. Therefore, it is an important task to decide, for any given data set, which method produces the best results (James et al., 2021).

Additionally, like James et al. (2021) explain, there are a number of very powerful tools at our disposal, like random forest, boosting, SVM, and NNs, to name a few; and then there are the linear models and their simpler variants. When faced with new data modeling and prediction problems, it might be tempting to always go for the trendy new methods. They usually give extremely impressive results, especially when the data sets are very large and can support the fitting of high-dimensional non-linear models. Nevertheless, if models can be produced with the simpler tools that perform as well, they will probably be easier to fit and understand, as well as less fragile, than the more complex approaches. Therefore, following the recommendation of James et al. (2021), the simpler models are tried as well as the trendy new ones, and only then a choice is made.

Also, as mentioned in the performance metrics' section, recall that the dependent variable is continuous and thus the problem is of the regression type. In general terms, suppose that a quantitative response Y and p different predictors, X_1, X_2, ..., X_p, are observed. It is assumed that there is some relationship between Y and $X = (X_1, X_2, ..., X_p)$, which can be written in the very general form

$$Y = f(X) + \epsilon. \tag{24}$$

There, f is some fixed but unknown function of X_1, X_2, ..., X_p, and ϵ is a random error term,

which is independent of X and has mean zero. In this formulation, f represents the systematic information that X provides about Y (James et al., 2021).

In many settings, like this thesis' one, a set of inputs X is readily available, but the output Y cannot be easily obtained. Since the error term averages to zero, Y can be predicted using

$$\hat{Y} = \hat{f}(X), \tag{25}$$

where $\hat{f}$ represents the estimate for f, and $\hat{Y}$ represents the resulting prediction for Y. The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities: the reducible and the irreducible errors. Regarding the former, generally, $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error. This error is reducible because the accuracy of $\hat{f}$ can potentially be improved by using the most appropriate statistical learning technique to estimate f. Meanwhile, regarding the latter, even if it were possible to form a perfect estimate for f, so that the estimated response took the form $\hat{Y} = f(X)$, the prediction would still have some error in it. This is due to the fact that Y is also a function of ϵ, which, by definition, cannot be predicted using X. The quantity ϵ may contain unmeasured variables that are useful in predicting Y or variation that cannot be measured. The variability associated with ϵ also affects the accuracy of our predictions. This is known as the irreducible error, because no matter how well f is estimated, the error introduced by ϵ cannot be reduced (James et al., 2021).

Having given an overview, the predictive models concretely implemented are the following. To begin with, a baseline model is implemented; in other words, a straw man: A simple and sensible prediction that can be used as a baseline for comparison (James et al., 2021). Specifically, as anticipated in the section of problem and objective, this baseline consists on predicting the historical median TROS associated with the brand author of the Tweet.

Then, various models related to linear regressions are implemented. A linear regression is a useful and widely used statistical learning method for predicting a quantitative response. The most common approach to fit this model is the least squares one (Jurafsky and Martin, 2021).

A simple linear regression is an approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes there is an approximately linear relationship between X and Y. Mathematically, this linear relationship can be written as

$$Y \approx \beta_0 + \beta_1 X. \tag{26}$$

The "$\approx$" can be read as "is approximately modeled as". Meanwhile, β_0 and β_1 are two unknown constants that represent the intercept (i.e., the expected value of Y when $X = 0$) and the slope (i.e., the average increase in Y associated with a one-unit increase in X) terms, respectively, in the linear model. Together, they are known as the model coefficients or parameters. Once the training data has been used to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, a prediction can be made computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{27}$$

where $\hat{y}$ indicates a prediction of Y based on $X = x$. In practice, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unknown. The data must be used to estimate the coefficients. The goal is to estimate them in a way such that the linear model fits the available data well, so that $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$. By far, the most common approach for measuring closeness involves minimizing the least squares criterion. This is the approach taken, which chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS (James et al., 2021).

As James et al. (2021) explain, this simple linear regression model can be extended to directly

accommodate multiple predictors. This can be done by giving each predictor a separate slope coefficient in a single model. In general, suppose that there are p distinct predictors. Then, the multiple linear regression model takes the form

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p, \tag{28}$$

where X_j represents the j^{th} predictor and β_j is interpreted as the average effect on Y of a one-unit increase in X_j, holding all other predictors fixed. The regression coefficients β_0, β_1, ..., β_p are estimated using the same least squares approach, obtaining the multiple least squares regression coefficient estimates. Given estimates $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$, predictions can be made using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p. \tag{29}$$

It is worth adding that, to include qualitative predictors, a dummy variable is incorporated for each of its possible values, except for one of them, which constitutes the baseline category. The decision of which category is the baseline is arbitrary and has no effect on the regression fit, but does alter the interpretation of the coefficients (James et al., 2021).

This multiple version, applied by J. Park et al. (2011), is the linear regression first implemented here. The one initially implemented only includes the main attributes without NA values (to avoid an important reduction of the sample due to missing values in just one or a few attributes, since the algorithm omits observations with at least one NA value). Then, another implemented one includes all attributes without NA values. Lastly, the final implemented one only considers the significant attributes according to the p-values retrieved by the previous models; i.e., the attributes whose β is significantly different from zero and, thus, which do influence the response variable[98].

The linear regression model (both the simple and the multiple one) assumes a linear relationship between the response and the predictors, but the true relationship may be non-linear. Therefore, the linear model can be directly extended to accommodate non-linear relationships, using a polynomial function

$$y_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_d x_i^d. \tag{30}$$

The coefficients here also can be estimated using least squares linear regression and, for a large enough degree d, a polynomial regression allows to produce an extremely non-linear curve. However, it is unusual to use a d greater than 3 or 4 because for large values of d, the polynomial curve can become overly flexible (James et al., 2021).

A polynomial regression is also implemented, using as a starting point the best of the three previous linear regressions in the validation set, which was the one using only the main attributes[99]. Taking into account what was mentioned in the previous paragraph, a d equal to 3 is chosen. To decide to which variables to apply the polynomial function of third degree, the figures shown in the exploratory data analysis are considered and attributes whose relationship to the TROS is likely non-linear are looked for. As a result, it is applied to the variables

- `original_created_at_day_number`,
- `original_created_at_day_hour`,
- `original_text_emotion_joy`,

[98] More on this later in this same section.
[99] More on this later in the next section.

- `original_text_emotion_trust`,
- `original_text_emotion_anger`,
- `original_text_emotion_disgust`,
- `original_text_emotion_fear`, and
- `original_text_emotion_sadness`.

An alternative is a *piecewise* polynomial regression, which involves fitting separate low-degree polynomials (typically, cubic) over different regions of X, instead of a high-degree polynomial over the entire range of X. These polynomials are fitted under the constraint that the fitted curves must be continuous, as well as their first and second derivatives at the corresponding knots (i.e., the points where the coefficients change). This model can also be fitted using least squares and is known as (cubic) splines. In practice, it is common to place the knots in a uniform fashion, by specifying the desired degrees of freedom and then having the software automatically place the chosen number of knots at uniform *quantiles* of the data, and to use CV to objectively determine that number (James et al., 2021).

A piecewise polynomial regression is implemented, as well. The starting point used is also the best of the previous regressions, which at this point was the polynomial one. Then, the piecewise polynomial function is applied to the same variables as for the polynomial regression, and the knots are placed according to the respective first, second, and third quartiles.

As seen in the section of data, many are the attributes in this problem, both main and control ones. Generally, adding additional attributes that are truly associated with the response will improve the fitted model, in the sense of leading to a reduction in the test set's error. However, adding noise features will increase the *dimensionality* of the problem, exacerbating the risk of overfitting without any potential upside in terms of improved test set's error. Therefore, using more attributes is a double-edged sword: It can lead to improved predictive models if those attributes are relevant to the problem in question, but will lead to worse results if they are not. Even if they are relevant, the variance incurred in fitting their coefficients may outweigh the reduction in bias that they bring (James et al., 2021).

For example, in the multiple regression setting with p predictors, it is needed to ask whether there is no relationship between the response and the predictors. To answer this question, a hypothesis test is used. The null hypothesis H_0: $\beta_1 = \beta_2 = ... = \beta_p = 0$ is tested versus the alternative hypothesis H_a: at least one β_j is non-zero. This hypothesis test is performed by computing the F-statistic, whose formula is

$$F = \frac{TSS - RSS}{p}. \tag{31}$$

When there is no relationship between the response and the predictors, one would expect the F-statistic to take a value close to 1. Contrarily, if H_a is true, then F is expected to be greater than 1. For any given value of n and p, any statistical software package can be used to compute the p-value associated with the F-statistic using this distribution. Based on this p-value, it can be determined whether or not to reject the H_0. The p-value is the probability of observing any number equal to $|t|$ or larger in absolute value, assuming $\beta_j = 0$. It is interpreted as follows: A small p-value indicates that it is unlikely to observe such a substantial association between the predictors and the response due to chance, in the absence of any real association between the predictors and the response. Therefore, if the p-value is small, then it can be inferred that there is an association between the predictors and the response. So, if the p-value is small enough,

the null hypothesis is rejected (i.e., it is declared that a relationship exists between X and Y). Typical p-value cutoffs for rejecting the null hypothesis are 1, 5 or 10%. If it is concluded on the basis of that p-value that at least one of the predictors is related to the response, then it is natural to wonder which are the "guilty" ones. Usually, the response is associated with only a subset of the predictors. The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is known as variable selection (James et al., 2021).

In this context of many attributes, it is necessary to carry out selection or regularization to reduce the models' variance and increase the predictive quality. Various are the tools available. One is computing the variance inflation factor (VIF). It implies calculating the $p * p$ correlation matrix between attributes and discarding those with a VIF higher than 5. Nevertheless, this tool is not practical for a medium or large number of attributes and does not capture correlation between three or more (James et al., 2021).

A second available tool is represented by the different automatic selection methods. There are mainly three. The first one is the exhaustive approach, which explores all the possible models and so is computationally very expensive. In fact, it is recommended for when p is at most around 35. The second one is the forward approach, which reduces the number of tried models by keeping attributes once they are decided to be added. Thus, it is a greedy approach and the winner is not necessarily the optimum. The third approach is the backward one. Instead of starting with a model with no attributes and adding them gradually, it starts with a model with all attributes and discards them gradually. Once an attribute is discarded, it is no longer considered in the rest of the models. The forward and the backward approaches are useful for when p is big (but not greater than n) and they can be combined, by computing both, comparing their corresponding winners, and keeping the one with the highest performance in the validation set (James et al., 2021). This last combination approach is the one here employed. It is worth adding that, before starting applying each method, the NA values are replaced with the median or the mode of the variable, according to whether it is continuous or categorical, since these methods would otherwise directly omit observations with at least one NA value. Additionally, one-hot encoding[100] is applied to both the Boolean and the categorical variables.

A third group of tools is composed by the shrinkage methods. They require standardizing the variables and, contrarily to the automatic methods, imply that the coefficients' estimates can no longer be interpreted. One is the ridge regression, also known as L2 regularization; while the other is the lasso regression, also called the L1 regularization. The former generates weight vectors with very small weight, whereas the latter generates sparse solutions with some larger weights but many more weights set to zero. Consequently, the lasso regression leads to far fewer attributes (Jurafsky and Martin, 2021). Mathematically, they both imply adding a shrinkage penalty in the quantity to minimize and a hyper parameter λ (which is usually learned by CV) to control the impact of that penalty. In the ridge regression, that penalty implies β_j^2; while in the lasso regression, it implies $|\beta_j|$. Therefore, when $\lambda \to \infty$, the coefficients' estimates of the ridge regression approach to zero but none of them equals exactly zero (unless $\lambda = \infty$), thus including all the attributes in the final model. Meanwhile, when λ is sufficiently large, some coefficients' estimates of the lasso regression are forced to be equal to zero, thus performing variable selection.

[100] One-hot encoding is the process of creating dummy variables to handle a qualitative predictor. It implies generating as many columns as possible categories and placing, for each observation, a 1 in the position corresponding to its category, while zeros elsewhere (James et al., 2021). In this way, the resulting structure has only one 1 in each row and is mainly made up of zeros.

Additionally, in contrast to the coefficients' estimates in the ridge regression, the coefficients' estimates in the lasso regression do not change their sign when changing the λ (James et al., 2021).

Elastic nets seek to rescue the best of ridge and the best of lasso. They do so by adding to the objective function the weighted average of both types of penalties, which is then multiplied by λ. The weights are determined by the *hyperparameter* (i.e., a parameter that has to be tuned beforehand, instead of being learned during the model training) α, which can be learned using a validation approach. If α equals 0, then the elastic net is simply a ridge regularization; while if α equals 1, then it is a lasso one. Thus, any α between 0 and 1, none of the two extremes included, generate a combination of both types of penalties. The higher the α, the more importance is given to the lasso type while less to the ridge one.

