# Discovering Trends in Brand Interest through Topic Models

Diana Lopes-Teixeira[1], Fernando Batista[1,2] and Ricardo Ribeiro[1,2]
[1] *Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal*
[2] *L2F, INESC-ID Lisboa, Lisboa, Portugal*

Keywords: Topic Modeling, Topics Evolution, LDA, Preprocessing, Brand Interest.

Abstract: Topic Modeling is a well-known unsupervised learning technique used when dealing with text data. It is used to discover latent patterns, called topics, in a collection of documents (corpus). This technique provides a convenient way to retrieve information from unclassified and unstructured text. Topic Modeling tasks have been performed for tracking events/topics/trends in different domains such as academic, public health, marketing, news, and so on. In this paper, we propose a framework for extracting topics from a large dataset of short messages, for brand interest tracking purposes. The framework consists training LDA topic models for each brand using time intervals, and then applying the model on aggregated documents. Additionally, we present a set of preprocessing tasks that helped to improve the topic models and the corresponding outputs. The experiments demonstrate that topic modeling can successfully track people's discussions on Social Networks even in massive datasets, and capture those topics spiked by real-life events.

## 1 INTRODUCTION

The rapid growth of Internet has led to the growth of social media websites like Twitter, a micro-blogging platform launched in 2006. In social media websites people share diverse aspects of their life and talk about events happening that they are aware of. Thus, these websites produce tremendous amounts of data that can be used in many ways. For instance, it can be used to track emerging events, to discover trending topics, or to evaluate consumers' satisfaction toward a product in the market. Topic Modeling is amongst the Text Mining techniques applied to exploit Twitter data. However, performing Topic Modeling tasks in short messages, such as those available on Twitter, differs from performing them in longer documents, such as academic abstracts or newspaper articles. This is mainly because Topic Models infer topics based on the co-ocurrence of words in documents. Short messages limit this ability. Aggregating Twitter posts generated richer documents from which we can learn better topic models.

In this work, we focus on brand interest on Twitter. Our goal is to understand what people say about brands, and how that changes over time, and to point tendencies on those changes. In order to overcome the document length disadvantage, we present a pooling technique that consists in grouping together tweets by day and by brand to create longer documents that are going to be used to train our Topic Model. We also show that performing specific preprocessing steps has impact on the quality of the output of a Latent Dirichlet Allocation (LDA) (Blei et al., 2003) Topic Model, applied on the Twitter posts aggregations, written in Portuguese.

The remainder of this paper is organized as follows: Section 2 describes the related work; Sections 3 describes the dataset; Section 4 describes the Preprocessing and Topic Model Training; Section 5 presents the analysis and discussion of results, and Section 6 draws the major conclusions, presents the limitations of the current work, and proposes a set of tasks to perform as future work.

## 2 RELATED WORK

LDA, an unsupervised probabilistic model, models documents as distributions over topics, with topics being represented as distributions over words (Blei et al., 2003). This model has been applied on long documents such as academic abstracts (Moro et al., 2015), mid-sized documents such as customers' reviews (Calheiros et al., 2017), and short documents such as microblogging posts (Paul and Dredze, 2014). For the last one mentioned, some aggregation methods to reduce the length and sparseness disadvan-

Lopes-Teixeira, D., Batista, F. and Ribeiro, R.
Discovering Trends in Brand Interest through Topic Models.
DOI: 10.5220/0006936202450252
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 245-252
ISBN: 978-989-758-330-8

tages have been applied, resulting in longer pseudo-documents (Hong and Davison, 2010; Mehrotra et al., 2013).

LDA was used in (Hong and Davison, 2010) to evaluate the differences between topics learned from messages from the same user aggregated into a single profile scheme and topics learned by the aggregation of the user profiles, which in turn resulted from the aggregation of messages from the same user. Their results show that both approaches generated topics substantially different, meaning that topics learned using different strategies of data aggregation differ from each other. They also demonstrated the length of the documents influences the effectiveness of trained topic models, namely, a better model can be trained by aggregating short messages.

Another application of LDA was conducted in (Alvarez-Melis and Saveski, 2016), in which tweets belonging to the same conversation were grouped, with each group of related tweets corresponding to a single document. They evaluated whether the proposed technique outperforms alternative schemes. The resulting topics performed better than those derived by hashtag-based pooling.

