ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

# Spatio-Temporal Distribution Analysis of Brand Interest in Social Networks

**Diana Von-Haff Lopes Teixeira**

A dissertation submitted in partial fulfillment of the requirements for the degree of **Master in Information Systems Management**

**Supervisor:**
Fernando Manuel Marques Batista, Assistant Professor, PhD
ISCTE-IUL – Instituto Universitário de Lisboa

**Co-Supervisor:**
Ricardo Daniel Santos Faro Marques Ribeiro, Assistant Professor, PhD
ISCTE-IUL – Instituto Universitário de Lisboa

October, 2018

# Resumo

Actualmente, plataformas como Twitter e Facebook fazem parte do dia-a-dia de muitas pessoas e são usadas por milhões de utilizadores. Nestas plataformas, denominadas Redes Sociais, os utilizadores partilham informações incluindo opiniões, sentimentos, experiências e pensamentos. A plataforma Twitter, em particular, é usada para partilhar diversos tópicos, que podem incluir discussões sobre marcas, seus produtos e/ou serviços. O presente estudo analisa como o interesse numa marca é reflectido na Rede Social Twitter e apresenta uma metodologia que permite utilizar o Twitter como fonte de informação para monitorizar o que os utilizadores dizem acerca de determinadas marcas. O interesse numa marca pode ser definido como o nível de interesse que um indivíduo tem por uma marca, e o nível de curiosidade que um indivíduo tem e que o leva a aprender mais acerca dessa marca. Neste estudo, o número de tweets publicados é usado para medir o interesse nas marcas escolhidas. A metodologia seguida baseia-se na data em que o tweet foi publicado, localização, e número de publicações, para efectuar uma análise espaço-temporal.

Adicionalmente, apresenta-se uma *framework* que possibilita a exploração de um vasto conjunto de dados, com o objectivo de revelar padrões latentes, bem como analisar o interesse nas marcas seleccionadas, usando o Twitter como fonte dados. Para o efeito, aplicou-se *Topic Modelling*, uma técnica de *Text Mining* bastante utilizada para descobrir tópicos em texto não estruturado. Algoritmos de *Topic Modelling* têm sido amplamente utilizados para monitorizar eventos e tendências e descobrir tópicos em áreas como educação, marketing, saúde, entre outras. A *framework* consiste em treinar o modelo de tópicos LDA (Latent Dirichlet Allocation) usando tweets agrupados (considerando determinado critério) e posteriormente aplicar o modelo treinado noutro conjunto de tweets agrupados (considerando outro critério). Descreve-se um conjunto de tarefas de pré-processamento dos dados que ajudaram a melhorar o desempenho dos modelos, a obter melhor resultados e, consequentemente, a efectuar uma melhor análise. As experiências revelam que através de *Topic Modelling* é possível rastrear discussões de utilizadores de Redes Sociais durante um longo período de tempo, e capturar alterações relacionadas com acontecimentos reais.

# Abstract

Social Networks applications such as Facebook and Twitter became part of many people's lives and are used daily by millions of users. In such platforms, users share their emotions, opinions, experiences, and thoughts. Twitter, in particular, is used to discuss diverse topics, including brands, their products and services. In this thesis, we analyse how brand interest is reflected on Twitter and how this platform can be used to monitor what people say about specific brands, as an indicator of brand interest. Brand interest can be defined as the level of interest one has in a brand, and the level of curiosity one has to learn more about a brand. For this work, the volume of tweets is used as a measure of brand interest. Our methodology is based on time, location, and the number of brand-related tweets to perform a spatio-temporal analysis.

Additionally, we propose a framework for discovering latent patterns (topics) from a large dataset of grouped short messages to analyse brand interest, using Twitter as a data source. We applied a well-known Text Mining technique called Topic Modelling, which is an unsupervised learning technique used when dealing with text data, useful to uncover topics in a collection of documents. This technique provides a convenient way to retrieve information from unstructured text. Topic Modelling tasks have been applied to track events/trends and uncover topics in domains such as academic, public health, marketing, and so forth. The framework consists of training LDA (Latent Dirichlet Allocation) topic models on aggregated tweets, and then applying the model on different documents, also composed by grouped Twitter posts. Furthermore, we describe a set of pre-processing tasks that helped to improve the performance of topic models, enabling us to obtain a better output, thus performing a better analysis of it. The experiments demonstrated that Topic Modelling can successfully track people's discussions on Social Networks even in massive datasets such as the one used in the current work, and capture those topics spiked by real-life events.

# Palavras Chave
# Keywords

## Palavras chave

Modelagem de Tópicos

Evolução de Tópicos

Interesse na Marca

Análise Espaço-Temporal

Análise de Tópicos

Redes Sociais

## Keywords

Topic Modelling

Topics Evolution

Brand Interest

Spatio-Temporal Analysis

Topic Analysis

Social Networks

# Agradecimentos
# Acknowledgements

A realização desta dissertação de mestrado contou com importantes apoios e incentivos sem os quais a sua conclusão não se teria certamente realizado, e aos quais serei grata para todo o sempre.

Ao Professor Doutor Fernando Batista, pela sua orientação, compreensão, amabilidade, apoio, e disponibilidade, por todo o conhecimento que transmitiu, pelas sugestões, opiniões e críticas construtivas, total prontidão e colaboração a solucionar problemas e dúvidas que surgiram ao longo da realização do presente trabalho, e acima de tudo por todo o incentivo e bom humor.

Ao Professor Doutor Ricardo Ribeiro, pela sua co-orientação que em nada deixou a desejar, compreensão, amabilidade, apoio, e disponibilidade, por todo o saber que transmitiu, pelas sugestões, opiniões e críticas construtivas, total prontidão e colaboração a solucionar problemas e dúvidas que surgiram ao longo da realização do presente trabalho, e acima de tudo por todo o incentivo e bom humor.

À minha irmã Sónia de Carvalho que apostou em mim sempre, que me ajudou e incentivou a concluir este desafio. À minha irmã Bernadia Lopes Teixeira por todo apoio, incentivo, e suporte emocional.

Ao meu companheiro Miguel Prata, por todo incentivo, força, compreensão, e suporte emocional.

Ao tio João Firmino Almeida Henriques, que esteve sempre atento e disponível, dando sempre um incentivo para que não deixasse passar esta oportunidade.

Às minhas tias Marília Lopes Teixeira e Domingas Lopes Teixeira, por todo carinho e apoio, sempre.

Ao meu amigo Luati Roque Fontes, por todo incentivo e encorajamento.

Ao Miguel Freire por todo força, incentivo, e compreensão.

À minha amiga Sandra Caeiro, pelo incentivo e interesse que sempre demonstrou quando o tema fosse "conclusão da dissertação".

Ao Francisco Baptista e Inês Lameiras, pelo apoio e disponibilidade.

Por último, mas não menos importante, dirijo ao meu pai Augusto Lopes Teixeira e à minha mãe Marcília Von Haff, a quem devo a vida, por serem os principais responsáveis pela pessoa em que me tornei. A eles dedico este trabalho.

Lisboa, 26 de Outubro
Diana Von Haff Lopes Teixeira

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| ATAM | Ailment Topic Aspect Model |
| CTM | Correlated Topic Model |
| HDP | Hierarchical Dirichlet Processes |
| IE | Information Extraction |
| IR | Information Retrieval |
| LDA | Latent Dirichlet Allocation |
| LSA | Latent Semantic Analysis |
| NLP | Natural Language Processing |
| PLSA | Probabilistic Latent Semantic Analysis |

# 1 Introductionf

In the recent years, the rapid growth of the Internet has led to the growth of social media websites like Twitter, a micro-blog platform launched in 2006. In social media websites people share diverse aspects of their lives, and communicate about current events that they are aware of. Due to the fact that these platforms became so popular, a great number of companies also utilize them to promote their products and services. These websites produce tremendous amounts of data that can be used in various ways. For instance, it can be used to track emerging events, to discover trending topics, to evaluate consumers' satisfaction toward a product in the market, and so on.

Several researchers have examined Twitter from different perspectives. Predicting box-office revenues (Asur and Huberman, 2010), attempting to predict elections outcomes (Tumasjan et al., 2010), measuring the effects of Word-of-Mouth on movie sales (Rui et al., 2013), tracking real-world events (Lau et al., 2012), detecting illegal online sales (Mackey et al., 2017), identifying public health-related issues (Kaveri and Maheswari, 2017; Paul and Dredze, 2014), are some examples of potential exploitation of data available on Twitter. Topic Modelling is among the Text Mining techniques applied to exploit Twitter data. However, performing Topic Modelling tasks in short text, such as those available on Twitter, differs from performing them in longer documents, such as academic abstracts or newspapers. This is mainly because in such posts, underlying topics are usually conveyed using short messages. Aggregating similar Twitter posts enrich the content of a single document from which the LDA can learn a better topic model. In this matter, several studies have recently applied Topic Modelling on Twitter Data using pooling techniques to overcome the length disadvantage (Alvarez-Melis and Saveski, 2016; Hong and Davison, 2010; Mehrotra et al., 2013; Quan et al., 2015). Nonetheless, few have focused on pooling Twitter posts based on a combination of Geo-location (country), brand, within a time span, to create longer documents in order to perform a Topic Modelling tasks for Brand Interest tracking purposes.

Machleit et al. (1990) define "brand interest" as the level of interest a consumer has in the brand and the level of curiosity he/she has to learn more about the brand. The authors also stated that a high levels of brand interest may lead a consumer to try the brand or to search for

more information regarding the brand. Even though brand interest effect on purchase intentions is not significant (Machleit et al., 1993), keeping brand interest is relevant because:

i) high levels of brand interest may lead a consumer to try the brand again and reassess his/her beliefs, which may reduce switching behaviors, as consumers tend to switch among familiar brands in an effort to relieve boredom (Machleit et al., 1990);

ii) consumers like familiarity but are interested in novelty (Sung et al., 2016).

## 1.1 Motivation and Goals

Our goal is to create a system capable of depicting what people discuss about on Social Networks, and how it changes over the time, using Portuguese written posts as data. In other words, we aim to derive insights from large volumes of unstructured text, through unsupervised machine learning. Nowadays, many companies value the information like the one our system is able to provide, for purposes of Marketing or simply to track what catches people attention and what does not, regarding their products and services. It is, for instance, a great alternative to questionnaires, as people usually share their opinions and thoughts in a more spontaneous and genuine way when using Social Media platforms such as Twitter and Facebook. In this manner, one is able to discover and track what people think about any given subject, brand, event, and so on, by following the proposed framework, using publicly available data.

Our motivation emerged from the fact that, to the best of our knowledge, there are only few studies analysing Portuguese written Social Network data for the purpose of tracking brand interest for such a wide time span. Moreover it is an interesting area with a great potential, and the benefits companies and individuals could obtain from it are promising. Some possible ways of benefiting from this system are:

- It enables companies to discover, track, and analyse purchasing behaviours, in order to find out which products are sold the most and adjust their business according to this trend.
- Track and monitor online reputation is important, as we live in a Social Media era, meaning that anybody can share anything, and one single person has the power of influencing many other people. Thus, monitoring what people say about a company might be crucial. In this manner, companies who want to monitor and improve the online presence of their business might also benefit from this framework.
- Individuals can also benefit from this system. Following this framework, one can discover, for instance, why the shares of a given company has risen or fallen, or why they would possibly rise or fall, which could help individual investors.

## 1.2 Research Questions

Findings of this study are expected to help brands to understand how consumers' interest in any given brand vary over the time and from place to place, and adjust their strategies accordingly. More specifically, this research attempts to answer the following research questions:

**RQ1:** How can text mining techniques be applied to detect hidden patterns in user's interest in a specific brand?

**RQ2:** How to unveil brand interest changes over a period of time using text mining?

**RQ3:** How can text mining techniques be applied to discover seasonal brands?

**RQ4:** Which text mining techniques can detect changes in brand interest from one place to another?

## 1.3 Contribution

Taking into consideration the fact that most of the existing works were elaborated for English and other languages, and that for Portuguese there is much less studies, we decided to create a system that could analyse brand interest using Twitter public available information. Our contribution in the area of research is as follows:

- A consistent framework that performs Topic Modelling to uncover latent patterns on unstructured and short Portuguese written messages, on massive datasets.

- A set of Portuguese lexicon that comprises verbs, adverbs, conjunctions, and common social media slang and abbreviations was created, and will be made available in order to enrich the existing Portuguese stop words list, which we found out very incomplete for the present study.

- This framework, which a front-end development is in plan for future work, simplifies the process of extracting insights from people's conversation available on Social Networks platform, using massive datasets. Moreover, it can be adjusted to track discussions and topics on datasets of any size, by simply adjusting the time span.

## 1.4 Methodology

This thesis presents a Text Mining analysis of Twitter Posts, regarding 10 brands with a Twitter account, providing that each brand has 1 million followers, and that at least 5 thousand of tweets mentioning this brand are available.

The data collected refers to a time span of four years (2014 to 2017). The tweets are grouped in longer documents. Each document is a result of the aggregation of tweets posted during a week or month, in any given country, towards a specific brand. The LDA algorithm is applied. The final number of topics was defined after several iterations, as for this algorithm the number of topics had to be known in advance.

## 1.5 Structure of the Document

This study is organized as follows. Chapter 2 provides a brief literature review on Topic Modelling as a Text Mining technique, mentioning some of the most common algorithms applied. Moreover, we enumerate some studies that exploited Twitter as a source of data, for Topic Modelling tasks. Chapter 3 describes the composition of the dataset, and the steps and tasks conducted to prepare its usage. In Chapter 4 the system architecture is presented, along with the spatio-temporal analysis conducted. Chapter 5 comprises the experiments conducted and the results analysis, which is divided in to a brand-based analysis and a country-based analysis. Finally, Chapter 6 presents the conclusions, implications and limitations of the present study, and directions for future research are proposed.

# Literature Review 2

The goal of the current study is to uncover patterns in brand interest over the time applying Topic Modelling techniques. In this chapter a brief review of relevant subjects such as Topic Modelling along with LDA algorithm and Pooling Techniques is going to be presented. Moreover, some studies using Twitter as a data source for tracking events through Topic Modelling are also presented.

## 2.1 Topic Modelling

Text Mining, which is the process of extracting meaningful, nontrivial information and knowledge from unstructured text (Tan et al., 1999), has gained a lot of attention in the past years. The increasing interest in this field is due to the enormous volume of text data generated every day in social networks, news outlets, blogs, healthcare insurance data, customer reviews, and so on. Because machines lack the ability to process and perceive unstructured text as humans do, proper techniques and algorithms are applied with the purpose of discovering underlying patterns. Some of the algorithms and techniques commonly applied in Text Mining for analysing text include Information Retrieval (IR), Information Extraction (IE), Text Summarization, Natural Language Processing (NLP), Sentiment Analysis, Supervised Learning Methods, and Unsupervised Learning Methods (Gupta et al., 2009). Unsupervised Learning Methods are techniques applied with the purpose of finding hidden structures and patterns out of unmarked data. Contrarily to supervised learning techniques, these techniques do not need any marked training data, overcoming the disadvantage of manual effort needed to label training data. Topic Modelling is a widely used unsupervised learning technique used when dealing with text data. It is a type of statistical model used to represent latent patterns, called topics, in a collection of documents (corpus). In these probabilistic models, latent patterns are usually presented as multinomial distributions over words, based on the assumption that each document of the collection can be described as a mixture of latent topics (Allahyari et al., 2017). This technique provides a convenient way to retrieve information from unclassified and unstructured text. It considers that a topic contains a cluster of words that frequently occur together. A Topic

Model can connect words with similar meanings and distinguish between usage of words with multiple meanings.

Some examples of Topic Modelling methods are Hierarchical Dirichlet Processes (HDP), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM) and Topic Evolution Model (Alghamdi and Alfalqi, 2015). Topic Modelling tasks have been performed for tracking events/topics/trends in different domains such as academic, public health, marketing, news, business management, and so on. The usefulness of such techniques has been broadly stated in the literature. For this purpose, researchers have been collecting data from different sources: social media platforms, academic papers, news articles, health/medical records, customer reviews platforms, etc. For example, applying Topic Modelling to analyse texts describing firms' business, Shi et al. (2016) built a scalable business proximity metric based on the outcome of such analysis. This measure can be used, for instance, to pinpoint potential competitors, partners, and alliance or acquisition targets. Calheiros et al. (2017) performed a Sentiment Classification of Consumer-Generated Online Reviews. They surveyed the polarity of costumers' sentiments regarding specific aspects of a hotel unit, using Topic Modelling. For business purposes, such analysis can be used to identify which aspects of the business need improvement. Using Text Mining and Topic Modelling, Amado et al. (2017) performed an assessment of the literature with the purpose of identifying research trends on Big Data, in the Marketing field. Their results demonstrated that the interest on Big Data in Marketing have been increasing over the years, which means that Topic Modelling can be indeed applied to analyse trends over time. Similarly, but related with public health issues, Wang et al. (2016) applied Topic Modelling on literature to identify mental disorder (depression) and substance use among adolescents, as well as what the authors have called "hot" and "cold" topics. Paul and Dredze (2014) described a statistical Topic Modelling framework for identifying public health information from health-related Twitter posts. Using both a state-of-the-art Topic Model and what they called "Ailment Topic Aspect Model", a novel framework they proposed, they found that both the unsupervised models could automatically discover word clusters that meaningfully corresponded with real-world events. Kaveri and Maheswari (2017) proposed a framework for health-related topics recommendation, also based on Topic Modelling, for identifying public health-related topics and themes in twitter posts. Their results indicated that the model discovered most of the health-related issues, as well as relevant topics regarding the tobacco use.

Latent Dirichlet Allocation (LDA) is an unsupervised generative probabilistic model, which generates mixtures of latent topics from a collection of documents, where each mixture of topics produces words from the collection vocabulary with certain probabilities. A distribution over topics is first sampled from a Dirichlet distribution, and a topic is picked based on it. It models each document as a distribution over topics, with topics represented as distributions over words.

LDA considers that each document contains words from multiple topics; the proportion of each topic in each document is different, yet the topics are the same for all documents (Blei et al., 2003). LDA has been widely used in the past few years. Experiments have been performed using long documents such as academic abstracts, not so long documents such as customers' reviews, and very short documents such as micro-blogging posts. For the last one, some aggregation techniques to minimize the length and sparseness disadvantages have been applied, resulting in longer pseudo-documents. For instance, Moro et al. (2015) analysed literature in Baking and Business Intelligence domains from 2002 to 2013, with the purpose of identifying new trends research and possible gaps that could lead to research directions for future work. Several terms and topics grouping articles were found as a result. Another application of LDA was performed by Hong and Davison (2010), who evaluated the differences between topics learned by what they have called MSG scheme (messages from the same user aggregated into a single profile) and topics learned by the User scheme (aggregation of the user profiles, each one resulting from the collection of messages from the same user). Their results demonstrated that both produced substantially different topics, meaning that Topics learned using different strategies of data aggregation differ from each other. They also demonstrated that the length of the documents influences the effectiveness of trained topic models; namely, a better model can be trained by aggregating short messages. Alvarez-Melis and Saveski (2016) applied LDA on related Twitter posts (i.e. tweets belonging to the same conversation), with each group of related tweets corresponding to a single document, to evaluate whether the proposed technique would outperform alternative schemes. The resulting topics performed better than those derived by hashtag-based pooling. Calheiros et al. (2017) uncovered the customers' perceptions and feelings using customers' reviews regarding a single hotel, crossing semantics of sentiment polarity and hotel domain. The results demonstrated that LDA can be applied so that one can understand the strengths and weaknesses of a specific hotel unit. Hu et al. (2012) jointly demonstrated the topics of specific events as well as its associated tweets, while performing an event segmentation, with the event consisting of several paragraphs, each one of them discussing a particular set of topics. They assumed that an event, or a segment of it, can impose topical influences on the related tweets, resulting either in general topics, which are constant during the event, and specific topics, which are related to specific segments of the event. Role Discovery, Emotion Topic, automatic essay grading, and anti-phishing, are some examples of the applications and models based on the LDA method (Alghamdi and Alfalqi, 2015).

## 2.2 Pooling Techniques

Twitter posts presents some challenges due to their sparseness, as the short documents (posts) might not contain enough data to establish satisfactory term co-occurrences. Although, LDA have been proved to produce good results when applied to long documents corpora, such as news articles (Zhao et al., 2011), and academic abstracts (Yau et al., 2014), they often produce less coherent results when the application is performed on posts from micro-blogging platforms such as Twitter posts. This is due to the sparse nature of tweets, and due to the sparsity of short documents in general. Therefore, in order to alleviate the disadvantages caused by the short documents' sparseness, several pooling schemes to group together tweets into individual documents have been proposed, so that the LDA performance is improved without having to modify its basic machinery.

Examples of these techniques are hashtag-based aggregation (Mehrotra et al., 2013; Steinskog et al., 2017), user-based aggregation (Hong and Davison, 2010), or even user-to-user conversation aggregation (Alvarez-Melis and Saveski, 2016). A Topic Model based on self-aggregation was also presented by Quan et al. (2015), which is based on the assumption that each text snippet is sampled from a long pseudo-document. A tweet-pooling technique (treating each Twitter post as a single document) was also proposed in the work of Zhao et al. (2011), based on the assumption that a single tweet is usually about a single topic. The researchers proposed a technique for modeling topics in short texts (like tweets) called Biterm Topic Model (BTM), in which the topics are learnt by modeling the generation of word co-occurrence patterns in the whole corpus, rather than modeling the generation of word co-occurrence patterns using a single document, as in LDA algorithm.

## 2.3 Twitter Data as source for Trend Detection

In the past few years, Twitter data has drawn attention among researchers, as it can be exploited, using proper techniques and approaches, to track events, trends, sentiments, and relevant topics, in many domains. In the work of Lau et al. (2012), the authors proposed a Topic Modelling based methodology to analyse Twitter trends. Their approach enabled them to identify:

i) Fine-grained insights into the nature of the event;

ii) Shifts in the Topic Model, enabling the detection of emerging events in the data stream, which replace rare topics as they fade away;

iii) Popular topics in particular locations.