These three shrinkage methods are implemented. For the three of them, the hyperparameter λ is tuned and, for the elastic nets, the hyperparameter α is also tuned. The data are processed in the same way as with the automatic selection methods: The NA values are replaced by the median or the mode of the variable, and one-hot encoding is applied to both the Boolean and the categorical attributes.

The last group of tools has to do with principal components analysis (PCA). It is a dimension reduction technique for regression; it is a popular approach for deriving a low-dimensional set of attributes (i.e., the principal components) from a large set of variables, loosing as little information as possible. It is used only with quantitative variables and, before generating the principal components, if variables are measured in different units, they should be standardized by, for instance, dividing them by the standard deviation. In this way, they are all on the same scale, avoiding the scale to have an effect on the principal components obtained. Having standardized if necessary, the first principal component direction of the data is that along which the observations vary the most. The second principal component is a linear combination of the variables that is uncorrelated with the first principal component (i.e., its direction must be perpendicular to the first principal component) and has the largest variance subject to that constraint. And so on with the following principal components. How many principal components to retain is a decision that can be made following the accumulated variance criteria (i.e., as many as those who achieve to capture between 70 and 90% of the variance) or looking at a scree plot (which, in this case, shows the variance captured by the number of principal components) and following the elbow criteria (James et al., 2021).

The principal components regression (PCR) approach involves constructing the first M principal components, Z_1, ..., Z_M, and then using these components as the predictors in a linear regression model that is fitted using least squares. The main idea is that usually a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response. By estimating only $M << p$ coefficients, overfitting can be mitigated. It is worth noting that, although PCR provides a simple way to perform regression using $M < p$ predictors, it is not a feature selection method. This is due to the fact that each of the M principal components used in the regression is a linear combination of all p of the original attributes. Consequently, while PCR usually performs quite well in many practical settings, it does not result in the development of a model that relies upon a small set of the original features. In this sense, PCR is more closely related to the ridge regression than to the lasso one (James et al., 2021). Additionally, like the ridge and lasso regressions, the estimated coefficients cannot be interpreted.

In the PCR approach, the directions are identified in an unsupervised way: The response Y is not used to help determine the direction of the principal components. Therefore, there is no guarantee that the directions that best explain the predictors will also be the ones that best predict the response. A supervised alternative to PCR is partial least squares (PLS). Like PCR, PLS is a dimension reduction method that first identifies a new set of attributes Z_1, ..., Z_M which are a linear combination of the original ones and, then, fits a linear model via least squares using these M new attributes. However, unlike PCR, PLS identifies these new attributes in a supervised way: It makes use of the response Y in order to identify new attributes that not only approximate the old attributes well, but also that are related to the response and help explain it. To do so, when constructing the principal components, PLS places the highest weight on the variables that are most strongly related to the response. In practice, PLS usually performs no better than ridge regression or PCR: While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance, so the overall benefit of PLS relative to PCR can be a wash (James et al., 2021).

Both PCR and PLS are implemented. It is done with the NA values already replaced with the corresponding median or mode (to avoid observations with at least one NA value being directly omitted) and Boolean as well as categorical attributes transformed by being applied one-hot encoding. Also, for these methods, the variable `cluster` is not considered since, theoretically, it does not make much sense, as it would imply making a reduction of dimensionality of an already reduced dimensionality. Nevertheless, this variable was tried to be included just in case and what was found was that, in practice, it also made more sense to exclude it, since the performance on the validation set improved when this variable was not part of the input data. Furthermore, all considered attributes are scaled before the principal components are computed. Finally, the number of components considered is tuned for each of the two methods.

The following implemented model is a K-nearest neighbors (KNN) regression. The idea is to estimate the response for a non-seen instance using the responses of instances that are the closest to the non-seen instance. Specifically, given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that are the closest to x_0, represented by $\mathcal{N}_0$. It then estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$. In other words,

$$\hat{f}(x_0) = \frac{\sum_{x_i \in \mathcal{N}_0} y_i}{K}. \tag{32}$$

The choice of K has a drastic effect. A small value for K provides a very flexible fit, which will have a low bias but high variance. This variance is because the prediction in a given region is entirely dependent on just one observation. In contrast, larger values of K provide a smoother and less variable fit. The prediction in a region is an average of several points, and so changing one observation has a smaller effect. However, the smoothing may cause bias by masking some of the structure in $f(X)$ (James et al., 2021). Therefore, an intermediate K is usually looked for using a validation method. This is exactly what it is done here using the holdout set approach.

Additionally, KNN regression works better with continuous attributes but can deal with categorical ones applying them one-hot encoding, and all attributes should be standardized so that their variance does not affect the distance measure. What is decided here is to input to this model only continuous attributes to improve the chances of it working better, and to standardize these, subtracting the mean and dividing the result by the standard deviation of each corresponding variable; both the mean and the standard deviation calculated considering only the training observations to avoid data leakage.

It is worth adding that the decrease in performance as the dimension increases is a common problem for KNN. It results from the fact that, in higher dimensions, there is a reduction in sample size: When p is larger, there are higher chances that a given observation has no nearby neighbors. This is known as the curse of dimensionality. The K observations that are nearest to a given test observation x_0 might be very far away from x_0 in a p-dimensional space, when p is large, leading to a very poor prediction of $f(x_0)$ and thus a poor KNN fit. As a general rule, parametric methods (i.e., those that explicitly assume a parametric form for $f(X)$), like linear regression, tend to outperform non-parametric ones when there is a small number of observations per predictor (James et al., 2021). This is also why it is decided not to include categorical attributes, since doing so would imply having to apply them one-hot encoding and, thus, extremely increasing the number of attributes and, so, also the risk of the curse of dimensionality.

The next group of implemented models is related to tree-based methods. The process of building a regression tree has two main steps. First, divide the predictor space (i.e., the set of possible values for X_1, X_2, ..., X_p) into J distinct and non-overlapping rectangular regions, R_1, R_2, ..., R_J. That shape is chosen for simplicity and for ease of interpretation of the resulting predictive model. Second, for every observation that falls into the region R_j, make the same prediction, which is the mean of the response values for the train's observations in R_j (James et al., 2021).

The objective is to find regions R_1, R_2, ..., R_J (or, in other words, the decision rules) that minimize the RSS. Since it is computationally infeasible to consider every possible partition of the feature space into J rectangles, a top-down greedy approach is taken, known as recursive binary splitting. It is top-down because it starts at the top of the tree, at which point all observations belong to a single region, and then successively divides the predictor space; each split is indicated via two new branches further down on the tree. It is greedy because, at each step of the tree-building process, the best split at that particular step is made, instead of looking ahead and picking a split that will lead to a better tree in some future step (James et al., 2021).

This process continues until the RSS equals zero, so it is likely to *overfit* the data from the train set, leading to a poor test set performance. The maximum possible number of regions equals the number of observations; while if there is only one region, the prediction equals the mean of the response values corresponding to the train set. Since a neither very deep nor very shallow tree is good, a balance is needed. Thus, what is actually done is to grow a very large tree and then prune it back in order to obtain a *subtree* (James et al., 2021). To prune it, there are three *hyperparameters* that work as stopping criteria. First, the maximum depth of the tree (`max_depth`), which is the maximum number of nodes along the longest path from the root node down to the farthest leaf node. Second, the minimum split (`min_split`), which is the minimum number of observations in a decision node to make a split. Third, the minimum bucket (`min_bucket`), which is the minimum number of observations there must be in each leaf node for the split to be accepted and so it has to be lower than the `min_split`. The higher the `max_depth`, the deeper the tree; while the higher the `min_split` or the `min_bucket`, the shallower the tree.

A decision tree is implemented, considering both the main and the control attributes, since this method selects the important features by itself. First, a version with the default values of the hyperparameters is implemented and, then, several tuned ones modifying the `max_depth`, `min_split`, and `min_bucket` are implemented. It is worth clarifying that, before implementing the tuned versions, the default one is implemented to know whether the tuning makes sense or not.

It is worth adding that this algorithm's default action is to delete all observations for which y is

missing, but to keep those in which one or more predictors are missing. At a decision node, if an observation has an NA value for the predictor in the corresponding rule, then the algorithm defines surrogate variables[101], which are found by reapplying the partitioning algorithm without recursion to predict the two categories, corresponding to the branches of that decision node, using the other independent variables. The surrogates are ranked by error, and any surrogate variable which does no better than the blind rule "go with the majority" is discarded from the list. So, any observation which is missing the split variable is classified using the first surrogate variable, or if missing that, the second surrogate is used, and so on. If an observation is missing all the surrogates, then the blind rule is used.

Decision trees may outperform classical approaches, like linear regression, if there is a highly non-linear and complex relationship between the features and the response. Additionally, they can easily handle qualitative predictors without the need to create dummy variables (James et al., 2021) and are ideal in contexts of mixed variables. Furthermore, they are an embedded method: The attributes important to predict Y are selected by the learning algorithm itself. Finally, they are easy to interpret.

Nonetheless, trees can be very non-robust: A small change in the data can cause a large change in the final estimated tree. A solution is to produce multiple trees and then combine them to yield a single consensus prediction. This usually results in dramatic improvements in performance, at the expense of some loss in interpretation. To do this ensemble, there are several methods, but they all have in common being an approach that combines many simple "building block" models to obtain a single and potentially very powerful model. These simple building block models are also known as weak learners, since on their own they might lead to poor predictions (James et al., 2021).

One of these ensemble methods is bagging, also called bootstrap aggregation. It is a general procedure to reduce the variance of a statistical learning method and makes use of the fact that averaging a set of observations reduces the variance. Specifically, it implies generating B different bootstrapped training data sets (i.e., taking B samples with replacement from the training data set; each sample has the same size as the original training data set; inside each sample, there can be repeated observations; different samples can have common observations; and it is not equivalent to obtaining new information, since the independence assumption does not stand), training the method on the b^{th} bootstrapped training set in order to get $\hat{f}^{*b}(x)$, and averaging all the predictions to obtain

$$\hat{f}_{\text{bag}}(x) = \frac{\sum_{b=1}^{B} \hat{f}^{*b}}{B}. \tag{33}$$

Although bagging can improve predictions for many regression methods, it is particularly useful for decision trees. To apply bagging to regression trees, B regression trees have to be constructed, using B bootstrapped training sets, and the resulting predictions have to be averaged. These trees are grown deep and are not pruned. Thus, each tree has high variance, but low bias, and by averaging these B trees, the variance is reduced. By combining hundreds or thousands of trees into a single procedure, bagging has been shown to provide impressive improvements (James et al., 2021).

It is worth mentioning the following. On average, each bagged tree makes use of around two-thirds

[101] https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

of the observations. The remaining one-third of the observations not used to fit a given bagged tree are known as the out-of-bag (OOB) observations. The response for the i^{th} observation can be predicted using each of the trees in which that observation was OOB. This yields around $B/3$ predictions for the i^{th} observation. To obtain a single prediction for the i^{th} observation, these predicted responses can be averaged. In this way, an OOB prediction can be obtained for each of the n observations, from which the overall OOB error can be computed. This resulting OOB error is a valid estimate of the test error for the bagged model, since the response for each observation is predicted using only the trees that were not fitted using that observation (James et al., 2021). However, this approach cannot be used for this thesis' problem because it implies the risk of making a prediction using trees that have seen original Tweets posted later in time, than the original Tweet to predict for. So, instead, the previously explained validation set approach is the one employed.

When applying bagging to regression trees, the hyperparameters are two. First, B, which is the number of trees. Since it is not sensitive to overfitting, the maximum B possible should be chosen, bearing in mind the own computational restrictions. The idea is to increase B until the estimated empirical error flattens. Second, the complexity of the trees given, for instance, by n_{min}, which represents the minimal node size. This hyperparameter is sensitive to overfitting and thus should be optimized using the validation set.

Bagging is implemented, also considering both main and control attributes, as well as the default and several tuned versions modifying n_{min}. However, since this method does not allow NA values in the predictors, before implementing the model, the NA values are replaced with the median or the mode of the corresponding continuous or categorical features, respectively.

The main disadvantage of bagging is the following. Suppose that there is one very strong predictor in the data set, together with other moderately strong predictors. Then, most or all of the bagged trees will use that strong predictor in their top split. Therefore, all of the bagged trees will look quite similar, thus the predictions from the bagged trees being highly correlated. Averaging any highly correlated quantities does not lead to as large of a reduction in variance as doing so for many uncorrelated quantities. Consequently, in this setting, bagging will not lead to a substantial reduction in variance over a single tree (James et al., 2021).

