INSTITUTO
UNIVERSITÁRIO
DE LISBOA

What are People Talking About in your Region? Applying a Topic Modeling Approach to Portuguese Geolocated Tweets

Érica Sofia Palmeirim Santos Rosa

Master in Data Science

Supervisor:

Doctor Fernando Manuel Marques Batista, Associate Professor,
ISCTE – University Institute of Lisbon

Co-Supervisor:

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
ISCTE – University Institute of Lisbon

October, 2022

Department of Quantitative Methods for Management and Economics

Department of Information Science and Technology

What are People Talking About in your Region? Applying a Topic Modeling Approach to Portuguese Geolocated Tweets

Érica Sofia Palmeirim Santos Rosa

Master in Data Science

Supervisor:

Doctor Fernando Manuel Marques Batista, Associate Professor,
ISCTE – University Institute of Lisbon

Co-Supervisor:

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
ISCTE – University Institute of Lisbon

October, 2022

Acknowledgments

This master thesis wouldn’t been able to be successfully conclude on time without the help and the amazing support and effort of both supervisors, Fernando Batista and Ricardo Ribeiro, who have always given me the strength to keep working on this project nevertheless of the difficulty’s that came up into my way and help me went through them successfully, ending up completing this project with proud of myself and of the path I have done to conclude it. I would also like to thank them both for all the attention and time spent with this project with me during this journey and for accepting to guide me on this final project of my master’s degree.

I would also like to thank my mum Margarida and my mentor Sofia Tenreiro for never letting me give up on my professional ambitions and dreams.

Lisboa, 31 de Outubro
Érica Rosa

Resumo

Esta tese tem como principal objetivo identificar os tópicos acerca dos quais os portugueses falaram na rede social Twitter, durante os primeiros 6 meses de 2021, em cada região do país. Como fonte de dados, foi-nos possível obter, através da API do Twitter, uma base de dados de cerca de 1 milhão de tweets, escritos ao longo deste período, em todo o país. Tendo os dados disponíveis, foi nos possível, através da criação de um dicionário de palavras, atribuir a cada localidade do país mencionada na base de dados, uma região de NUTS nível 2, de forma a atribuirmos a cada Tweet apenas uma região por entre 5 regiões: Alentejo, Algarve, Centro, Lisboa ou Região Norte. De seguida, fomos analisar os modelos de modelagem de tópicos mais utilizados no momento atual e, em particular, quais os que têm demonstrado melhor performance quando aplicados a textos curtos, como acontece quando falamos de tweets. Após esta análise bibliográfica, optámos por aplicar à nossa base de dados, e avaliar a performace, dos modelos LDA- Latent Dirichlet Allocation e MM - Multinomial Mixture Model. Através da medição da coherência em ambos os modelos, conseguimos resultados mais satisfatórios na aplicação do modelo MM, selecionando então este modelo para aplicar à nossa base de dados. Com os tópicos já definidos e atribuídos a cada tweet, foi realizada uma análise por região e diária, dos tópicos mais referidos pelos portugueses. Conseguimos concluir que os temas mais falados em Portugal, considerando a amostra recolhida na rede social Twitter, são: a política, a religião e a fé, os jogadores de futebol e a comida e a cozinha. Por fim, fizémos então a análise de tópicos por região e por dia, por entre as nossas conclusões, concluímos que o tópico da comida e da cozinha se destacam no Algarve e no Norte, e que o tópico das eleições ganha predominância, no geral do país, entre o final do mês de Janeiro e meados do mês de Fevereiro.

Palavras chave

Modelagem de Tópicos; Agrupamento de texto curto; Mistura Multinomial de Dirichlet; Alocação de Dirichlet latente; Processo de Grupo de Filmes; Classificador Naive Bayes; Twitter em Portugal; Geolocalização de Tweets.

Abstract

The main objective of this thesis is to identify the topics that the Portuguese spoke about on the social network Twitter, during the first 6 months of 2021, in each region of the country. As a data source, we were able to obtain, through the Twitter API, a database of around 1 million tweets, written throughout this period, across the country. Having the data available, it was possible, through the creation of a dictionary of words, to assign to each locality of the country mentioned in the database, a region of NUTS level 2, in order to attribute to each Tweet only one region among 5 regions: Alentejo, Algarve, Centre, Lisbon or North Region. Next, we analyzed the most used topic modeling models at the moment and, in particular, which ones have shown better performance when applied to short texts. After this bibliographic analysis, we chose to apply to our database, and evaluate the performance, of the LDA- Latent Dirichlet Allocation and MM - Multinomial Mixture Model models. By measuring the coherence in both models, we achieved more satisfactory results in the application of the MM model, selecting this model to apply to our database. With the topics already defined and assigned to each tweet, an analysis was carried out by region and time period, of the topics most mentioned by the Portuguese. We were able to conclude that the most talked about topics in Portugal, considering the sample collected on the social network Twitter, are: politics, religion and faith, football players and food and cuisine. Finally, we then analyzed topics by region and by day, among our conclusions, was that the topic of food and cuisine stands out in the Algarve and in the North, and that the topic of elections gains predominance, in general in the country, between the end of January and the middle of February.

Keywords

Topic Modeling; Short text clustering; Dirichlet Multinomial Mixture; Latent Dirichlet Allocation; Movie Group Process; Naive Bayes Classifier; Twitter in Portugal; Tweets geo-location.

Contents

List of Figures

List of Tables

1 Introduction

During the past few years, our ability to deal with big data, together with the higher accessibility to the Internet, has created huge stores of digital data. Therefore, the challenge of finding and extracting relevant information began to stand out for the general market and to be one major investigation subject to data scientists. A need of having tools that can effectively extract and summarize the content had came up, and between those, is topic modeling, a method that allows us to extract hidden themes or topics in a large collection of documents.

Information appears stored in many forms in the digital world, and some of this information is stored as short text, that is usually used on social networks posts, like Tweets, where people share it is ideas, interests and opinions with only a few words. Short text stores can potentially provide us with interesting information about general topic, as the public opinion or some current trends, for instance, making this type of search an interesting one for the business or political companies. Unlike long text, one of the commonly challenges of applying topic models on short text, is the fact that it is made of just a few words, difficult the model's job of finding meaningful words to match with each topic found.

The Latent Dirichlet Allocation (LDA) model is one of the most popular topic models used nowadays by the data scientists. It makes the generative assumption that each document belongs to many topics. Simultaneously, the Multinomial Mixture (MM) model, assumes that a document can belong to one topic only, being used a lot for short text investigations. Considering this difference, we can intuitively assume that probably the MM model should perform better than the LDA when applying topic modeling to tweets.

The main purpose for this project is to figure out what were the Portuguese people talking about in tweet during the first 6 months of 2021 in each region. We were able to get some clear topics by region and consequently to note some marketing insights and ideas that could be applied by the companies and organizations that sell services or objects, these were mention by the end of this project. Another objective we consequently found for this project, was to compare LDA and MM model's performance when applied to short text database of Portuguese Tweets written in Portugal, using coherence as our main performance measure. According to our performance coherence results, MM model seems to perform better than the LDA model on short text, matching with the results found by the generality of the bibliography found on the subject.

1.1 Motivation

The main motivation for this work was to discover which topics the Portuguese people are talking about in each region of the country. To do that, we would need to choose a topic model to apply and to consequently study which model would be the best fit for our case scenario.

Between the majority of the data scientists that work with topic modeling, Latent Dirichlet Allocation (LDA) is probably the most popular topic model.This model has has proven over the last years, to have a very good performance when applied to long text documents, such as news articles or academic abstracts [1]. Nowadays, we verify a higher interest on doing short text analysis, since the digital user's comments on websites, social networks or at micro blogs, are written as short text mostly. This type of posts holds information that is potentially interesting for sentiment analysis [2], prediction purposes [3] or to product marketing [4]. Unlike long text, the fact of short text is build of just a few words, makes harder to found interesting words to describe the hidden topic within the corpus presented to the model. For this reason, we also found here as a point to work on, the study and testing of topic modeling models for short text.

1.2 Objectives

The main objective of this work is to investigate what topics are the Portuguese people talking about in each region of the country and what are the topic variations at the same region during the first semester of the year. To put this into practice, we needed to search about topic models and to apply them to our data to achieve our results. Therefore, as a second objective, we will have MM and LDA models comparison when applied to short text.

We could point out our objectives with the following questions we pretend to answer with this investigation project:

1. **What are the Portuguese people talking about in twitter? do the Portuguese people talk more regularly about specific topics? or have any specific topic that talked about during a particular time of the semester according to the region?**
2. **Will the MM model actually work better with our short text data, comparing to the LDA model?**

1.3 Contribution of this work

Comparison's between the application of the MM and the LDA models in short text is the thing that has not been done until 2015. So, we expect that our hypothesis will be sup-

ported through our experimentation and that this could provide researchers with one more research handling short text and also, on seeing whatever shortcomings that the Multinomial Mixture model may possess, we will be in a position to propose possible solutions and modifications that could be investigate in future studies.