Mehrotra et al. (2013) proposed, among others, a Temporal Pooling scheme to aggregate tweets into what the authors have referred to as macro-documents, based on the assumption that when important events occur, a great number of users start posting about the event within a short time span. As such, the authors pooled together tweets posted within the same hour. They found that such scheme can improve Topic Modelling on Twitter, without having to modify LDA machinery. Paul and Dredze (2014), for instance, aiming to characterize health information discussed on Twitter, presented a framework for discovering health-related topics, called ATAM. Their approach could indeed discover coherent clusters of tweets, and the authors were able to demonstrate that some of the clusters were correlated to temporal and geographical surveillance data. They identified, through the presented approach, seasonal temporal patterns and geographical trends. They found, for example, that exercise and obesity were significantly correlated to survey data in the United States.

# Dataset Description and Preparation

# 3

This chapter describes which criteria we based our collection process on, the dataset composition, and all the steps and tasks conducted before we could start our experiments. Moreover, we describe which tasks worked well for this study, and which tasks did not, leading us to skip them, in the scope of data preparation.

## 3.1 System Architecture

In this system, the data was extracted using the Twitter Real-time filter API[1], where a filter was firstly applied so that only the Portuguese written Geo-located tweets were caught. Next, we chose 16 clothing brands based on:

i) The number of followers the brand has;

ii) The number of tweets mentioning its label.

Next, a brand-related tweets selection process took place. The tweets considered relevant were those mentioning at least one of the 16 brands previously selected. After this phase, a pre-processing step was conducted, to prevent irrelevant tweets from being stored in the database. After removing irrelevant tweets, the tweets were concatenated to create larger pseudo-documents, based on three different time spans: day, week, and month. These documents were then imported to R[2], a software environment for statistical computing and graphics, in which they were pre-processed once again, and the Topic Models were trained/applied. The R packages used to achieve our results are the following:

i) Data pre-processing: Text2vec, Ptstem, and Rslp;

ii) Data modelling: LDA, RTextTools, SparseM, Tidyverse, TidyR, TM, and TopicModels;

[1] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
[2] https://www.r-project.org/

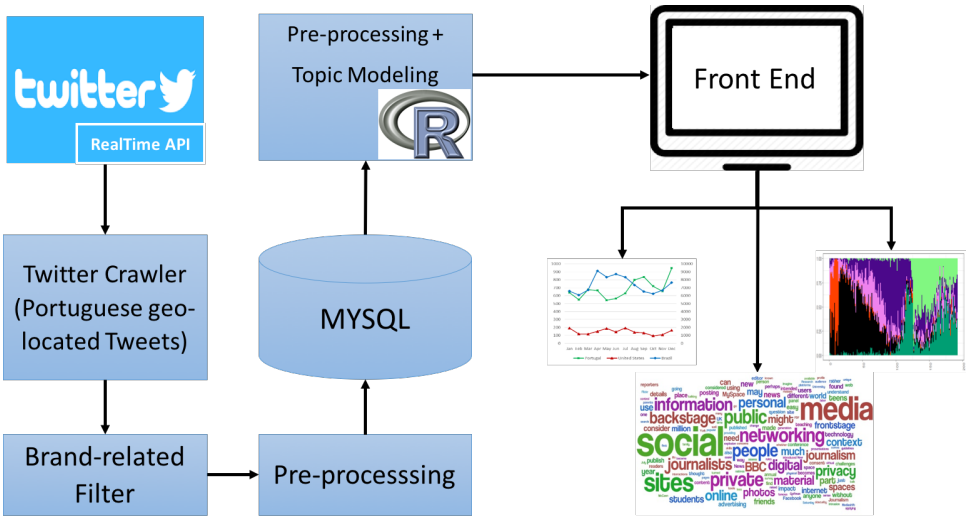



Figure 3.1: System Architecture

iii) Data visualization: Ggplot2, LDAvis, Pals, RColorBrewer, Reshape2, SnowballC, and Wordcloud.

## 3.2 Dataset Description

This study uses the same dataset used in the works presented by Lopes-Teixeira et al. (2018b) and by Lopes-Teixeira et al. (2018a), which consists of about 357,944 Portuguese Geo-located tweets, posted by 159,615 users from 206 countries across the world (according to the platform indication). The tweets were collected between May 2014 and November 2017, corresponding approximately to a four years' time span. Each post includes the metadata information as follows: user id, username, user description, country and city from which the tweet was posted, date and time, the tweet id, and the message content.

The dataset was filtered using a brand criteria, so that only tweets mentioning at least one of the 16 brands selected would be retained. The brands, which were picked based on the number of followers and the number of tweets, are the following: Adidas, Armani, Burberry, Christian Louboutin, Converse, Dolce & Gabanna, Gucci, Marc Jacobs, Michael Kors, Nike, Puma, Tommy Hilfiger, Valentino, Vans, and Victoria's Secret. Like in Lopes-Teixeira et al. (2018b) and in Lopes-Teixeira et al. (2018a), for the current study, only the top 10 brands were considered, which are the brands with more volume of posts in the dataset, as shown in Table 3.1.

In this study, the countries to be analysed are restricted to Brazil, Portugal and the United

Table 3.1: Brand-related Tweets

| Brand | Users | % | Tweets | % | Tweets/User |
|---|---|---|---|---|---|
| Nike | 68,098 | 42.66% | 126,427 | 35.32% | 1.86 |
| Adidas | 65,870 | 41.27% | 120,784 | 33.74% | 1.83 |
| Vans | 41,071 | 25.73% | 70,091 | 19.58% | 1.71 |
| Puma | 12,763 | 8.00% | 18,710 | 5.23% | 1.47 |
| Victoria's Secret | 9,574 | 6.00% | 12,642 | 3.53% | 1.32 |
| Gucci | 5,312 | 3.33% | 7,988 | 2.23% | 1.50 |
| Versace | 4,989 | 3.13% | 7,312 | 2.04% | 1.47 |
| Valentino | 3,924 | 2.46% | 6,083 | 1.70% | 1.55 |
| Converse | 4,893 | 3.07% | 5,975 | 1.67% | 1.22 |
| Michael Kors | 1,075 | 0.67% | 1,558 | 0.44% | 1.45 |

Table 3.2: Tweets per Country

| Country | Users | % | Tweets | % | Tweets/User |
|---|---|---|---|---|---|
| Brazil | 136,753 | 85.68% | 311,277 | 86.96% | 2.28 |
| Portugal | 10,645 | 6.67% | 28,691 | 8.02% | 2.70 |
| United States | 3,912 | 2.45% | 5,642 | 1.58% | 1.44 |
| Argentina | 1,254 | 0.79% | 1,374 | 0.38% | 1.10 |
| Spain | 1,052 | 0.66% | 1,292 | 0.36% | 1.23 |
| France | 973 | 0.61% | 1,309 | 0.37% | 1.35 |
| United Kingdom | 827 | 0.52% | 1,240 | 0.35% | 1.50 |
| Mexico | 740 | 0.46% | 921 | 0.26% | 1.24 |
| Italy | 654 | 0.41% | 990 | 0.28% | 1.51 |
| Indonesia | 444 | 0.28% | 701 | 0.20% | 1.58 |

States, as these three countries lead the list of the countries with more users, and Portuguese written tweets.

## 3.3 Dataset Preprocessing

Several pre-processing steps were applied to remove irrelevant tweets. For example, regarding "Valentino" brand, posts mentioning "Bobby Valentino" and "Valentino Rossi" were removed from the database, as well as all the tweets mentioning "Valentino" posted by users from Argentina, as for this particular country many instances of tweets mentioning "Valentino" were not referring to the brand; rather the posts were most likely about persons or pets. Similarly, tweets having the words "Valentino" and "Humoro" were also removed, as in these cases the users were not talking about the brand. Additionally, as long as no other brand have been mentioned in the post, tweets containing ice-cream-related words and the word "Valentino"

were also stripped out, as they were referring to an homonimous ice-cream shop instead of the brand. This was also done for tweets containing the word "Versace", as there is also an ice-cream shop with the same name. Tweets mentioning "Gucci Mane", "Gucci gang", and "Gucci fica bem com ela" (Gucci looks good on her) were filtered out. All the posts from a specific user from Indonesia were discarded, as such user presented an unusual volume of posts, and we have found that the corresponding account was used only for advertising purposes. Several other accounts used for the same purposes were also identified, mainly among the United States users. Posts uploaded by these users were discarded as well.

Table 3.1 shows the selected brands and that Nike and Adidas are two of the most well-known brands, being mentioned by the majority of the users in our database. As the country field appeared written in several different languages, we conducted a normalization step which consisted of defining a translation table where all the values were translated into English, except for "Cabo Verde", "Côte d'Ivoire", and "Costa Rica". Although, Hong Kong and Macao are currently provinces of China (officially the People's Republic of China), both were treated as separated regions, as they hold the statute of special administrative regions. Taiwan was also treated separately, even though this country is sometimes still considered as a province of China. A total of 86 tweets had the location filled with the hyphen mark. For some of them, the location was inferred from other Geo-located tweets posted by the same users who posted the non-located tweets. The ones that no other Geo-located tweets posted by the same user were found, were removed, as they were not valid for this analysis. All instances of "Michael Kors" and "Victoria's Secret" brands were concatenated, so that the words composing the brands' names could be considered as a single word as much as possibly counting as one. Regarding the brand "Converse All Star", we replaced the different ways people referred to this brand by the label first term, which is "Converse". The reason why we opted to proceed differently for Michael Kors and Victoria's Secret is that "Michael" and "Victoria" are terms that could easily be mistaken for people's names. So, we chose to concatenate them in oppose to have them split.

In order to apply Topic Modelling techniques, and to help mitigate tweets length disadvantage, all the tweets mentioning the same brand and posted during the same week were aggregated using a concatenation script. This step resulted in 1918 documents, posted over a total of 192 weeks, with an average of approximately 10 documents/week. Nonetheless, the model was trained using tweets grouped by day instead of tweets grouped by week. This was done because of the fact that tweets aggregated on a daily basis produced better topics than those resulting of tweets grouped by week. This is in line with the study presented in Mehrotra et al. (2013), in which the researchers grouped their tweets by time spans of 2 hours, based on the idea that a great number of users tweets more about events within a short time span. Nevertheless, we chose to widen our time span to 24 hours rather than using a lesser time unit, due to the size of our dataset (almost four years of data). After training our model on tweets grouped by day,

the model was applied on tweets grouped by week for tweets from Brazil, and by month for the other two countries analysed, resulting in obtaining clearer trends visualization. The reason why Portugal and the United States had a different criteria for tweets aggregation is due to the lack of data, i. e., there are several weeks with no Geo-located tweets, written in Portuguese mentioning the brands analysed.

Several other pre-processing steps had to be applied to the dataset so that more coherent and informative topics could be returned. Because it is common to find tweets containing URLs, slang, misspellings, and hashtags, the steps applied consisted of removing URLs, stop words, hashtags, punctuation, numbers, and white spaces. In order to retain a good vocabulary to represent the whole dataset, Term Frequency-Inverse Document Frequency (TF-IDF) weighting algorithm was applied. Also, terms that are not present in at least 2 pseudo-documents were not included in the vocabulary. This is achieved by removing sparse terms with a sparsity factor between 0.9947 and 0.99983, depending on the size of the dataset being analysed. The objective of this step is to avoid as much as possible that misspellings, which may occur only a few times, were caught by the TF-IDF weighting measure, without stripping away words with a low frequency rate that could, actually, be relevant. The vocabulary was restricted to 5000 words. Vijayarani et al. (2015) provided an overview of pre-processing tasks, and discussed what they have considered the three key steps of pre-processing, namely: Removing stop words, stemming and using TF-IDF weighting algorithms. Similarly, in the work presented by Srividhya and Anitha (2010), the researchers evaluated several pre-processing techniques and analysed the effect of such pre-processing tasks on text classification using machine learning algorithms. Along with the pre-processing steps aforementioned, other pre-processing steps were applied to the dataset. These steps consist of removing adverbs, cardinal numbers, ordinal numbers, conjunctions, verbs, slang and abbreviations widely used in social Networks. Verbs expressing some kind of willingness to have or purchase brand items, or demonstrating brand liking/loving, were kept. Brand names composed by two words (e.g. Michael Kors) had its name concatenated, so that the TF-IDF algorithm, which was applied to create the vocabulary, could handle all the occurrences properly.

The influence of the pre-processing steps on Topic Modelling output can be assessed comparing Table 3.3 and Table 3.4. As can be observed, more informative topics were produced when the pre-processing tasks were conducted. Taking for instance Topics 1 and 4 from Table 3.3, apart from brand name, only 3 out of 10 terms can be considered informative.

Topics produced from non-processed text have several irrelevant words such as “http” (URL prefix), stop words, and the name of the brand itself. As the Term-Frequency (TF) weighting measure was applied instead of the TF-IDF algorithm (for testing purposes), the topics are mainly composed by non-relevant terms, such as “com” (with) and “esse” (that), and the brand

Table 3.3: Top 5 Nike topics from non-processed data

| Topic | Terms |
|---|---|
| 1 | Adidas http meia meu nao nike que tenis uma vou |
| 2 | Com era meu nike nos que tenis https sem pes |
| 3 | Adidas com meu nao nike que tenis uma vou https |
| 4 | Adidas com comercial esse http nao nike pra propaganda que |
| 5 | Com mais meu nao nike pra que tem vou quero |

Table 3.4: Top 5 Nike topics from pre-processed data

| Topic | Terms |
|---|---|
| 1 | Pés adidas comprar quero air chinelo querendo comprei loja max |
| 2 | Adidas quero comprar air comprei loja chinelo queria casaco boné |
| 3 | Camisa adidas comprar quero inter air corintthians camisas uniforme queria |
| 4 | Comercial adidas propaganda meia brasil bota joga copa canela fenomenal |
| 5 | Meia bota shox adidas canela quero joga comprar air mola |

itself. In this case, this is not so informative when each brand is being analysed separately. Also, as stop words are frequent words throughout the dataset, and they were not removed, all the topics produced contain several stop words. Finally, the number of topics were chosen after several iterations, considering both the information each topic conveys and the clearness of the resulting topics evolution plot. As our goal is to observe changes over time, with the purpose to pinpoint trends, it is important that the resulting charts don't contain "noisy" topics, i. e., topics whose trend could barely be tracked.

Despite the application of stemming algorithms being among the important tasks of pre-processing presented by Vijayarani et al. (2015), such task resulted in some relevant words being incorrectly stemmed. For instance, the term "copa", which refers to the World Cup Championship, was incorrectly converted to "cop", as can be observed in Table 3.5. As such, we chose to skip this step.

Table 3.5: Top 5 Puma topics from stemmed vocabulary

| Topic | Terms |
|---|---|
| 1 | disc cop adidas novo nike quero tenis gira catraca camisa |
| 2 | camisa adidas uniforme disc nike cola cop novo chuteira |
| 3 | tenis quero cop chinela rihanna fenty novo colecao adidas bts |
| 4 | rihanna quero tenis disc adidas cop novo nike cara lindo |
| 5 | disc adidas cop mizuno novo camisa nike quero janeiro clube |

## 3.4 Summary

The importance of pre-processing tasks in Natural Language Processing for Portuguese written documents was emphasized in this chapter. The experiments demonstrated that pre-processing steps do have impact on the quality of the topics resulting from documents written in Portuguese. We obtained more informative topics when the documents were previously processed. Tasks such removing URL's, removing stop words and choosing the representation vocabulary based on TF-IDF can mitigate common issues that reduce the coherence of the topics. Results demonstrated that this framework can be followed to obtain coherent Portuguese written topics, enabling one to get insights about people's conversations/discussions on Social Networks.

# 4 Spatio-Temporal Analysis

In this chapter, a framework that uses Twitter as a data source to perform spatio-temporal analysis of brand interest is described. Brand Interest, which can be defined as the level of interest one has in a brand and the level of curiosity one needs to learn more about a brand (Machleit et al., 1990), is measured through the number of tweets, as more interesting brands are more likely to be talked about (Berger and Milkman, 2012).

## 4.1 Geographical Analysis

As can be observed, users from Brazil posted much more tweets than both users from Portugal and from the United States. Regarding users from Portugal, a possible reason is that Twitter is much more popular in Brazil than it is in Portugal. In fact, Twitter occupies the 6th position in the ranking of Social Networks with more browsers[1], while Brazil is the country with more users outside the United States[2]. Moreover, the population of Brazil is much larger than the population of Portugal. This disparity in the volume of posts was the reason why tweets from Portugal and from the United States had a different aggregation criteria than the one applied to tweets from Brazil. The way the tweets were aggregated is explained in detail in Chapter 3. This disparity also had impact on the number of topics produced for each country.

A geographical analysis was then conducted in order to observe the spatial distribution of brand interest, by measuring the number of brand-related tweets generated in specific locations (on country level). Figure 4.2 shows the distribution of tweets (top) and the distribution of users (botton) worldwide, revealing that more tweets were produced by users from America and Western Europe. Comparing both maps, we can see that the distribution of tweets and users follow a similar pattern. The maps were generated through an online map generating tool[3].

[1] tek.sapo.pt/noticias/internet/artigos/portugueses-continuam-a-gostar-do-facebook-e-a-ligar-pouco-ao-twitter
[2] https://www.omnicoreagency.com/twitter-statistics/
[3] http://www.openheatmap.com

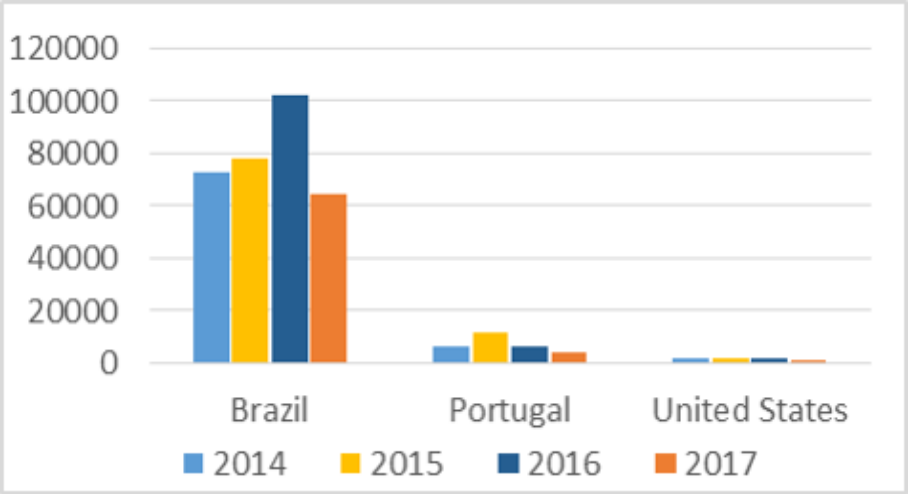



Figure 4.1: Tweets posted per year.

## 4.2 Temporal Analysis

The goal of temporal analysis is to track changes in the volume of tweets mentioning the brands and related terms over time. Figure 4.3 presents an overview of Brand-related tweets produced in Brazil, Portugal and USA. Concerning users from Brazil, it reveals that some labels have similar trend patterns. For instance, Nike and Adidas presented a similar trend behaviour, even though Adidas reached almost 9000 tweets on April 2016. Victoria's Secret differs from the rest of the brands, as its trend follows a seasonal pattern, with peaks occurring near the end of the year. Vans started with a high number of tweets, but the interest decreased over the time.

Concerning users from Portugal, the interest in Adidas, Nike, Vans, and Victoria's Secret presents, on a smaller scale, a similar trend pattern comparing to the trend of users from Brazil. Adidas and Nike follow similar trend behaviours, and the tweets about Vans decreased over the time as well. Victoria's Secret interest also presents a seasonal behaviour.

Concerning the United States users, the Adidas brand interest trend presents the same behaviour of Brazil and Portugal trends. Vans presents ups and downs, but its decreasing trend remained.