Random forest deals with that kind of setting, sampling not only observations, but also features. It overcomes the problem by forcing each split to consider only a subset of the predictors. As a result, on average, $(p - m)/p$ of the splits will not even consider the strong predictor, and so other predictors will have a higher chance. It can be thought of as *decorrelating* the trees, thus making the average prediction of the resulting trees less variable and more reliable. More specifically, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically $m \approx \sqrt{p}$ is chosen. So, when building a random forest, at each split in the tree, the algorithm is not allowed to consider a majority of the available predictors. Bagging is a particular case of random forest where $m = p$ (James et al., 2021).

Regarding the value selection for the hyperparameters, as with bagging, random forests do not overfit if B is increased. So, in practice, a sufficiently large value of B is used to settle the error rate. Meanwhile, the value of m should be small to be helpful when having a large number of correlated predictors (James et al., 2021). In regression problems, as a general rule, $m = p/3$ and $n_{min} = 5$. Also, m and some hyperparameter related to the complexity of the trees, like n_{min},

can be learned using the validation set.

Like with bagging, random forest is implemented considering both main and control attributes, as well as the default and several tuned versions, replacing the NA values with the median or the mode according to the type of the feature. However, in this case, in the tuned versions, different values are tried for not only n_{min}, but also m.

A different ensemble method is boosting. Like bagging, boosting is a general approach that can be applied to many statistical learning methods for both regression and classification (James et al., 2021). Given this thesis' problem, it is next explained applied to the context of regression decision trees.

While bagging and random forest ensemble complex models to reduce their individual variance, boosting ensembles simple and interdependent models to obtain a complex estimator. Interdependent is said since the trees are grown sequentially: Each tree is grown using information from previously grown trees, so the construction of each tree strongly depends on the trees that have already been grown. Also, in contrast to bagging and random forest, boosting does not involve bootstrap sampling; instead, each tree is fitted on a modified version of the original data set (James et al., 2021).

Concretely, according to James et al. (2021), the boosting algorithm for regression trees is the following. Firstly, set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set. Secondly, for $b = 1, 2, ..., B$, repeat the following three steps. First, fit a tree $\hat{f}^b$ with d splits ($d + 1$ terminal nodes) to the training data (X, r). Second, update $\hat{f}$ by adding in a shrunken version of the new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$. Third, update the residuals: $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$. Thirdly, and lastly, output the boosted model, which is

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x). \tag{34}$$

In this way, the boosting approach learns slowly. Each of the trees can be rather small, with just a few terminal nodes, determined by the hyperparameter d in the algorithm. By fitting small trees to the residuals, $\hat{f}$ is slowly improved in areas where it does not perform well. The shrinkage parameter λ, which is another hyperparameter, slows the process down even further, allowing more and different shaped trees to attack the residuals. Generally, statistical learning approaches that learn slowly tend to perform well (James et al., 2021).

So, in summary, boosting has three hyperparameters. First, the number of trees B. In contrast to bagging and random forest, boosting can overfit if B is too large, but this overfitting generally occurs slowly if at all. Second, λ, which is a small positive number that controls the rate at which boosting learns. In other words, it determines when the model starts overfitting. If λ is small, then the model starts doing so later; if it is big, then sooner. Typical values are 0.01 or 0.001, and a very small λ can require using a very large value of B to achieve a good performance. Third, the number d of splits in each tree, which controls the complexity of the boosted ensemble. In contrast to random forest, in boosting, since the growth of a particular tree takes into account the other trees that have already been grown, smaller trees are typically sufficient. The value of each of these three hyperparameters is selected according to which one generates the "best" performance on the validation set (James et al., 2021).

When implementing boosting, first, logical attributes are converted into numeric ones, so that they can be considered by the method and, then, both the default and several tuned versions are

implemented. These tuned versions imply trying different values of the three aforementioned hyperparameters: B, λ, and d.

It is worth adding that this algorithm handles NA values in the following way. Observations are divided into left, right, and missing splits. If the observation is in the missing split, then the algorithm applies a surrogate split method like the one used by the classic decision tree method, previously explained.

Another ensemble method that uses decision trees as its building blocks corresponds to Bayesian additive regression trees (BART). It is related to both random forest and boosting: Each tree is built in a random manner as in the former, and each tree tries to capture signal not yet accounted for by the current model as in the latter. The main novelty of BART is the way new trees are generated. Following James et al. (2021), let K denote the number of regression trees, and B the number of iterations for which the BART algorithm will run. The notation $\hat{f}_k^b(x)$ represents the prediction at x for the k^{th} regression tree used in the b^{th} iteration. At the end of each iteration, the K trees of that iteration are summed:

$$\hat{f}^b(x) = \sum_{k=1}^{K} \hat{f}_k^b(x), \quad b = 1, 2, ..., B. \tag{35}$$

Still following James et al. (2021), in the first iteration of the BART algorithm, all trees are initialized to have a single root node, with $\hat{f}_k^1(x)$ equal to the mean of the response values divided by the total number of trees. In subsequent iterations, BART updates each of the K trees, one at a time. In the b^{th} iteration, to update the k^{th} tree, the predictions are subtracted from each response value, from all but the k^{th} tree, to obtain a partial residual for the i^{th} observation, $i = 1, ..., n$. Instead of fitting a fresh tree to this partial residual, BART randomly chooses a perturbation to the tree from the previous iteration from a set of possible perturbations, favoring those that improve the fit to the partial residual.[102] To this perturbation, there are two components: What might be changed is the structure of the tree by adding or pruning branches, or the prediction in each terminal node of the tree. Consequently, the output of BART is a collection of prediction models.

The first few of these prediction models are generally thrown away, since models obtained in the earlier iterations (known as the burn-in period) tend not to provide very good results. So, let L denote the number of burn-in iterations. To obtain a single prediction, the mean is computed after the L burn-in samples (James et al., 2021).

Consequently, when applying BART, what must be selected is the number of trees K, the number of iterations B, and the number of burn-in iterations L. Generally, large values are chosen for K and B, while moderate values are chosen for L. For example, $K = 200$; $B = 1,000$; and $L = 100$. BART has been shown to perform very well with minimal tuning (James et al., 2021). BART is implemented, both the default and several tuned versions modifying K, B, and L.

It is worth adding that BART handles missing values in the predictors, applying hot decking imputation, through which each missing value is replaced with an observed response from a "similar" unit[103].

[102] The reason behind this method's name is that BART can be seen as a Bayesian approach to fitting an ensemble of trees: Each time a tree is randomly perturbed to fit the residuals, a new tree is being drawn from a posterior distribution.

[103] https://cran.r-project.org/web/packages/BART/BART.pdf.

One last ensemble method is XGBoost. There exists a generalization of boosting named Gradient Boosting Machine (GBM) that allows performing classification, regression, and ranking. XGBoost is a better formalization and implementation of GBM. In fact, XGBoost is a very important algorithm in both the academia and the industry. It is one of the most powerful and flexible models in machine learning.

Seven are the main hyperparameters in XGBoost[104]. First, `nrounds` which refers to the number of trees. Second, `max_depth` which indicates the maximum depth of the trees. Third, `eta` which is what previously was presented as λ; it is the learning rate, which can go from 0 to 1. Fourth, `gamma` which refers to the minimum error reduction to generate a cut. Fifth, `colsample_bytree` which indicates the proportion of variables to sample and consider in each tree. Sixth, `min_child_weight` which is the minimum number of observations in the children to consider a cut; it is like `min_bucket`, previously described in the context of a single decision tree. Seventh and last, `subsample` which refers to the proportion of observations to consider in each tree. XGBoost can do overfitting. Generally, it does more overfitting the higher the `nrounds`, `max_depth`, `eta`, `colsample_bytree`, or `subsample`; while the lower the `gamma` or `min_child_weight`.

XGBoost is also implemented. Like with the previous models, the default version as well as several tuned ones are implemented. These tuned versions imply trying different values for each of the seven aforementioned hyperparameters, employing a random grid search approach, except for `nrounds` which is controled through early stopping. This is a regularization technique that consists on analyzing the evolution of the error in the incremental sequence of ensemble models and stopping the training after n number of iterations in which the error on the validation set did not get reduced. Lastly, it is worth mentioning that XGBoost considers NA values as "missing"[105].

Moving on, the next group of implemented models is related to SVM, like the ones applied by L. Liu et al. (2018), Romão et al. (2019), and Zohourian et al. (2018). In the context of regression, the objective of an SVM is to find a hyperplane[106] to fit the data. This fit must be in a way such that the distance of the hyperplane to each training point is not greater than ϵ, which is a very small number chosen before training the model; i.e., a hyperparameter. However, if ϵ is too small, then the solution might not exist. The problem consists of a convex optimization problem to learn the parameters β; in other words, the objective is to minimize the coefficients. This model can be relaxed by introducing slack variables. For any value that falls outside of ϵ, its deviation from the margin can be denoted as ξ. By introducing the slack variables, when learning the model's parameters, some observations are allowed to be at a distance greater than ϵ from the regression function. The amount they are allowed depends on the hyperparameter C, which determines the trade-off between the flatness of f and the amount to which deviations larger than ϵ are tolerated. If C is increased, then the slope is variable and the variance is higher. Meanwhile, if C is decreased, then the slope is small and the bias is higher. Furthermore, the model can be turned non-linear by introducing K_σ. The hyperparameters ϵ, C, and σ can be learned using the validation set.

To implement SVM, first, all non-numeric attributes are turned into numeric ones. To do so,

[104] https://xgboost.readthedocs.io/en/latest/parameter.html.

[105] https://cran.r-project.org/web/packages/xgboost/xgboost.pdf.

[106] In two dimensions, a hyperplane is a flat one-dimensional subspace; thus, a line. Meanwhile, in three dimensions, a hyperplane is a flat two-dimensional subspace; so, a plane. In general terms, in a p-dimensional space, a hyperplane is a flat subspace of dimension $p - 1$ (James et al., 2021).

one-hot encoding is applied to the factor ones, and the Boolean values `TRUE` and `FALSE` are respectively converted into 1 and 0. Also, since this models' implementation does not handle NA values, they are replaced with the train median of the corresponding attribute. Then, the attributes are scaled considering only the train set, and a Gaussian kernel[107] is chosen because it adapts well to all contexts. Finally, the default version is implemented, as well as several tuned ones modifying ϵ and C.

The final group of implemented models is related to NNs (recall that these initials stand for neural networks). Deep learning is a very active area of research in the machine learning and, more broadly, the artificial intelligence communities. The cornerstone of deep learning is the NN, whose name originally derived from thinking of its hidden units as analogous to neurons in the brain (James et al., 2021).

Following James et al. (2021), an NN takes an input of p variables $X = (X_1, X_2, ..., X_p)$ and builds a non-linear function $f(x)$ to predict the response Y. What differentiates NNs from the previous models is the particular structure of the model. The NN model has the form

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X) = \beta_0 + \sum_{k=1}^{K} \beta_k g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j). \tag{36}$$

It is built in the following way. First, the K *activations* A_k, $k = 1, ..., K$, in the hidden layer are computed as functions of the input features $X_1, ..., X_p$,

$$A_k = h_k(X) = g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j), \tag{37}$$

where $g(z)$ is a non-linear activation function specified in advance. These K activations from the hidden layer then feed into the output layer, resulting in

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k A_k. \tag{38}$$

All the parameters $\beta_0, ..., \beta_K$ and $w_{10}, ..., w_{Kp}$ need to be estimated from the data. So, in words, the model derives new features from the inputted ones by computing different linear combinations of X and then passes each through an activation function $g(.)$ to transform it. The final model is linear in these derived variables (James et al., 2021).

Regarding the activation function, it is essential for it to be non-linear, since without this property the model would collapse into a simple linear model in $X_1, ..., X_p$ and would not be able to approximate (almost) all functions. The universal approximation theorem states that a feed-forward network with only one hidden layer is enough to approximate, with an arbitrary precision, any function with a finite number of discontinuities, *as long as the activation functions in the hidden neurons are non-linear*. The non-linear activation functions are the ones that allow the model to capture complex non-linearities and interaction effects (James et al., 2021).

Previously, the *sigmoid* function used to be the chosen one. However, when x tends to either of the two infinities, the gradient of this function tends to zero. The multiplication of small gradients due to the chain rule (which is part of the *backpropagation* algorithm used to calculate

[107] A kernel is a function that quantifies the similarity of two observations (James et al., 2021).

the gradient of a function when fitting an NN) vanishes the gradient. Consequently, the preferred choice in modern NNs is the rectified linear unit (ReLU) activation function, which takes the form

$$\text{ReLU}(x) = \max(0, x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise.} \end{cases} \tag{39}$$

Therefore, if what flows through the network are positive values, there is a gradient signal and the problem of the vanishing gradient is avoided. However, this does not apply for the negative values: The gradient is 0 for the negative xs and this can lead to a process known as Dying ReLU. Therefore, the Leaky ReLU activation function was created. It avoids the Dying ReLU problem by generating gradients different from zero also for the negative values of x. Formally,

$$\text{LeakyReLU}(x) = \begin{cases} 0.01x & \text{if } x <= 0 \\ x & \text{otherwise.} \end{cases} \tag{40}$$

Finally, the parametric ReLU is the general version of the Leaky ReLU, where 0.01 can instead be any α value (different from zero). In this thesis, the Leaky ReLU is chosen, since it deals with the vanishing gradient problem, as well as with the Dying ReLU problem, and avoids having to decide the value of α.