There are just a few papers done with Portuguese tweets written in the Portuguese language, and, from what we found, neither of them had the purpose to identify topics according to the location of where they were written. With this in mind, we will offer with this project the interesting insights and conclusions about the different subjects and the quantity of mentions that they have in each region in Portugal.

1.4 Dissertation structure

This dissertation structure is as follows:

Chapter 2 introduces various existing topic models in more detail.

In Chapter 3 we do a further investigation about the Multinomial Mixture model (MM). Here we study with more detail the assumptions and the generative process of the model, since is the one model we end up choosing to apply to our scenario case and get our final topics results.

After that, in Chapter 4, we explain how have we done our data preparation and what hyper parameters have we chosen to apply to the LDA and MM models. Coherence values for both models and first topic results are also obtained in this chapter. For the MM coherence values calculation two different methods options were applied, being those 'c_v' and 'u_mass' methods, the conclusions were taken about these two methods application.

The exploratory analysis of the database used in the project was done in Chapter 5, the analysis was done with the purpose of taking already the conclusions about the tweets by region and by month.

In Chapter 6, we present the experiments that were performed, as well as our final results.

Finally, in Chapter 7, we present our conclusions and our discussion of the future work that could be done in the subject.

Concepts and related work

2

In this chapter we will do a general explanation about what is topic modeling, why is used and how this clustering technique is done to identify the topics in documents. We will also explain the of the most used topic modeling models and the specific case scenario of short text when applying topic modeling techniques. Finally, we will explain our problem statement for this project to proceed with all this statements clarify and well identified.

2.1 Topic modeling

With the technology advances, our capability of collecting and storing information digitally had an huge increase. Information is nowadays stored in many forms, such as articles, web pages, micro blogs such as the social networks or scientific articles. Such collections of electronic information continue growing at very high rates, which results in massive stores of data. As a consequence, it becomes increasingly difficult finding and extracting relevant information from these sources. This situation created the need of developing efficient learning techniques that could enable us to extract and understand the information available within these sources [5].

A lot of already existing learning techniques, have been proposed for the analysis of large document collections, and they are usually categorized as unsupervised or supervised techniques. When talking about supervised learning context, documents are classified into predefined classes. Unfortunately, we usually have not any prior knowledge about the documents information, which makes such techniques difficult to be applied. In these cases, we use unsupervised learning techniques instead. These techniques allow the user to do a document classification without using any predefined categories or labels, doing a subjective classification.

Clustering is an unsupervised technique that, when applied to documents, group the documents by it is similarity and separates them from those that reveal different topics within it is content [6]. Each document is represented as a high-dimensional vector of word frequencies (the bag-of-words representation) and standard vector-based clustering techniques, such as k-means and agglomeration clustering, are applied to the documents, to allow this similarity grouping by. One of these methods weakness is that it is not capable of

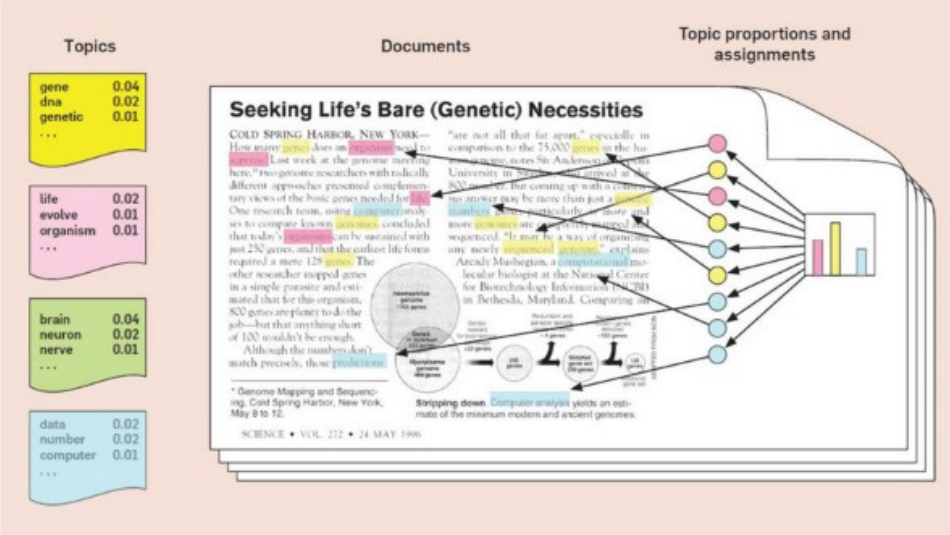

Figure 2.1: LDA illustration model (Blei et al., 2012)

knowing what topic each cluster specifically represents. Additionally, clustering techniques make the assumption that each document can only belong to one topic [7]. Nevertheless, considering long text, it would be more realistic to assume that a document has within more than one topic with different weight percentages.

Topic modeling it is a specific text clustering technique that is, by definition, a text mining technique that allows us to find the underlying hidden topics or themes within a large amount of documents, by clustering them based on thematic similarity. This topic classification assumes that each document has been created through a generative process, which can be seen as an imaginary process by which documents are assumed to have been created [8]. There are different types of topic models, which have different assumptions, a few of them assume that documents contains just one topic, while others, assume that one document can present a variety of different topics within.

2.2 Visualizing topic models

To present topic model concept and provide a better understanding of it, we present in Figure 2.1 an overview of a specific topic model, the Latent Dirichlet Allocation [8]. This model assumes that a document can contains multiple different topics in different weights, as we can see in the left side of the Figure, where we see different topics that contains different principal words within, with different percentage weights each. The presented example is part of an article titled as Seeking Life's Bare (Genetic) Necessities, which is about the determination of the number of genes that are required to the survival of a living organism.

Looking at the three top words in each topic, we can see that the yellow topic is char-

acterized by the words gene, DNA and genetic, which make easy to identify that the topic mention here is genetics. Using the same logic, we can characterize the pink, green and blue topics, which match with the evolutionary biology topic, nervous system and data analysis, respectively.

Assuming that each word belongs to at least one topic, each word in the document presented in Figure 2.1 can be highlighted with each color, according to the topic to which it belongs. Following this logic, the words genes and genomes were highlighted in yellow, whereas the words like computer and predictions were highlighted in blue. For this document, which is a small one, words could be highlighted manually, as we can see they were, but supposing that each word, excluding non-informative words like and, but or if, were highlighted according to it is topic, the document would contains different topics in various proportions, which are represented by the histogram on the right of Figure 2.1.

When we run a topic model, the outcome for each topic will be always a set of words with different weighs, which requires that after that, a person needs to look at the words found by the model, and then decide which name should be attribute to each of the topics found, considering it is principal words.

For this specific example, we only looked at the LDA model, but other probabilistic topic models could have been applied here, and they would work in a similar way, depending on it is generative processes and model assumptions. As previously mentioned, topic models don't assume having any previously knowledge about the subjects presented in each document. These models consider that the inferred hidden structure of the data will represent it is thematic structure. As a consequence, topic modeling is a tool that can be interesting for information retrieval, classification or data exploration, talking about large data scales, that would be unrealistic to do manually.

2.3 Approaches for topic modeling

This subsection presents the most relevant topic models used nowadays in the subject's field of study.

2.3.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis assumes that the documents are represented as high dimensional vectors of word frequencies, while the entire collection of documents, which is called corpus, represents the document word matrix. This analysis [9] works by performing a dimension reduction on the document's frequency vector, by projecting the vector to a lower dimensional latent vector space, that captures the all the vectors variety found in the corpus [10]. By applying the Singular Value Decomposition (SVD) on the document word matrix, is possible built a linear mapping. The dimension's reduction allows to reduce the sparsity

of the vectors, and will do the mapping of the terms with similar meanings, grouping them into vectors with the same direction in the latent space [11]. As a result, the analysis find meaningful relationships between thematically similar documents, even when they do not contains any common words [10]. However, and despite of the fact that this analysis has also been proven to be applied successful in many fields, it presents the disadvantage of lacking a solid statistical foundation [10]. Considering this weakness, [10] design a new other model, that is based on LSA, but in which we have a statistical basis, with a clearly defined generative model for the data, which was called the Probabilistic Latent Semantic Analysis. This model would use a standard statistical model selection and fitting procedures [11], which provides the statistical foundation that was needed to support the previously model.

2.3.2 Gamma-Poisson model

The Gamma-Poisson (GaP) model is a probabilistic model proposed by [12]. This model allows us simultaneously improve the search and retrieval of information from the corpus, and to do the topic identification and clustering of the documents, depending on it is thematic content similarity [12]. According to [12], this model assumes that a document word matrix, $F = (F_{ij})$, containss the number of occurrences of a word i in a document j, and that for this matrix, each element, ij, denotes the probability of word i in a topic j.

The Gamma distribution is seen as a very flexible distribution, since the model can take on varying shapes. For this reason, we can expect that for small sizes input data, the parameters of this distribution will have a low accuracy. As a consequence, this model reveals to be preferable to be used in long text documents analysis.

2.3.3 Latent Dirichlet Allocation (LDA)

The Latent Dirichlet Allocation [8] model is likely one of the most popular probabilistic topic models used nowadays by the data scientists around all world. The model assumes that the corpus presented containss a certain number of topics, T , and that each document is formed through a generative process as follows [8]:

1. The model randomly choose T topic distributions: t Dirichlet( ).

   It is considered a word matrix $= (1; 2; :::; T)$, in which each elements, ij, will represent the probability of the i^{th} word belonging to the j^{th} topic.