Figure 4.4 depicts the evolution of Adidas. In January 2016, Adidas launched a new campaign to revive the line of the Adidas, Original[4]. This might explain the increase in interest registered then, which achieved its highest point on April 2016, regarding the users from Brazil. Also, the brand announced its new product on March 29th, which was then launched in April. The 2016 UEFA European Championship, held in France from June to July, might also be the

[4]https://www.forbes.com/sites/willburns/2016/01/31/adidas-originals-pushes-counter-culture-with-new-advertising/#55a605904bd5

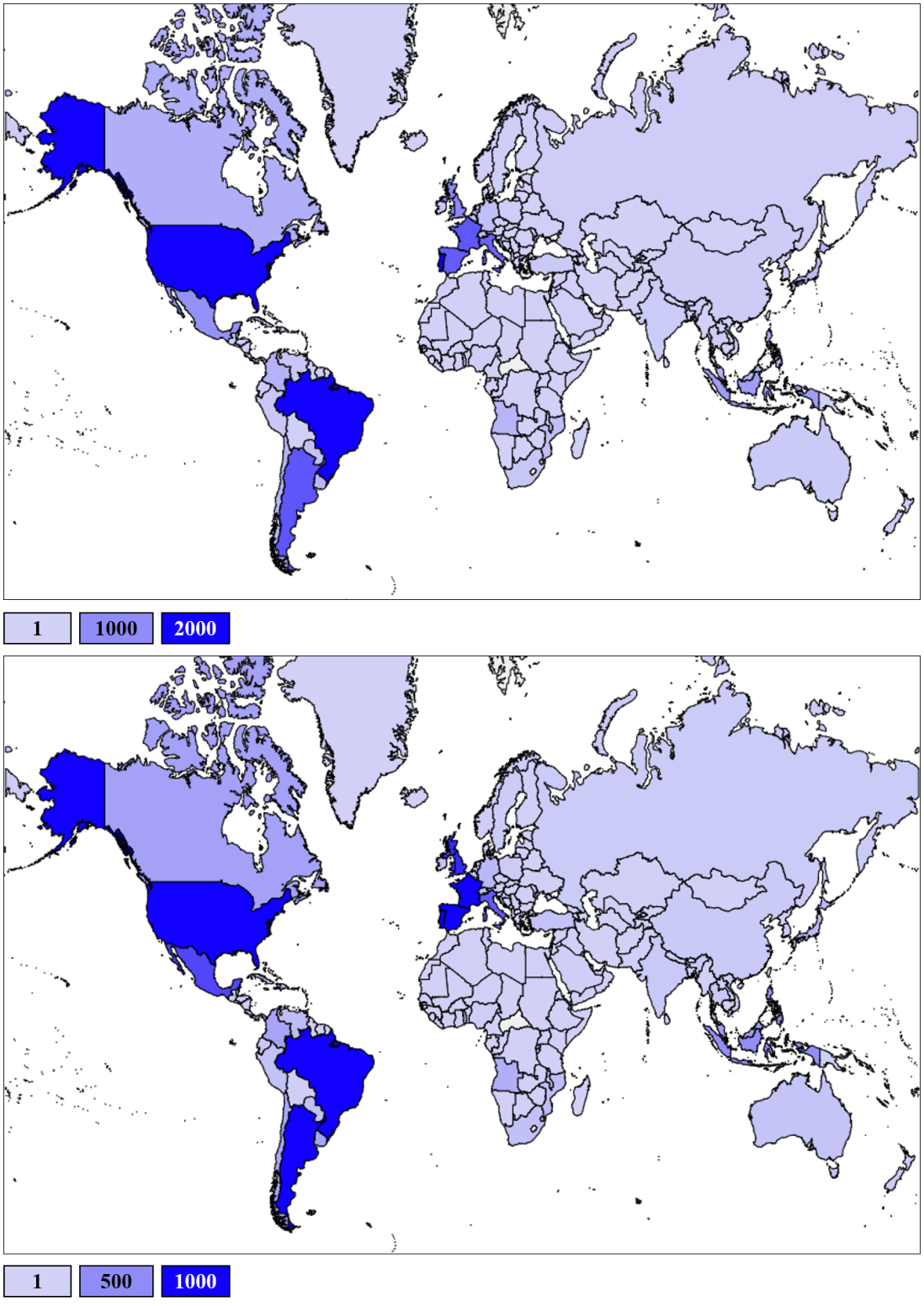
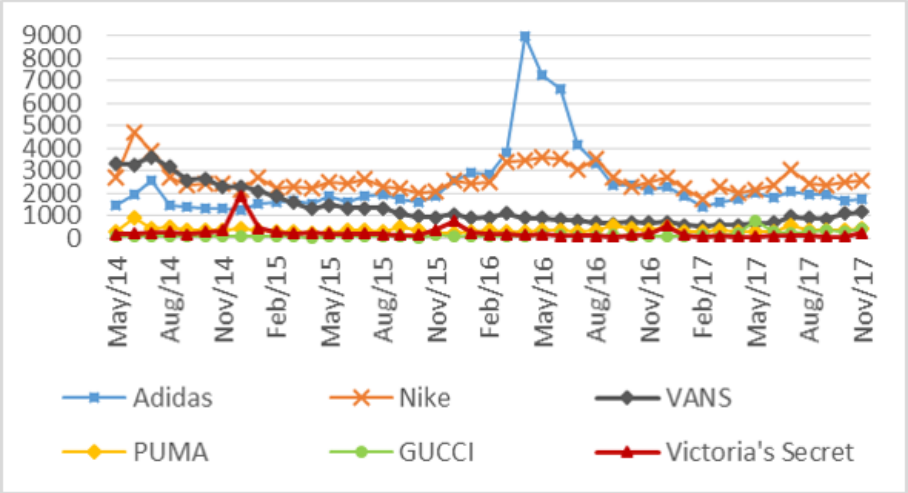
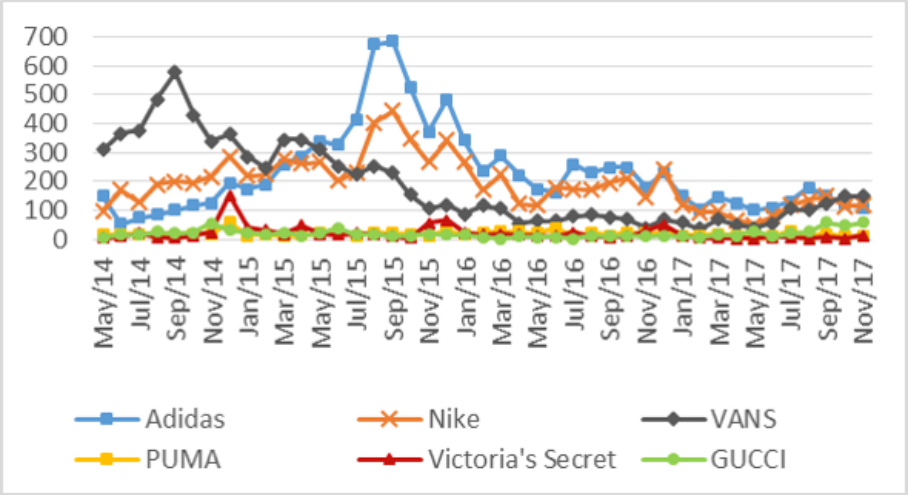



Figure 4.2: Distribution of tweets (top) and users (bottom) worldwide.

Brand-related tweets from Brazil




Brand-related tweets from Portugal




Brand-related tweets from the United States

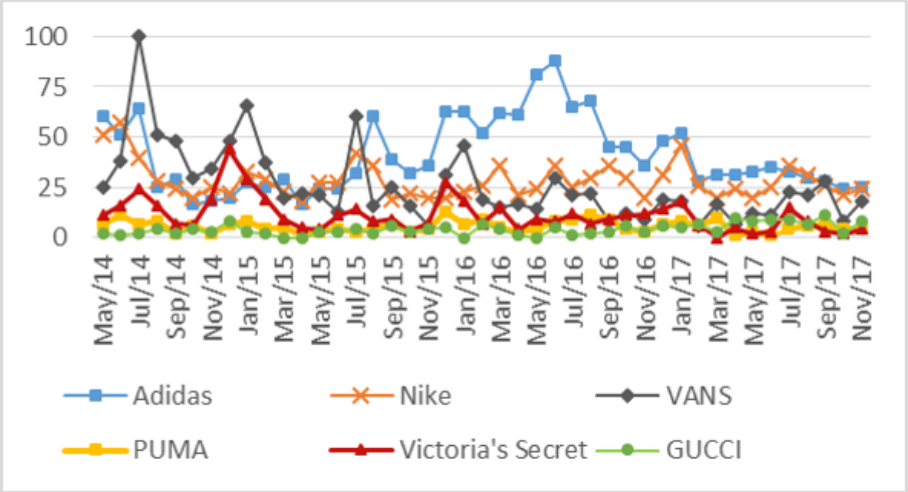



Figure 4.3: Brand-related tweets from Brazil, Portugal and USA.

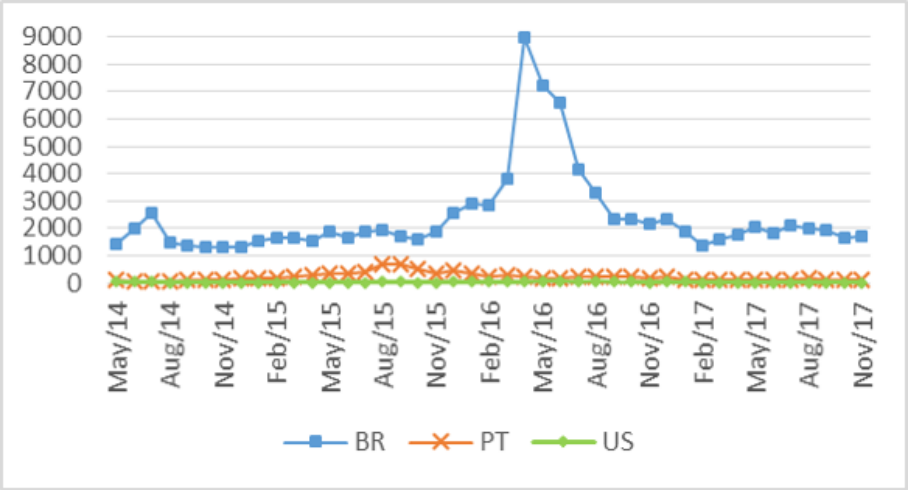



Figure 4.4: Adidas tweets time evolution

reason for the increase on tweets from April to June. The number of posts decreased considerably from June to July, which might be related to the EURO 2016 end. The tweet below shows a user comparing Adidas and Nike uniforms during EURO 2016 Championship:

- *"Esses novos uniformes dos clubes europeus estão muito bonitos Adidas mandando ver e deixando a Nike pra trás"* / These new European team uniforms are very beautiful, Adidas is leaving Nike behind.

In Figure 4.5, we can observe that in June 2014 Nike posts registered an increase. This might be due to the its World Championship campaign for 2014[5]. The fact that the World Cup 2014 occurred in Brazil might also have had an influence, as Football is very popular in Brazil. In 2015, the number of tweets registered another increase, from March to June. This is in line with Nike Spring/Summer collections launch, in March[6]. It might also be the reason why, for users in Portugal, the number of tweets increased during the same period. The Christmas season, along with the Nike's Women's Campaign, launched in January 2016, might be a reason for to the increase in the number of posts registered from December 2015 to February 2016. The "Unlimited" campaign for Rio 2016, launched in July, along with its follow-up ad called "Unlimited You", released in August, might have influenced the increase on brand interest around that period. Nike launched Signature Neymar (a well-known football player) football boots in July 2017. This launch might explain why the posts increased in the same period. December 2016 also registered an increase on brand interest, which can be related to the Christmas season. Tweets below show which (Nike) aspects users were discussing during Euro 2016:

[5]https://news.nike.com/news/nike-launches-winner-stays-second-film-in-the-risk-everything-football-campaign

[6]https://news.nike.com/news/nike-unveils-new-spring-collections-connects-with-65-million-women-across-global-digital-community

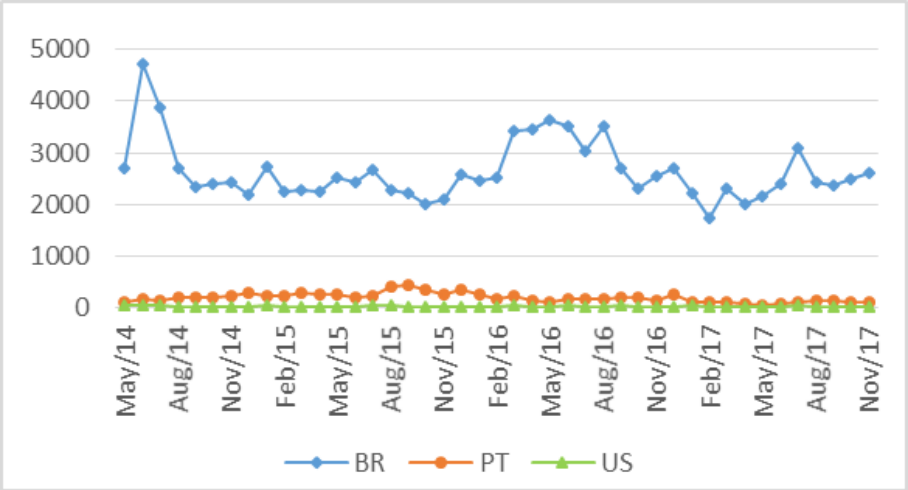



Figure 4.5: Nike tweets time evolution

- *"Adoro mt o trailer a publicidade da Nike para o mundial"* / I love the Nike ad trailer for the World Championship very much.
- *"Uniformes da Nike nessa #EURO2016 são todos iguais. Alguns são horríveis"* / Nike uniforms for #EURO2016 are all the same. Some of them are awful.

Vans announced the official re-launch of the Vans Women's Apparel and Accessories Collection, along the spring 2014 campaign video. In the same year, Vans partnered with Captain Fin to launch their Fall 2014 Collection. The brand also launched a clothing line in August 2014. In May 2015, Vans launched their first-ever Skateboarding Video. As the number of tweets was decreasing month after month, this event might explain why it stop dropping, and even increased in Brazil and Portugal (Figure 4.6). The brand 50th anniversary in March 2016 is a possible reason for the increasing in brand interest at the same period.

Victoria's Secret presents a seasonal trend behaviour, as the posts starts to rise in November, achieving its highest point in December, starting to decrease in January. This trend can be verified in Figure 4.7. Although the data from December 2017 is absent in this study, an increase on the number of tweets can be observed in November 2017, which is in line with the behaviour verified in the previous years. This trend might be due to the brand's annual fashion show, which is run in this time of the year since 2001. Figure 5.6 depicts it quite well, as the topic representing this scenario presents ups and downs in line with the trend showed in Figure 4.7. This trend is also observable on users from Portugal and the United States, as shown in the following two tweets:

- *"O que os outros vêem em dezembro: ferias. O que eu vejo: Natal, férias e Victoria's Secret Fashion Show!"* / What other people see in December: Vacations. What I see:

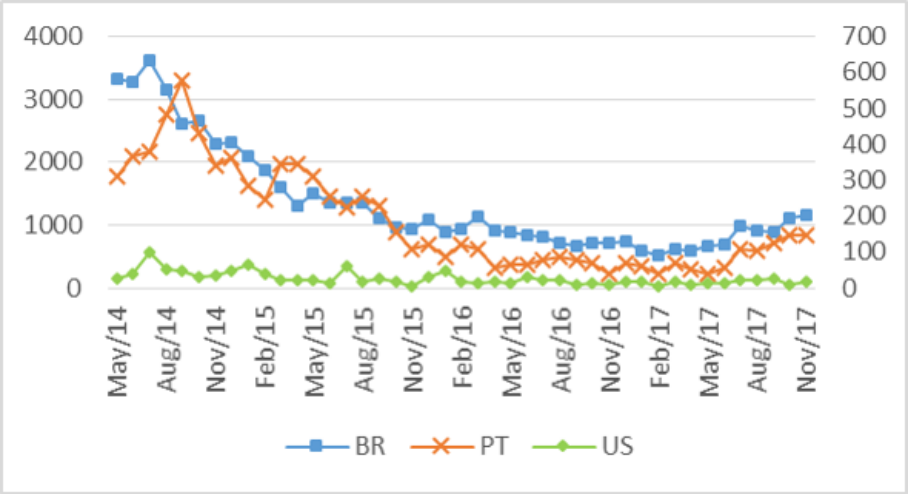



Figure 4.6: Vans tweets time evolution

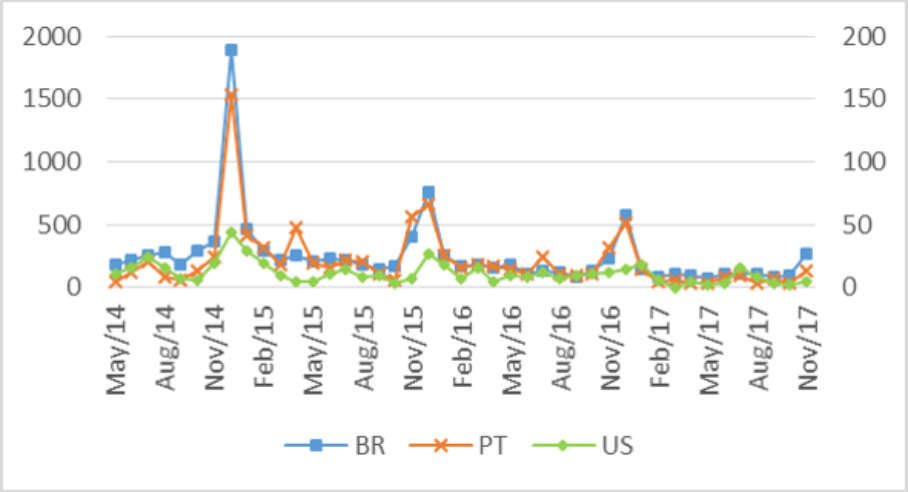



Figure 4.7: Victoria's Secret tweets time evolution

Christmas, vacations and Victoria's Secret fashion show

- *"Adorava estar no Victoria's Secret Show"* / I'd love to be in Victoria's Secret Show.

Donatella Versace's presence at the Versace for Riachuelo fashion show, during São Paulo Fashion Week, is likely the reason why Versace tweets increased, back in November 2014 (Figure 4.8). The absence of changes in Portugal and United States might be due to the brand Riachuelo not being popular outside Brazil.

- *"Hoje eu vi as peças da Versace pra Riachuelo e meu Deus, que coleção linda!"* / Just saw Versace for Riachuelo clothes and God, what a beautiful collection!"

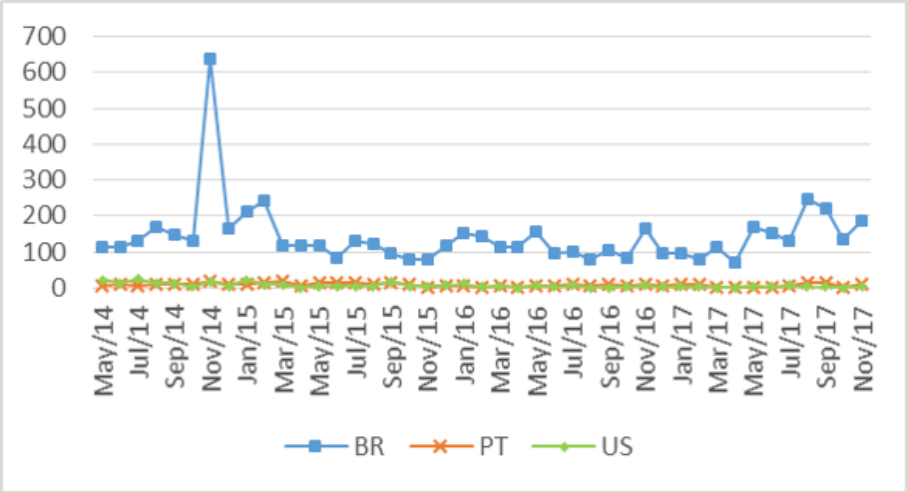



Figure 4.8: Versace tweets time evolution

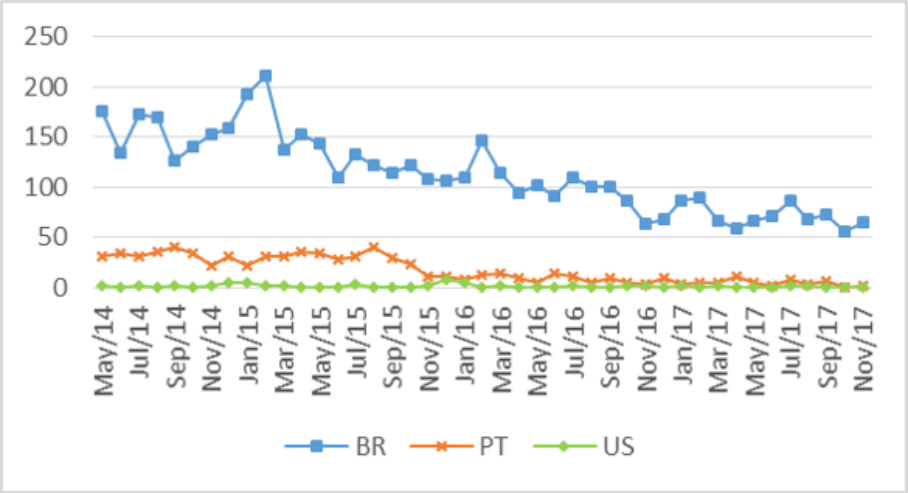



Figure 4.9: Converse tweets time evolution

- *"Eu sou fã de @Versace mas esse vestido que a Donatella usou no MET apelou né?."* / I'm a fan of @Versace but this dress Donatella wore at the MET is too much.

Converse launched a Spring collection back in March 2014, followed by the launch of a new design of the shoes right in the following year[7]. Both events might explain the rise in the number of posts in Brazil and Portugal from 2014 to 2016. The Valentine's Day Collection launch might explain the increase in posts from Brazil (Figure 4.9), in February 2016.

The first rise in the number of Gucci posts (Figure 4.10 ), in June 2016, is likely to be related to the Gucci fashion show that took place on June 2nd. The second peak, registered in May 2017, is in line with the Gucci Pre-Fall 2017 Campaign, which draw a lot of attention

[7]https://news.nike.com/chuck-taylor-all-star

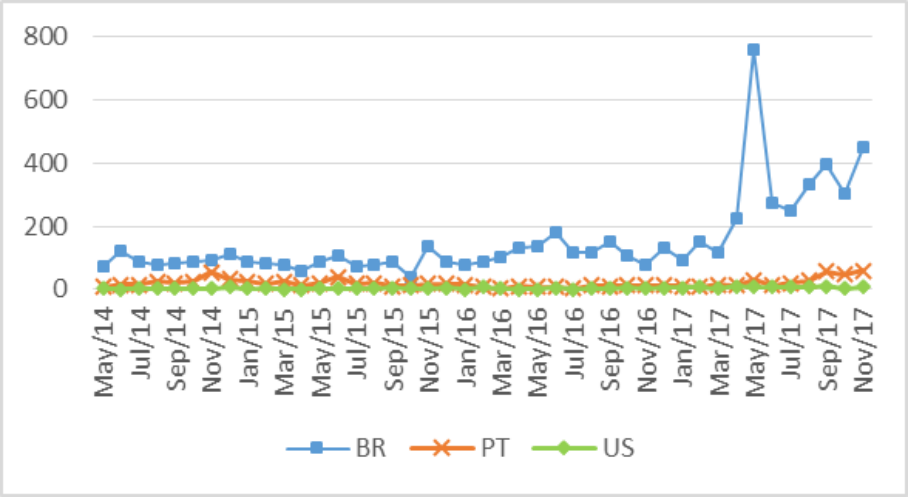



Figure 4.10: Gucci tweets time evolution

for featuring exclusively Black Models, and for having several 1960s influences. The Gucci designer Alessandro Michele's first women fragrance launch might have helped to maintain the brand interest[8]. Regarding users from Portugal, an increase in the number of tweets occurred in November 2014, which might be due to the holydays. From September to November 2017, another increase in the number of posts is verified, which might be due to the holidays season and the Gucci's fragrance launch.

Figure 4.11 demonstrates that the countries analysed, have different brand interest patterns, and periods. Brazil, for instance, has more tweets from April to July, then it rises in December, most likely due also to the Christmas. The United States, on the other hand, had more posts in January, May, July, and December. The increase on tweets in July might be related to the Independence Day. In December and January, might be related to Christmas day. The Mother's Day and The Memorial Day, both in May, might explain the increase in tweets. Portugal brand interest is higher during August and September, which is the summer and vacation season. December also registered an increase in posts, which might be related to the Christmas season.