When the response is quantitative, like in this thesis' problem, when fitting the NN to estimate the unknown parameters, what is usually minimized is the squared-error loss. The problem is nonconvex in the parameters. To overcome this problem and reduce the chances of doing overfitting, two general strategies are employed. First, slow learning: The model is fitted in a slow iterative fashion, using gradient descent. Second, regularization: Penalties are imposed on the parameters, usually lasso or ridge[108] (James et al., 2021).

The idea of gradient descent is the following. Suppose all the parameters are represented in one long vector θ. Then, the objective is

$$R(\theta) = \frac{\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2}{2}. \tag{41}$$

So, start with a guess θ^0 for all the parameters in θ and set $t = 0$. Iterate the following steps until the objective fails to decrease: Find a vector δ that reflects a small change in θ, such that $\theta^{t+1} = \theta^t + \delta$ reduces the objective; i.e., such that $R(\theta^{t+1}) < R(\theta^t)$; and set $t \leftarrow t + 1$ (James et al., 2021).

To find the directions to move θ so as to decrease the objective $R(\theta)$, the gradient of $R(\theta)$ is calculated, evaluated at some current value $\theta = \theta_m$, which is the vector of partial derivatives at that point:

$$\nabla R(\theta^m) = \left.\frac{\partial R(\theta)}{\partial \theta}\right|_{\theta=\theta^m}. \tag{42}$$

This gives the direction in θ-space in which $R(\theta)$ increases most rapidly. Thus, the idea of gradient descent is to move θ a little in the opposite direction:

$$\theta^{m+1} \leftarrow \theta^m - \rho \nabla R(\theta^m), \tag{43}$$

[108] More on this later in this same section.

where ρ is a small value referring to the learning rate. The calculation of this vector of partial derivatives is quite simple, even for complex networks, thanks to the chain rule of differentiation. The act of differentiation assigns a fraction of the residual to each of the parameters via the chain rule and this process is known as backpropagation (James et al., 2021).

Gradient descent usually takes many steps to reach a local minimum. In practice, this process can be accelerated by, for instance, when n is large, instead of considering all n observations, sampling a small fraction or *minibatch* of them each time a gradient step is computed. This process is known as stochastic gradient descent (SGD) and is the state of the art for learning deep NNs (James et al., 2021).

The higher the number of units per layer, the more important is regularization to avoid overfitting. One form of regularization used in the context of NNs is weight decay. It is also known as L2 or ridge regularization, which was previously explained. To recall, it generates weight vectors with very small weight, by adding a shrinkage penalty in the quantity to minimize and a hyperparameter λ to control the impact of that penalty. In this way, weight decay stops the model from giving too much importance (i.e., weight) to a feature or neuron. That penalty implies β_j^2. Therefore, when $\lambda \to \infty$, the coefficients' estimates of the ridge regression approach to zero but none of them equals exactly zero (unless $\lambda = \infty$), thus including all the attributes in the final model (James et al., 2021). Additionally, an excessive λ tends to lead to *underfitting*, while one extremely near 0 tends to lead to overfitting. Thus, a medium one is looked for using, for instance, a grid search mechanism together with a validation set.

So, briefly, there are a number of choices that have an effect on the performance of an NN. For instance, the number of units per layer. However, the modern thinking is that the number of units per hidden layer can be large, and that overfitting can be controlled via regularization. Therefore, one of the main choices that has an effect on the performance of an NN is the λ regularization tuning parameter in the case of ridge regularization. These choices can make a difference. Finer tuning and training of a network can reduce the error, but the tinkering process can be tedious and can result in overfitting if done carelessly (James et al., 2021).

Consequently, when designing the implemented NN with a single fully-connected hidden layer, λ is tuned, as well as the number of units in that layer. It is worth adding that, before inputting the attributes to this model, factor and Boolean attributes are transformed into numerical ones, through one-hot encoding and binary representation, respectively. Also, to avoid the gradient descent algorithm having to give bigger steps in some directions relative to others only due to the attributes' different scales, the features are standardized (subtracting the mean and dividing the result by the standard deviation of each corresponding variable, both the mean and the standard deviation calculated considering only the training observations to avoid data leakage). Finally, since the default action to handle NA values is for the procedure to fail[109], NA values are replaced with the corresponding train median.

Modern NNs usually have more than just one hidden layer and, generally, many units per layer. Despite what the universal approximation theorem states, the learning task of discovering a good solution is made much easier with multiple layers, each of modest size. In this way, through a chain of transformations, the network is able to build up quite complex transformations of X that finally feed into the output layer as features. Each element $A_k^{(1)}$ feeds to the second hidden layer L_2 via the matrix of weights $\mathbf{W}_2$ (James et al., 2021).

[109] https://cran.r-project.org/web/packages/nnet/nnet.pdf.

Here, regularization is essential to avoid overfitting. One form of regularization used in the context of NNs and not previously described is dropout. It is a relatively new and efficient form of regularization, in which the idea is to randomly remove a fraction ϕ of the units in a layer when fitting the model. To avoid the model assigning too much responsibility to a single neuron or feature, in each iteration of the gradient descent algorithm, neurons are ignored in the forward pass during training, randomly with a probability $1 - p = \phi$. In practice, dropout is achieved by randomly setting the activations for the dropped-out units to zero, keeping in this way the architecture intact (James et al., 2021).

An NN with more than a single fully-connected layer is also implemented. Specifically, one with two hidden layers, both with Leaky ReLU as their activation function and with dropout as the chosen regularization technique. Additionally, the number of units is tuned, as well as ϕ, for each layer. It is worth mentioning that the data are applied the same transformations as with the single hidden layer model before being inputted. So, among other things, the NA values are replaced by the corresponding train median. Finally, 32 and 10 are chosen as the batch size and the number of epochs (i.e., the number of times the full training set has been processed), respectively, since they are the "usual" values.

And what about long short-term memory? LSTM is a variant of recurrent neural networks (RNNs) (Jurafsky and Martin, 2021). In this study, LSTMs are not implemented because, as seen in the Deep Learning course that belongs to the master's degree in question, this kind of architecture is nowadays not used as much, given transformers. Just in case, recall that transformers were previously mentioned in this thesis, when explaining and justifying the selected method to calculate the sentiment of the replies and Quote Tweets. Additionally, going back to LSTMs, these models take a long time to train and, thus, make exploring different architectures and optimizing parameters an extremely tedious task (James et al., 2021).

So, in summary, several predictive models are implemented, based on different learning algorithms, including: multiple linear, polynomial, and piecewise polynomial regressions; PCR; PLS; KNN; decision tree; random forest; boosting; SVM; and NN.

After having implemented all those predictive models, given the results, it was decided to implement some more models, as a check. Concretely, five more categories of models were implemented, for each of which it was tried the default, the already tuned, and a new tuned version of the two best models at the moment, which were the tuned boosting and, then, the tuned random forest.

The first category is made up of models whose only predictors are the 53 token variables created from the bag of words of the original Tweets' text. Meanwhile, the second category is of models whose predictors are the result of a bag of words of also the original Tweets' text, but considering all tokens of at least 4 characters and present at least 10 times, having cleaned the text in the same way as with the previous bag of words (i.e., the same cleaning process as for EmoLex, but removing only the `@` instead of also the name of the account mentioned and also removing the stop words and all types of punctuation). These first two categories represent a rather simple bag-of-words model.

The third category groups models that, before being trained, the selection method known as Boruta is applied to all the available attributes, except the ones that resulted from the

bag of words, which are not considered in these models[110]. The Boruta technique consists on the following. For each attribute, a corresponding "shadow" attribute is created by shuffling the values of the original attribute across the observations. Then, a random forest model is implemented[111], using all attributes, and the importance for each of them is computed. The importance of a shadow attribute can be different from zero only due to random fluctuations, and so the importance of each of these shadow attributes is used as a reference to decide which of the original attributes are truly important. Attributes that have a significantly worse importance than shadow ones are consequently dropped, while attributes that are significantly better than shadow ones are confirmed. This procedure is repeated to obtain statistically valid results. The algorithm stops when only confirmed attributes are left or the maximum number of indicated iterations has been reached[112]. In this way, Boruta's main objective is to find all attributes for which their correlation with the response variable is higher than that of the random attributes (Kursa and Rudnicki, 2010). For this thesis' case, from the 54 inputted attributes, Boruta discarded 4 of them and confirmed the remaining ones. Specifically, it discarded the attributes `account_description_emotion_anger`, `account_description_emotion_disgust`, `original_possibly_sensitive`, and `original_reply_settings`.

The fourth category is of models which use the 50 attributes previously confirmed important by Boruta, together with the 53 token variables (the same as the ones used in the first category).

Finally, the fifth category is made up of models which have available all the main and control attributes stated in Table 14 (i.e., Summary of Variables), except for `original_brand` and `original_handle`, to see whether the same or even better performance can be achieved if the brand is not so explicitly included in the inputted features.[113]

5. Results

5.1. Performance on Train and Validation Sets

Table 15 shows each of the selected performance metrics on both the train and the validation sets and displays the implemented models ranked according to their RMSPE on the validation set, in ascending order (since this metric represents an error and, thus, should be minimized).

[110]Applying Boruta to all the available attributes, including the 53 that resulted from the bag of words, and tuning both a random forest and a boosting model was tested, but the performance did not improve. It is also worth clarifying that applying Boruta to all the available attributes excluding those 53, but including all the tokens found with the bag of words used in the second category of models was not tested, because Boruta's algorithm has a time complexity of approximately $O(P * N)$, where P and N are, respectively, the number of attributes and observations (Kursa and Rudnicki, 2010). Thus, it would have been extremely time-consuming, given the large number of attributes and observations.

[111]Like with the previously implemented random forest models, the employed approach is not the OOB one, but instead the validation set one, to avoid the risk of making a prediction using trees that have seen original Tweets posted later in time, than the original Tweet to predict for. Additionally, the Boruta technique is implemented considering only the train set and dividing this set in the following way: approximately, the first 80% to train the model, while the final 20% to validate it.

[112]If the algorithm stops because the maximum number of indicated iterations has been reached, some attributes may be left without a decision and they are claimed tentative. One can tune different parameters to end up also getting a decision for these attributes (Kursa and Rudnicki, 2010). However, none of those was necessary in this thesis' case, since the algorithm stopped because only confirmed attributes were left.

[113]Applying Boruta to this group of attributes and, then, tuning both a random forest and a boosting model was also tested, but the performance did not improve.