2. For each document, $d = (w_1; w_2; :::; w_N)$, presented in the corpus:

   (a) The model will randomly choose a distribution over topics: d Dirichlet( ).

   (b) For each of the N words, w_n, in a document d:

      i. The model will randomly choose a topic z_n : Multinomial(d).

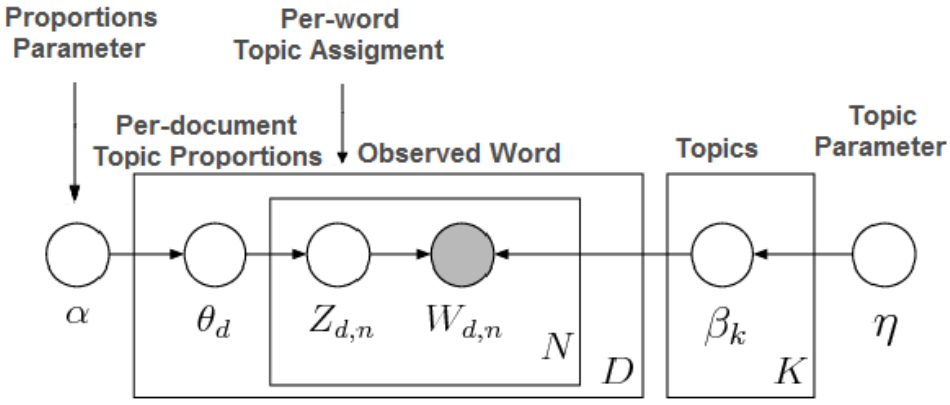

Figure 2.2: LDA graphical model

ii. The model will randomly choose a word w_n: Multinomial(z_n).

In Figure 2.2 we can see this process represented graphically.

Considering that this model makes the generative assumption that each document containss multiple topics in varying proportions, we can expect it will work much better with long text documents comparing to the short text ones, where we probably would have just one or two topics. This intuitive prediction will be tested within this research project, and the results will be presented further in this document.

2.3.4 Multinomial Mixture model (MM)

The Multinomial Mixture model, as the name reveals, assumes a multinomial distribution over words, meaning that the corpus is modeled as a mixture of multinomials. Looking at the amount of available literature, the Multinomial Mixture model appears to be a less popular probabilistic generative model, probably due to the fact that it makes the assumption that in each document in a corpus, we can only found a single topic. However, even this assumption is not well seen in many situations, this could be actually a strength in short text scenario cases, where it is reasonable to assume that a few little words would just hide only one specific topic. As this model will be one of the main focus of our work, it will be discussed it with more detail in Chapter 3.

2.4 Challenges related with short text documents

Nowadays, it is clear the huge growing of the Internet data in general, including a lot of online publishing, instant messaging, e-commerce and social media contents. With the increasing of the internet access in the last few years, an uncountable number of textual data stores have been created, and it is expected that this number will continue exponential rising. As a consequence of this situation, short text information content available start to

have a research interest to analyze the people opinions and interests, revealed in a few words in the Internet. Examples of short text include posts on Twitter, called tweets, which can not have more than 140 letters on it. The challenge found when handling with short text by data scientists was work with just a few words, meaning that there could not be enough words to attribute a topic model to each text.

During the last few years, it is easy to verify that we had a significant increase in the amount of short text available, mostly at social networks like Linked In, Facebook and Twitter. Such data stores usually contains relevant and potentially valuable information, that is interesting to the business companies, and for data and users control, as for instance the time when the posts were created, since these websites are commonly used to communicate breaking news, to witness accounts [13], an also to share ideas [14]. As a consequence, with the purpose of getting interesting information from these data stores, having efficient techniques to extract this information became increasingly relevant, especially for short text.

At the moment, the application of information retrieval tools on tweets has become a subject of much interest, since they can provide us interesting information as the indication of current trends, public interests or public reactions to the breaking news [14]. As an example of this, we have the study made on [15]. This study do the relation between tweets and the results of the Irish General Election, the results shown that the tweets had predictive qualities.

As mentioned before, due to it is sparsity and it is limited information content, with just a few words in each document, short text presents challenges when applying traditional topic models, as the LDA, for example. However, a lot of investigation is being done with the purpose of finding techniques that would work effectively with short text, making this model an interesting one to continue studying about in the future.

3 Proposed approach

In this chapter we will study with more detail the Multinomial Mixture Model, the one approach chosen by us, due to the above mention explanations, to apply to our database and to get the topic results that will be analyze to take final conclusions and insights according to this project's objectives.

To study this model we will explain the Movie Group Process, the one process used by this short text model, and also the Dirichlet Multinomial Mixture model, which expressions allowed, together with the already existing Naive Bayes Classifier, to get to this Multinomial Mixture model, known nowadays for it is good performance when dealing with short text documents.

3.1 Movie Group Process

To explain the Movie Group Process we will use an analogy to help us understand both the short text clustering problem and the MM model logic, the one chosen by us to be applied within this project.

Let's imagine that a teacher during a movie discussion, aims to group his students in several groups. He wants the students from the same group to have watched similar movies, so they would have more possible content to discuss. To assure that, the teacher asks the students to write down a list of the films they have seen, during a few minutes. As the students have a short time to do it, they wouldn't write a long a list and would probably end up writing down movies they have recently seen or it is favorites movies. With each student having his list of movies, the teacher would be able to build groups of students with similar interests and similar movie lists.

Explaining more formally the Movie Group Problem, the input in this example would be the D students (documents) and each student (document) is being represented by a short list of movies (words), with the main objective of clustering the students (documents) into several groups, in a way that students (documents) with similar movies (words) would be at the same group. The number of distinct movies (words) is defined as V, which is very large due to the sparse characteristic of short text, while the average number of words ($\bar{L}$) in each short text will be small instead.

As common similarity-based models used for text clustering, we found K-means and HAC, which usually represent the documents using the Vector Space Model (VSM). This model gets into consideration that each document (student) is represented by a vector of length V , and that each vector is made of the weight of it is corresponding word (e.g., $TF - IDF$). Considering the sparsity found in short texts, most words at the documents have $TF = 1$, meaning that TF is almost useless in the representation of short texts, being each vector characterized only by IDF.

We can imagine that the teacher from the previously analogy invites the students into a huge event and randomly assigns the students to K teams. Then she asks the students to choose again a table. We can then expect that the students will choose a table according to the following rules:

Rule 1: Choosing a table (cluster) with more students (documents);

Rule 2: Choosing a table (cluster) whose students (documents) share similar movies lists (words).

As this process goes on, the tables (clusters) will grow larger and other clusters, with less importance, will disappear. Finally, we can expect that only a part of the tables (clusters) will remains having students (documents), and the students (documents) in each table (clusters), will share similar movie lists and interests (words).

The Movie Group Process (MGP) can be considered to be equivalent to our collapsed Gibbs Sampling algorithm, used at the Multinomial Mixture model (MM) working process.

3.2 Multinomial Mixture (MM)

In this Section, we will introduce the Multinomial Mixture (MM) model.

MM is a probabilistic generative model for documents, which respects two major assumptions about the generative process:

1. Documents are generated by a mixture model;
2. A one-to-one correspondence is used between mixture components and clusters.

When generating a document d, MM model first selects a mixture component, called a cluster k, according to the mixture weights of all the mixture components, $p(z = k)$. After that, the document d is generated by the selected cluster by the probability distribution: $p(d|z = k)$. Therefore, we will be able to characterize a document d, as the sum of the total probability over all the clusters, respecting the expression shown in equation 3.1.

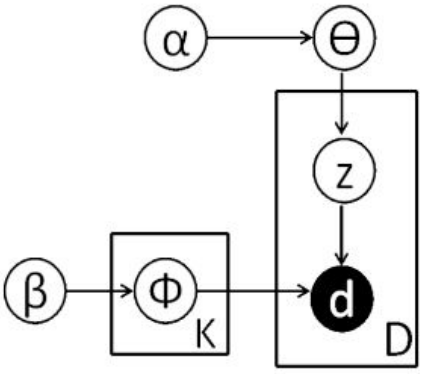

Figure 3.1: Graphical model of DMM

$$p\left(d\right) = \sum_{k=1}^{K} p(d|z=k)p(z=k) \tag{3.1}$$

In the previously equation, K represents the number of clusters. We need then to define $p(d|z=k)$ and $p(z=k)$. MM model makes the Naive Bayes assumption that the words in the document are generated independently when the document's cluster label k is recognized already, and the probability of a word is independent of it is position within the document. Then, the probability of a document d, generated by cluster k, can be derived as shown in the next equation 3.2.

$$p(d|z=k) = \prod_{w \epsilon d} p(w|z=k) \tag{3.2}$$

The above equation has derived from the following assumptions:

- $p(w|z=k) = p(w|z=k, \Phi) = \Phi(k,w)$, where $w = 1, ..., V$ and $\Sigma(w)\Phi(k,w) = 1$;
- $p\left(\Phi|\beta\right) = Dir(\Phi\left(k\right)|\beta)$, with Φ and β as vectors;
- $p(z=k) = p(z=k|\Theta) = \theta(k)$, where $k = 1, ..., K$ and $\Sigma(k)\theta(k) = 1$;
- $p(\Theta|\alpha) = Dir(\theta|\alpha)$, with θ and α as vectors.