## 4.3 Summary

This chapter describes a framework capable of tracking brand interest, both geographically and temporally, using Twitter as data source. The experiments demonstrated that the system was able to detect variations in brand interest, and that these variations are likely to be related to real-life situations, which is in line with previous studies. Moreover, the system depicted differences

[8]https://www.vogue.com/article/gucci-bloom-fragrance-perfume-launch-alessandro-michele

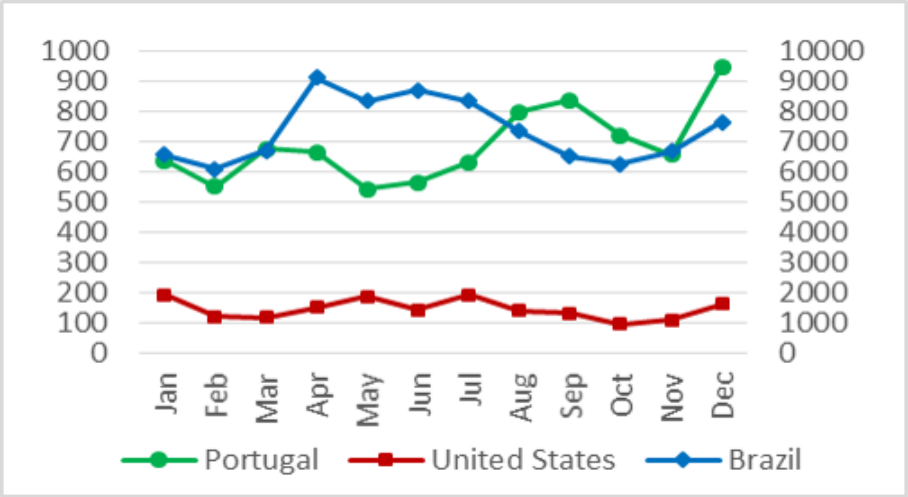



Figure 4.11: Average tweets posted monthly (Brazil in secondary axis)

in the way users from Brazil, Portugal and the United States shared their opinions and thoughts. The three countries present ups and downs in the number of tweets in different periods of time. Nonetheless, they all had an increase in brand interest around December, followed by a decrease that commences in January.

# 5 Topic Trend Analysis

This chapter describes the experiments performed, as well as the analysis conducted based on the experiments results. The results described hereby comprise experiments conducted on brands both individually and together. Country-based experiments and results analysis are also presented in this chapter.

## 5.1 Brand-based Analysis

In line with the work presented by Lopes-Teixeira et al. (2018b), it can be observed that brand interest, i.e. the volume of brand related posts, changed over the time. It presents ups and downs, and several peaks could be related to real-world events, as some previous studies demonstrated (Mehrotra et al., 2013; Paul and Dredze, 2014). The experiments also depicted that each brand presents its own pattern of brand interest, and in some cases, some brands presented similar brand interest patterns.

Comparing both Figure 5.1 and Figure 5.5, it can be observed that it is also possible to track trends from day to day for some brands. For instance, this approach worked quite well for Nike, Adidas, and Puma, as both week and day charts are very similar, but did not work so well for other brands such as Versace, i. e., less clear trends visualization as can be observed in Figure 5.2. Therefore, we adopted to perform a week-based analysis for each brand.

Figure 5.3 shows that the first weeks have more shares about Topics 1 and 4. Topic 4, which is about Nike, “propaganda”/ “comercial” (commercial), “camisa” (shirt), “copa” (world cup), “chuteira” (football boots), and “Messi” (the football player), are clearly related to the Football World Championship that took place in Brazil, which spiked brand interest regarding sport brands, such as Adidas, Nike and Puma, during the championship period, back in 2014 (Lopes-Teixeira et al., 2018b). Terms such as “chuteira” (football boots) and “camisa” (shirt), might explain why the topic did not vanish after the World Cup end, as it is also related to the Football itself. Topic 2 is also Football-related, as it mentions “Flamengo” and “Palmeiras”, which are Brazilian Football teams, “shirt” (camisa) and “tenis” (sneaker). Topic 1, Topic 3, and Topic 5

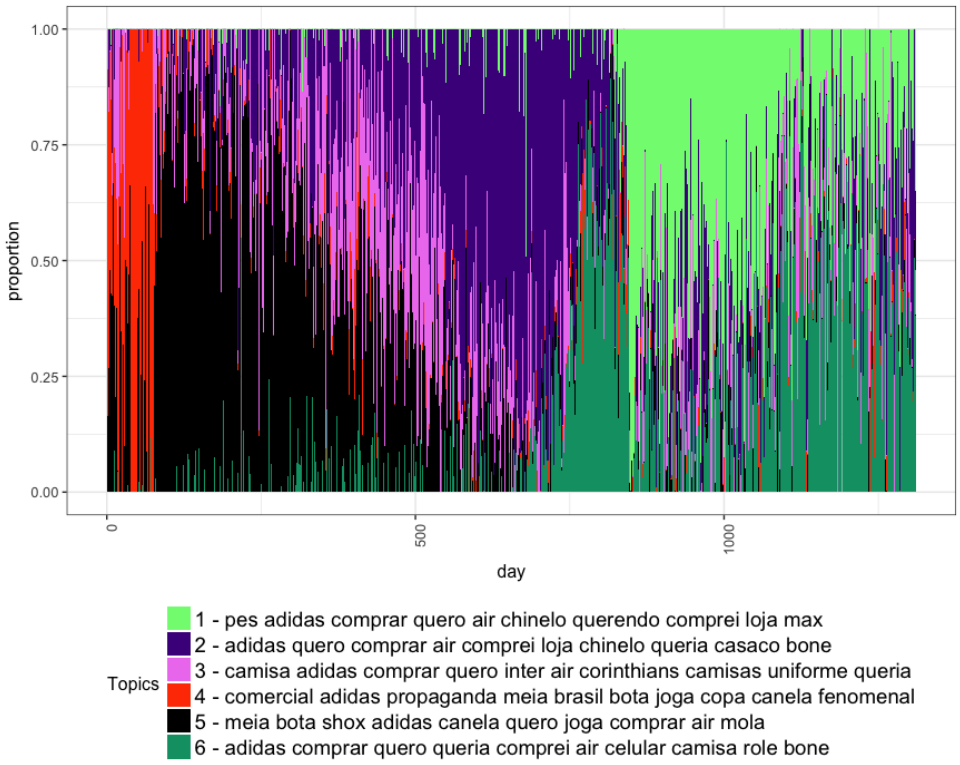



Figure 5.1: Nike topics daily evolution

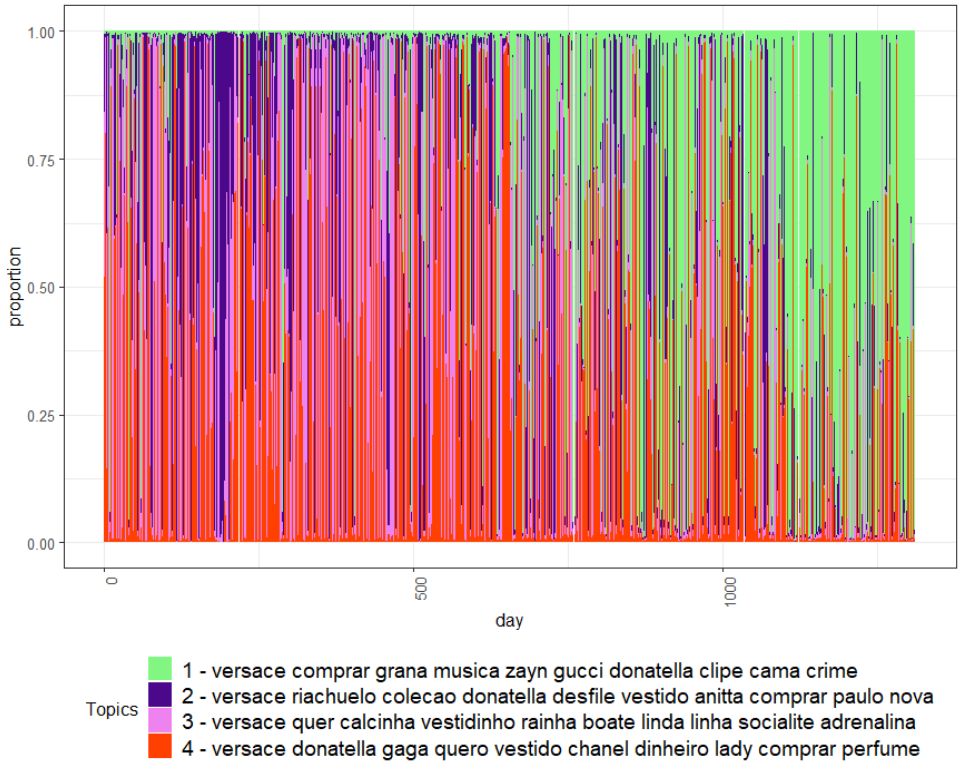



Figure 5.2: Versace topics daily evolution

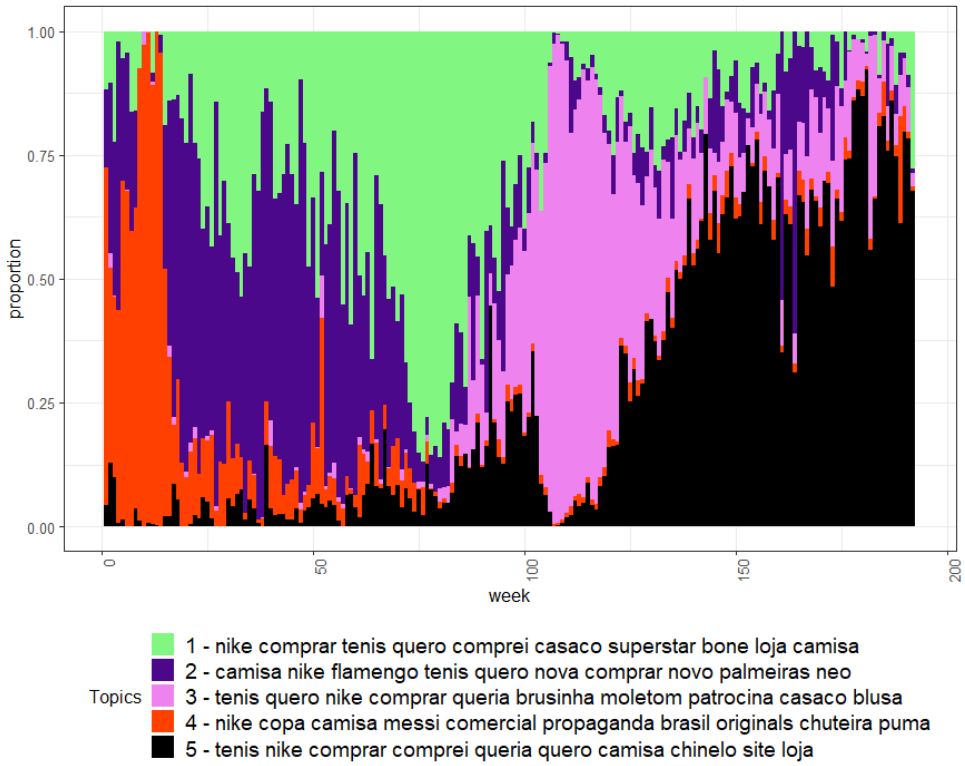



Figure 5.3: Adidas topics weekly evolution

express the intention of purchasing new items: “novo”/“nova” (Portuguese for new), “camisa” (shirt), “ténis” (sneakers), “moletom” (pullover/sweater), “chinelo” (flip-flops/slippers), “boné” (bonnet), “casaco” (jacket), and so on. A possible reason for Nike and Puma being present in topics from Adidas data might be due to sport brands being very often subjects of comparison.

Figure 5.4 illustrates how Topic 3, which is about “camisas” (shirt), “uniformes” (uniforms), Nike, Adidas, “Disc” (Puma sneakers), and “copa” (World Cup), starts with a high proportion but by the end its relevance decreases. This predominance is also related to the World Football Championship that took place in Brazil, back in 2014. In fact, in line with Lopes-Teixeira et al. (2018b), the three sports brands Adidas, Nike, and Puma, present a spike in the beginning of the chart, which is related to the Football World Cup event. As it is a football-related topic, it did not fade completely, due to the fact that football championships unfold almost during a whole year. Topic 4, which mentions “tenis” (sneakers), “Disc” and “Mizuno” (Puma sneakers), “camisa” (shirt), “Adidas” and “Nike”, along with the term “comprei” (I bought), has a higher proportion from the beginning until the middle of the dataset, losing strength afterwards. In the fall of 2015, the first sneaker of Rihanna’s collaboration with Puma was released, which sold out online with the pre-sale launch. Over the next two years, Rihanna also released several other merchandise, which were all met positively by both critics and buyers. In 2016, Rihanna debuted her first clothing line in collaboration with Puma. In the

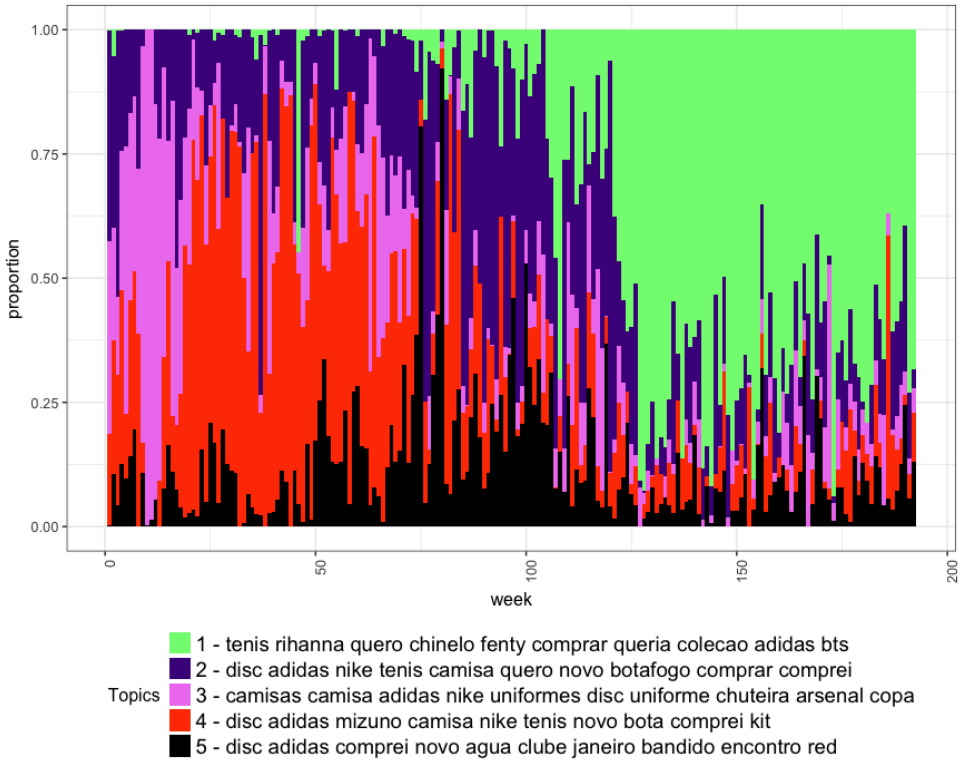



Figure 5.4: Puma topics weekly evolution

spring of the same year, the second collection was also unveiled. In Autumn 2017, the debut of their autumn collection was presented. The chart shows that the evolution of Topic 1 is in line with these events. Topic 5, in which figure terms like “disc”, “Adidas”, “novo” (new), and “comprei” (I bought), is present almost from the beginning until the end of the plot.

Similarly to Adidas topics, Nike topics also contain the term “Adidas” as we can see in Figure 5.5, demonstrating that these brands are mentioned in the same document several times. The wish of purchasing is common to almost every topic, and it is shared across the weeks. What distinguishes them from one another are, in essence, the items which is the object of their desire. Topic 6, for instance, is about “camisa” (shirt), Air (sneakers), “boné” (bonnet), while the first topic mentions “Max” (sneakers), “boné” (bonnet), and “chinelo” (slippers). Clearly, this indicates that, for this brand, the items users are interested in changed over the time.

The third topic is related to Football equipment items such as “uniforme” (uniform) and “camisas” (shirts), and two Brazilian Football teams (Corinthians and Inter). This topic is present in almost every documents, which might be due to the Brazilian Football Championship, which occurs throughout the year. The next topic, which is about the World Cup, can also easily be spotted in the chart. It can be observed that this topic was more discussed in the early weeks of the dataset, then its proportion decreased considerably as the time went by. Topic 5, in which

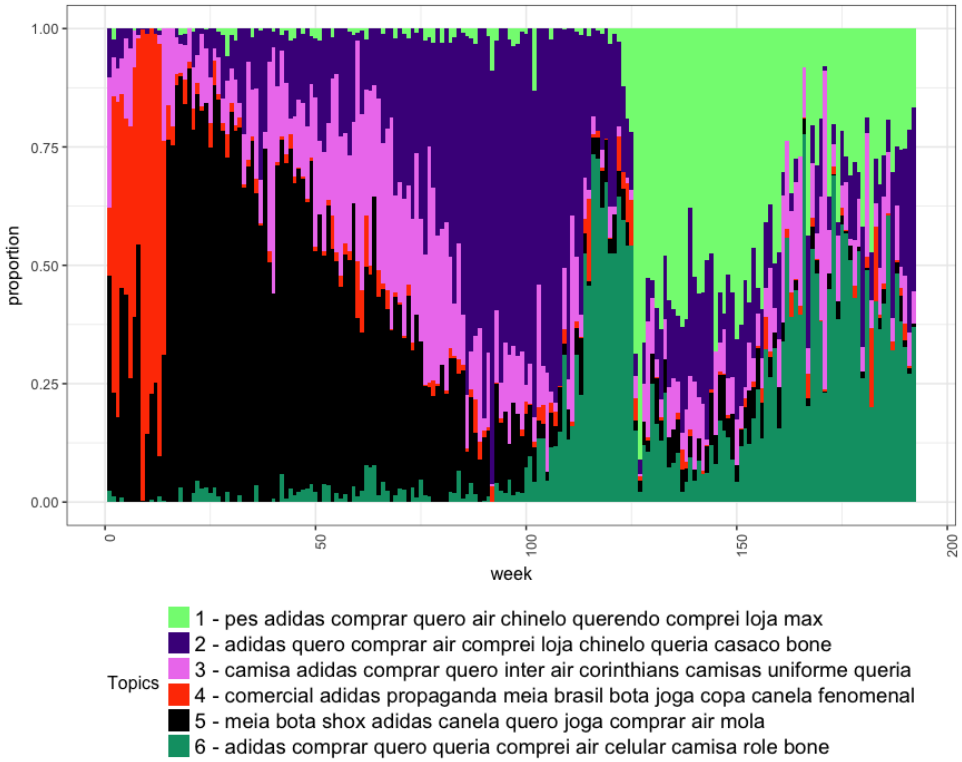



Figure 5.5: Nike topics weekly evolution

figure the terms "Shox" and "Air" (Nike sneakers), "bota" (boots), and "meia" (socks), was discussed from the beginning to the middle of the dataset, then it loses relevance. This is in line with the launch of Nike Spring/Summer collection, which occurred around the first semester of 2015 (Lopes-Teixeira et al., 2018b).

- *"Adoro mt o trailer a publicidade da Nike para o mundial"* / I love the Nike ad trailer for the World Cup very much.

- *"Esse comercial da nike ta oh ???? Que venha Copa Nike"* / This Nike commercial is lit! Let the world cup begin.

In Figure 5.6 it can be observed that the last topic for Victoria's Secret, which is essentially about their annual fashion show, presents a seasonal behaviour. This trend was also highlighted in the work of Lopes-Teixeira et al. (2018b). The same behavior can be observed in Figure 4.7, suggesting that this topic is responsible for the brand seasonal pattern. The other two topics don't behave the same way, rather, the second topic, which is about "roupa" (clothes), "morango" (strawberry), "baunilha" (vanilla), and chantilly scented body lotions, is more talked about in the first half of the dataset. The first topic, on the other hand, presents ups and down from the

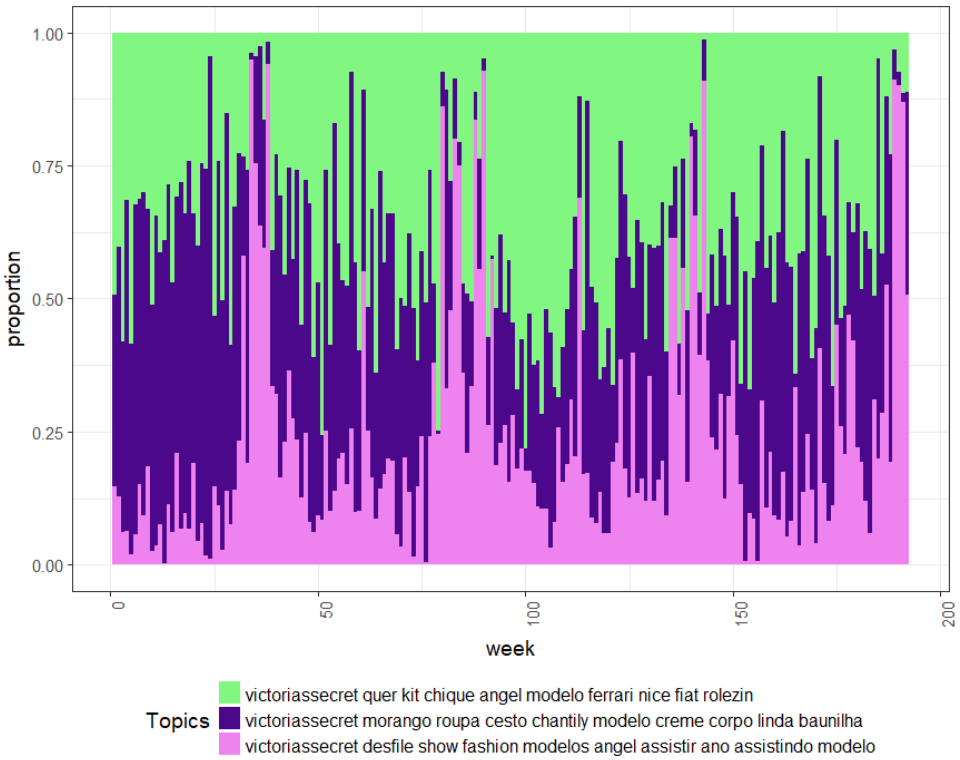



Figure 5.6: Victoria's Secret topics weekly evolution

beggining until the end of the dataset. Among all the brands analysed in this study, Victoria's Secret is the only one presenting a topic with a clear definitive seasonal trend behaviour.

Topic 1, depicted in Figure 5.7, is not quite related to the brand. In fact, this topic is related to a song from a Brazilian singer, in which brands like Armani, Oakley, Lacoste are also mentioned. The name "Harry" also figures in this topic, referring to the former member of a British boy band, Harry Styles, whose appreciation for the Gucci brand culminated in him being the new face Of Gucci's Tailoring Collection. Similarly, although mentioning fashion brands such as Gucci, Versace, and Prada, Topic 3 is also related to a Brazilian song.