Table 15: Performance Metrics on Train and Validation Sets and Ranking of Models

Rank	Model	RMSPE		RSE		Adjusted R^2	
		Train	Validation	Train	Validation	Train	Validation
1	Boruta & 53 tokens new tuned boosting	6.66	7.76	0.04	0.04	0.36	0.28
2	Tuned boosting	6.55	7.77	0.03	0.04	0.38	0.28
3	Boruta & 53 tokens new tuned random forest	6.39	7.77	0.03	0.04	0.42	0.26
4	Boruta & 53 tokens already tuned boosting	6.55	7.77	0.03	0.04	0.38	0.28
5	Tuned random forest	6.49	7.78	0.03	0.04	0.40	0.26
6	Tuned XGBoost	6.47	7.79	0.03	0.04	0.37	0.18
7	Boruta only new tuned boosting	6.68	7.79	0.04	0.04	0.36	0.29
8	No brand new tuned random forest	6.63	7.79	0.03	0.04	0.37	0.25
9	Boruta only already tuned boosting	6.56	7.80	0.03	0.04	0.38	0.29
10	Tuned bagging	6.64	7.81	0.03	0.04	0.37	0.25
11	Boruta & 53 tokens already tuned random forest	6.49	7.81	0.03	0.04	0.40	0.25
12	No brand already tuned random forest	6.49	7.81	0.03	0.04	0.40	0.25
13	Boruta only new tuned random forest	6.21	7.82	0.03	0.04	0.46	0.27
14	No brand new tuned boosting	6.57	7.82	0.03	0.04	0.37	0.26
15	Tuned decision tree	6.89	7.83	0.04	0.04	0.32	0.25
16	Boruta & 53 tokens default boosting	6.83	7.84	0.04	0.04	0.33	0.27
17	No brand already tuned boosting	6.63	7.84	0.04	0.04	0.36	0.26
18	Boruta only already tuned random forest	6.50	7.85	0.03	0.04	0.40	0.27
19	Splines	6.85	7.86	0.04	0.04	0.33	0.25
20	Lasso	6.85	7.87	0.04	0.04	0.33	0.29
21	Boruta only default boosting	6.87	7.87	0.04	0.04	0.33	0.28
22	Polynomial	6.85	7.88	0.04	0.04	0.33	0.26
23	Elastic	6.80	7.88	0.04	0.04	0.34	0.29
24	Default boosting	6.83	7.88	0.04	0.04	0.33	0.26
25	Main linear	6.87	7.89	0.04	0.04	0.32	0.26
26	Default decision tree	6.97	7.91	0.04	0.04	0.30	0.23
27	Tuned BART	6.66	7.91	0.04	0.04	0.36	0.28
28	Forward	7.00	7.92	0.04	0.04	0.30	0.26
29	No brand default boosting	6.90	7.92	0.04	0.04	0.31	0.24
30	Default BART	6.68	7.94	0.04	0.04	0.36	0.27
31	PCR	6.98	7.95	0.04	0.04	0.28	0.17
32	Default random forest	3.09	7.97	0.02	0.04	0.87	0.26
33	Boruta & 53 tokens default random forest	3.10	7.98	0.02	0.04	0.86	0.25
34	Default XGBoost	6.21	7.99	0.03	0.04	0.42	0.14
35	PLS	6.88	8.00	0.04	0.04	0.31	0.19
36	Tuned SVM	6.65	8.00	0.04	0.04	0.31	0.11
37	Boruta only default random forest	3.10	8.01	0.02	0.04	0.87	0.26
38	No brand default random forest	3.12	8.01	0.02	0.04	0.86	0.25
39	Backward	7.07	8.02	0.04	0.04	0.29	0.25
40	Ridge	6.86	8.03	0.04	0.04	0.30	0.18
41	Default bagging	2.98	8.10	0.02	0.04	0.88	0.24
42	Default SVM	6.34	8.17	0.03	0.04	0.37	0.07
43	Multi-layer NN	6.84	8.17	0.04	0.04	0.28	0.11
44	New tokens new tuned random forest	5.51	8.26	0.03	0.04	0.43	-1.48
45	New tokens already tuned random forest	6.27	8.31	0.03	0.04	0.26	-1.52
46	Single-layer NN	6.68	8.33	0.04	0.04	0.34	0.10
47	New tokens default random forest	3.79	8.34	0.02	0.05	0.73	-1.54
48	New tokens new tuned boosting	7.15	8.36	0.04	0.05	0.01	-1.59
49	New tokens already tuned boosting	7.30	8.42	0.04	0.05	-0.03	-1.64
50	Baseline	7.33	8.45	0.04	0.05	0.20	0.11
51	New tokens default boosting	7.61	8.59	0.04	0.05	-0.12	-1.75
52	53 tokens new tuned random forest	8.12	9.10	0.04	0.05	0.06	-0.02
53	Tuned KNN	8.10	9.11	0.04	0.05	0.06	0.04
54	53 tokens default random forest	8.11	9.11	0.04	0.05	0.06	-0.02
55	53 tokens already tuned random forest	8.11	9.11	0.04	0.05	0.06	-0.02
56	53 tokens default boosting	8.18	9.16	0.04	0.05	0.03	-0.03
57	53 tokens new tuned boosting	8.19	9.16	0.04	0.05	0.03	-0.03
58	53 tokens already tuned boosting	8.20	9.17	0.04	0.05	0.03	-0.03
59	Default KNN	7.12	9.80	0.04	0.05	0.30	-0.07
60	Significant linear	6.86	10.80	0.04	0.06	0.33	-0.44
61	All linear	6.85	11.63	0.04	0.06	0.33	-0.71

Note. The models are ranked according to their RMSPE on the validation set.

In general terms, in Table 15, it can be seen that the RSE is lower than the RMSPE. Additionally, it can be seen that, in some cases, especially in those at the bottom of the ranking, the adjusted R^2 is negative. It is worth clarifying that this happens when the R^2 is small, relative to the ratio of parameters to cases. In other words, the model is over-parameterized, and the results might be improved by increasing the sample size.

Meanwhile, in Table 15, it can also be seen that the winner is the boosting model which is tuned particularly for the attributes selected by Boruta plus the 53 token variables; in other words, this model is the one with the lower RMSPE on the validation set. As James et al. (2021) anticipated, this statistical learning approach, that learns slowly, ended up performing best. Concretely, this model's RMSPE on the validation set is 7.76: 0.69 percentage points less than the baseline's RMSPE on the validation set; in other words, a reduction of 8.17%. It is worth adding that 49 out of 60 (i.e., around 81.67%) of the implemented models ended up performing better than the baseline.[114]

Also, in Table 15, it can be seen that PLS performed no better than PCR. As James et al. (2021) suggest, although the supervised dimension reduction of PLS probably reduced the bias, it probably also increased the variance, making the overall benefit of PLS relative to PCR be a wash.

Furthermore, the top 14 models (i.e., around 23.33%) correspond to tuned ensemble methods based on decision trees, while classical approaches, such as linear regression, are lower in the ranking. Thus, following James et al. (2021), there is probably a highly non-linear and complex relationship between the features and the response. This was kind of anticipated during the exploratory data analysis, by the predominantly low r obtained between each continuous attribute and the TROS.

Finally, as suggested by James et al. (2021), by combining several trees into a single procedure, the tuned version of bagging achieved improving the performance of the tuned one of decision trees.

5.2. Winning Model's Feature Importance

On which attributes did the winning model focus on? A collection of ensemble trees is much more difficult to interpret than a single tree. Nevertheless, an overall summary of the importance of each predictor can be obtained, using the RSS (since the problem in question is of the regression type). Specifically, what can be recorded is the total amount that the RSS is decreased due to the splits over a given predictor, averaged over all B trees. A large value indicates an important predictor (James et al., 2021).

Figure 44 is a graphical representation of the importance of each variable relative to the others. It can be seen that, in the winning model, the attribute that generated the largest mean decrease in RSS was, by far, `original_brand`. Is there any intuition about why it is like this? Going back to the revised literature, there are several ideas that together can make up an intuition. Bazi et al. (2020) identify, as one of the reasons why users engage with luxury brands on social media, the relationship between a customer and a brand, represented through the love for the brand. Meanwhile, Ramadan et al. (2018) establish that one of the types of online luxury followers is

[114] If the baseline would have been the mean, instead of the median, historical TROS associated with the brand author of the Tweet, then almost half of the implemented models would have ended up performing better than this alternative baseline, since its RMSPE on the validation set is 7.93.

the image seeker, who loves luxury brands. Similarly, de los Santos (2009) considers that one of the types of luxury clients is constituted by what the author calls "people who know" and these people feel a genuine connection to brands. Therefore, a possible intuition is that some people engage with a luxury fashion brand's Tweet because of the love they feel for the brand itself (Bazi et al., 2020), beyond the characteristics of the Tweet per se, and in fact, these people are so many that they even constitute a type of follower (Ramadan et al., 2018) and customer (de los Santos, 2009). It is worth adding that this can be seen as in line with some of the discoveries made by Y. Choi et al. (2021), who study the Big 4 Fashion Weeks held in 2019. Concretely, with the findings that most of the keywords that appear at London Fashion Week are related to (British) fashion brands and designers and that the top keywords mentioned at Paris Fashion Week are mostly related to fashion brands themselves.

Figure 44: Feature Importance According to the Winning Model

Note. This figure considers the best model, according to the RMSPE on the validation set, fitted on the train set.

Moving on, in Figure 44, it can also be seen that many of the attributes that ended up being relevant for the model in question were a result of the feature engineering process; in the end, many of the created attributes were useful.

Furthermore, it is worth noting that, in contrast to Cuevas-Molano et al. (2021) who find a low influence of temporal factors (on engagement with Instagram posts of Spanish brands from different sectors), several attributes related to the original Tweet's time of creation ended up being important for predicting the TROS. Specifically, the attributes related to the original Tweet's time of creation that can be seen in Figure 44 are

- `original_created_at_hour`,

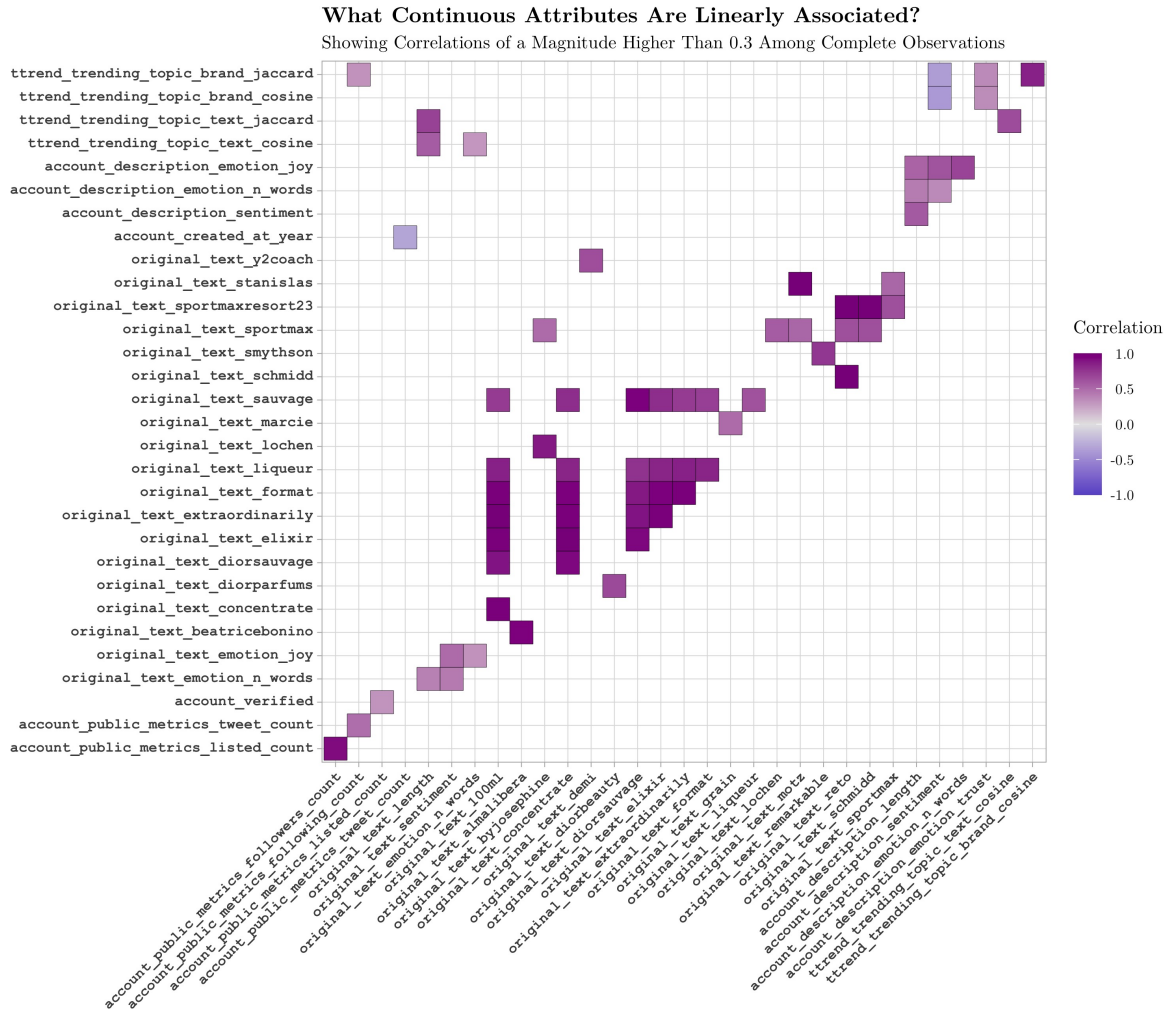
- `original_created_at_day_number`,
- `original_created_at_minute`,
- `original_created_at_day_week`,
- `original_created_at_month`,
- `original_created_at_second`,
- `original_created_at_year`, and
- `original_created_at_weekday`.

If the relative importance of these attributes is added, then the temporal factors as a whole have a relative importance of 3.025%, which makes them third in the ranking.

It is worth adding that, when looking at Figure 44, it should be borne in mind that if two variables are correlated, then one of them might not have been given importance (or not all of it), since (part of) its information was already being included through the importance given to the other one. For instance, regarding categorical variables, in Figure 44, it can be seen that `original_brand` appears, but not `original_handle`. Among these two attributes there is what in the relational model of databases is known as functional dependency: Applied to this case, if two observations have the same brand, then they also have the same handle. Furthermore, in this case, the functional dependency also applies the other way around: If two observations have the same handle, then they also have the same brand. Then, the model did alright in giving importance to just one of the two.