The graphical model final result of the MM can be seen in Figure 3.1 to have a clearly idea of the models structure.

3.3 Meaning of Alpha and Beta

In this Section, we will try to explore the meaning of α and β , with the help of the Movie Group Process (MGP), previously explain in Section 3.1.

From the equation represented in 3.3, we can see that α is related with the probability of a student (document) of choosing a table (cluster). If we set $\alpha = 0$, a table (cluster) will never be chosen by the students once it gets empty, because the first part of equation will be zero. When α gets larger, the probability of a student of choosing an empty table (cluster), will also gets larger.

$$p(z_d = z|z_{\neg d}, d) \propto \frac{m_z, \neg d + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d}(n_z^w, \neg d + \beta)}{\prod_{i=1}^{N_d}(n_z, \neg d + V\beta + i - 1)} \tag{3.3}$$

We can also see that β is at the second part of the equation, which is related to the MGP's second rule, about choosing a table (cluster) whose students (documents) share similar interests (words). Then, if we set $\beta = 0$, a student (document) will never choose a table (cluster), since it's movie list (words) contains one single movie that his not at his movie list (words). We can see this is not reasonable, since other movies (words) of the students list (document) could appear many times in that table (cluster), and he may share many similar interests with the students of that same table (cluster). MM assumes that we have the same α for all tables (clusters) and the same β for all movies (words). The same α for all tables (clusters) implies that different tables (clusters) are equally important, simultaneously, the same β for all movies (words) implies that different movies (words) are equally important. We should then give less emphasis on too popular movies (words that appear in too many documents). To achieve this goal, we will give a larger β for these less important words.

3.4 Relationship with Naive Bayes Classifier

The conditional distribution $p(z_d = z|z_{\neg d}, d)$, with z and d vectors, presented in the equation 3.4, represents the Naive Bayes Classifier (NBC). For this equation, we can say that a document d corresponds to a cluster z ,with the largest conditional probability of $p(z_d = z|z_{\neg d}, d)$, the one chosen to sample a cluster z from the conditional distribution in the MM model. This allows this model to avoid falling into a local minimum, which is a common problem at this type of algorithms.

While the documents are grouped and split into clusters at the MM model process, a Naive Bayes Classifier (NBC) is learning. Then, each time a new document arrives, we can classify it to one of the already existing clusters, with the Naive Bayes Classifier, and update the classifier (update z, m_z, n_z, and n_z^w). The conditional distribution $p(z_d = z|z_{\neg d}, d)$ presented in the equation 3.3is equivalent to the Bayesian Naive Bayes Classifier (BNBC). This means that BNBC over emphasizes words that appear more than once in a text document, which means that if a word (movie) w appears twice in a document d (student's movie list), the contribution of w for the bellow equation is $(n_z, \neg d + \beta)^2$, while the contribution of w at equation 3.4, will be $(n_z, \neg d + \beta)(n_z, \neg d + \beta + 1)$. This difference is greater when

a word appears more frequently in a document. However, this is a good property to the clustering problem, since words in a document tend to appear in bursts, meaning that if a word appears once, it is more likely for them to appear again. The conditional distribution $p(z_d = z|z_{\neg d}, d)$ in the equation 3.3, simultaneously, can give the model words that appear multi-times in a document more emphasis, and allows MM to capture a bigger amount of words.

$$p(z_d = z|z_{\neg d}, d) \propto \frac{m_z, \neg d + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_z^w, \neg d + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_z, \neg d + V\beta + i - 1)} \tag{3.4}$$

Data preparation and defining topic modeling parameters

4

This chapter explains how have we done the data preparation of ours database before applying both the Latent Dirichlet Allocation model and also the chosen topic modeling approach. We will describe the chosen hyper-parameters to apply these models, as well as the performance evaluation methods applied to evaluate the models performance results. Finally, we will present the coherence results and due to those, we will announce our final decision about the number of clusters we will use to get our final clusters results by region.

4.1 Database pre-processing

Before applying the two adopted topic modeling models, some pre-processing should be done so the results of the models could be improved and be according with our study of understanding which topics were the Portuguese people talking about in each group of regions, grouping them by NUTS level II, which would mean splitting the tweets in 5 region groups: Lisbon, Algarve, Alentejo, Center and North regions.

After that, we drop all the columns that weren't needed anymore from the database and we then stayed with a database made of only three columns, being those "created"(date of the tweet's creation), "NUTS2" and "text" (tweet's content). In this database we convert "created" column into a date time format and created a list with all the content of the "text" column to group all the tweet's content.

Finally, we convert to lowercase all the tweets and also applied to them the stemmer stemming for Portuguese language, to reduce inflected words to it is word stem and improve the coherence whenever we apply the models, and we extract symbols, links, punctuation and stop words from the text. the stop words were removed manually and the with the nltk function for Portuguese language.

With all this first data preparation work done, we were ready to use the data as an input to apply the topic modeling models and get the topics mention in each NUTS Level II region in Portugal during the first six months if the 2021 year.

4.2 Applying Latent Dirichlet Allocation model (LDA)

This subsection presents the hyper-parameters chosen to apply the topic modeling model, including the number of clusters chosen to apply.

4.2.1 Getting the topics

After pre-processing the data we start by applying the LDA model to the texts, we have done the tokenization (that was needed to apply this model) and after we use the nltk to create the bigrams and the trigrams from the list of texts we had. After that, we were ready to apply the LDA model function to obtain the corpus from this model.

With the corpus created we were able to create a dictionary of words and then it's matrix. With these two inputs created, the dictionary and the matrix, we had the conditions to apply the tf-idf model, short for term frequency–inverse document frequency, which is a numerical statistic that is intended to reflect the importance of a word in a document of the corpus, using a weighting factor. After running the tests, we decided to choose a minimum of 15 docs. The output was then used to run the LDA model. The hyper-parameters values that represent our final choice according to the topics output were the following:

- corpus = doc_term_matrix,
- id2word = dictionary,
- num_topics = 20,
- passes = 50,
- decay = 0.9,
- alpha = 'auto',
- eta = 'auto'

These were chosen with the purpose of getting clusters with a group of words consistent as part of one unique and clear topic, nevertheless, evaluating the topics found by the model, the model seemed to be capturing good words but with no common sense between each other has a group of the same cluster, making it difficult to assign a topic to each cluster or to accept the model as a good one to apply in this scenario case.

As we have seen in Section 2.3.3, LDA model is a model that takes into account that we have several topics into one document and that is not as good as we had also seen in Section 2.4 so we can say that these results were expected, the model is mixing words from different topics in each cluster.

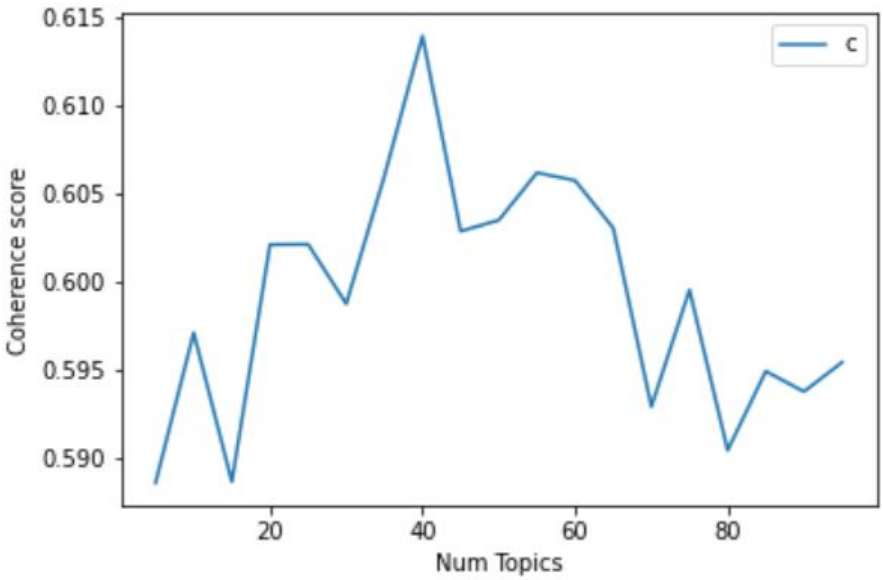

Figure 4.1: LDA model coherence scores

4.2.2 Measuring topics coherence

After analyzing the results, we calculate the coherence properly with a function. We were able to get the results seen in the Figure 4.1, where we can see that we have a coherence distribution between 0.58 and 0.62 depending of the number of clusters, having the higher value for 39 clusters (topics).

After getting these results, we applied again the LDA model but this time with “num_topics = 39”, to try it for the number of clusters with the higher coherence score. We were able to verify that the results were similar, we have good words with a good content, which is probably the reason for us to get this satisfactory coherence, but it is possible to verify that the words at the same cluster were from different topics, as for instance the word “Liverpool” (topic sports or futebol), “tribunal” (topic justice) and “rato” (topic animals), there were all present in the model as part of the same topic.