By the end of the chart, we can easily spot Topic 2, which is about Kim Taehyung's (from the South Korean boy band "Beyond The Scene") appreciation to Gucci clothes; his appreciation to the brand resulted in it to be noticed/talked about in 2016. Topic 4 is composed by terms such as "Gucci", "cinto" (belt), "tenis" (sneakers), "bolsa" (bag), "roupas" (clothes), "dinheiro" (money), and another *haute couture* brand, "Chanel". This topic has its proportion increased from the middle to the end of the chart. Topic 5 mentions other two fashion brands (Prada and Chanel) along with the terms "tenis" (sneakers), "suítes" (suite), "óculos" (glasses), "desfile" (fashion show), and "quero" (I want). The relevance of this topic is higher by the middle of the chart.

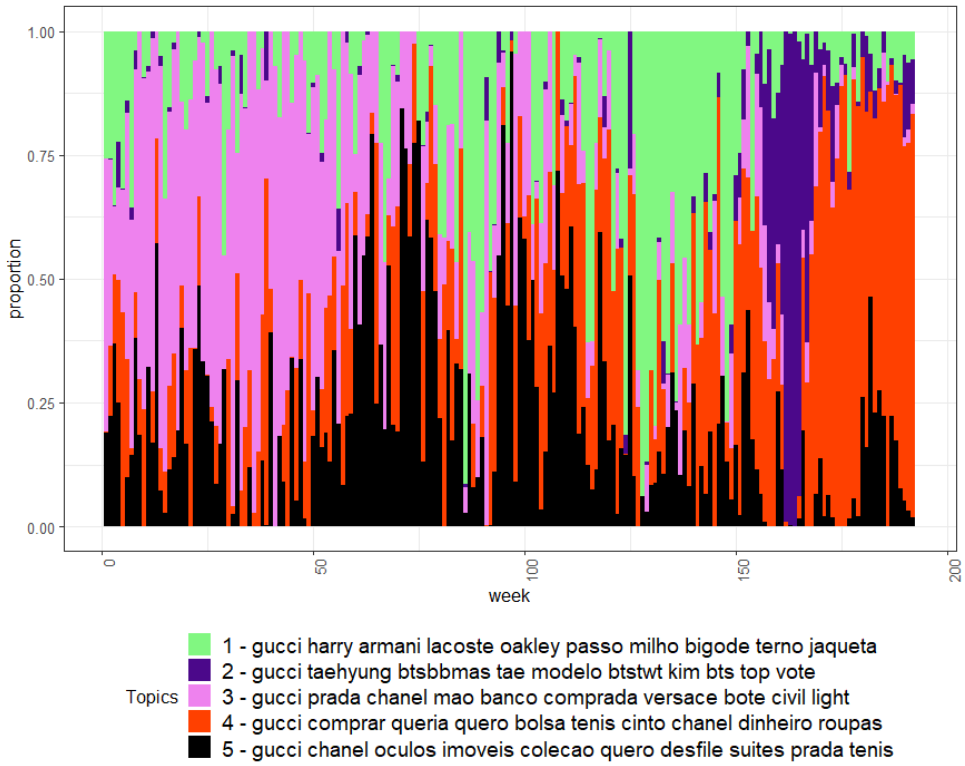



Figure 5.7: Gucci topics weekly evolution

Figure 5.8 shows that Topic 2, which is about a collaboration between Versace and Riachuelo, has an unusual proportion somewhere before week 50. This high proportion is in line with Lopes-Teixeira et al. (2018b), coinciding with the fashion show in which Riachuelo presented its Versace collaboration collection, back in November 2014. As this topic is also about "desfile" (fashion show), Donatella Versace, "roupas" (clothes), and Anitta (a Brazilian singer), the topic did not disappear completely. In fact, it presents some ups and downs that are likely related to the brand fashion shows carried out every year. The first topic, which seems to talk about haute couture brands, as it mentions Gucci and Versace, "grana" (a Brazilian Portuguese slang, which means money), is what people discussed more in the later weeks. Before this topic appeared, people were also talking about "Vestidinho" (short dress), "socialite", "boate" (nightclub), "rainha" (queen), "linda" (beautiful). This topic is related to (the lyrics of) a Brazilian song, rather than the brand itself.

Figure 5.9 shows that Converse All Star topics tend to refer Vans, maybe because people perceive these two brands as similar brands, or people tend to compare both very often. It also happens to topics from Vans, as we can see in Figure 5.11. Converse items discussed comprise "tenis" (sneakers), "roupa" (clothes), colors such as "azul", "branco", and "preto" (Portuguese for blue, white, and black), and brands Like Adidas, Nike, and Vans. We can notice that the first topic is steady, while the other two topics switched positions: the second one started with more

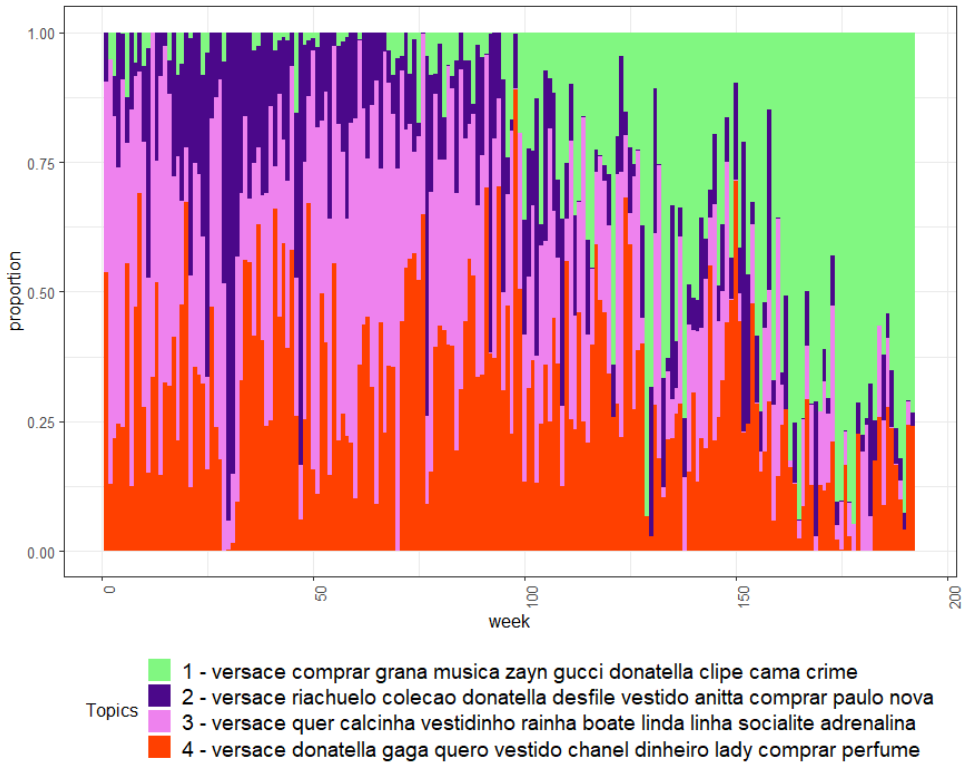



Figure 5.8: Versace topics weekly evolution

relevance, but as the time went by, its proportion decreased, giving place to third one.

Figure 5.10 shows that topics about Michael Kors revolve mainly around three items, which are “relógios”, “bolsa”/“mala”, and “óculos” (watch, bag, glasses), along with some compliments like “lindo” (beautiful), and a purchase intention sharing depicted by the terms “quero”/ “queria” (I want/ I wanted) and “comprar” (to buy). We can observe that, overall, the second topic, which is about “mala”, “bolsa”, “relógio”, Lacoste, and (Hugo) Boss, is the one with less proportion in comparison with the other two topics. The white line in the chart is due to lack of (Portuguese written) tweets about Michael Kors.

The first topic in Figure 5.11 depicted quite well the increase in popularity of Vans Old Skool sneakers, which started back in 2015. As it started to gain more proportion, discussions about the second topic, which is about the wish of purchasing new items like “ténis” (sneakers) and “moletom” (sweater), started to decrease more and more. The last topic, on the other hand, achieved higher proportions after week 50.

Figure 5.12 shows that the first topic revolves around the semi-annual fashion event of Paris Fashion Week. This topic also mentions the former boy band leader Zayn Malik, which was present at the fashion event, back in March 2017. As this fashion event is semi-annual, several increases of the first topic proportion can be spotted in the chart. Topic 2 refers to Valentino

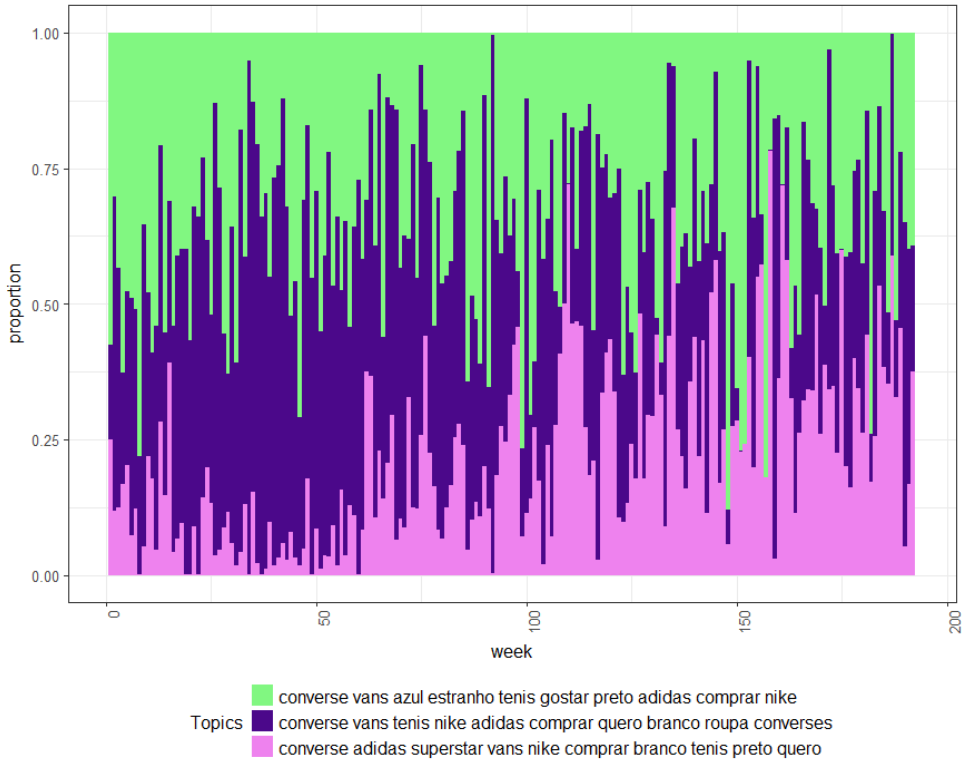



Figure 5.9: Converse All Star topics weekly evolution

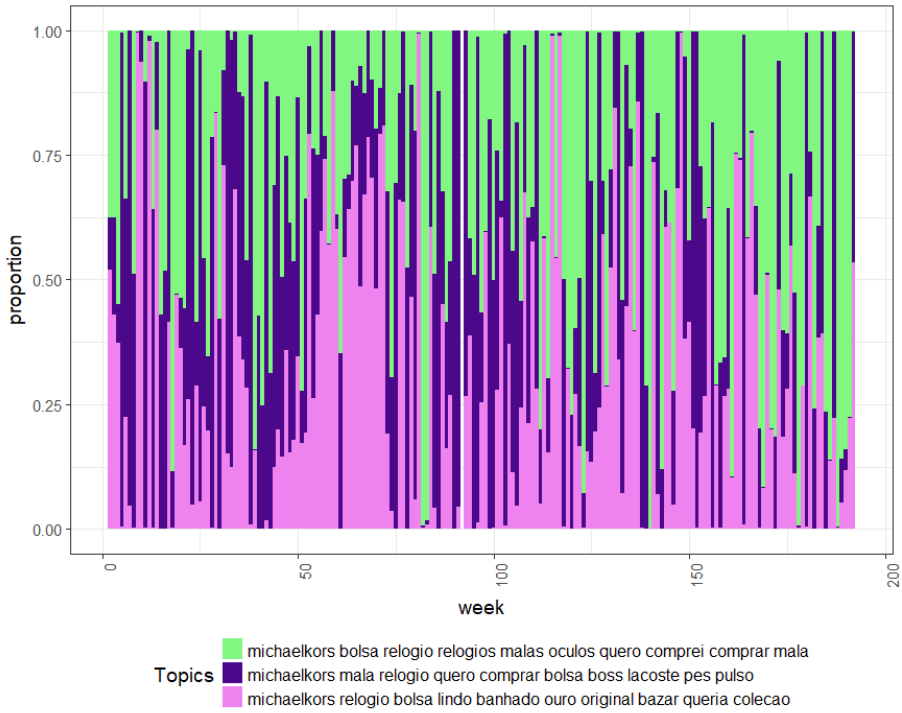



Figure 5.10: Michael Kors topics weekly evolution

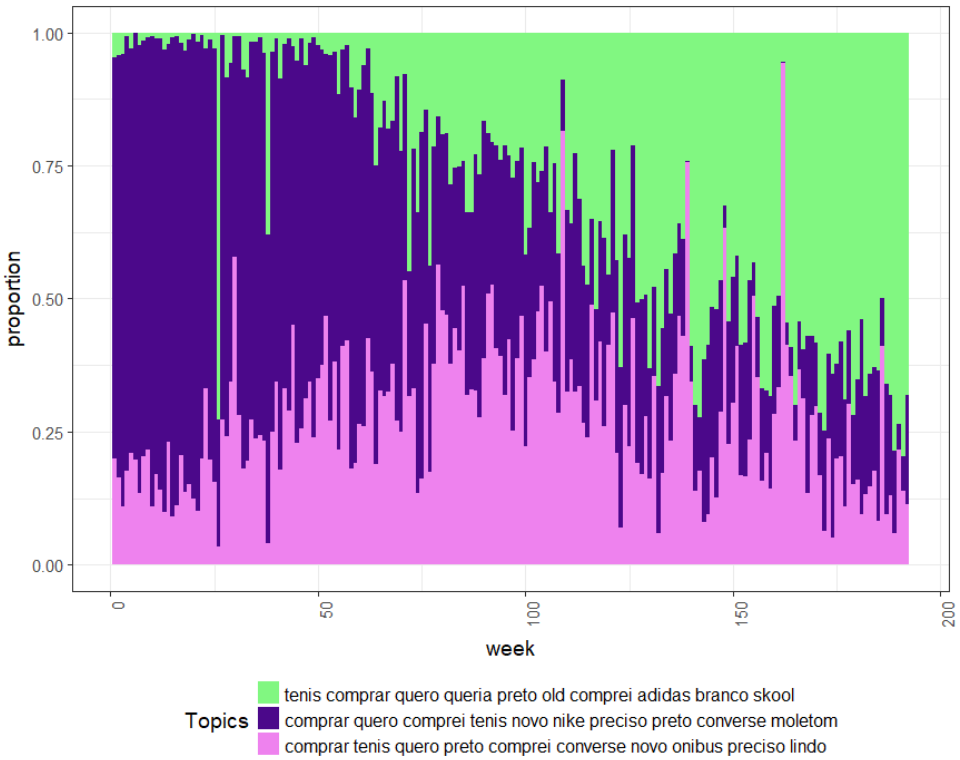



Figure 5.11: Vans topics weekly evolution

similar brands such as Dior and Chanel, and Prada, and to the Puerto Rican singer Ricky Martin.

Topic 3 has almost nothing to do with the brand; rather it seems to be about MotoGP, the sport motocross event, as it refers to “motogp”, “motogpnosportv”, “ganhar” (to win), “seguidores” (followers), and Valentino (Rossi), which is a professional motorcyclist. Topic 4 mentions Valentino Khan (an american DJ and producer), along with the terms “vestido” (dress), “bolsa” (bag), “linda” (beautiful), “quero” (I want), and “comprar” (to buy).

Figure 5.13 shows that there are three kinds of topics: the ones that emerge and vanish within a few weeks, the ones that emerge and stay longer than the former but eventually also disappear, and those discussed throughout the whole dataset. Taking for instance the topic about the World Cup “Adidas copa nike camisa Messi Brasil sport final originals alemanha” (seventh topic, in green), that started to be discussed in the early weeks, then disappears. Comparably, the first topic (in mint green) and the 36th topic (in green, bottom left), mentioning “Nike”, “Adidas”, “comercial” (commercial), “Brasil” (Brazil), “jogo” (match) “copa” (World Cup) and “Neymar” (the Brazilian football player), are also about the World Cup, and they also appear and disappear at the beginning of the chart. The same occurs with the 33rd topic (in pale pink, bottom right), which is about Kim Taehyung’s appreciation for Gucci clothes. Another example for fast-vanishing topics is the fourth topic (in bright orange), which can be spotted somewhere

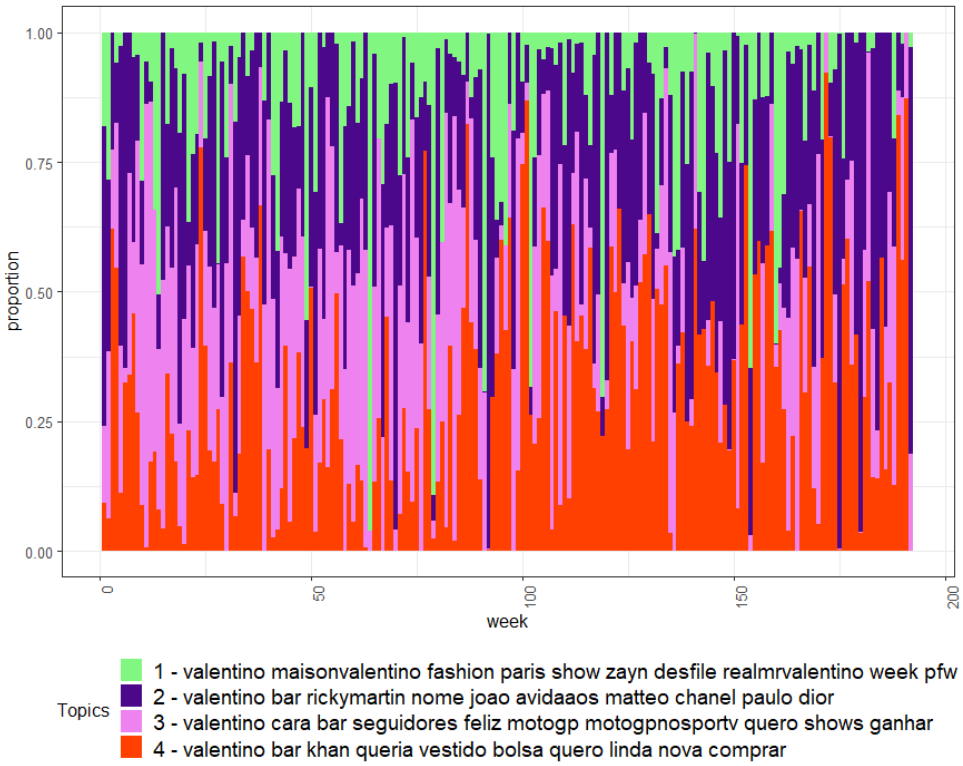



Figure 5.12: Valentino topics weekly evolution

between 100th and 150th week.

An example of topics lasting more than the ones above-mentioned is Topic 9 (in olive green), which is about buying sneakers from sport brands - “neo” (from nike) and “mizuno” (from Puma) - and shirts, that only disappears by the middle of the dataset. Topics 14, 15 and 18 (in pink, light blue, and gray, respectively) are also examples of topics that emerge and are discussed for several weeks. The fifth topic (in black), which is about Michael Kors items such as “bolsa”(bag), “relógio”(watch), and “óculos”(glasses), is talked about throughout the whole dataset. Similarly, the 26th topic (in green), about “Gucci”, “cinto” (belt), “óculos”(glasses) and other *haute couture* brands, is constant during the weeks. Victoria’s Secret Fashion Show topic (number 24, in brown), is also constant throughout the weeks, but its proportion fluctuates in a seasonal way.

Several iterations with different number of topics were conducted. We performed experiments with 30, 40, 50, and 60 topics for the whole dataset. Charts with both 50 and 60 topics were difficult to perform trend analysis on, so we chose 40 topics iteration, as it produced good (informative) topics and enabled us to conduct a better analysis comparing to the remaining iterations output. Figure 5.14 shows the 50 topics experiment output.