Meanwhile, as to continuous attributes, to see whether there is any attribute that does not appear in Figure 44 when it should due to the aforementioned problem, the correlation matrix is calculated for the continuous attributes. Figure 45 shows the results. In it, it can be seen that correlations of a magnitude greater than 0.3, considering complete observations, are mostly positive. Also, many of those correlations are of tokens that probably appear together in the original Tweet's text. For example, `sauvage` with `100ml` and `diorsauvage`, as well as `diorparfums` with `diorbeauty`. Besides, the tokens `diorsauvage` and `100ml` were given at least some importance by the winning model, but not `sauvage`, due to the aforementioned situation. Other instances in which the importance of an attribute is probably masked due to the presence of *collinearity* are the following: `ttrend_trending_topic_brand_jaccard` with `account_public_metrics_following_count`; `account_description_emotion_joy` with `account_description_emotion_n_words`; `account_created_at_year` with `account_public_metrics_tweet_count`; and `account_verified` with `account_public_metrics_listed_count`.

Figure 45: Correlation Matrix of Continuous Attributes



Note. This figure considers the data inputted to the best model to be first fitted on.

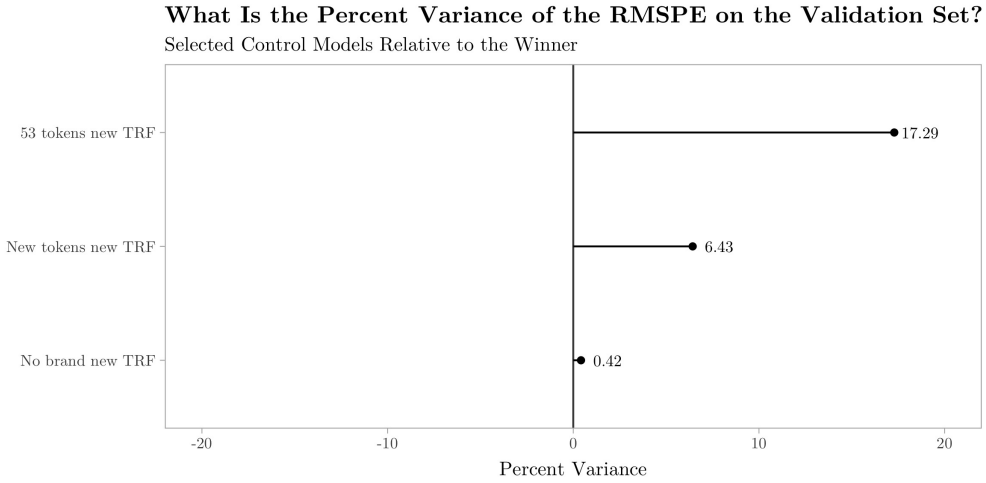
5.3. Comparison With Selected Control Models

Before moving on, let's go back to the fact that the winning model gave `original_brand`, by far, the greatest importance. The winner before implementing the five additional categories of models, which was the tuned boosting, also gave that attribute the highest relevance and, indeed, an extremely similar magnitude. In fact, it was due to this that those additional models were implemented, especially the ones that are inputted only token attributes, as well as the ones that are not inputted `original_brand` nor `original_handle`. So, advantage is taken of the fact that these special models have already been implemented to now compare their RMSPE and their feature importance against the ones of the winner, in order to know what the models focus on to predict the TROS when attributes explicitly related to the brand are not available. Specifically, to make this comparison clearer, the best model of each of these groups is selected: "53 tokens new TRF", "New tokens new TRF", and "No brand new TRF". Figures 46, 47, and 48 show the results.

In Figure 46, it can be seen that, as expected, the RMPSE of each of the selected control models is higher (thus, worse) than the winner's one. However, the percent variance corresponding to the model "No brand new TRF" is considerably small. This shows that the TROS can still be predicted very well without having available attributes explicitly related to the brand.

Nonetheless, it must be pointed out that the third facet in Figure 47 illustrates that the three most relevant attributes for this model are characteristics of the account author of the original Tweet, rather than characteristics of the original Tweet per se.

Figure 46: Comparison of RMSPE



So, what would a model focus on to predict the TROS if no attributes related to the account or any other entity other than the original Tweet were available? The first two facets in Figure 47 answer this question. It can be observed that most of the important features are of tokens present in the original Tweet's text that, in one way or another, indicate the brand. For example, the most important attribute for "53 tokens new TRF" was `original_text_pattinson` (recall what was pointed out in the exploratory data analysis: Robert Pattinson, in February 2023, starred in the campaign for the perfume Dior Homme Sport); while, the one for "New tokens new TRF" was directly `original_text_louisvuitton`.

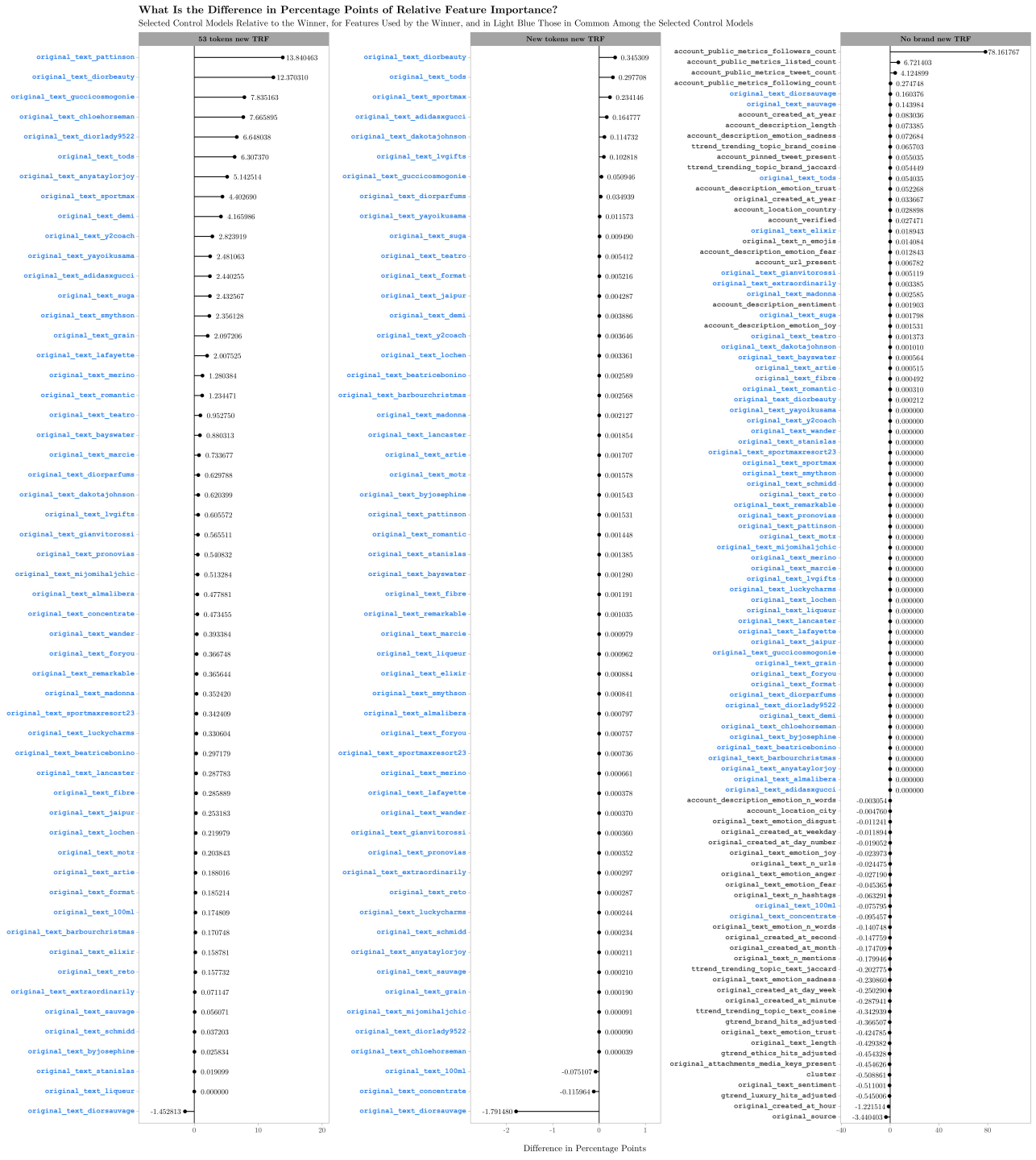
Nevertheless, some of the important features are of tokens more neutral to the brand, as well as not extremely usual in this domain[115], and so they can help make a difference. For instance, for model "53 tokens new TRF": `romantic`, `wander`, `foryou`, `remarkable`, `luckycharms`, and `fibre`. Whereas, for model "New tokens new TRF": `global`, `ambassador`, `actress`, `appearance`, `shoulder`, `award`, `custom`, `kiss`, `boutique`, `edition`, `black`, and `echo`. It is worth noting that the tokens `foryou` and `custom` relate to the importance of customization for the luxury audience (de los Santos, 2009). Meanwhile, the tokens `ambassador` and `actress` have to do with the attitudinal psychological process known as source attractiveness, which refers to the source's perceived social value. This quality can emanate from the person's physical appearance, personality, social status, or the person's similarity to the receivers. Additionally, these two tokens are associated with the fact that celebrities appeal to a common reference group (Vinerean and Opreana, 2019), and the fact that celebrity endorsement improves the perceived relevance of content, which is one of the motivations behind customers engaging with luxury brands on social media platforms (Bazi et al., 2020).

[115]Examples of extremely usual tokens in this domain are `wear`, `show`, `campaign`, `house`, `much`, `dress`, `look`, `collection`, `discover`, `fashion`, `collaboration`, `fall`, `print`, `design`, `paris`, and `live`.

Figure 47: Feature Importance According to the Selected Control Models

Finally, as to those attributes present in both the winner and any of the selected control models, Figure 48 shows how much more or less importance those control models gave them, relative to the winner.

Figure 48: Comparison of Feature Importance



It can be seen that the model "53 tokens new TRF" gave much more importance to the tokens `pattinson` and `diorbeauty`. In contrast, the model "New tokens new TRF" gave each of the attributes in common with the winner quite similar importance; the only token with a higher difference is `diorsauvage`: Concretely, this control model assigned it less importance than the winner. Lastly, the model "No brand new TRF" gave far more importance to the attribute

`account_public_metrics_followers_count`; indeed, like Figure 47 shows, this attribute ended up being this model's most important one.

5.4. Performance on Test Set

Many models have already been trained on the train set, it has been seen how they perform on the validation set, and the model with the best performance on the latter set has been identified, that being "Boruta & 53 tokens new tuned boosting". Thus, now, that model is trained again, but this time with both the train and the validation sets' data. Having this final model at hand, what now is seen is how this model predicts on the test set, as an estimation of how it would perform in production.

Table 16 shows the results. In it, it can be seen that the final model improved the three performance metrics in respect to the baseline: It lowered the RMSPE and the RSE, while it raised the adjusted R^2. Regarding particularly the RMSPE, the final model achieved a reduction of 0.79 percentage points; in other words, a decrease of 9.03%.

Table 16: Performance Metrics on Test Set

	RMSPE	RSE	Adjusted R^2
Baseline	8.76	0.05	0.03
Final model	7.97	0.04	0.22

It must be clarified that the information about the Google Trends is from 4 days prior, while the information about the account posting the original Tweet and the Twitter Trends is from the beginning of the day on which the original Tweet is posted. However, in production, it could be asked to predict the TROS for an original Tweet to be posted two weeks from now. In that case, the aforementioned information could not be from 4 days prior or at the beginning of the day on which the original Tweet is posted, because those days have still not happened and, so, that data still do not exist. At best, the Google Trends' information could be from 4 days prior and the account's and Twitter Trends' information could be from the beginning of the day, relative to the day in which the prediction request is made. In this way, the most recent available information would be used. Consequently, it must be noted that the higher the time distance between the moment in which the prediction request is made and the moment in which the original Tweet would be posted, the higher the potential vagueness of the information regarding Google Trends, the author account, and Twitter Trends and, thus, probably also of the resulting prediction. Nonetheless, it must also be pointed out that this probability is kind of mitigated by the fact that, as seen in the feature importance analysis, the three most important predictors for the winning model (i.e., `original_brand`, `original_source`, and posting time attributes as a whole) are characteristics of the Tweet itself, instead of being of Google Trends, the author account, or Twitter Trends.

5.5. Comparison Between Brands

Finally, the performance on the test set is analyzed by brand, to see whether there are any differences. First of all, in the test set, there are original Tweets from 79 of the 93 brands in the train or validation sets. Those 14 remaining brands are Balenciaga, Bally, Barbour, Blumarine, CHANEL, Goyard, HELMUT LANG, Maison Margiela, Marchesa, Monique Lhuillier, Moschino, PAIGE, Theory, and Yohji Yamamoto.