In (Hu et al., 2012), the researchers modeled the topics of specific events as well as their associated tweets, while performing event segmentation, with an 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.

The researchers in (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 starts 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 modeling on Twitter, without having to modify LDA machinery.

Twitter posts presents some challenges due to sparseness, as 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. 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, several pooling schemes to group together tweets into longer 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 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.

## 3 DATASET

This study uses a dataset previously used in (Lopes-Teixeira et al., 2018), consisting of about 357944 geolocated tweets, written in Portuguese, posted by 159615 users from 206 countries across the world (according to the platform indication), collected between May 2014 and November 2017, covering 192 consecutive weeks, and corresponding approximately to a four years time span. Each tweet includes the meta-data 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.

To the collecting process, a brand filter was applied, so that only tweets mentioning at least one of the 16 brands selected would be retained. The brands, which were selected based on the number of followers and the number of tweets, are the following: Adidas, Nike, Vans, Puma, Victoria’s Secret, Gucci, Valentino, Versace, Converse, Michael Kors, Burberry, Marc Jacobs, Armani, Tommy Hilfiger, Christian Louboutin, and Dolce & Gabanna. As in (Lopes-Teixeira et al., 2018), for this study, we are only considering the top 10 brands, which are the brands with more tweets in the dataset. Additional processing steps were applied to remove irrelevant tweets. For instance, 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. The last step was needed because the word “Valentino” is commonly mentioned in posts from Argentina, but they were most likely referring to a person or to pets with the same name. Tweets having the words “Valentino” and “Humoro” were also removed, as in these cases the users were not talking about the brand.

Table 1: Database properties.

| Brand | Users | Tweets | Tweets/User |
|---|---|---|---|
| Nike | 68098 | 126427 | 1,86 |
| Adidas | 65870 | 120784 | 1,83 |
| Vans | 41071 | 70091 | 1,71 |
| Puma | 12763 | 18710 | 1,47 |
| Victoria's Secret | 9574 | 12642 | 1,32 |
| Gucci | 5312 | 7988 | 1,50 |
| Versace | 4989 | 7312 | 1,47 |
| Valentino | 3924 | 6083 | 1,55 |
| Converse All Star | 4893 | 5975 | 1,22 |
| Michael Kors | 1075 | 1558 | 1,45 |

Similarly, 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, as they were referring to an ice-cream shop named Valentino. The same was done for tweets containing the word "Versace", as there's also an ice-cream shop with this 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 number of posts, and we've found out that the corresponding account was used solely for advertising purposes. There were several other accounts used for the same purposes, mainly amongst United States users. Tweets posted by these users were discarded as well.

Table 1 shows user and tweet statistics for the selected brands, revealing 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. Finally, a total of 86 tweets had the location filled with the hyphen mark. To some of them, the location of another tweet posted by the same user was assigned to the tweet with no location. The ones that no other located tweets posted by the same user were found, were removed, as they were not valid for this analysis. In the dataset, all the instances of the brands Michael Kors, Converse, and Victoria's Secret were concatenated, so that the words composing the brands' names could be considered as a single word, thus counting as one.

In order to perform topic modeling tasks, all the tweets mentioning the same brand and posted during the same week were aggregated using a concatenation script. This step resulted in a dataset composed by 1918 documents, posted over a total of 192 weeks, with an average of approximately 10 documents/week.

## 4 PREPROCESSING AND TOPIC MODEL TRAINING

(Vijayarani et al., 2015) provides an overview of preprocessing tasks, and discussed what they've considered the three key steps of preprocessing, namely: Removing stop words, stemming and using TF-IDF weighting algorithms. Similarly, in (Srividhya and Anitha, 2010) the researchers evaluated several preprocessing techniques and analyzed the effect of such preprocessing tasks on text classification using machine learning algorithms.

Our experiments apply a set of preprocessing steps to the dataset, so that more coherent and informative topics could be produced. 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 whitespaces. In order to retain a good vocabulary to represent the whole dataset, Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme was applied. Also, terms that are not present in at least 0.1% of the documents, which corresponds to approximately two documents, were not included in the vocabulary. 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 important. This step resulted in removing words occurring less than four times in the whole dataset. The vocabulary was restricted to 5000 words.