In Figure 5.15 we can observe that the 50 terms considered the most relevant may differ

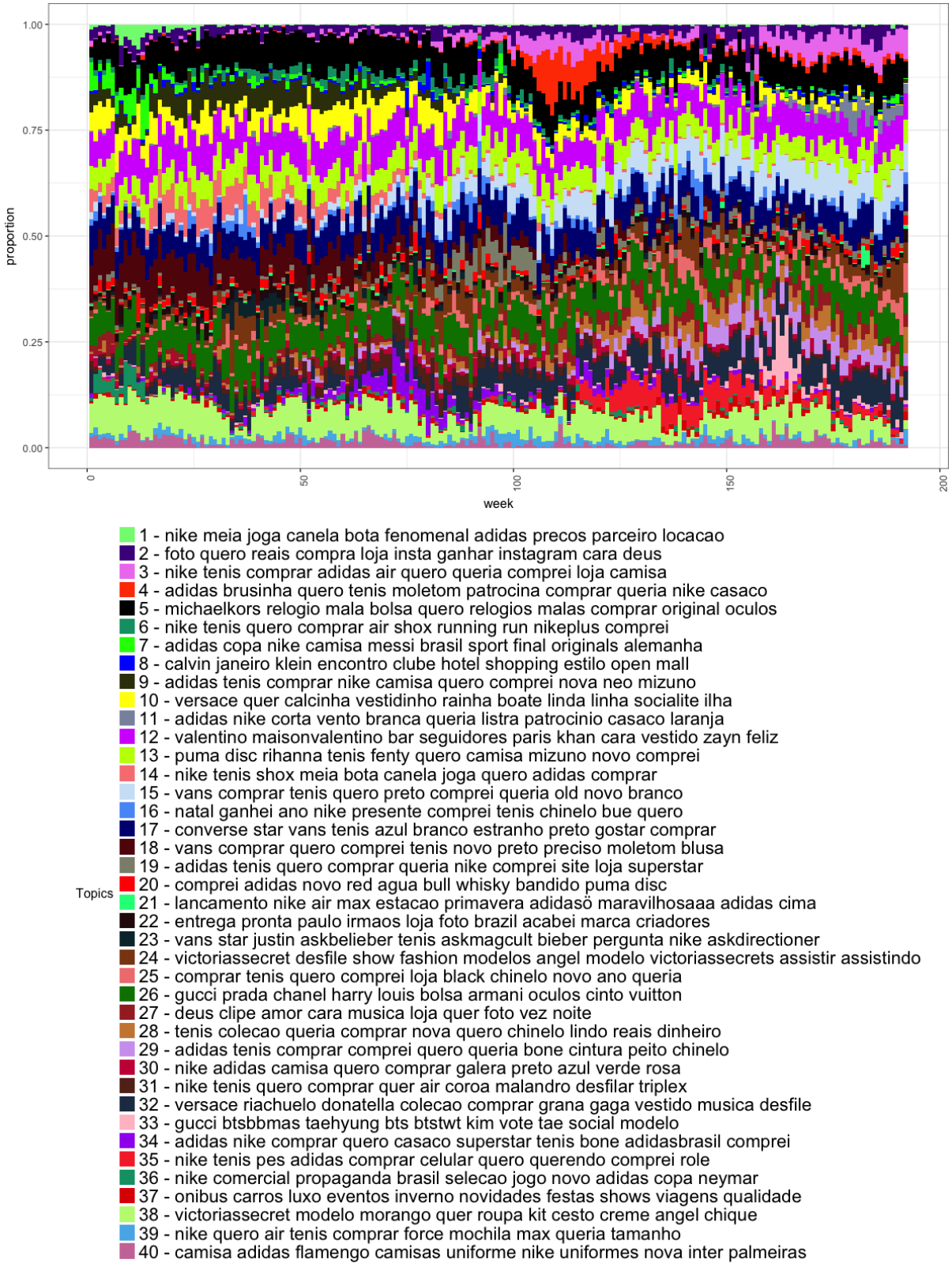



Figure 5.13: All brands 40 topics weekly evolution

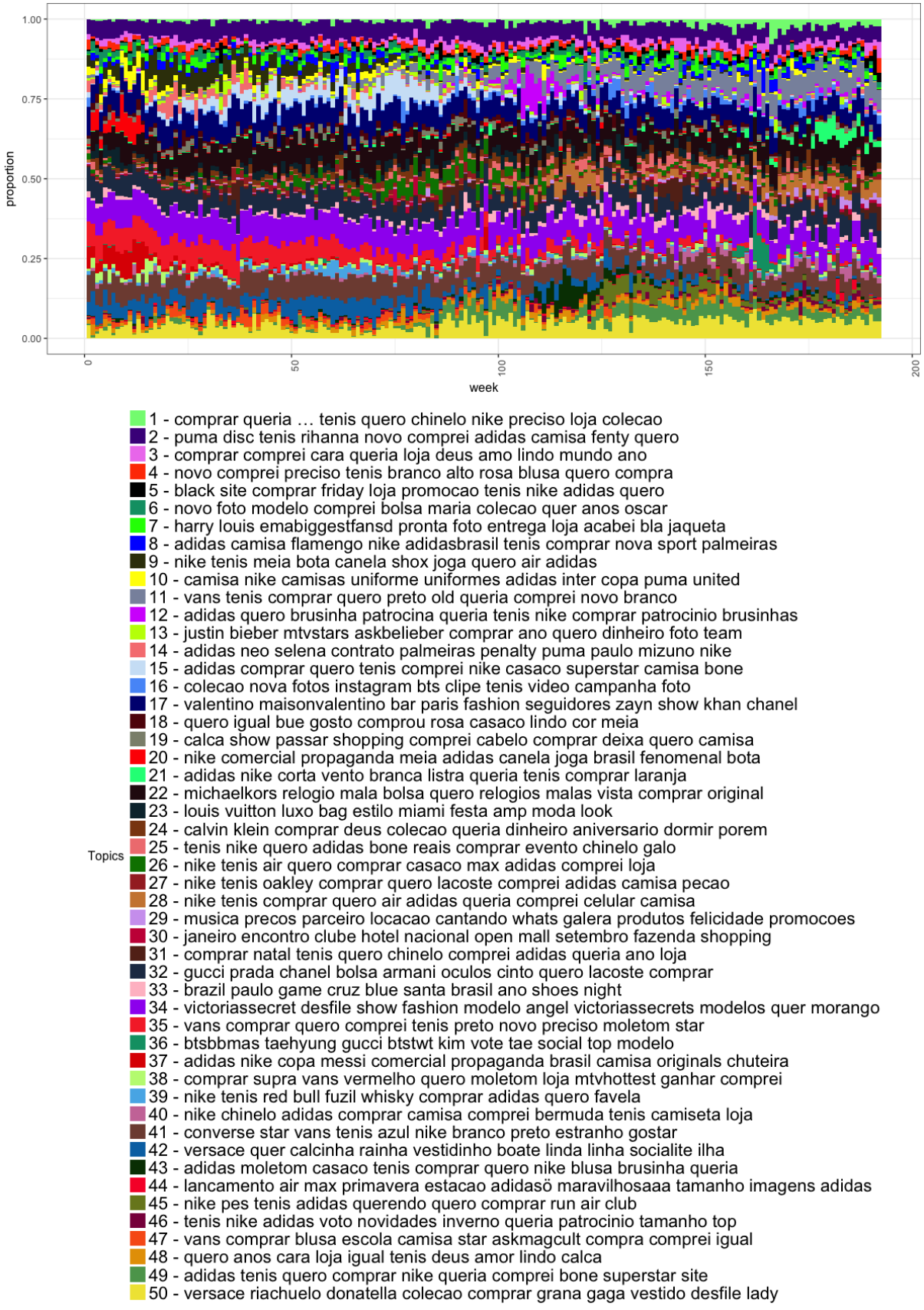



Figure 5.14: All brands 50 topics weekly evolution


Figure 5.15: Vans, Visctoria's Secret, Nike and Gucci topics

from brand to brand. For instance, while for Victoria's Secret we have "desfile" (Portuguese for fashion show), "show", "fashion", "modelos" (models), "angels", "corpo" (body), for Gucci we have "bolsa" (bag), "caro" (expensive), "cinto" (belt), "óculos" (glasses), "clipe" (video clip), "preço" (price), "dinheiro" (money), and so on. It is noticeable that the common aspect shared by Gucci lovers are the cost of its items, for example. Furthermore, identical brands to Gucci are also widely discussed, as we can easily see Versace, Prada, Armani, Louis Vuitton presented in Figure 5.15.

This means that sometimes the aspects discussed are very particular to the brand we chose to analyse. Nevertheless, these aspects might also be very similar if we spot light on similar brands. For instance, Nike, Adidas and Puma do share some terms considered the most relevant ones, as mentioned above. This might be so because they are sport brands, and sell similar items. This might also the reason they are mentioned in other topics of another sport brand, as referred before.

Nevertheless, this aspects might also be very similar if we compare similar brands, as depicted in Figure 5.16. For instance, Nike and Adidas do share several terms considered the most relevant ones. This might be so because they are sport brands, and sell similar items. This might also be the reason Nike is mentioned on Adidas topics and vice-versa very often, as referred before. The fact that both topics revolve around the World Championship, in which both brands were sponsors, might also play a role, in this particular case.

## 5.2 Country-based Analysis

The first topic (in gray) present in Figure 5.17, illustrates users interest ("gosto"/I like, "comprar"/to buy) in Vans and Nike sneakers ("ténis"). This topic is more evident around weeks 20 and 75. Topic 4 (in brown), which is about wanting/buying ("quero"/ "preciso"/ "comprar"/ "comprei") Vans "ténis" (sneakers), "moletom" (sweater), and "blusa" (tops), has more proportion until week 50. Topic 8 (in green) is about Nike "Shox" and "Air" sneakers,


Figure 5.16: Nike and Adidas World Cup topics

and other items such as “meia” (socks) and “bota” (boots). It also mentions “joga” (it/he/she plays), which might indicate that the World Championship was also discussed, which in turn might explain the decreasing in the topic proportion by the week 50. Topic 9 (in royal blue, in the center) depicts Brazilian users sharing about their interest in Adidas items such as sneakers, tops, and pullovers, jackets. It also mention Nike. Topic 11 (in bright t green), can be spotted in the beginning of the chart, and refers to the World Cup along with “chuteira” (football boots), “propaganda” (comercial), Adidas, Nike and Messi (the football player). Similarly, topic 23 (in mauve), by the beginning of the plot, is about the World Cup as well. This topic is common to the three sport brands analysed in this study (Adidas, Nike, and Puma), as shown in Figure 5.3, Figure5.5, and Figure 5.4. Topic 16 (in dark brown) revolves around Rihanna Fenty and Puma collaboration, which started back in 2015.

By the middle of the chart, at the bottom, we can also spot Topic 34 (in green), which is about some Nike and Adidas items such as “tenis” (sneakers) and “boné” (bonnet). Topic 38 (in indigo blue) is a very discreet topic that presents a rise in its proportion a little before each mark off of the chart. As it contains the words “natal” (Christmas), “present” (gift), “ganhei” (I got), “ano” (year), “amigo” (friend), “comprei”(I bought), it can be concluded that this topic has a seasonal trend behaviour. Topic 40 (in light green, in the bottom) depicted very well that Vans “Old Skool” sneakers became very popular back in 2016, and remained so during 2017.In fact, the topic even has some granularity about the most common set of colors for these sneakers, which are black (“preto”) and white (“branco”).

As mentioned in Lopes-Teixeira et al. (2018b), users from Brazil present a sharing behaviour very different from users from Portugal and from the United States. This difference can clearly be observed comparing the graphics presented in Figure 4.3. In fact, it is hard to spot trends in Figure 5.19, and even more difficult to spot trends in Figure 5.21, as the number of

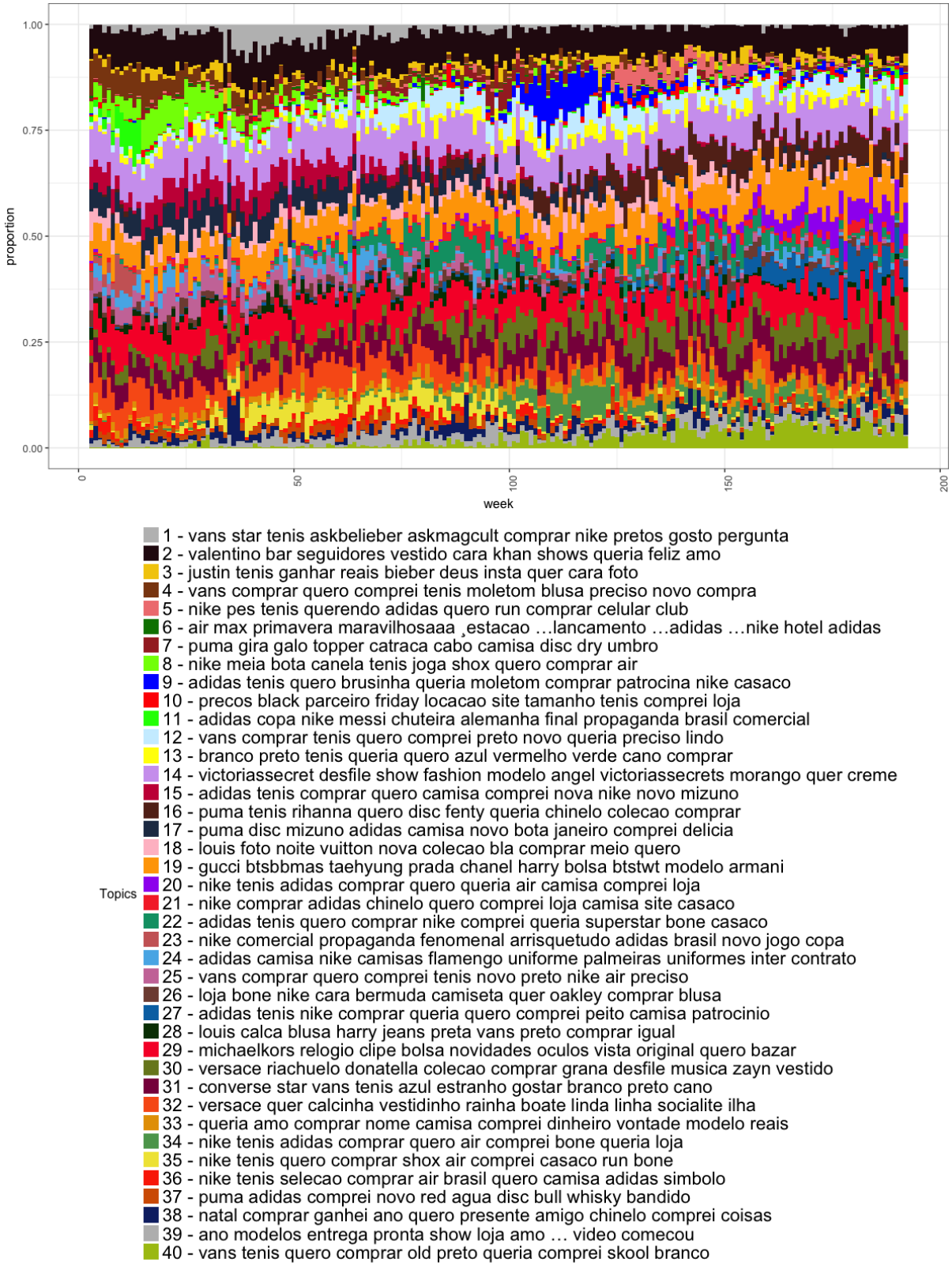



Figure 5.17: Brazil topics weekly evolution

posts is considerably lower for these two countries. For Portugal and the United States, there are weeks with no tweets in Portuguese mentioning the brands analysed in this study, therefore the charts present gaps. To overcome this difficulty, which resulted in gaps in the charts when performing a weekly analysis, the tweets had to be grouped by month rather than by week, so that we could perform a clear trend analysis. It worked for users from Portugal, but for users from United States the chart still presents a few gaps. Nonetheless, the trend analysis could be performed.

The first topic in Figure 5.18 is related to the *haute couture* such as brands Gucci, Prada, Armani, Chanel, and Versace. Topic 7 (in bright green) is a topic discussed frequently, as can be observed as it was spoken about in every document. The usage of the topic fluctuates. A Vans "Old Skool" sneakers related topic (in mint green, by the bottom) can be spotted more prominently roughly between the 30th and the 35th months, in line with the trend also identified in Figure 5.17. At bottom of the chart (in pink), we can spot a Victoria's Secret related topic, which refers to "morango" (strawberry) and "baunilha" (Vanilla) - most likely related to the body lotion product -, "kit", "angel", "modelo" (model), and "roupa" (clothes). This topic is discussed almost every month. The Victoria's Secret Fashion Show topic (in light blue, by the middle) also presents a seasonal behaviour, i.e., it appears and disappears several times throughout the whole chart. Topic 17 (in indigo blue) depicts the time Rihanna and Puma's collaboration started, back in 2015. The frequency of the topic increased roughly around that time.

Topic 17 (in indigo blue) depicts the time Rihanna and Puma collaboration started, back in 2015. The topic frequency increased roughly around that time. Michael Kors items such as "relógio" (watch) and "mala" (bag), along with the brands (Hugo) "Boss" and "Lacoste", are present in Topic 20 (in red), which is discussed almost in every document. Comments shared about Paris Fashion Week, depicted in the Topic 19 (in gray), also appear and disappear several times. This trend is likely related to the calendar of the event itself, which occurs several times a year. The last topic, which contains terms such as "Vestidinho" (short dress), "socialite", "boate" (nightclub), "rainha" (queen), and "linda" (beautiful), is related to (the lyrics of) a Brazilian song, rather than the brand itself.

Because the documents are result of pooling tweets together based on month, the topics presented in Figure 5.18 tend to be less specific than the ones depicted in Figure 5.17, for example. The number of topics could be increased, but as many topics were very similar, we opted to fix the number of topics in 26, even though there are still some similar topics left. This also helped to achieve a better trend analysis.

Figure 5.20 shows that Michael Kors-related topic, which is the third one (in fuchsia), is the kind of topic discussed very often, as it is present in almost every document. In the other

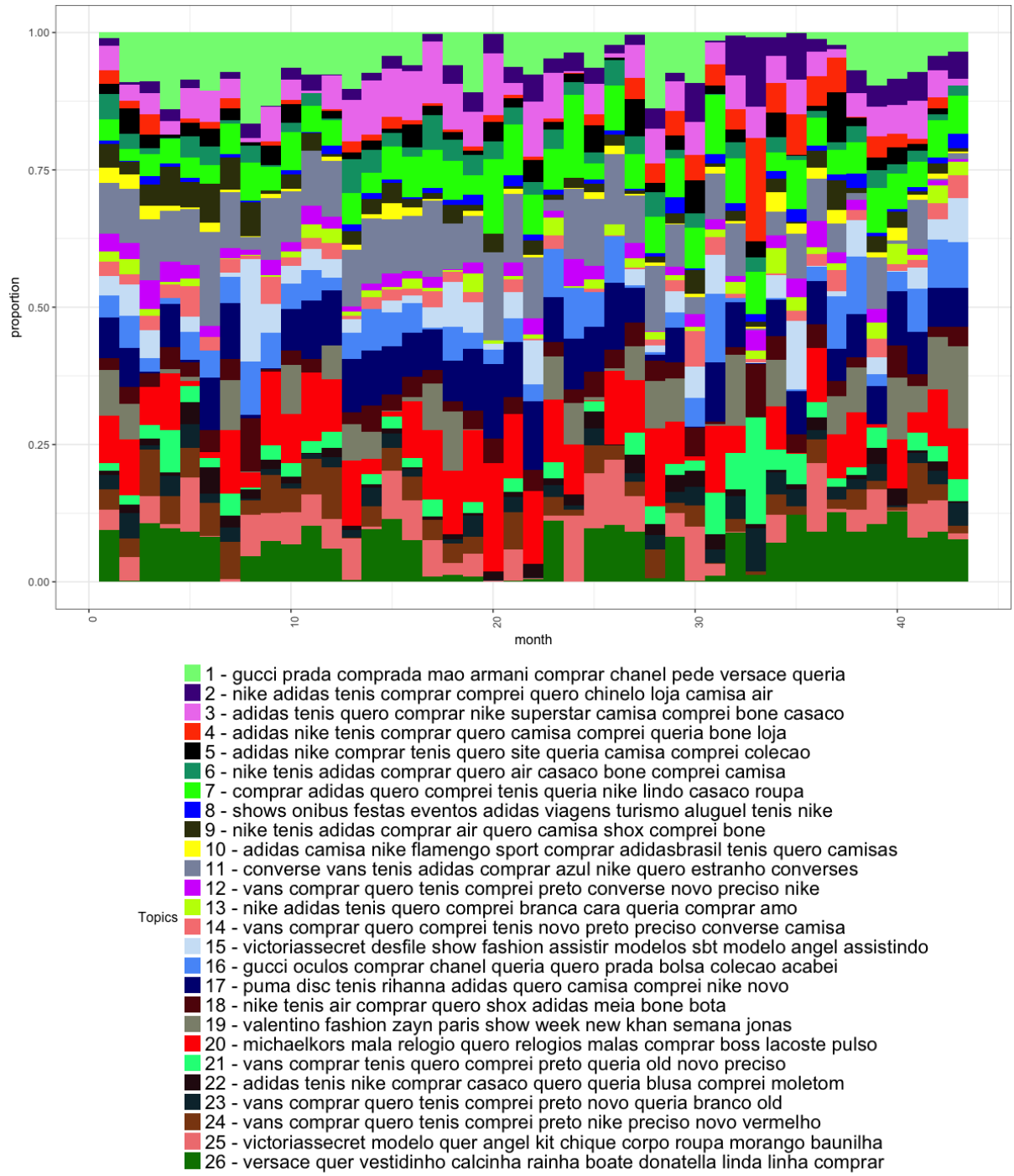



Figure 5.18: Portugal topics monthly evolution

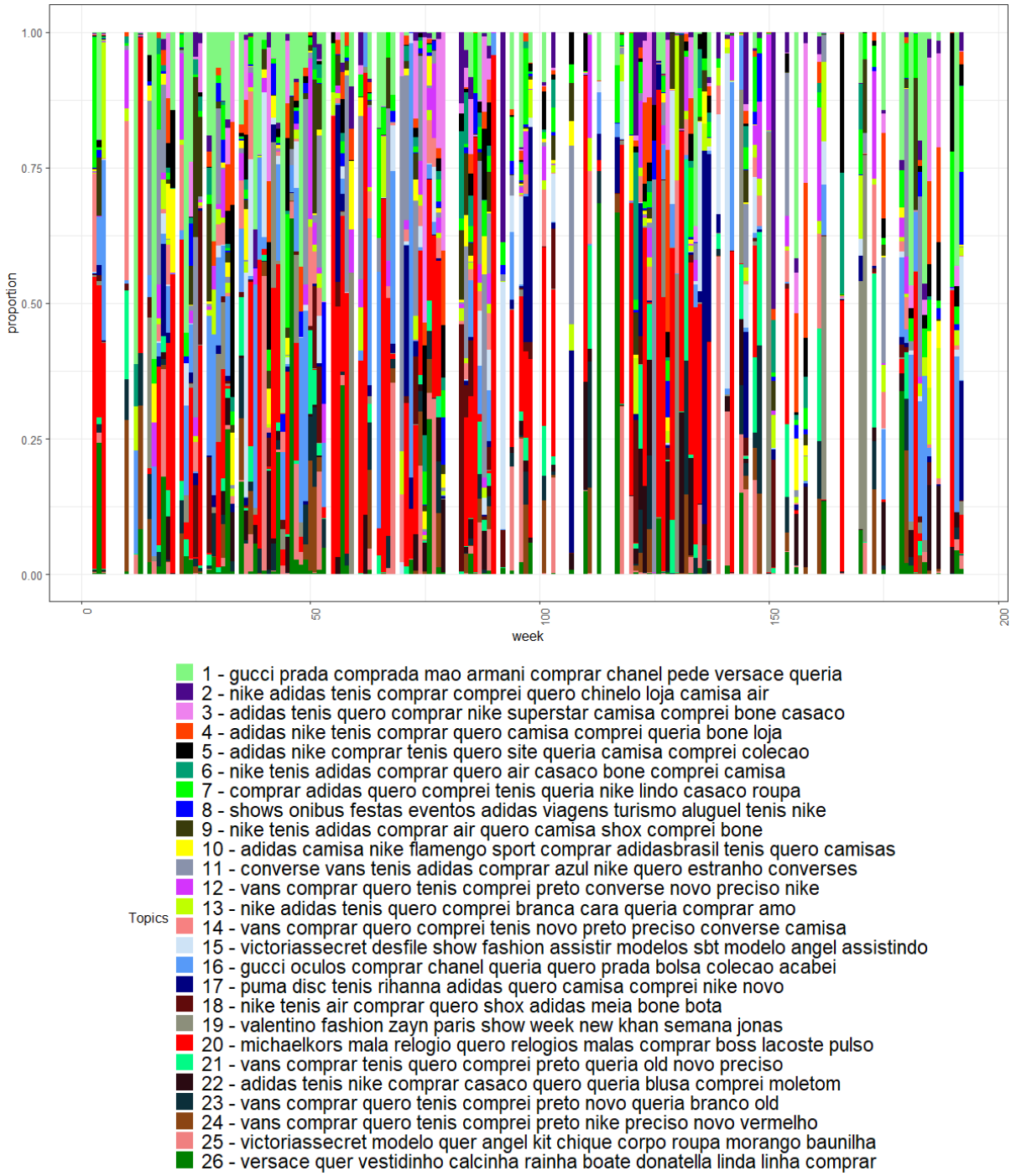



Figure 5.19: Portugal topics weekly evolution

hand, we have Topic 4 (in orange) being discussed only a few times, with more prominence around month 17. Topic 5 (in black), which contains terms such as “Vestidinho” (short dress), “socialite”, “boate” (nightclub), “rainha” (queen), and “linda” (beautiful), is related to (the lyrics of) a Brazilian song, as mentioned before. Topic 12 (in purple) revolves around *haute couture* brands like Gucci, Prada, Dior, and Chanel. The Victoria’s Secret Fashion Show topic, which is the sixteenth (in blue), also presents a seasonal behaviour, i.e., it appears and disappears several times throughout the whole chart. This topic is discussed/shared by users from the three countries. Another Victoria’s Secret related topic (in royal blue), which refers to “morango” (strawberry) and “baunilha” (Vanilla) - most likely related to the body lotion product -, “kit”, “angel”, “modelo” (model), and “roupa” (clothes), can be spotted several times. Topic 15 (in light blue) is a mixture of subjects: it mentions “Maison Valentino” (the *haute couture* brand) and “vestido” (dress), but it also mentions “Valentino” and “ganhar” (to win), giving a hint that this topic might also be related to Moto GP championship. The last topic (in red, by the bottom), which is a sneaker launch related topic, can be sporadically spotted in the chart.