Then, as Table 17 shows, for more than half of the brands, the RMSPE of the final model for the particular brand is lower than the one in general (i.e., 7.97). Additionally, it is worth noting that a lower RMSPE for a brand is not necessarily associated with a higher number of original Tweets from this brand on the data with which the final model was trained (i.e., the train set plus the validation set).

Table 17: RMSPE on Test Set by Brand

Brand	n	n%	RMSPE
René Caovilla	59.00	0.65	1.49
Sportmax	70.00	0.78	1.55
Aquazzura	7.00	0.08	1.56
Lafayette148 NewYork	153.00	1.69	1.56
Margaret Howell	16.00	0.18	1.57
Ettinger	78.00	0.86	1.66
Temperley London	18.00	0.20	1.70
Mugler	13.00	0.14	1.79
Judith Leiber	53.00	0.59	2.05
Proenza Schouler	25.00	0.28	2.11
TUMI	22.00	0.24	2.51
Manolo Blahnik	38.00	0.42	2.52
Chloé	142.00	1.57	2.56
Herno	40.00	0.44	2.66
Paco Rabanne	28.00	0.31	2.77
Smythson	97.00	1.07	2.99
ERDEM	115.00	1.27	3.16
LANVIN	2.00	0.02	3.51
ZIMMERMANN	80.00	0.89	3.62
Johnstons of Elgin	81.00	0.90	3.74
Etro	261.00	2.89	3.95
CELINE	195.00	2.16	4.30
Acne Studios	150.00	1.66	4.34
MARNI	71.00	0.79	4.38
Gianvito Rossi	69.00	0.76	4.93
Globe-Trotter	31.00	0.34	5.03
KENZO	286.00	3.17	5.31
Axel Arigato	25.00	0.28	5.53
Roger Vivier	10.00	0.11	5.59
Mulberry	120.00	1.33	5.63
Canada Goose	79.00	0.88	5.96
Needle & Thread	2.00	0.02	6.02
Belstaff	24.00	0.27	6.14
RIMOWA	108.00	1.20	6.22
Max Mara	94.00	1.04	6.46
Gucci	366.00	4.05	6.55
Coach	199.00	2.20	7.06
MCM	5.00	0.06	7.12
Tory Burch	41.00	0.45	7.15
Victoria Beckham	93.00	1.03	7.22
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
LOEWE	194.00	2.15	7.24
Louis Vuitton	253.00	2.80	7.29
Hermès	28.00	0.31	7.33
Carolina Herrera	245.00	2.71	7.37
Vivienne Westwood	62.00	0.69	7.69
FERRAGAMO	17.00	0.19	7.91
Georges Hobeika	79.00	0.88	7.93
TOM FORD	37.00	0.41	7.99
Vera Wang	1.00	0.01	8.04
Miu Miu	205.00	2.27	8.12
Jean Paul Gaultier	107.00	1.19	8.14
Armani	97.00	1.07	8.18
Longchamp	46.00	0.51	8.42
Stella McCartney	152.00	1.68	8.43
Christian Louboutin	95.00	1.05	8.48
Michael Kors	109.00	1.21	8.58
Thom Browne	114.00	1.26	8.60
Paul Smith	99.00	1.10	8.70
DSQUARED2	53.00	0.59	8.71
Prada	326.00	3.61	9.50
Valentino	385.00	4.26	9.52
Balmain	46.00	0.51	9.59
Givenchy	32.00	0.35	9.61
Alexander McQueen	271.00	3.00	9.66
Oscar de la Renta	67.00	0.74	9.71
Dior	549.00	6.08	9.81
Saint Laurent	147.00	1.63	9.81
Jenny Packham	9.00	0.10	10.11
Missoni	101.00	1.12	10.19
Dolce & Gabbana	179.00	1.98	10.24
Fendi	103.00	1.14	10.24
Anya Hindmarch	38.00	0.42	10.28
Versace	28.00	0.31	10.42
Ralph Lauren	209.00	2.32	10.43
Jimmy Choo	128.00	1.42	10.62
Burberry	9.00	0.10	12.92
ELIE SAAB	95.00	1.05	13.18
Moncler	112.00	1.24	14.12
Tod's	41.00	0.45	15.06

Note. n and n% are calculated from the data used to train the final model, that being the train set plus the validation set.

Furthermore, compared to the performance of the baseline on the test set by brand, the final model achieved a median decrease in the RMSPE of 0.11 percentage points; in other words, a median reduction in the RMSPE equal to 2%.

6. Conclusions

6.1. Brief Work Summary

Given the importance of social media (Chen, 2021; Vinerean and Opreana, 2019) and luxury fashion (Bazi et al., 2020; Vinerean and Opreana, 2019), as well as of the need for these two to nurture each other (Bazi et al., 2020; Chen, 2021; Cuevas-Molano et al., 2021; Eastman et al., 2018; Tack et al., 2020; Vinerean and Opreana, 2019; Zohourian et al., 2018), this thesis' research question was whether it is possible to predict the reaction a post will generate in the audience of luxury fashion brands. More specifically, whether it is possible to predict the reaction that a post *on Twitter* will generate in the audience of luxury fashion brands *the day it is posted*.

To answer this question, first of all, an extensive literature review was done, regarding both domain and methodology; the unit of analysis was established; and the brands were selected. Then, the concept of reaction, which is the dependent variable, had to be defined.

To do so, a composite index was created and named TROS (i.e., Tweet reaction overall score), which is made up of 16 indicators related to likes, Retweets, and sentiment and emotion of replies and Quote Tweets. This definition is new and more comprehensive, relative to existing ones in both the academic and business fields; can be used to calculate other descriptive statistics, like a brand's average TROS in its Twitter profile, which could be a new KPI; can be molded to measuring a specific desired kind of reaction, by modifying the weights and, thus, prioritizing a subset of indicators over the rest; and can be adapted to other types of social media posts, by slightly adjusting only a few of its indicators. Indeed, the TROS, created in this thesis, represents a solid and relevant contribution for both academia and business.

Having defined the dependent variable, it was followed by collecting the data, exploring them, and engineering features. Finally, the chosen methodology was implemented and the results were fully analyzed.

6.2. Lessons

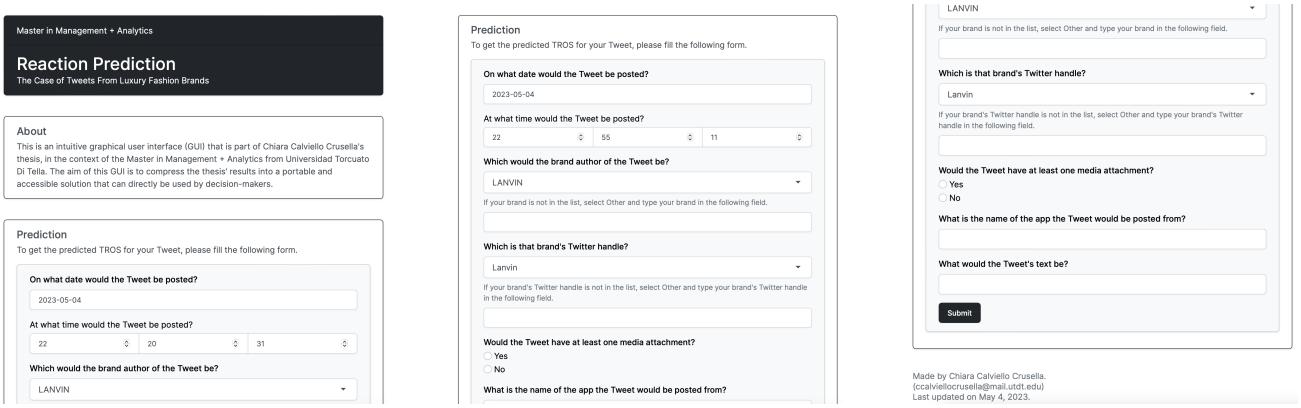
Mainly, it was discovered that, in line with the hypothesis, it is indeed possible to predict the reaction that a post on Twitter will generate in the audience of luxury fashion brands the day it is posted. Several models were implemented, based on distinct learning methods, and they presented different results. Around 81.67% of these models performed better than the baseline and the top 23.33% corresponded to tuned ensemble methods based on decision trees. Concretely, the winning model was a tuned boosting with Boruta previously applied to a group of its inputted attributes. This model, compared to the baseline, achieved a reduction of the RMSPE on the validation set of 8.17%, while on the test set of 9.03%. Therefore, it was learned not only that the TROS can indeed be predicted, but also that several different learning methods can improve the performance of a baseline, being ensemble decision trees at the top of the list.

Furthermore, it was found that attributes related to the brands or posting time are especially important. It was also found that attributes of tokens like `foryou`, `custom`, `ambassador`, and `actress` are important as well. This is in line with the reviewed literature: Customization is of the uttermost importance for the luxury audience (de los Santos, 2009); and celebrity endorsement improves the perceived relevance of content, which is one of the motivations behind customers engaging with luxury brands on social media platforms (Bazi et al., 2020).

Consequently, when designing a post, a luxury fashion brand should pay special attention to

aspects like the brands mentioned in the Tweet's text; as well as the exact time of publication; signs of customization; and the references to actresses, actors, celebrities, or brand's ambassadors. Different Tweet options can be designed (modifying especially those aspects), the TROS can be predicted for each of them, and they can be ranked according to the obtained results, thus assisting decide which of them to post. This process can help decision-makers validate posts before publishing them, and therefore adapt and improve posts, as well as the marketing strategy in general. Additionally, this process could be made easier through an intuitive graphical user interface (GUI) in which the user enters a Tweet's text, together with some other few arguments, and the corresponding predicted TROS is returned, maybe even allowing the user to add this scenario to a scenarios' database to make comparisons and download for deeper analysis. Although the final model was inputted 104 attributes, the user would have to complete only 7 fields, since the rest can be automatically calculated from the entered text, handle, or date in the back end. To better illustrate this idea, Figure 49 shows an example of how the input part of the GUI could look like. This GUI is built using the Shiny R package[116]. In brief, in this way, the results can be compressed into a portable and accessible solution that can directly be used by decision-makers.

Figure 49: Example of GUI



6.3. Limitations and Future Research

Despite its relevant contributions, this thesis is not free from some limitations, which represent threats to validity. The first one corresponds to the ability to generalize beyond the fashion luxury sector. Although the results are useful, they cannot be directly extrapolated to other sectors within the fashion industry or (even less) to different industries. For the results of this thesis to be of a general nature, they first need to be confirmed in other contexts.

The next two limitations (the second and the third ones in the limitations' enumeration) are about the TROS and show that this new composite index could be improved. Firstly, regarding the indicators $\frac{L_i}{F_i}$ and $\frac{RT_i}{F_i}$, the division by the number of followers is done to control for the community size of each brand since, as previously mentioned, it is what is done in the professional practice (Cuevas-Molano et al., 2021) and, on average, there is a linear dependency between the total number of likes received by a post and the author's current number of followers (Vassio et al., 2022). However, in the end, the number of followers is a proxy for the number of times the post has actually been seen. This last piece of information was available for this thesis only on Twitter, but not through its API, given the author's access permits, since it is a non-public

[116] https://shiny.rstudio.com.

metric[117]. Thus, if it were to be collected, it would have had to be done manually by the author on its own for more than eleven thousand Tweets. Consequently, it remains to be done by future research the replacement of number of followers by number of views, to better capture how many times a post has actually been seen.

Secondly, regarding the replies to and the Quote Tweets of the original Tweets, their corresponding indicators included in the TROS capture their sentiment and emotion through their text, but they do not consider their own number of likes or Retweets. All replies and Quote Tweets are given the same importance when calculating their general sentiment and emotions, but one that has a higher number of likes or Quote Tweets probably expresses an opinion shared by more users and, so, should probably be given more weight[118]. Therefore, future works could try implementing this change; for instance, by calculating, instead of the median, an average weighted by the result of dividing the sum of likes and Retweets of that reply or Quote Tweet by the total number of likes and Retweets received by all that original Tweet's replies or Quote Tweets.

Then, a fourth limitation concerns the original Tweets collected. As mentioned in a footnote placed in the data collection's section, some of the downloaded original Tweets were actually destined to particular users or with the exact same text to others posted by the same account at the exact same time. Thus, before computing any calculus, these original Tweets were discarded, together with their corresponding replies and Quote Tweets, if they had any. To automatically filter this kind of original Tweets out, all the patterns or rules found on the go were used, like the text containing the phrase "Reply #stop to unsubscribe" or "Opt out by replying" or two or more Tweets from the same brand posted at the exact same time with the exact same text. By exploring the train set, it can be said that these filtering criteria achieved removing many (or even most) of them. However, some few were left, like those in which there is a tiny but irregular posting time difference, since that irregularity impedes creating a rule and, thus, obliges to identify them manually. Consequently, future research could try implementing a more robust (but probably still imperfect) automatic filtering criteria or, if the resources are available, a manual check.