Along with the preprocessing steps previously explained, several other preprocessing steps were applied to the dataset, namely: removing adverbs, cardinal numbers, ordinal numbers, punctuation, conjunctions, social networks common slang and abbreviations, and verbs. Verbs expressing some kind of willing to acquire/buy 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 were applied to create the vocabulary, could handle all the occurrences properly. Additionally, only terms being present in at least two documents were consi-

Table 2: Top 5 Puma topics from stemmed text.

| # | Terms |
|---|---|
| 1 | camisa uniforme adidas nike disc cola chuteira jogos arsenal novo |
| 2 | disc adidas cop novo nike bandido agua tenis whisky red |
| 3 | tenis quero rihanna cop fenty novo colecao adidas bts |
| 4 | disc cop adidas camisa quero nike mizuno novo tenis bota |
| 5 | disc rihanna adidas gira novo camisa cop paulo catraca tenis |

Table 3: Top 5 Nike topics from preprocessed text.

| # | Terms |
|---|---|
| 1 | adidas comprar quero comprei loja air queria chinelo bone casaco |
| 2 | meia canela botajoga adidas comercial camisa shox quero propaganda |
| 3 | adidas pes quero comprar celular querendo camisa bone comprei role |
| 4 | adidas quero comprar air shox bone comprei camisa quer coroa |
| 5 | compra adidas quero shox air comprei camisa fuzil mola red |

Table 4: Top 5 Nike topics from unprocessed text.

| # | 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 |

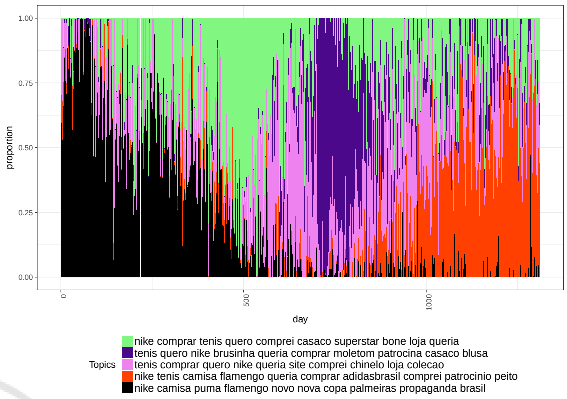

Figure 1: Adidas topics daily evolution.

dered (sparse factor $> 0.99983$), thus whipping out misspellings words that could be interpreted by TF-IDF as low frequency relevant words.

Although mentioned in (Vijayarani et al., 2015) as an important preprocessing task, we opted not to use stemming, as it does not work well for Portuguese. For instance, the word ”copa” (/cup), which refers to the Football World Championship, was reduced to ”cop”, and figures in almost every topic of the Top 5. Table 3 shows that more informative topics can be produced when preprocessing tasks are applied. For example, apart from the brand name, only 3 out of 12 terms can be considered informative in topics 1 and 6 from Table 2. Table 4 shows that topics produced from unprocessed texts contain 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 one, which reduces the importance of non-relevant words appearing frequently, all the topics begin with the name of the brand, which is not informative as each brand has been evaluated separately. Also, as stop words are frequent words throughout the dataset, and they were not removed, all the topics produced contain several stop words.

In order to limit the number of documents used training our topic model, we have created documents that aggregate groups of tweets, either on a daily or in a weekly basis, and we have concluded that documents grouping a day of tweets produce better results. This is in line with the work presented in (Mehrotra et al., 2013), in which the researchers grouped their tweets by time spans of 2 hours, based on the assumption that a great number of users posts about happenings within short time spans. Nonetheless, in order to obtain clearer trends visualization, the models were applied to documents that aggregate longer time span, usually one week of data. Figures 1 and 3 were created using the same model, but the former groups tweets by day, while the later groups tweets by week. It can be observed that both figures show the same clear trends over time. While such approach works well for Adidas, as both week and day charts are very similar, it does not work so well for Versace, where Figure 2 shows a less clear trends visualization than Figure 7. Therefore, we adopted to perform a week-based analysis for every brand.