We were able to see and prove with these tests that LDA is a very good topic modeling model, but is a better model to apply for long text scenario cases, since for short text we will have just one topic for each document and the results will not be the best ones for that specific scenario cases, since the LDA model statement is that each document as more than one topic mention on it.

Hereupon, we decided to apply next the model we saw from previously literature that was one of the best ones, to apply in short text scenario cases as ours, the Multinomial Mixture Model, with the purpose of getting better topics and see which topics are being mentioned in each region in Portugal and if there is any pattern among all the regions.

4.3 Applying Multinomial Mixture model (MM)

This subsection presents our coherence results for two different applied methods and our conclusions about it.

4.3.1 Running the model and creating a function

Taking into consideration the fact that the Multinomial Model would be a good chance for us to get good results for our study, we applied this model as well to our pre-processed data.

We start by testing that the model function was working well by choosing only 2 and 3 clusters, and after seeing that we were getting topics with sense, we decided to run the model to a list of clusters. The number of clusters in that list were 5,10,15,20,25,30,35,40,45 and 50.

With the purpose of creating an efficient function, we create a code that allowed us to run the model, display the number of documents per topic and the most important clusters (by the number of documents inside) and it is top words, and also to display the coherence, all of this in one function were we end up also creating an excel file that allowed us to get all this information together into one excel file to all the clusters numbers (k) that the list containsed.

4.3.2 Coherence methods

We found two possible methods to apply the coherence, which are the c_v and the u_mass methods, which we will explain bellow in this subsection.

C_V Coherence Score is one of the most popular coherence metrics. It creates content vectors of words using it is co-occurrences and, after that, calculates the score using normalized point wise mutual information (NPMI) and the cosine similarity. This metric is popular because it's the default metric in the Gensim topic coherence pipeline module but even the author of this metric does not recommend using it because some associated inconsistencies where found when using it, and due to the latge amount of time needed to run this method when talking about applying it to big data.

UMass Coherence Score defines the score to be based on document co-occurrence of two words v_i and $v_j$that appear together in the corpus, defined as:

$$C_{UMass}(v_i, v_j,) = log\frac{D(v_i, v_j) + 1}{D(v_j)} \tag{4.1}$$

where $D(x, y)$ represents the counting of the number of documents containsing the words x and y, while $D(x)$ represents the counting of the number of documents containsing x. The UMass metric computes these counts over the original corpus used to train the topic model, rather than using an external corpus.

4.3.3 Applying C_V Coherence Score

After applying the function we got all the topics and all the coherence for all the clusters in the list. Many topics had good top words with a clear association to a topic, like Futebol or the Covid topic for example, so the model was working properly and seemed to have been a good option to apply for our scenario case. About the coherence, in a first test, we choose to display them with the coherence hyper parameter equals 'c_v', it was a good option but we run into a blocking situation when we realized that it was taking a long time to get any results, and after running and getting the coherence for this list of clusters, we have not been able to get any other coherence properly.

After doing the research we saw that this option was better but that we could use 'u_mass' to get also valid coherence values in a much faster way and that the researchers have realized that for same situation and data, when using u_mass to calculate coherence, there was a peak, and then it trended down, while for c_v, the values had a monotonous increase.

In Table 4.1 we can see the coherence results reached applying the coherence = 'c_v' method.

As we can see we had the highest value for 50 clusters and then also a peak for 35 clusters, making an interesting study to see if we had even highest values for 34 or 36 clusters for instance.

Due to the long time needed to calculate again the coherence but this time for 34 and 36 clusters, we decided to apply the 'u_mass' method instead to see the results quicker and check if the best option to proceed would really be using 50 clusters, since it was the scenario case with the highest coherence score and a reasonable one.

4.3.4 Applying the UMass Coherence Score

We apply the model for all the same number of clusters, but this time including 34 and 36 clusters, with 'u_mass' method, so we could see if the coherence values are really the best for 50 clusters or not. We can see the coherence values results for this method in Table 4.2, which we can see bellow. Differently from 'c_v' method, in which coherence values are presented between 0 and 1, in the 'u_mass' method, values are presented between -14 and 14, where zero is the perfect coherence.

Table 4.1: Coherence results applying ’c_v’ method

Nr Clusters	**Nr Docs Per Topic**	**Most Important Cluster**	**C_V Coherence Value**
5	[8234 72286 33346 40690 72927]	[4 1 3 2 0]	**0,320767051787827**
10	[6500 33919 15311 23386 19019 38068 22657 30788 14173 23662]	[5 1 7 9 3 6 4 2 8 0]	**0,311957320429606**
15	[20857 24706 11775 6489 6320 24918 12217 20068 8225 1038 14135 10170 31918 20164 14483]	[12 5 1 0 13 7 14 10 6 2 11 8 3 4 9]	**0,337725199135929**
20	[9113 10357 950 10203 15858 12643 25992 10014 13214 19977 11235 13120 6915 16674 5917]	[6 9 15 13 4 8 11 5 10 1 3 7 18 0 12]	**0,347137597821128**
25	[7462 9825 4612 5726 12871 5333 4358 8689 7147 9150 4039 11712 10504 6919 13716]	[20 21 14 18 4 22 24 11 12 1 23 9 7 0 8]	**0,326053674818176**
30	[9010 3644 5060 4589 5057 7141 18035 3052 8369 7976 6651 826 2047 5371 13647]	[6 24 14 20 29 25 21 17 0 15 26 8 28 9 5]	**0,364255810985571**
35	[4417 4635 4118 3704 11665 4917 795 4463 8732 3885 10419 6071 4871 7070 4940]	[26 28 15 4 10 24 29 19 8 17 23 25 13 16 20]	**0,378186020433627**
40	[9413 6945 2892 6167 6754 4925 1210 1879 8041 5422 6277 12025 3251 8442 777]	[15 11 25 23 20 0 31 13 37 8 28 1 4 35 17]	**0,297237755198277**
45	[3626 4745 2357 1307 2892 3967 6235 1967 9612 1973 6456 3254 1076 2503 9295]	[37 21 41 19 8 35 14 42 43 32 15 23 33 10 31]	**0,367889123090435**
50	[3051 7577 3960 6916 6233 3615 4981 2321 8213 12401 2394 2826 2756 3042 5831]	[35 9 48 45 8 29 1 3 40 30 47 27 22 4 14]	**0,391407612370588**

Table 4.2: Coherence results applying ’u_mass’ method

Nr Clusters	**Nr Docs Per Topic**	**Most Important Cluster**	**U_Mass Coherence Value**
5	[155518 403771 51375 149880 223113]	[1 4 0 3 2]	**-4,33779250307134**
10	[110750 126811 35591 67146 132106]	[9 7 4 1 0]	**-4,33779250307134**
15	[89983 111688 95254 61631 76894]	[6 8 1 2 0]	**-4,40442466005629**
20	[52508 44862 30655 17672 38979]	[16 14 10 9 17]	**-4,58241436548261**
25	[2831 55178 16730 46439 56094]	[13 21 15 12 19]	**-4,32285400610071**
30	[34671 45387 1597 33375 63162]	[21 4 29 19 17]	**-4,37893087744378**
34	[26814 42677 66176 16056 12739]	[13 2 24 20 1]	**-4,2333334996707**
35	[29223 20357 27731 24280 26846]	[25 29 7 11 22]	**-4,53650110078158**
36	[20523 19406 26957 48253 39265]	[17 34 12 3 13]	**-4,46897904967742**
40	[19014 13704 38413 13469 18450]	[11 21 23 16 22]	**-4,31361219472428**
45	[20828 3577 19163 40735 25021]	[7 31 43 3 42]	**-4,28345426162979**
50	[15160 38435 7716 28284 12368]	[13 24 37 1 27]	**-4,36995066184924**

As we can see in Figure 4.3, and as we had suspected, the coherence for 34 clusters seems to be, at least considering ’u_mass’ method coherence results, the best coherence for this model, since it is the closest to zero value, therefore, we seem to have found our best fit to proceed with the study and get the topic results for the best coherence score number of clusters we found between 5 and 50 clusters.

5 Exploratory Data Analysis

This chapter presents an exploratory analysis results. Before applying any change or topic model to our database, we have done the primary analysis to it with the purpose of understanding a little bit more the data we were dealing with. In this Section we will present the following analysis results: Tweets by month, Tweets by region, Tweets by month and region and Users by region.

5.1 Database

The database used to do this project was provided by the supervisor of this master thesis project, six databases were provided with Portuguese language tweets written in Portugal, each one from one month, between January and June 2021. All databases were combined to one, giving us a total of 1030334 lines with a count of 937164 tweets, written in Portuguese language in a Portuguese region between January and June 2021.

Our Database is made of 8 columns, being those created with the date of the tweet's creation, the id of the tweet, the place_name with the name of the region where the tweet was written, the place_type, the tweet's text, and then we have the user's fields, which are the user, the user_desc and the user_id. The only field where we have null values is the user_desc, all other fields have all the values in the database.