Even though the documents were produced by grouping tweets monthly, the chart still presents three gaps, which means that no tweets written in Portuguese mentioning one of the chosen brands were uploaded for these months. The number of topics was fixed in 20 for the same reasons indicated previously.

These topics were visualized using LDAvis package, which shows how similar or distinct the topics are (intertopic distance), and the relative sizes of the topics. Figure 5.22 shows the output of the topic model using this visualization, and it shows the words most strongly associated with each topic. The occurrence of specific words within each topic can also be visualized by selecting a word instead of a topic. Analyzing Figure 5.22we can observe that from Topic 1 to Topic 8 topics are somewhat overlapped or very close to each other. The topics started to differentiate from one another from Topic 9, even though there are still some topics very close to each other, taking for instance topics 10 and 16 or and topics 7, 12, and 15. We can also observe that there are topics next to each other but with no overlapping occurring, for example Topics 6 and 13, or Topics 18 and 24. For this study, we experimented with $\lambda$ set to 0.34 as in the study of Sievert and Shirley (2014), but ended up setting $\lambda = 0.32$ instead, for the sake of the terms considered relevant.

Topic 1 is the largest topic in the data from users from Portugal, and is comprised of people talking about “Converse”, “Old (Skool)” - by Vans, “camisa”/ “camiseta” (shirt), “cinto” (belt), “malas” (bags), “oferecer” (giving), “quero” (I want), “campera” and “Chiado” (which are Portuguese Malls), “flux”, “sneakers”, “modelo” (model), and so on. The closest topics to Topic 1 are Topic 2, which is people sharing about “Vans”, “mala” (bag), “blusas” (top/blouse), “Adidas”, “moletom” (Brazillian Portuguese for pullover), and so forth; And Topic 5, which

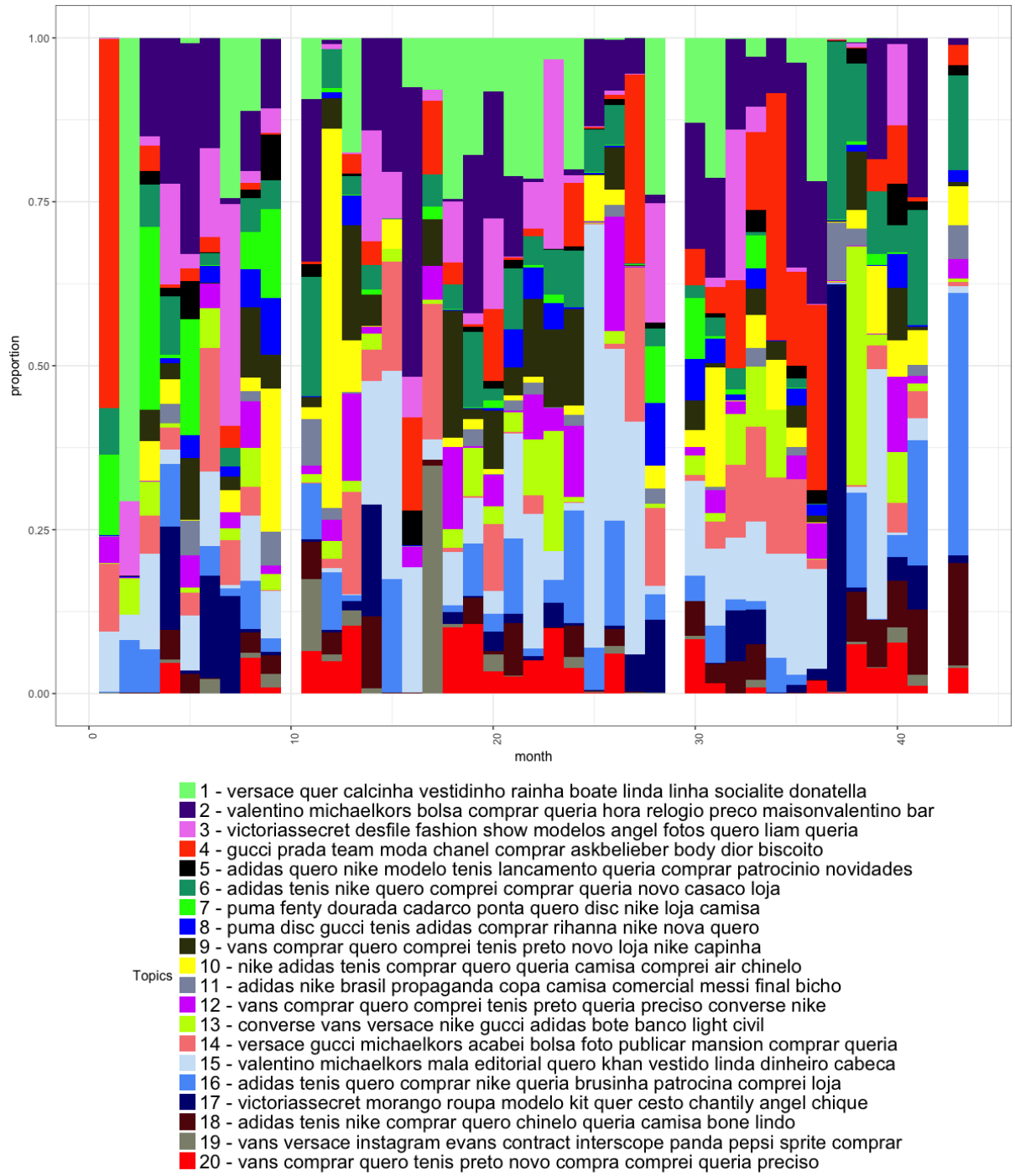



Figure 5.20: The United States topics monthly evolution

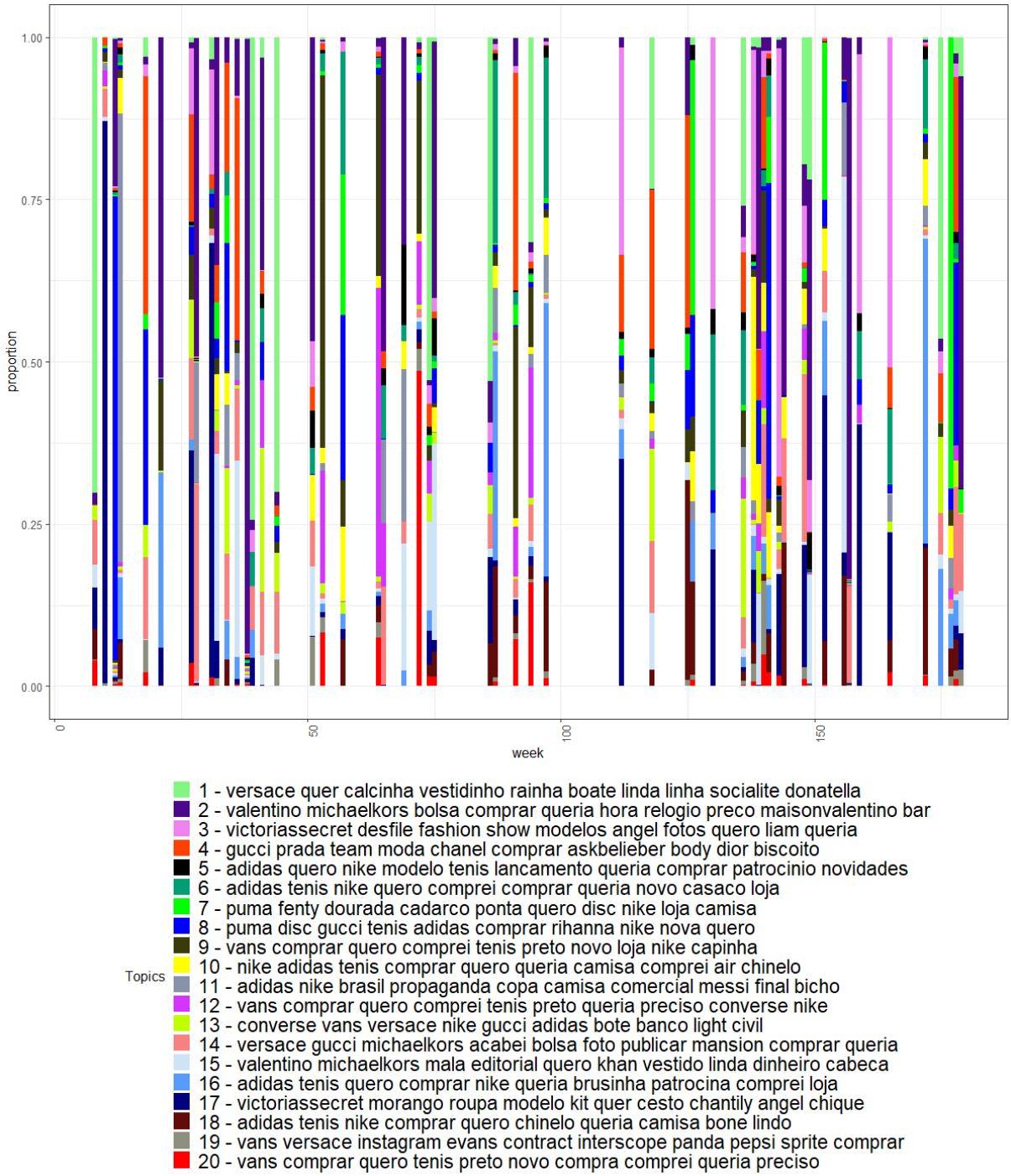



Figure 5.21: The United States topics weekly evolution

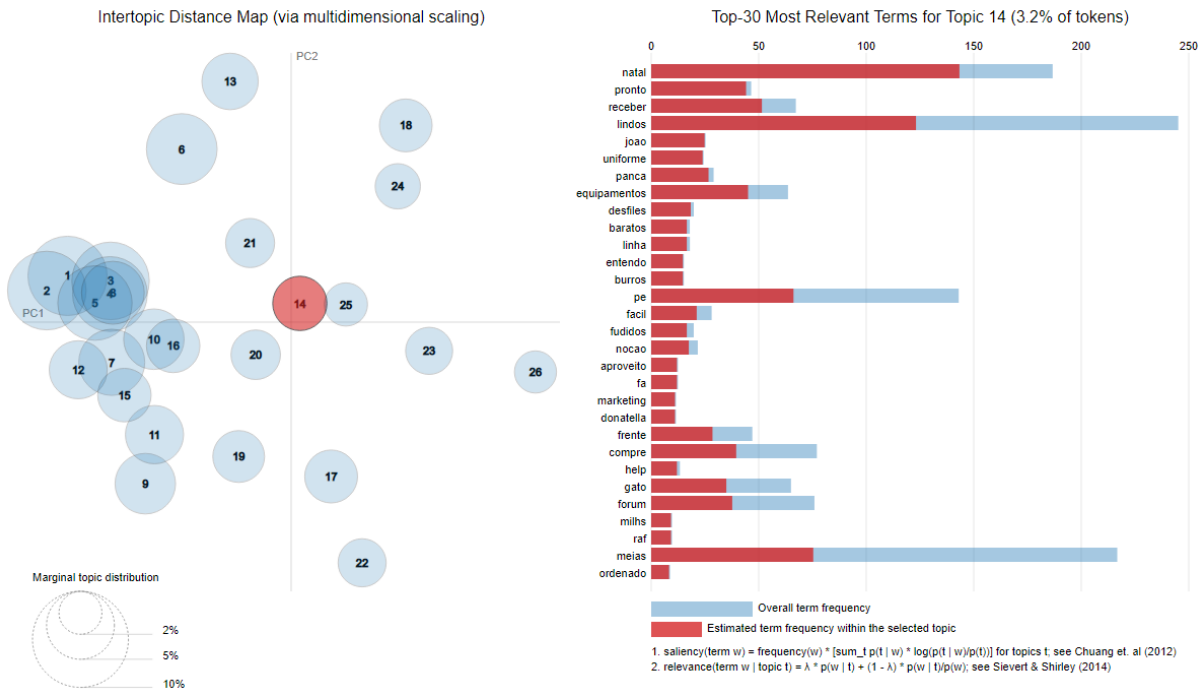



Figure 5.22: Topics visualization for Portugal

is comprised of posts about clothes (trousers/pants, jackets, socks), (Old) “Skool”, “modelos” (models) and so on. Words expressing sentiments are also part of these topics, such as “happy” in Topic 5, “love” in Topic 2, and “incrível” (unbelievable) in Topic 1. The term “quero” (Portuguese for “I want”) is one of the most relevant words for Topic 1, suggesting that this topic could also be interpreted as people sharing about wanting to acquire certain items.

The term “Colombo” (a well-known Portuguese mall), present in this topic, is one of the most relevant words for Topic 7, followed terms such as “sapatos” (shoes), “skate”, “Jordans”, “run”, “balance”, “Ericeira” (a Portuguese sport items store), indicating that this topic could be somehow sport-related and/or related to sport items.

Terms such as “Colombo” (from Topic 7), “Chiado” (from Topic 1), and “Primark” (Topic 25), which can be seen in Figure 5.25 are only present in (the top 30 terms of) topics discussed by users from Portugal. This could somehow suggest the country where the topics were being discussed, if the country were unknown.

Observing 5.26 we can see that the first 10 topics are much less overlapped than they are in Figure 5.22. We can also observe that there are many topics close to each other but with no overlap occurring, for instance Topics 12, 13, and 17, or Topics 8 and 10. On the other hand, we have Topic 7 overlapping Topics 3 and 6, and Topic 14 overlapping Topics 18 and 19. The fact that the number of topics are fixed in 20 rather than 28 might have influenced in less overlapping