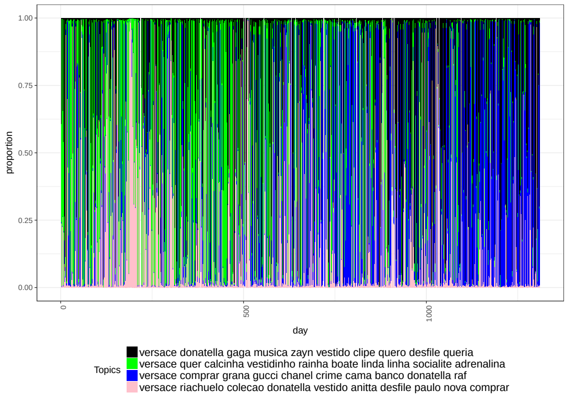

Figure 2: Versace topics daily evolution.

Concerning the number of topics being considered, that number was picked after several iterations, considering both the information each topic conveyed and the clearness of the resulting plot. As our goal is to observe changes over time in order to point trends, we have tried to reduce the overlapping topics, which result into more noisy plots.

## 5 ANALYSIS

In line with the work presented in (Lopes-Teixeira et al., 2018), it can be observed that brand interest, i.e. the volume of posts mentioning the brands over analysis, changed over the time. It shows ups and downs, and several peaks could be related to real-world events, as other studies have demonstrated (Mehrotra et al., 2013; Paul and Dredze, 2014).

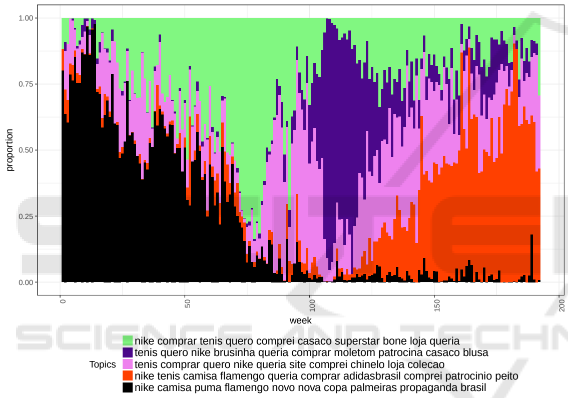

Figure 3: Adidas topics weekly evolution.

Figure 3 shows that the first weeks have more shares about the last two topics. The last one, which is about Nike, ”comercial”/”propaganda” (commercial), ”Brasil” (Brazil), ”copa” (world cup) 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, in 2014 (Lopes-Teixeira et al., 2018). The first, second, and third topics express the intention of purchasing new items: “novo”/“nova” (new), “camisa” (shirt), “tênis” (sneakers), ”moletom” (pullover), ”chinelo” (flip-flops/slippers), ”boné” (bonnet), and so on. The fourth topic, in which figure the name of two Brazilian Football teams (Flamengo and Palmeiras), had more shares until roughly the 50th week. This might be, in part, related to the two matches in which these two teams faced each other, more specifically in May 2014 and September 2014. A possible reason for Nike and Puma being present in topics from Adidas data might be due to sport brands being very often subject of comparison.

> Não resisti entrei no site da adidas e comprei a tal camisa edicao limitada de 300 dilmas do Flamengo / *Couldn't resist, I accessed Adidas website and bought that 300 Flamengo limited edition shirt*
>
> 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*

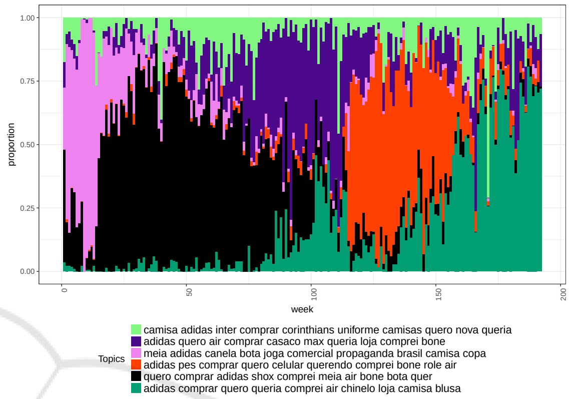

Figure 4: Nike topics weekly evolution.