Furthermore, we can see that we are representing in these six months database of tweets, 2168 distinct place names, 4 distinct place types and 18364 distinct users.

5.2 General analysis

This Section do a further investigation in the database so we could be aware of our data before applying any pre-processing or any model to the data, so we can see after in the results Section if the results make sense according to our initial analysis.

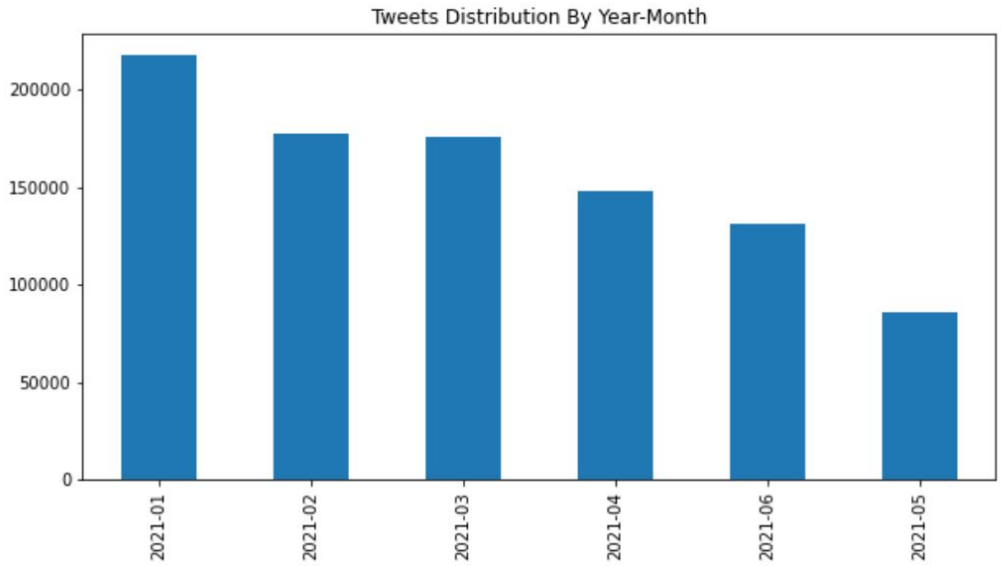

Figure 5.1: Tweets distribution by year-month

5.2.1 Tweets by month

The first analysis made was the amount of tweets that was done in each month, in Figure 5.1 we can see the month distribution of tweets in a descending order. We realized that the majority of tweets was written in January and in February and March, existing a decrease of tweets written in May to almost a half, and starting to increase again in June.

5.2.2 Tweets by region

Secondly, we went to analyze the number of tweets by place name, we select the 20 regions with more tweets written there during these six months. In Figure 5.2 we can observe that Lisbon, Sintra and Porto regions were the three regions with more tweets written there during this time period, values combine make up about 17% of all the tweets of the database.

5.2.3 Tweets by month and region

Another analysis done was the number of tweets wrote by month and by place for the top five places with more tweets written in all the database. We were able to realize by this analysis that most of the database's tweets were written in Lisbon and Sintra, and that the majority of the tweets were written in January for all regions. Between January and May we can see that the count of tweets decrease for all regions, having an increase in June.We could also observe that Lisbon gets the almost the double of tweets in June, comparing to each of the remaining regions. We can observe these results below in Figure 5.3, which

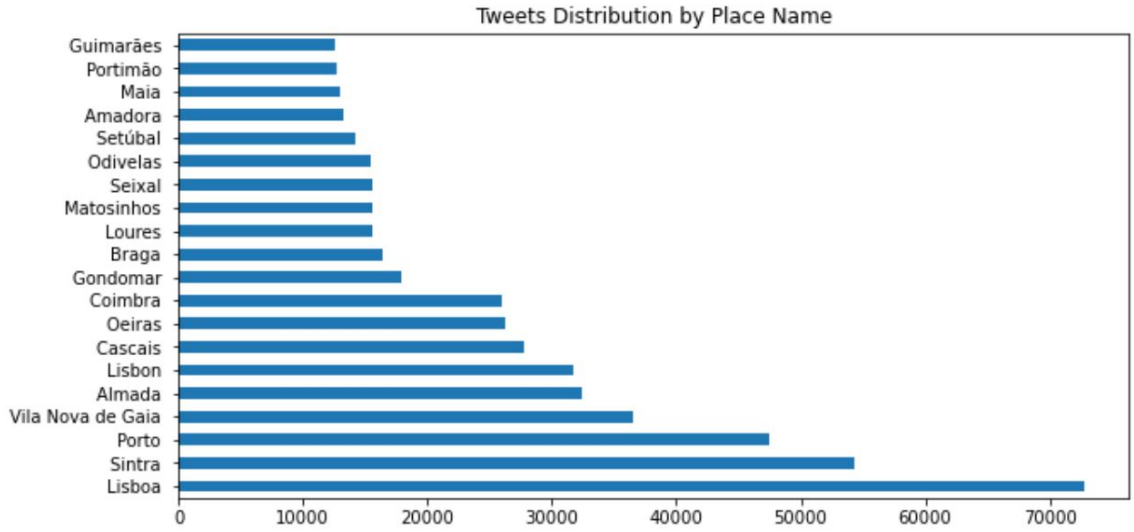

Figure 5.2: Tweets distribution by place

shows the five top places with more tweets written, by month, for the 6 months time period included in our database, from January to June 2021.

5.2.4 Users by region

Finally, the last analysis done was the number of users by region, we have analyze the percentage of users within the regions with more than 1% of the users that wrote the tweets during these six months among the 2168 regions included in the database. We could observe that Lisbon was the one place with more users writing tweets, with 7,76% of the database's users in the city. Sintra and Porto were the other two city places with more users writing tweets there during this time period.

Simultaneously, we could also analyze that 37% of the regions in the database had less that 1% of the database's users there.

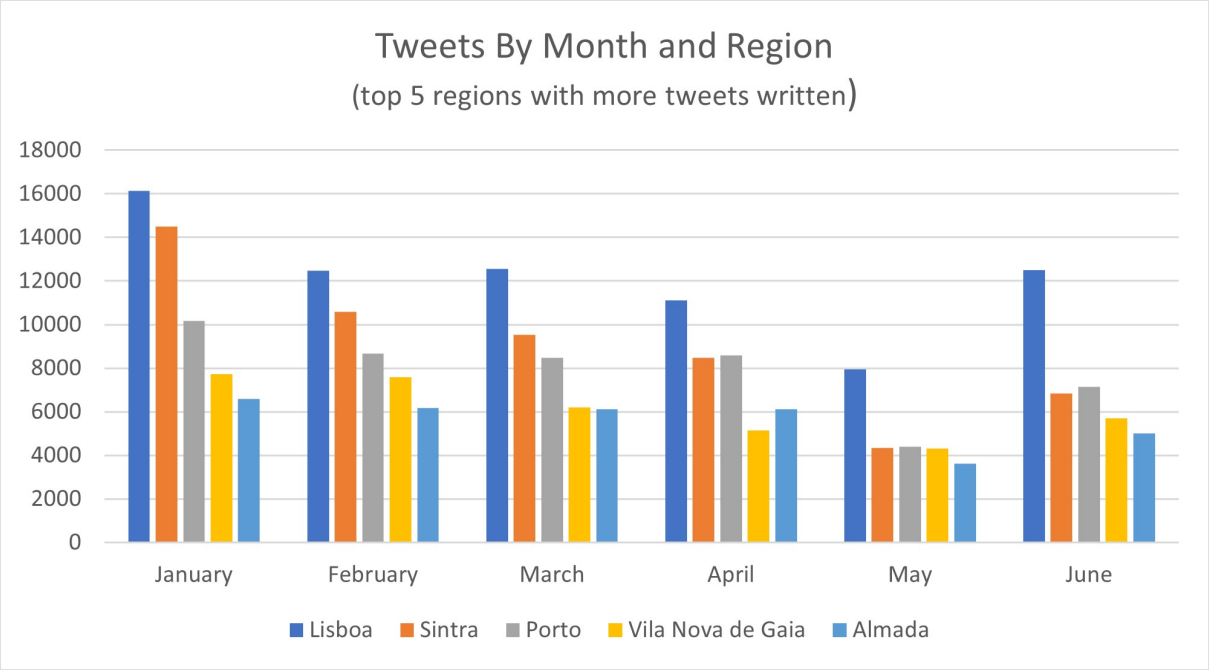

Figure 5.3: Tweets distribution by month and place

Results based on Topic modeling 6

This chapter presents our topic modeling results by region and also point out the most relevant topics we found in each region. We will present, in Section 6.2, the studies results about the Twitter in Portugal, and we will try to characterize the highlighted topics and find the subject topic that seems to match with each of these topics. Finally, we will identify possible market business chances, according to all the previously findings and analysis done after applying the topic modeling to our database.

6.1 Topic results for 34 clusters

This subsection will observe and interpret our final results, obtained by running the Multinomial Mixture model for 34 clusters, best fit topic model and number of clusters combination found for our scenario case, after studying the models, running the tests and calculating an evaluating the coherence scores.