A fifth limitation refers to the fact that the volume of collected data can be considered rather small for applying some more "complex" learning methods, like NNs. Therefore, an even bigger data set could be collected and it could be analyzed whether NNs or other methods' performance for this thesis' problem improves. Additionally, having this bigger data set would enable as well to deal with a different problem: Given a desired TROS (or tuned version of the TROS), which should precisely be the original Tweet's text to achieve it? This would be a complex NLP problem, requiring a generative model, which tends to demand more observations. Future research could try to solve it.

Coming back to the problem dealt with in this thesis and moving on to the sixth limitation, during the exploratory data analysis, it was observed that original Tweets with higher TROS tended to mention a famous Asian person. None of the actual features is able to capture this aspect; not even the ones created as a result of the bag of words from the original Tweet's text, since the minimum frequency for a token to be considered is, in this case, 150 because if it was

[117] https://developer.twitter.com/en/docs/twitter-api/metrics.

[118] The number of replies or Quote Tweets is not mentioned, because of the problems they imply, which were described in the dependent variable's section; nor their sentiment or emotion are mentioned because if not, by following this same logic, when calculating them, their own number of likes and Retweets would also have to be considered, thus turning into an infinite loop.

any lower, then extremely more and irrelevant tokens would be included. Consequently, future research could try capturing the aspect in question so that models can have it as an available input.

Then, extracting the sentiment or the emotion from a piece of text is a sub problem inside this thesis' main problem, which is dealt with enough to be able to construct some of the TROS' indicators, as well as some of the features from the given attributes `original_text` and `account_description`. Thus, future research could address this further, to see whether a new different technique can improve the results of the extraction.

Finally, features related to images, videos, or files in graphics interchange format (GIF) attached to the original Tweets are not included in the analysis. This introduces a bias, since the textual content is systematically being favored over the aforementioned (audio)visual contents. Future research could try to solve the problem of extracting attributes from these (audio)visual contents, to input them to a model and see whether this helps improve the performance of predicting the TROS. Furthermore, analyzing these (audio)visual contents would be especially important if the future research were to be conducted about a social media platform like Instagram. If that were to be the case, then the opportunity could be seized to adapt the TROS to posts on Instagram.

References

Al Rawashdeh, H. M. (2017). *Use data mining techniques to predict users' engagement on the social network posts in the period before, during and after Ramadan* (Master's thesis). The British University in Dubai. https://bspace.buid.ac.ae/bitstream/handle/1234/1246/2015128050.pdf?sequence=3&isAllowed=y

Athwal, N., Istanbulluoglu, D., & McCormack, S. E. (2018). The allure of luxury brands' social media activities: A uses and gratifications perspective. https://core.ac.uk/download/pdf/199219076.pdf

Bazi, S., Filieri, R., & Gorton, M. (2020). Customers' motivation to engage with luxury brands on social media. *Journal of Business Research*. https://eprints.ncl.ac.uk/file_store/production/263725/F158539B-0793-4829-B675-706EBEB6DA3D.pdf

Becker, K., Lee, J. W., & Nobre, H. M. (2018). The concept of luxury brands and the relationship between consumer and luxury brands. *Journal of Asian Finance, Economics and Business*, *5*(3), 51–63. https://doi.org/10.13106/jafeb.2018.vol5.no3.51

Caïs, C. (2021). Is sustainability the next frontier for luxury brands? [Online; accessed January 12, 2023]. https://www.forbes.com/sites/forbesagencycouncil/2021/11/24/is-sustainability-the-next-frontier-for-luxury-brands/?sh=384d2c1b96b5

Chen, Y. (2021). A social media mining and ensemble learning model: Application to luxury and fast fashion brands. *Information*, *12*(4), 149. https://doi.org/10.3390/info12040149

Choi, Y., Yoon, S., Xuan, B., Lee, S. T., & Lee, K. (2021). Fashion informatics of the Big 4 Fashion Weeks using topic modeling and sentiment analysis. *Fashion and Textiles*, *8*(1), 1–27. https://doi.org/10.1186/s40691-021-00265-6

Choi, Y. K., Seo, Y., Wagner, U., & Yoon, S. (2020). Matching luxury brand appeals with attitude functions on social media across cultures. *Journal of Business Research*, *117*, 520–528. https://doi.org/10.1016/j.jbusres.2018.10.003

Cuevas-Molano, E., Matosas-López, L., & Bernal-Bravo, C. (2021). Factors increasing consumer engagement of branded content in Instagram. *IEEE Access*, *9*, 143531–143548. https://doi.org/10.1109/ACCESS.2021.3121186

de los Santos, J. M. (2009). *Capturing and maintaining the essence of luxury in the dynamic global marketplace* (Master's thesis). University of Southern California. https://digitallibrary.usc.edu/asset-management/2A3BF1SN4HU3?FR_=1&W=738&H=834

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing*, *34*(6), 480–488. https://qa.reach-latam.com/inbound/fb/sentimentr/inst/articles/Dhaoui2017.pdf

Dixon, S. (2022). Twitter: Number of monetizable daily active users worldwide 2017-2022 [Online; accessed December 5, 2022]. https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/

Dubois, D., Jung, S., & Ordabayeva, N. (2021). The psychology of luxury consumption. *Current Opinion in Psychology*, *39*, 82–87. https://doi.org/10.1016/j.copsyc.2020.07.011

Eastman, J. K., Iyer, R., Shepherd, C. D., Heugel, A., & Faulk, D. (2018). Do they shop to stand out or fit in? The luxury fashion purchase intentions of young adults. https://www.researchgate.net/profile/C-Shepherd/publication/323073217_Do_they_shop_to_stand_out_or_fit_in_The_luxury_fashion_purchase_intentions_of_young_adults/links/5bb76a5492851c7fde2f0c80/Do-they-shop-to-stand-out-or-fit-in-The-luxury-fashion-purchase-intentions-of-young-adults.pdf

Fumagalli, E. (2021). Ethical consumerism and glass box branding: When companies' actions speak louder than words. In *Sage business cases*. SAGE Publications: SAGE Business Cases Originals. https://doi.org/10.4135/9781529753127

Godey, B., Manthiou, A., Pederzoli, D., Rokka, J., Aiello, G., Donvito, R., & Singh, R. (2016). Social media marketing efforts of luxury brands: Influence on brand equity and consumer behavior. *Journal of Business Research*, *69*(12), 5833–5841. https://doi.org/10.1016/j.jbusres.2016.04.181

Graziani, L., Melacci, S., & Gori, M. (2019). Jointly learning to detect emotions and predict Facebook reactions. *arXiv preprint arXiv:1909.10779*. https://arxiv.org/pdf/1909.10779.pdf

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for Twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11. https://doi.org/10.3390/mca23010011

Hemantha, Y. (2020). Retaining the cachet of luxury fashion brands on social media through storytelling and narratives. *IUP Journal of Brand Management*, *17*(3), 23–37. https://www.proquest.com/docview/2464177780/fulltextPDF/B2CFC667FFB94E56PQ/1?accountid=133671

Hogg, T., Lerman, K., & Smith, L. M. (2013). Using stochastic models to predict user response in social media. *2013 International Conference on Social Computing*, 63–68. https://arxiv.org/pdf/1308.2705.pdf

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1). https://doi.org/10.1609/icwsm.v8i1.14550

Jahn, B., Kunz, W., & Meyer, A. (2013). The role of social media for luxury brands — Motives for consumer engagement and opportunities for business. https://www.researchgate.net/profile/Werner-Kunz-3/publication/272236813_The_Role_of_Social_Media_for_Luxury-Brands_-_Motives_for_Consumer_Engagement_and_Opportunities_for_Businesses/links/59db93df0f7e9b1460fbcf43/The-Role-of-Social-Media-for-Luxury-Brands-Motives-for-Consumer-Engagement-and-Opportunities-for-Businesses.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* [Second edition on website at https://hastie.su.domains/ISLR2/ISLRv2_website.pdf]. Springer.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* [Third edition draft on website at https://web.stanford.edu/~jurafsky/slp3/].

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of statistical software*, *36*(11). https://doi.org/10.18637/jss.v036.i11

León-Sandoval, E., Zareei, M., Barbosa-Santillán, L. I., & Falcón Morales, L. E. (2022). Measuring the impact of language models in sentiment analysis for Mexico's COVID-19 pandemic. *Electronics*, *11*(16). https://doi.org/10.3390/electronics11162483

Liu, L., Dzyabura, D., & Mizik, N. (2018). Visual listening in: Extracting brand image portrayed on social media. *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/WS/AAAIW18/paper/viewFile/17094/15551

Liu, X., Shin, H., & Burns, A. C. (2021). Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing. *Journal of Business Research*, *125*, 815–826. https://doi.org/10.1016/j.jbusres.2019.04.042

Mohamaddoust, R., Mohammadzadeh, J., Khalilian, M., & Nikravanshalmani, A. (2021). Measuring and analyzing charisma on Twitter using combination weighting and TOPSIS

method. *International Journal of Nonlinear Analysis and Applications*, *12*(Special Issue), 1143–1158. https://ijnaa.semnan.ac.ir/article_5594_f75a26c37f5b2e84f5da0ce7576d2667.pdf

Muñoz, M. M., Rojas-de-Gracia, M., & Navas-Sarasola, C. (2022). Measuring engagement on Twitter using a composite index: An application to social media influencers. *Journal of Informetrics*, *16*(4). https://doi.org/10.1016/j.joi.2022.101323

Oliveira, M., & Fernandes, T. (2022). Luxury brands and social media: Drivers and outcomes of consumer engagement on Instagram. *Journal of Strategic Marketing*, *30*(4), 389–407. https://doi.org/10.1080/0965254X.2020.1777459

Pantano, E., Giglio, S., & Dennis, C. (2019). Making sense of consumers' tweets: Sentiment outcomes for fast fashion retailers through Big Data analytics. *International Journal of Retail & Distribution Management*, *47*(9), 915–927. https://eprints.mdx.ac.uk/25365/1/Sentiment%5C%20Analysis_fast-fashion-twitter%5C%28rev1%5C%29.pdf

Park, J., Song, H., & Ko, E. (2011). The effect of the lifestyles of social networking service users on luxury brand loyalty. *Journal of Global Scholars of Marketing Science*, *21*(4), 182–192. https://doi.org/10.1080/21639159.2011.9726521

Park, M., Im, H., & Kim, H. (2020). "You are too friendly!" The negative effects of social media marketing on value perceptions of luxury fashion brands. *Journal of Business Research*, *117*, 529–542. https://doi.org/10.1016/j.jbusres.2018.07.026

Ramadan, Z., Farah, M. F., & Dukenjian, A. (2018). Typology of social media followers: The case of luxury brands. https://laur.lau.edu.lb:8443/xmlui/bitstream/handle/10725/9828/Post-print%5C%20%5C%28Typology%5C%29.pdf?sequence=2%5C&isAllowed=y

Ratnakumar, D. (2021). *Analysing the brand popularity of Sri Lankan banks on social media engagement predict the most effective post time* (Master's thesis). Informatics Institute of Technology. http://dlib.iit.ac.lk/xmlui/handle/123456789/1072

Romão, M. T., Moro, S., Rita, P., & Ramos, P. (2019). Leveraging a luxury fashion brand through social media. *European Research on Management and Business Economics*, *25*(1), 15–22. https://doi.org/10.1016/j.iedeen.2018.10.002

Salamander, G. (2018). Your Tweet got ratioed, what next? [Online; accessed December 5, 2022]. https://eclincher.com/your-tweet-got-ratioed-what-next/

Tack, I., De Veirman, M., & Hudders, L. (2020). Building a luxury brand on Instagram: The case of Delvaux. *Marche et organisations*, *37*(1), 55–71. https://www.cairn.info/revue-marche-et-organisations-2020-1-page-55.htm?ref=doi

Vassio, L., Garetto, M., Leonardi, E., & Chiasserini, C. F. (2022). Mining and modelling temporal dynamics of followers' engagement on online social networks. *Social Network Analysis and Mining*, *12*(1), 1–17. https://doi.org/10.1007/s13278-022-00928-2

Vinerean, S., & Opreana, A. (2019). Social media marketing efforts of luxury brands on Instagram. https://web.archive.org/web/20200925140158id_/http://marketing.expertjournals.com/ark:/16759/EJM_714vinerean144-152.pdf

Wang, Y. (2022). A conceptual framework of contemporary luxury consumption. *International Journal of Research in Marketing*, *39*(3), 788–803. https://doi.org/10.1016/j.ijresmar.2021.10.010

Zohourian, A., Sajedi, H., & Yavary, A. (2018). Popularity prediction of images and videos on Instagram. https://www.researchgate.net/profile/Alireza-Zohourian/publication/325833366_Popularity_prediction_of_images_and_videos_on_Instagram/links/6193a5913068c54fa5efa037/Popularity-prediction-of-images-and-videos-on-Instagram.pdf