The first topic in Figure 4 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 World Cup topic, which is the third, can also be spotted in the plot. It can be observed that this topic was more discussed in the early weeks of the dataset, then its proportion decreased as the time went by. Similarly to Adidas topics, Nike topics also mention Adidas, demonstrating that these brands are mentioned in the same document several times. The topic in which figure the terms “Shox” and “Air” (Nike sneakers), ”boné” (bonnet), and ”bota” (boot) was discussed from the beginning to the middle of the set of weeks, then they faded. This is in line with the launch of Nike Spring/Summer collection, which occurred around the first semester of 2015 (Lopes-Teixeira et al., 2018). The desire of purchasing is common to almost every topics, and it's shared across the weeks. What distinguishes them are in essence the items that are object of desire. Topic 5, for instance, “camisa” (shirt), Air (sneakers), “blusa” (blouse/top), while the second 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.

- “Adoro mt o trailer a publicidade da Nike para o mundial”. “I love very much Nike ad trailer for World Cup”.
- “Esse comercial da nike ta oh ???? Que venha Copa Nike”. “This Nike commercial is lit! Let the world cup begin”

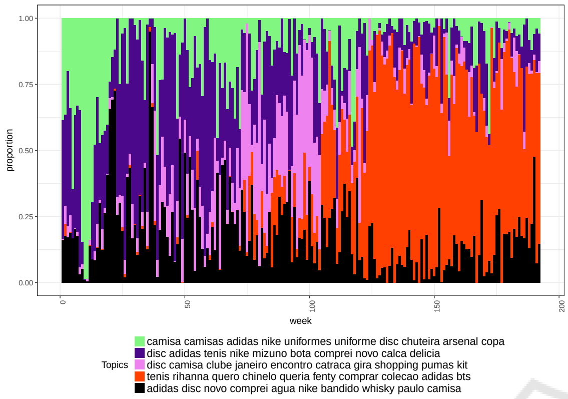

Figure 5: Puma topics weekly evolution.

Figure 5 illustrates how the first topic, which is about “camisas” (shirts), “uniformes” (uniforms),”chuteira” (football boot), ”copa” (Wolrd Cup), Nike, Adidas, “Disc” (Puma sneakers), starts with a high proportion but by the end of the set of weeks it is almost not discussed. The high proportion of this topic is most likely due to the Football Championship that took place in Brazil, back in 2014. The second topic, which mentions ”camisas” (shirts), ”tenis” (sneakers), ”Mizuno” and ”Disc” (Puma sneakers), ”calça” (trouser/pants), along with the brands Adidas and Nike, follows the trend of the first topic, being more discussed until the middle of the dataset, also losing relevance from that point until the end. 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, which were all met positively by both critics and buyers. In 2016, Rihanna debuted her first clothing line in collaboration with Puma. In the 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 fourth topic evolution is in line with these events, as terms such as ”Fenty”, ”Rihanna”, ”coleção” (collection),”queria” (I wanted), ”comprar” (purchase), ”tenis” (sneakers), and so on, can be spotted in this topic.

The last topic, in which figure the terms “tenis” (sneakers), “Mizuno” (Puma sneakers), “camisa” (shirt), Adidas, Nike, along with the word “comprei” (I bought), has a higher proportion from the beginning until the middle of the dataset, losing strength afterwards.

- “que adidas o que sua louca eu quero um creeper da puma q a dona rihanna fez”. “What Adidas?! I want a Puma Creeper made by Lady Rihanna.”

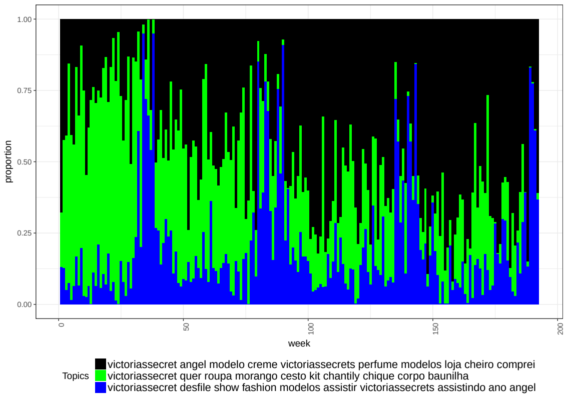

Figure 6: Victoria’s Secret topics weekly evolution.