The graphical results interpretation will allow us to answer the missing questions mentioned in our Objectives Section 1.2, being those: **What are the Portuguese people talking about in twitter? do the Portuguese people talk more regularly about specific topics? and are they different according to the region?**

6.1.1 Tweets by month and region

After running the MM model for 34 clusters we were able to finally get the quantity results. We start by doing a quick analysis of tweets percentage in each region divided by month, which we can see below in Figure 6.1, as we can see in the visual, almost all regions, excluding Algarve, had the higher amount of tweets in January. We could also observe that the number of tweets is decreasing near the summer for all the regions in general besides Algarve, since is where the people use to be a lot on June, since is a vacations period, as well as the North.

Let’s now analyze our topic results by region.

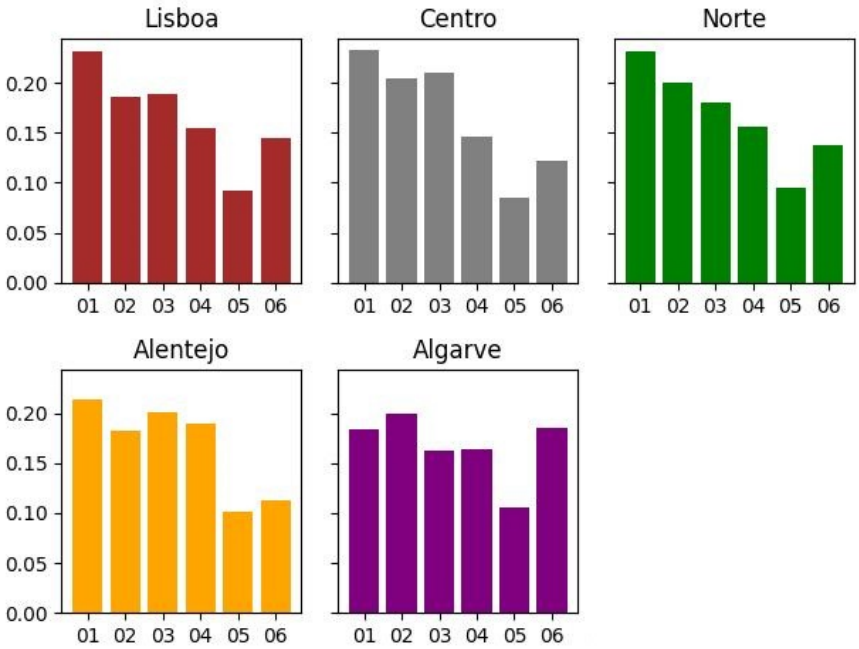

Figure 6.1: Percentage of tweets by month and region for a 34 clusters model

6.1.2 Tweets by topics and region

One of our major objectives for this project, was to understand what the Portuguese people are talking about in each region in Portugal. We can see our topic results by region bellow in Figure 6.2.

These were our first insights about the principal topics talked about in each region:

Topic 1 is one of the most talked about topic in Alentejo and Lisbon and the second most talked about topic in all other regions;

Topic 8 is one of the most talked about in Algave, and has a non relevant frequency when talking about all other regions;

Topic 10 reveals less relevance in Alentejo region, comparing to all other regions;

Topic 18 reveals more presence in the North, comparing to other regions;

Topics 22 gets the third position for the North, Center and Algarve regions, and the fourth position for Lisbon and Alentejo regions;

Topics 23 is the most talked about topics in Lisbon and in the North and shows a lot of less evidence in Alentejo, comparing to all other regions;

Topics 24 reveals much more relevance in Lisbon comparing to all other regions;

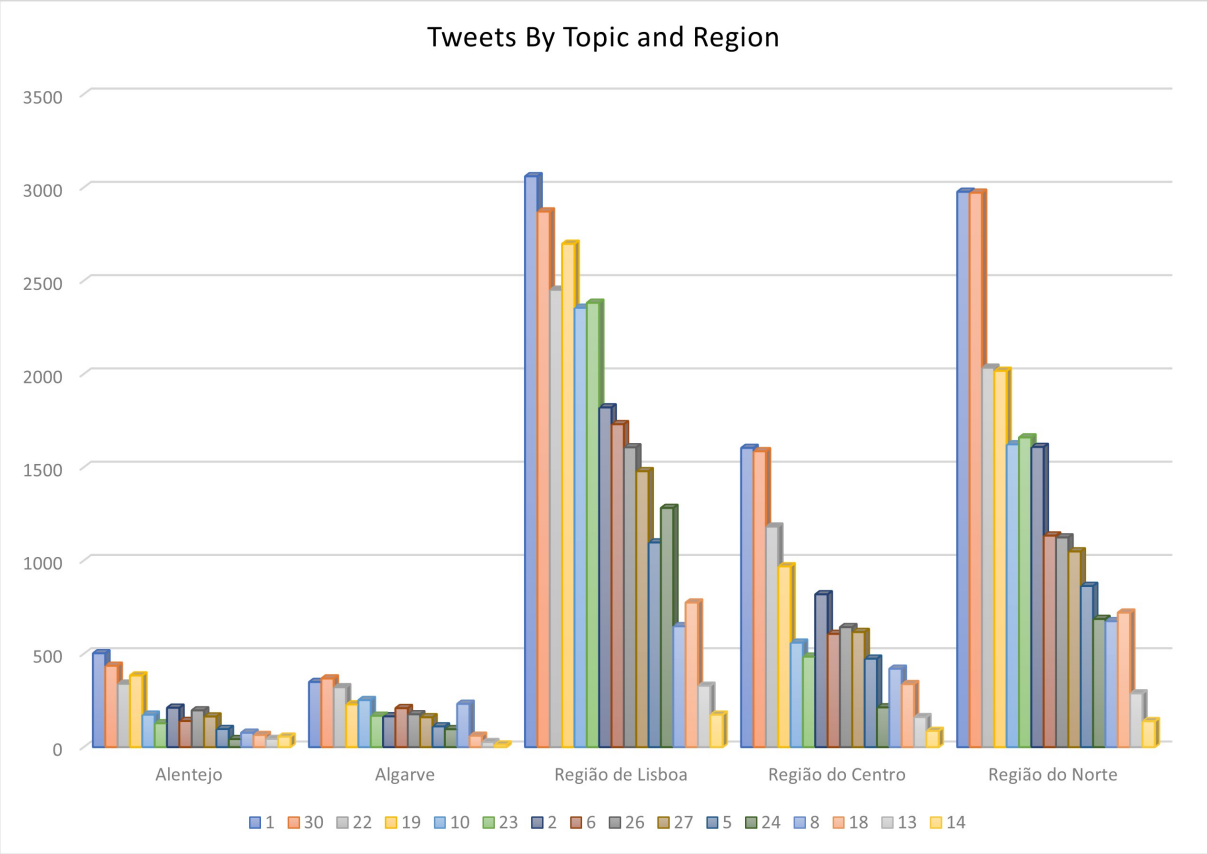

Figure 6.2: Tweets by topic for a 34 clusters model by region

Topic 30 is the most talked about topic in Algarve and is the second most talked about topic in all regions.

To better characterize each topic, we present Table 6.1 with the top words and it is count for each of these topics and the topic name attributed according to the terms interpretation.

In Chapter 7, in Section 6.2.2 we will try to identify specifically what are the topics and the specific subjects that people are talking about in each region.

6.2 Portuguese tweets written in Portugal

The graphical results interpretation allowed us to answer the missing questions mentioned in our Objectives Section 1.2, being those: **What are the Portuguese people talking about in twitter? there is it a different variety of subjects when we change from one region to another? do the Portuguese people talk more regularly about specific topics? and are they different according to the region?**

Table 6.1: The most representative terms for the most representative topics found in each region and the topic name chosen according to these terms and to the texts found in each of the topics IDs.