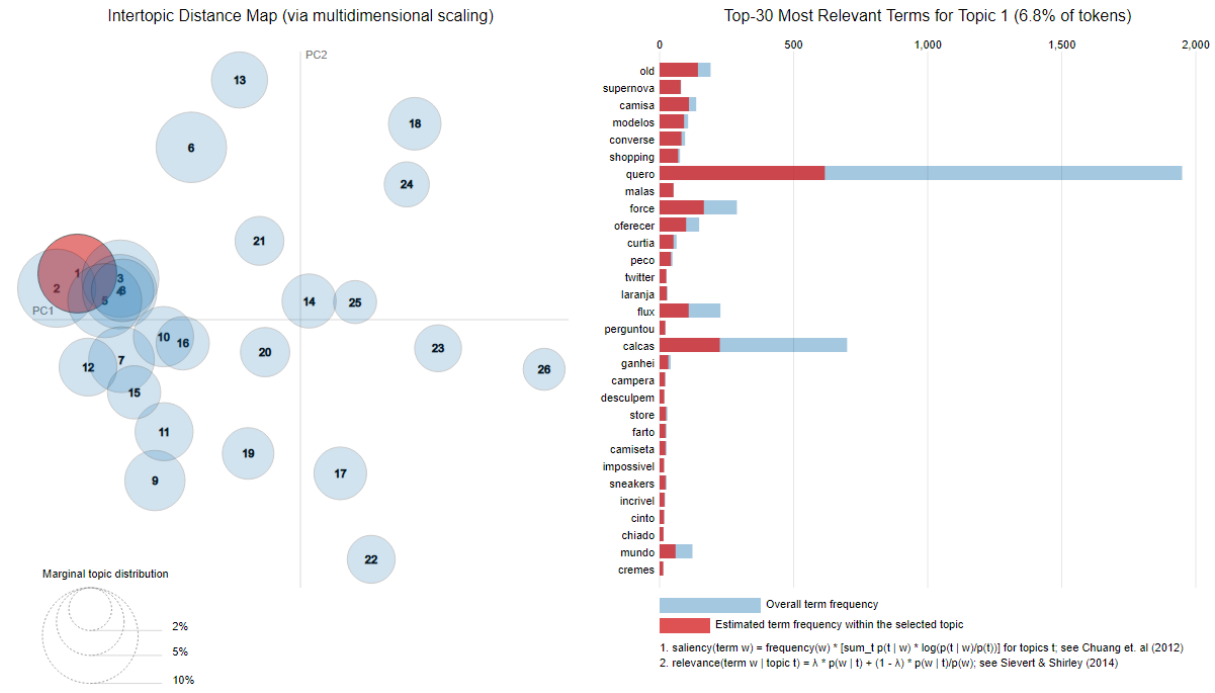



Figure 5.23: Topics visualization for Portugal

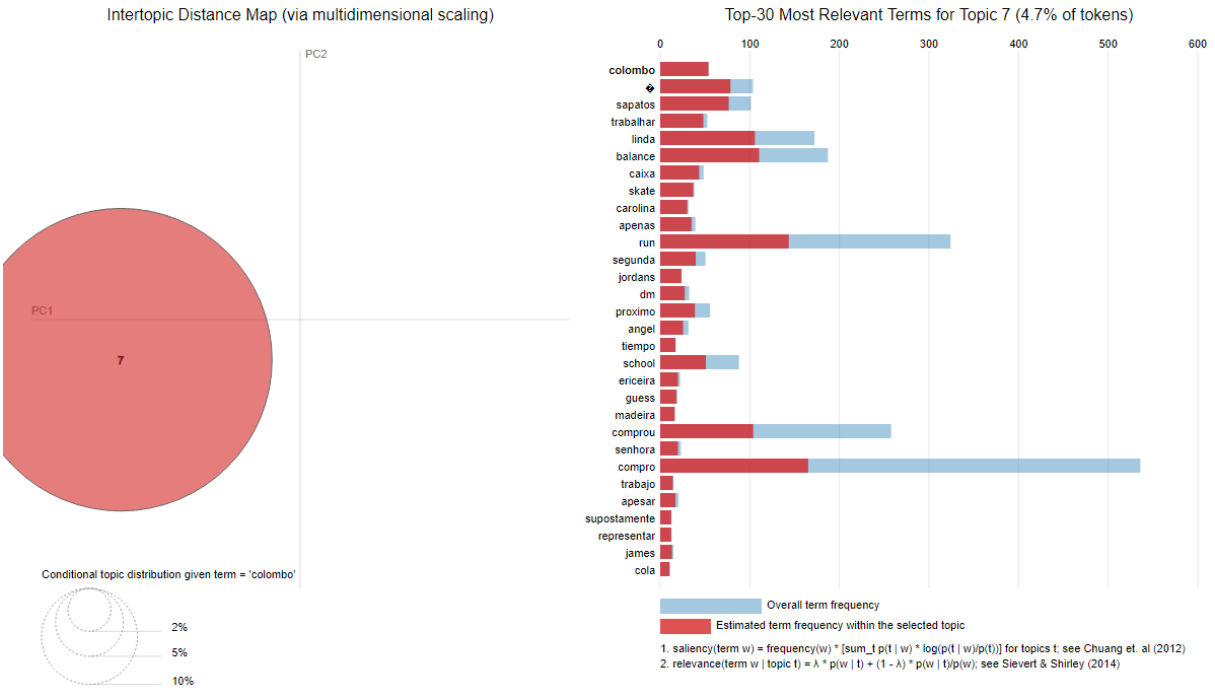



Figure 5.24: Topic 7 visualization

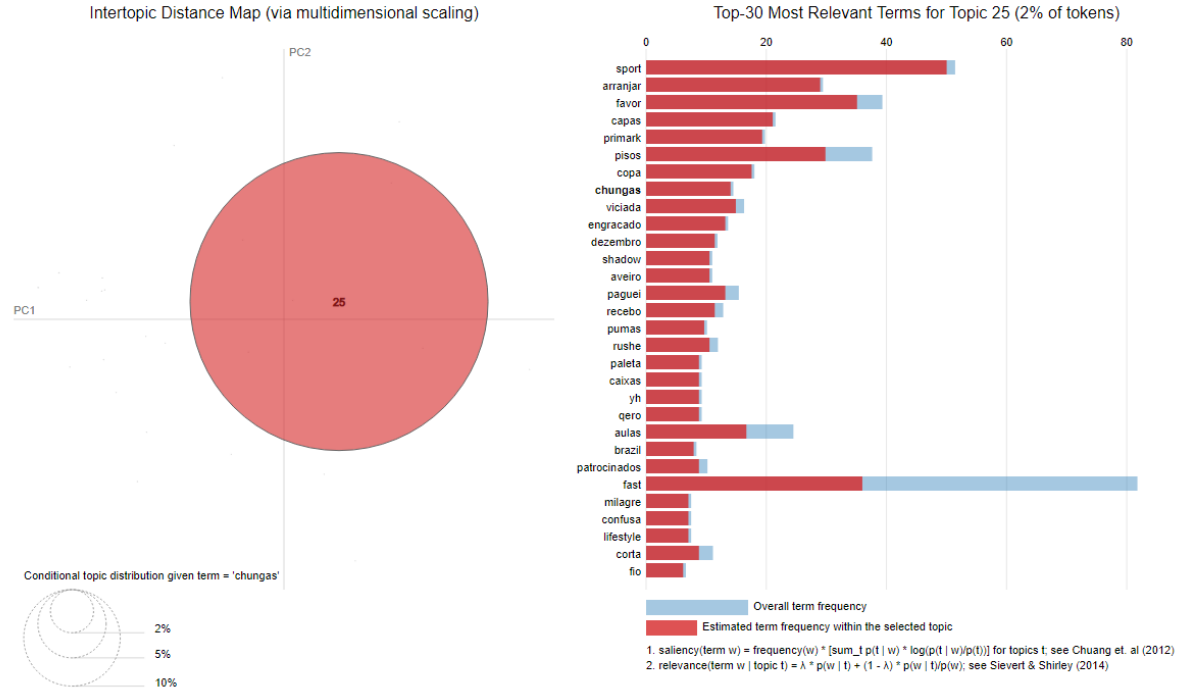



Figure 5.25: Topic 25 visualization

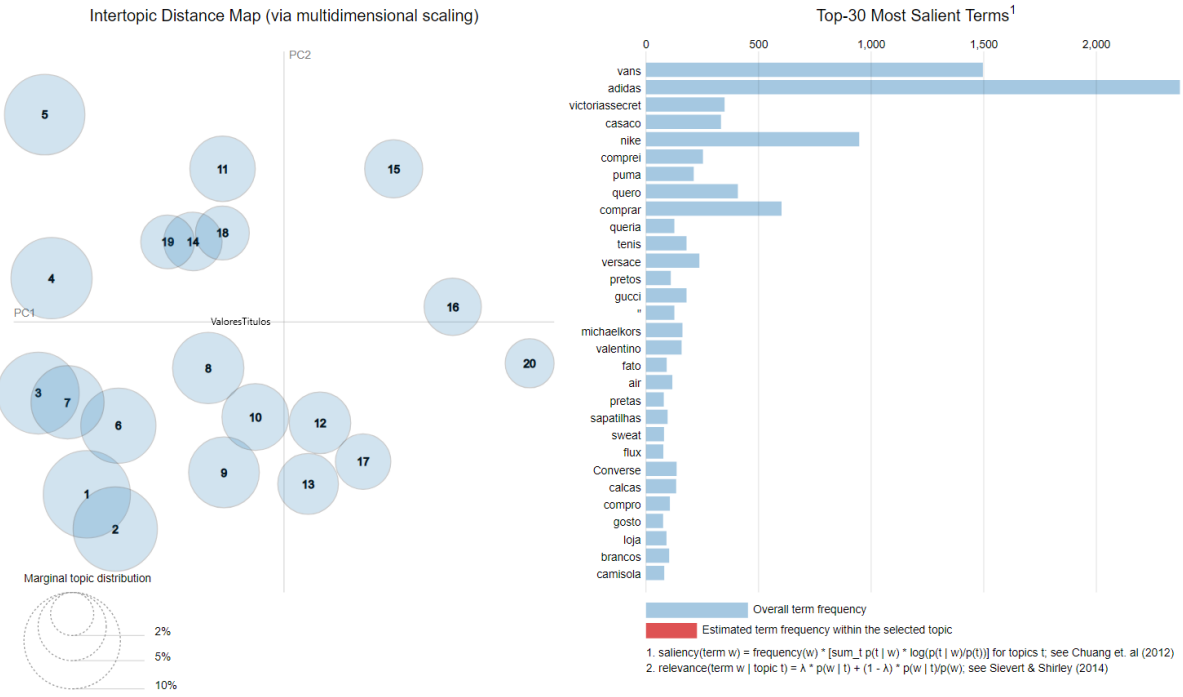



Figure 5.26: US Topics visualization

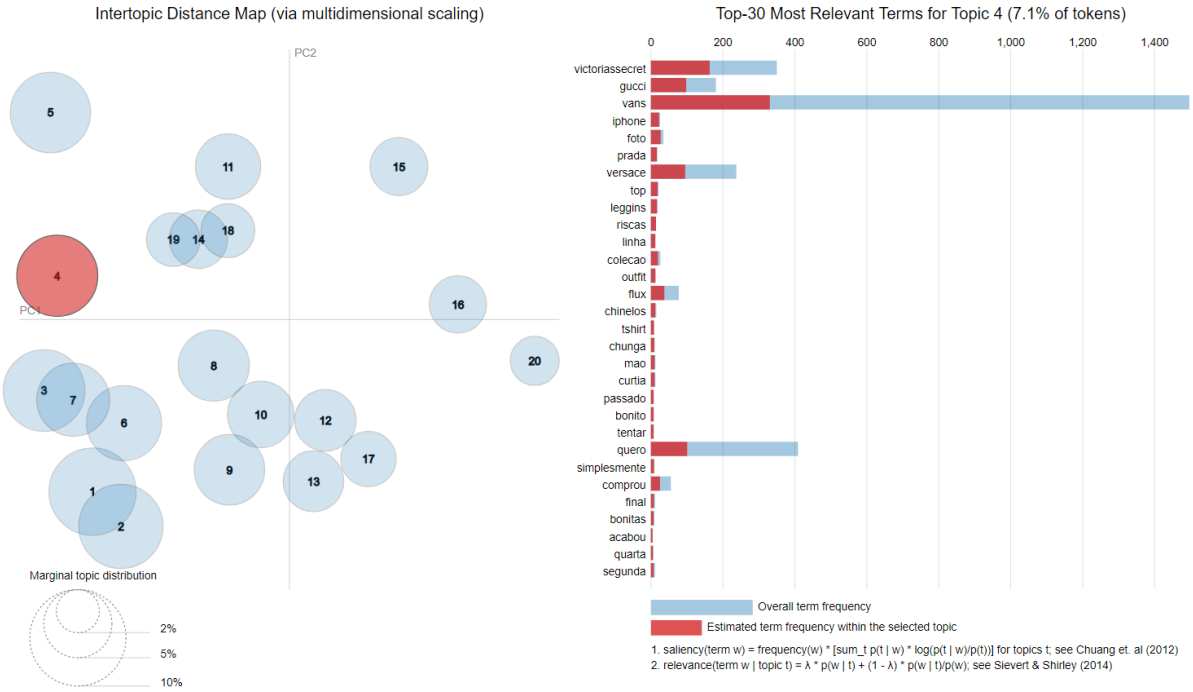



Figure 5.27: Topic 4 visualization

occurring.

Because the data was aggregated monthly rather than weekly, topics for the United States users are more diverse. Taking for instance Topic 4 presented in Figure 5.27 we can observe that brands such as Victoria's Secret, Vans, Versace, Gucci, and Prada, were discussed in any given month. A diversity of items were also the subject of discussion, such as "chinelos" (slippers), leggings, top, t-shirt, flux (Adidas sneakers), and "coleção" (collection). Iphone (Apple smartphone) and "foto" (photo) are in the Top 5 most relevant words for Topic 4, suggesting that this topic could also be interpreted as people sharing photos of their items, and that most likely these photos were taken using iphones.

Comparing the topics in Figure 5.28 we can observe that some terms have more or less relevance depending on the country. For instance, it is noticeable that the term "quero" (I want) is the most relevant term in all the three topics, regarding users from United States, while for users from Portugal, the term "Comprar" (to buy) comes first and the term "quero" (I want) comes in second. For users from Brazil, "comprar", "quero", and "tenis" (sneakers) comes at the same position of the ranking in the first topic, "comprei" (I bought) is the most relevant term in the second topic, while for the third topic "comprar" is the most relevant word.

Moreover, we can notice that the items discussed also vary from country to country. Taking for example the word "moletom" (Brazilian Portuguese for pullover/sweater), it shows up with a considerable relevance in the first topic, as well as in the second topic, although with less

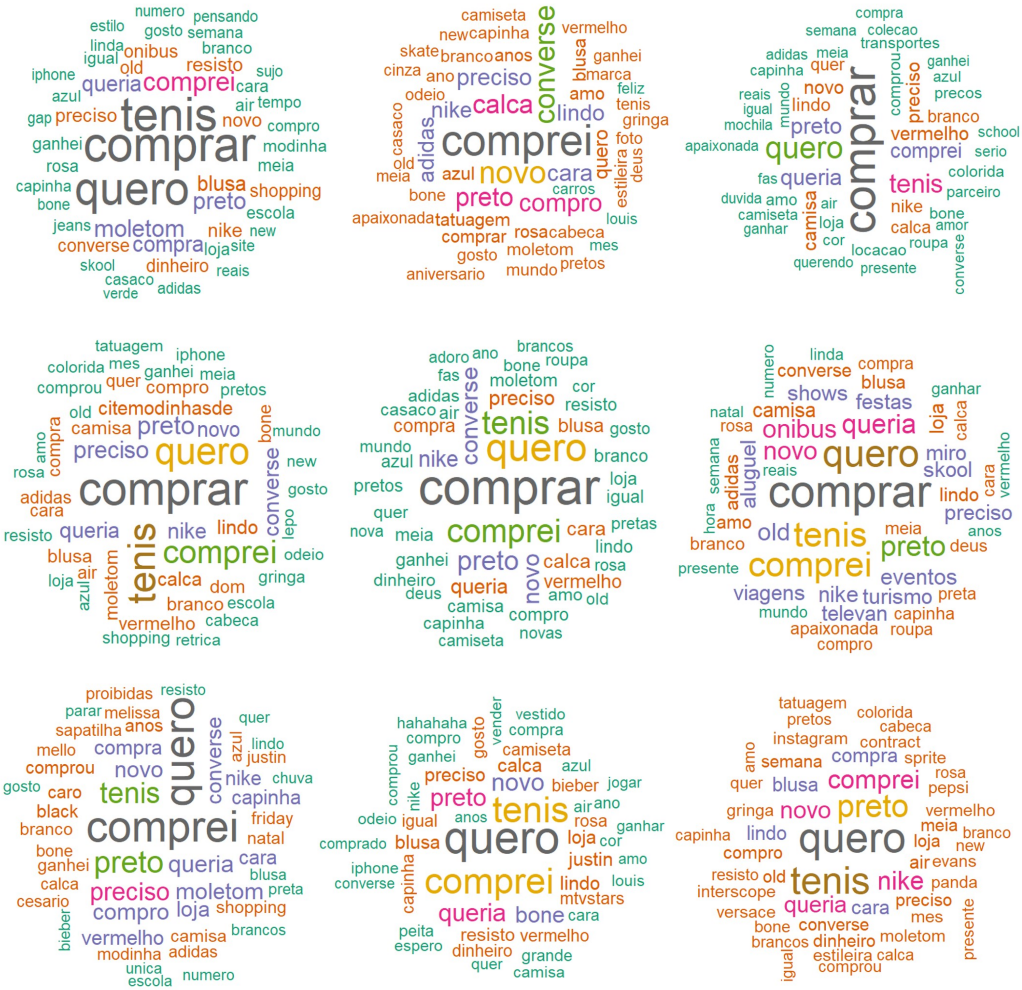



Figure 5.28: Vans Topics from Brazil (top), Portugal (middle), and US (bottom).

significance then in the previous topic. Users from Brazil seemed to show their feelings towards brands/brand items a little more vigorously than users from the United States . For instance, we can spot the word “amo” and “apaixonada” (Portuguese for “I love”, and “in love (with)”) in two out of the three topics presented, and the term “amor” (love) in the last one. The term “amo” also present in two of the three topics, but with less importance.

Another observation is that the word “comprei” (I bought) has more relevance in the topics discussed by users from the United States, and the term “comprar” (to buy) can barely be spotted. This may suggest that users from the United States share more about what they have done than what they want to do. On the other hand, we have the term “comprar” (to buy) as the most relevant term in the topics discussed by users from Portugal, followed by the term “quero” (I want). This may indicate that users from Portugal share a lot about their intentions towards something and a little less about what they have done, as the term “comprei” (I bought), also among the most relevant terms from users from Portugal, have less relevance than the two

terms mentioned before. Figure 5.28 depicts very well what Figure 5.11 portrays, which is the fact that Vans topics are very similar to each other, and their relevance is balanced from the beginning to the end of the chart.

## 5.3 Summary

In this chapter, we demonstrated that aggregating tweets routinely to train the Topic Model, and then applying the model on tweets grouped using a wider time span can produce informative topics. What is more, we demonstrated that topics that people discuss/share opinion about change over time, and from one country to another, and that several real-life events caused brand interest to increase. Also, in the charts presented we can observe that each brand had different brand interest pattern. For instance, Victoria's Secret presented a seasonal trend behaviour, while other brands such as Adidas presented an unpredictable increasing in brand interest.

# 6 Conclusions and Future Work

This work proposed a framework for tracking brand interest, both geographically and temporally, using Twitter as data source. The experiments, also presented by Lopes-Teixeira et al. (2018b), demonstrated that the system is able to detect variations in brand interest, and that these variations are likely to be related to real-world events, which is in line with previous studies. Moreover, the system depicted differences in the way users from Brazil, Portugal and the United States shared their thoughts and feelings. The three countries present ups and downs on the number of posts in different periods of time, as can be observed in Figure 4.11. Nonetheless, they all share a trend: an increase on the posts around December, followed by a decrease that starts in January.

This study also demonstrates that aggregating tweets based on the day (they were uploaded), and by brand, in order to train the Topic Model and then applying the model on tweets grouped using a wider time unit can produce coherent and informative topics. We also demonstrated that topics that people discuss/share opinion about change over time, and from one country to another. Moreover, findings of this study demonstrated that several real-life events caused brand interest to increase, which is in line with the work presented in Lopes-Teixeira et al. (2018b) and in Lopes-Teixeira et al. (2018a), and several previous studies. For instance, commercials, product launches, and events can lead to emerging of new topics, which may result (or not) in older topics fading. Additionally, in the charts presented we can observe that each brand had different brand interest patterns, which was also stated in the study of Lopes-Teixeira et al. (2018b). For example, Victoria's Secret topic about their fashion show comes and goes several times, i.e., it presents a seasonal trend pattern.

These findings can help brands to understand how consumers' interest changes and what events caused those changes to happen, using Social Networks as a data source. Moreover, brands can have insights of what aspects/items people like or dislike and what kind of sentiments are associated to those aspects/items, enabling them to adjust their strategies accordingly.

The importance of pre-processing tasks in Natural Language Processing was emphasized in this work. The experiments demonstrated that pre-processing steps do have impact on the quality of the topics resulting from documents written in Portuguese. We obtained more informative

topics when the documents were previously processed. Tasks such removing URL's, removing stop words and choosing the representation vocabulary based on TF-IDF can mitigate common issues that reduce the coherence of the topics. Results demonstrated that this framework can be followed to obtain coherent Portuguese written topics, enabling one to get insights about people's conversation/discussions on Social Networks.

Trying to achieve our goals, we have faced some limitations:

- Data collection phase: although brand filters were applied during this phase, many tweets considered irrelevant for this study were still present in the dataset. Even though we tried to later remove these irrelevant tweets, some tweets not related to the brands remained in the dataset. It is also important to highlight that the data collected is biased because, i) we chose an online social media platform to know what people discussed (which automatically excludes offline discussions); ii) we chose Twitter as data source; iii) we selected the Top 3 countries with more tweets uploaded to perform our analysis. We mitigate these limitations by extending our time span to almost four years.

- Data preparation phase: the complete removal of irrelevant terms such as stop words, numbers, and conjunctions was not completely successful. This was due to a lack of a thorough list of Portuguese stop words. In order to mitigate this limitation, an extensive Portuguese lexicon containing verbs, adverbs, additional stop words (available stop words lists for Portuguese are still very short), conjunctions and so on was created in order to improve Topic Modelling results. Some social media slang and abbreviations, misspelled words, and other words such as "ahahahah", which is a word used to mimic laugh, could not be completely removed. To help reduce the occurrence of irrelevant words, we applied TF-IDF algorithm, so that misspelled words occurring less than three times (in a document) could be dropped. This step helped, but there were still terms considered irrelevant for this study that remained, most likely because they occurred more than two times. The stemming algorithm did not suit well for this work, as terms considered relevant were being incorrectly stemmed. In this manner, we were looking for a Lemmatization algorithm for Portuguese, but with no success. Moreover, we think that as both algorithms wipe out the inflectional forms of the terms, topics could be less interesting if we applied one of them. As such, we opted to maintain the original vocabulary.

- Although the "remove punctuation" function was applied, some punctuation marks can still be seen in the topics (as terms). The encoding also presented an issue. As result, some unknown characters can be seen when visualizing the topics using LDA-VIS package.

- Topic Modelling phase: because users from Portugal and United States posted considerably less then users from Brazil did, the aggregation criteria had to be adjusted ac-

cordingly, as mentioned before. While tweets from users from Brazil were grouped by week, tweets from Portugal and from United States were grouped by month. Even though tweets were aggregated together by month, the chart obtained using United States documents still came out with two gaps, meaning that there are two months for which we don't have United States Portuguese written tweets, from United States, about any of the brands selected. In order to set a number of topics that suit this work, several iterations were required, as for this study the HDP algorithm was not used to automatically infer the number of topics. This might be a path to follow in future work.

Future work includes performing experiments applying other Topic Modelling algorithms in order to evaluate which one would perform better for this type of analysis, on large datasets. Applying HDP algorithm to infer the number of topics prior to LDA application would also be an interest direction for future research. Building a front-end application capable of performing the analysis described hereby would also be a good path for future work. An objective metric, like a mathematical equation, to evaluate how good the topics produced using this framework could also be a good direction for future research. Furthermore, a connection between the proposed framework and a Decision Support System, could also be interesting.

Another path for future work could be including Sentiment Analysis in the framework, in order to uncover good or bad sentiments expressed during online discussions, and how these sentiments change over the time, and from brand to brand.

# Bibliography


Alghamdi, R. and Alfalqi, K. (2015). A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1).

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). Text summarization techniques: A brief survey. *arXiv preprint arXiv:1707.02268*.

Alvarez-Melis, D. and Saveski, M. (2016). Topic modeling in twitter: Aggregating tweets by conversations. *ICWSM*, 2016:519–522.

Amado, A., Cortez, P., Rita, P., and Moro, S. (2017). Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*.

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 492–499. IEEE Computer Society.

Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of marketing research*, 49(2):192–205.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Calheiros, A. C., Moro, S., and Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, 26(7):675–693.

Gupta, V., Lehal, G. S., et al. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1):60–76.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM.

Hu, Y., John, A., Wang, F., and Kambhampati, S. (2012). Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *AAAI*, volume 12, pages 59–65.

Kaveri, V. V. and Maheswari, V. (2017). A framework for recommending health-related topics based on topic modeling in conversational data (twitter). *Cluster Computing*, pages 1–6.

Lau, J. H., Collier, N., and Baldwin, T. (2012). On-line trend analysis with topic models:\# twitter trends detection topic model online. *Proceedings of COLING 2012*, pages 1519–1534.

Lopes-Teixeira, D., Batista, F., and Ribeiro, R. (2018a). Discovering trends in brand interest through topic models. In *KDIR'2018 - 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.

Lopes-Teixeira, D., Batista, F., and Ribeiro, R. (2018b). Spatio-temporal analysis of brand interest using social networks. In *CISTI'2018 - 10th Iberian Conference on Information Systems and Technologies*.

Machleit, K. A., Allen, C. T., and Madden, T. J. (1993). The mature brand and brand interest: An alternative consequence of ad-evoked affect. *The Journal of Marketing*, pages 72–82.

Machleit, K. A., Madden, T. J., and Allen, C. T. (1990). Measuring and modeling brand interest as an alternative ad effect with familiar brands. *ACR North American Advances*.

Mackey, T. K., Kalyanam, J., Katsuki, T., and Lanckriet, G. (2017). Twitter-based detection of illegal online sale of prescription opioid. *American journal of public health*, 107(12):1910–1915.

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892. ACM.

Moro, S., Cortez, P., and Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent dirichlet allocation. *Expert Systems with Applications*, 42(3):1314–1324.

Paul, M. J. and Dredze, M. (2014). Discovering health topics in social media using topic models. *PloS one*, 9(8):e103408.

Quan, X., Kit, C., Ge, Y., and Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *IJCAI*, pages 2270–2276.

Rui, H., Liu, Y., and Whinston, A. (2013). Whose and what chatter matters? the effect of tweets on movie sales. *Decision Support Systems*, 55(4):863–870.

Shi, Z., Lee, G. M., and Whinston, A. B. (2016). Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quarterly*, 40(4).

Sievert, C. and Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Srividhya, V. and Anitha, R. (2010). Evaluating preprocessing techniques in text categorization. *International journal of computer science and application*, 47(11):49–51.

Steinskog, A., Therkelsen, J., and Gambäck, B. (2017). Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 77–86.

Sung, B., Vanman, E. J., Hartley, N., and Phau, I. (2016). The emotion of interest and its relevance to consumer psychology and behaviour. *Australasian Marketing Journal (AMJ)*, 24(4):337–343.

Tan, A.-H. et al. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, pages 65–70. sn.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, 10(1):178–185.

Vijayarani, S., Ilamathi, M. J., and Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16.

Wang, S.-H., Ding, Y., Zhao, W., Huang, Y.-H., Perkins, R., Zou, W., and Chen, J. J. (2016). Text mining for identifying topics in the literatures about adolescent substance use and depression. *BMC public health*, 16(1):279.

Yau, C.-K., Porter, A., Newman, N., and Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3):767–786.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*, pages 338–349. Springer.