Figure 6 shows that the last topic for Victoria’s Secret brand, which is essentially about their annual fashion show, presents a seasonal behavior, which reflects the trend pointed in (Lopes-Teixeira et al., 2018). The other two topics do not follow this trend, rather, the second topic, which is about clothes, Victoria’s Secret kit, and strawberry and Chantilly scented lotions, is more talked about in the first set of weeks. The last topic, in the other hand, has its proportion increased in the second set of weeks.

- “Adorava estar no Victorias Secret Show”.“Id love to be on Victorias Secret Show”.
- “To vendo os desfiles da Victorias Secret amoo Meu sonho e ser modelo da Victorias Secret”.“Im watching Victorias Secret Fashion Show, love it My dream is to become a Victorias Secret model”.

Figure 7 shows that the first topic, which is about Riachuelo having a Versace collection, has an unusual proportion somewhere before the fiftieth week of the dataset. This high proportion is in line with (Lopes-Teixeira et al., 2018), coinciding with the fashion show in which Riachuelo presented its Versace collection, in November 2014. As this topic is also about “desfile” (fashion show), Donatella Versace, “roupas” (clothes), the topic never really fades away. In fact, it presents ups and downs that are most likely related to the brand fashion shows carried out every year. The third topic, which seems to talk about high couture brands, as it mentions Gucci, Chanel, “grana” (a Portuguese slang for money), is what people talked about in the late weeks. Before this topic showed up, people were talking about “Vestidinho” (short dress), socialite and “boate” (nightclub).

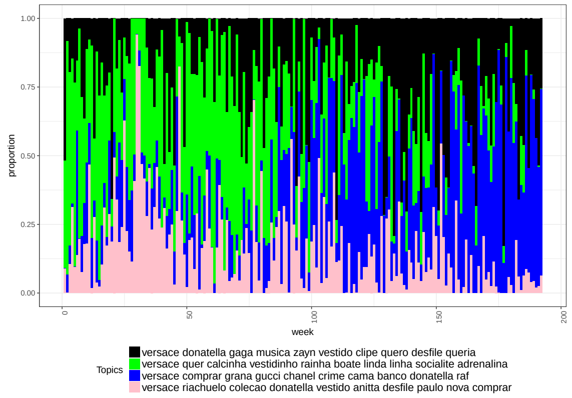

Figure 7: Versace topics weekly evolution.

- “Hoje eu vi as peças da Versace pra Riachuelo e meu Deus, que colecção linda”. “Just saw Versace for Riachuelo clothes and God, what a beautiful collection!”
- “Riachuelo fecha parceria com linha da Versace OMG ja querooo Versace tem que chegar na Riachuelo antes da primeira prova do enem”. “Versace starts a partnership with Riachuelo in Versace collection OMG I want it now. I hope it hits Riachuelo stores before the first national exam”.

The first topic illustrated in Figure 8 is about the semi-annual fashion event named Paris Fashion Week. This topic also mentions the former boys band leader Zayn Malik, who attended 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. The last topic captured shares about Valentino like brands such as Dior and Gucci and . Also, it mentions (Valentino) ”Khan”, which is a a well-known DJ and producer, and ”Ricky Martin”, a Puerto Rican singer. Topic 3 is composed by terms such as ”feliz” (happy), ”coleção” (collection), ”nova” (new), and ”modelo” (model). The last topic, though, has nothing to do with Valentino brand; rather it seems to be about a sport motocross event, as it mentions “motogp”, “ganhar” (to win), “seguidores” (/followers), and Valentino (Rossi), a professional motorcyclist.

The first topic in Figure 9 is composed by terms such as “Gucci”, “cinto” (belt), “tenis” (sneakers), ”coleção” (collection), ”bordado” (embroidered), and another *haute couture* brand ”Chanel”. This topic is present in most of the documents, having its proportion increased from the middle to the end of the chart. By the end of the chart, we can quickly spot Topic 3, which is about Kim Taehyung (from the South Korean boys band ”Beyond The Scene”) appreciation to Gucci clothes; his appreciation to the brand started to be noticed/talked about in 2016. Altough mentioning ”Gucci”, the fourth topic 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 last topic also mention other *haute couture* brands such as Prada, Chanel, and Louis (Vuitton), along with ”bolsa” (bag) and ”jaqueta” (jacket). The name ”Harry” also figures in this topic, refering to the former member of a British boys band, Harry Styles, whose appreciation to the Gucci brand culminated in him being the new face Of Gucci’s Tailoring Collection.

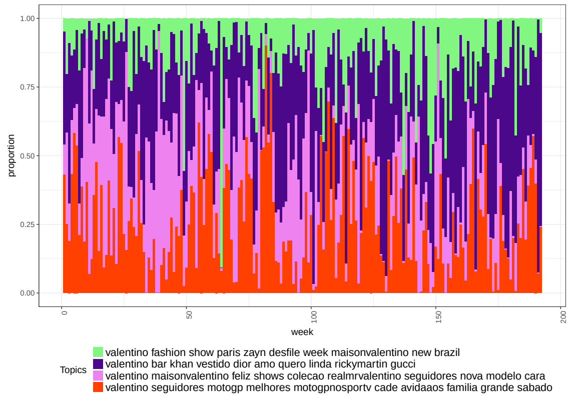

Figure 8: Valentino topics weekly evolution.

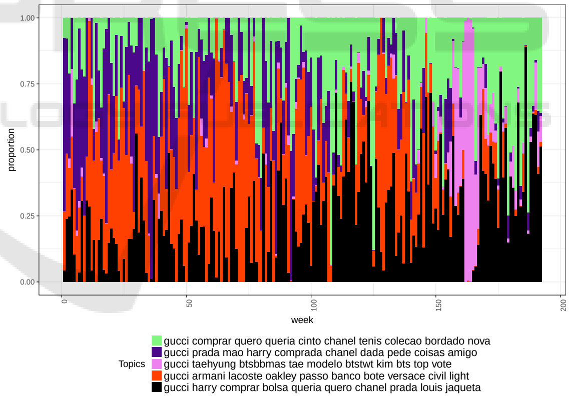

Figure 9: Gucci topics weekly evolution.

## 6 CONCLUSIONS AND FUTURE WORK

The current study demonstrates that grouping tweets based on the day they were uploaded, and by brand, to perform Topic Modeling tasks can produce coherent and informative topics. This study also shows that topics about what people discuss/share opinion and thoughts change over the time, and they can be related to real-life happenings, which is in line with the work presented in (Lopes-Teixeira et al., 2018).

For instance, commercials, products launches, and events can lead to emerging of new topics, which may result (or not) in older topics fading. Additionally, the plots presented show that each brand have different brand interest pattern, which was also stated in (Lopes-Teixeira et al., 2018). For example, Victoria's Secret topic about their fashion show comes and goes several times. Moreover, the importance of preprocessing in Natural Language Processing was emphasized. The experiments shows that preprocessing steps do have impact in the quality of the topics resulting from documents written in Portuguese. More elucidative/informative topics were produced when the documents were preprocessed. Tasks such removing URL's, removing stop words and choosing the representation vocabulary based on TF-IDF can avoid common issues that reduce the coherence of the topics. Results demonstrated that this framework can be followed to obtain coherent topics, enabling one to get insights about people's conversation/discussions on Social Networks.

Limitations of this study are related to the fact tweets frequently have slangs, hashtags with words concatenated, abbreviations and misspellings. Although the documents were preprocessed, not all the instances of this cases could be filtered out. Another limitation is that stop-words are still limited for Portuguese language. To overcome this, our own set of Portuguese words considered non-relevant for this study were created, so that meaningless words could be properly removed.

Future work includes applying another Topic Modeling algorithm in order to evaluate which one fits better for a large dataset. Discovering community patterns, i.e., how topics change from one community to another, is also a subject of future research.

## ACKNOWLEDGEMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

## REFERENCES

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

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.

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.

Lopes-Teixeira, D., Batista, F., and Ribeiro, R. (2018). Spatio-temporal analysis of brand interest using social networks. In *CISTI'2018 - 13th Iberian Conference on Information Systems and Technologies*.

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.

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.

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.

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.