Topic ID	**(Word, Count)**	**Topic Name**
1	[(’benfica’, 3601), (’sporting’, 3505), (’jogo’, 3244), (’porto’, 3148), (’futebol’, 2215), (’equipa’, 2077), (’clube’, 1978), (’ganhar’, 1892), (’jogos’, 1719), (’jogar’, 1583), (’liga’, 1512), (’anos’, 1486), (’melhor’, 1475), (’jogadores’, 1341), (’portugal’, 1321), (’vamos’, 1200), (’contra’, 1163), (’final’, 1087), (’campeonato’, 1068), (’champions’, 1011)]	Futebol finals games
8	[(’publicar’, 17113), (’foto’, 15334), (’acabei’, 15081), (’portugal’, 7254), (’acabou’, 2724), (’ajuda’, 2226), (’vídeo’, 2051), (’culinária’, 2033), (’tô’, 1607), (’luta’, 1514), (’rt’, 1507), (’contar’, 1505), (’gostaria’, 1505), (’apresentar’, 1496), (’canal’, 1365), (’porto’, 1352), (’puder’, 1231), (’lisboa’, 912), (’lisbon’, 894), (’onde’, 860)]	Food and cuisine posts
10	[(’presidente’, 875), (’portugal’, 694), (’costa’, 617), (’brasil’, 612), (’anos’, 521), (’país’, 478), (’governo’, 472), (’bolsonaro’, 447), (’ministro’, 409), (’pode’, 351), (’lula’, 344), (’lisboa’, 338), (’vergonha’, 336), (’sócrates’, 334), (’bom’, 327), (’porto’, 323), (’vieira’, 323), (’via’, 315), (’josé’, 315), (’contra’, 313)]	Presidents in Brasil and Lisbon
18	[(’comer’, 1978), (’bom’, 945), (’melhor’, 675), (’pão’, 666), (’almoço’, 641), (’café’, 617), (’vinho’, 588), (’água’, 576), (’leite’, 467), (’beber’, 466), (’gosto’, 454), (’queijo’, 452), (’chocolate’, 443), (’casa’, 435), (’arroz’, 419), (’bolo’, 411), (’fiz’, 407), (’vontade’, 389), (’jantar’, 378), (’comida’, 368)]	Food preferences
22	[(’bom’, 5981), (’boa’, 4633), (’amorzito’, 2614), (’noite’, 2076), (’vida’, 2018), (’tarde’, 1938), (’amor’, 1767), (’feliz’, 1630), (’parabéns’, 1214), (’melhor’, 1184), (’vamos’, 1118), (’ti’, 1082), (’obrigado’, 995), (’tão’, 919), (’amo’, 905), (’semana’, 904), (’és’, 885), (’outro’, 870), (’toda’, 795), (’tanto’, 758)]	Love
23	[(’feira’, 1253), (’vamos’, 1159), (’semana’, 999), (’amanhã’, 861), (’portugal’, 820), (’segunda’, 675), (’fim’, 667), (’bom’, 623), (’lá’, 617), (’sexta’, 614), (’bora’, 597), (’jogo’, 565), (’boa’, 488), (’final’, 460), (’grande’, 354), (’volta’, 339), (’melhor’, 333), (’vitória’, 319), (’obrigado’, 314), (’corrida’, 305)]	Sports
24	[(’tá’, 4536), (’to’, 2760), (’tô’, 2520), (’aí’, 1975), (’cara’, 1737), (’bom’, 1553), (’deus’, 1502), (’né’, 1328), (’kkkk’, 1318), (’kkkkk’, 1259), (’porra’, 1229), (’vida’, 1207), (’demais’, 1184), (’casa’, 1160), (’tava’, 1130), (’queria’, 1042), (’kkkkkk’, 1011), (’foda’, 1000), (’todo’, 971), (’pro’, 971)]	Electoral results opinions
30	[(’deus’, 2134), (’vida’, 963), (’amor’, 852), (’portugal’, 498), (’mundo’, 386), (’senhor’, 378), (’jesus’, 360), (’feliz’, 329), (’pois’, 307), (’coração’, 304), (’anos’, 294), (’bom’, 283), (’cada’, 268), (’pessoas’, 263), (’paz’, 262), (’fé’, 254), (’dias’, 252), (’onde’, 248), (’terra’, 246), (’todo’, 245)]	Faith

This subsection will present the answers we were able to get to this question according to our final results.

6.2.1 Twitter in Portugal

In Portugal, 63% of the population uses social networks (slightly more than the European average of 57%). Each of these Portuguese has, according to the 2021 edition of Marktest's study "The Portuguese and Social Networks", about six accounts on different platforms that, in most cases, they visit several times a day, at least one of which to share content.

Data Reportal count the minutes that Portuguese spend on social networks, and concluded that the Portuguese devote every day an average of 2 hours and 30 minutes to social networks. At the end of one year, we are talking about 32.850 minutes (almost 23 days) spent in likes, shares, friend requests, images and texts.

Among the most popular social networks are currently WhatsApp, used by 89% of internet users in Portugal, followed by Facebook, which, according to Marktest, is currently the social network with the highest abandonment rate by users, and finally Instagram. All of them platforms that belong to the Meta group.

Furthermore, conclusions from a 2019 study by Marktest indicate that 54.1% of Portuguese with a profile on social networks admit is to be a fan and follow brands, companies and other interest groups with a presence on Twitter, Instagram or Facebook, with the main reasons of either liking the brand or wanting to keep abreast of the news. According to the same survey, 47.2% of the Portuguese networks users have also the habit of following public Figures on these platforms, being Cristiano Ronaldo the most cited name for these situations.

According with Data Reportal studies, we can see that Twitter's impact in Portugal remains low, when comparing for instance to the US, where we know that most of the internet users uses Twitter. In the US, Twitter has been a favorite network since the beginning, because it's light, conversational (hashtags get thousands of strangers talking about certain topics) and doesn't require high speed connections, having also the possibility to send tweets by sms, without using mobile data. As the people were here, politicians went also to these platforms to get to the public.

According to a Data Reportal study, we were able to confirm that Twitter is used yet just by approximately 30% of the internet users, against approximately 80% of the users using the Meta apps. Additionally, studies revealed that Twitter in Portugal seems to be the youngest network, with 56% of the user's population under the 24 years old. This social network had an increase of 8 percentage points over last year and is expected to continue increasing in the next years among the young.

6.2.2 What are the Portuguese people talking about in each region?

This Section points out the 10 top topics found by the model in our database and we will try to characterize them by looking at the principal words presented in the texts that the model classified as part of that specific topic.

These were the Top 10 topics in it is top order: [13 2 24 20 1 6 5 14 19 26], with this number of documents inside each, respectively: [26814 42677 66176 16056 12739 38867 39374 16782 22046 32220].

Let's now try to get deeper conclusions about each of these topics:

Topic 13: this topic is the most talked about in the North and the less talked about in Lisbon, seems to be related mostly with messages, phone calls and the use of other apps like discord, this subject has not such interest in Lisbon probably because the utilization of sms's and phone calls or any other apps different from Facebook or Instagram, are not much used in the region;

Topic 2: this topic is clearly talking about the Covid time and the restrictions that came up with it, people are texting about how they need to be near it is homes and away from it is friends and how that remains to be a reality in our days after almost 1 year. This topic makes a lot of sense to be one of the topics found with more documents since ours database has data from the beginning of the 2021 year, when the pandemic situation had started about one year ago in Portugal;

Topic 24: this topic is all about politics and the electoral voting, people are manifesting it is political believes and it is opinions about the electoral results in 2021;

Topic 20: this topic has a lot of New Year messages, desires and the self questions or assumptions about how the new year will be for them and for the population in general;

Topic 1: this topic talks mostly about futebol games, and reveals also some comments about films or television shows, commenting the news and it is contents;

Topic 6: this topic is talking mostly about futebol, and criticizing some arbitration. We found also in this topic some politics discussions texts within this topic, most specifically talking about the foreign political situations, being names like Bolsonaro (Brazilian president) and Trump (United States president) mentioned with frequency. These two critical political scenario cases were discussed in all the world so it makes sense that these subjects were also discussed at the social media in Portugal. Furthermore, Twitter is known as as being one of the social media networks that mostly is used to talk about politics around the world, giving consistency to this statement;

Topic 5: this topic is talking mostly about going out to visit places, having a drink, taking some sun, mostly in Lisbon. Some texts about buying needs, mostly clothes, jackets

or training suit is were also found in this topic with less mentions. Buying furniture subject seems also to be included in this topic with even less evidence;

Topic 14: this topic is all about love and friendship messages, people seem to talk about missing or loving someone;

Topic 19: this topic includes texts about family memories, we have a lot of 'mum' or 'dad' mentions, as well as past timing mentions, as 'yesterday', what bring us associate this topic with family memories. Besides that we also found some mentions about entertainment television shows, mostly about the TVI and SIC (the two more seen Portuguese channels), TV shows, like The Mask (from SIC channel) and Big Brother (from TVI channel), which makes sense according to it is both huge popularity at the time, among the Portuguese spectators;

Topic 26: this topic talks about futebol, referring mostly futebol players names like 'ronaldo', 'santos' or 'palmeiras'.

After characterizing the 10 main topics found by the model, we took the other conclusions about the most mentioned topics in each region.

These were the other insights we were able to found between other topics:

Topic 22: is mostly talked about in Lisbon, and talks about love, using a lot words like: love, hapiness, thanks or congratulations;

Topic 23: is mostly talked about in Lisbon and talks about futebol games and other sports, like races;

Topic 10, also one of the most talked about in Lisbon, and one of the less talked about in the North, seems to be related with politics in Lisbon and Brazil, Portuguese people are mentioning a lot the presidents Bolsonaro and Lula names in these topic's texts;

Topic 30, the fourth most talked about in Lisbon and one of the less talked about in Alentejo, it's made mostly on texts about God, faith and peace, revealing that maybe Lisbon would be the region of the country with more religious people. A lot of comments and critics about politic and economic decisions were also found in this topic, which, since we have much more families living in Lisbon than we have in Alentejo, explains also why this subject is more talked about in Lisbon. We have also more workers in Lisbon and a youngest generation living there, generating more controversy among the region's population;

Topic 18, the most talked about in Alentejo, and one of less talked about in Lisbon, is clearly talking about food and food preferences, we see a lot of different types of food mentions and words like 'better', 'drink', 'eat' or 'taste' are the most mentioned in this topic, sustaining this interpretation;

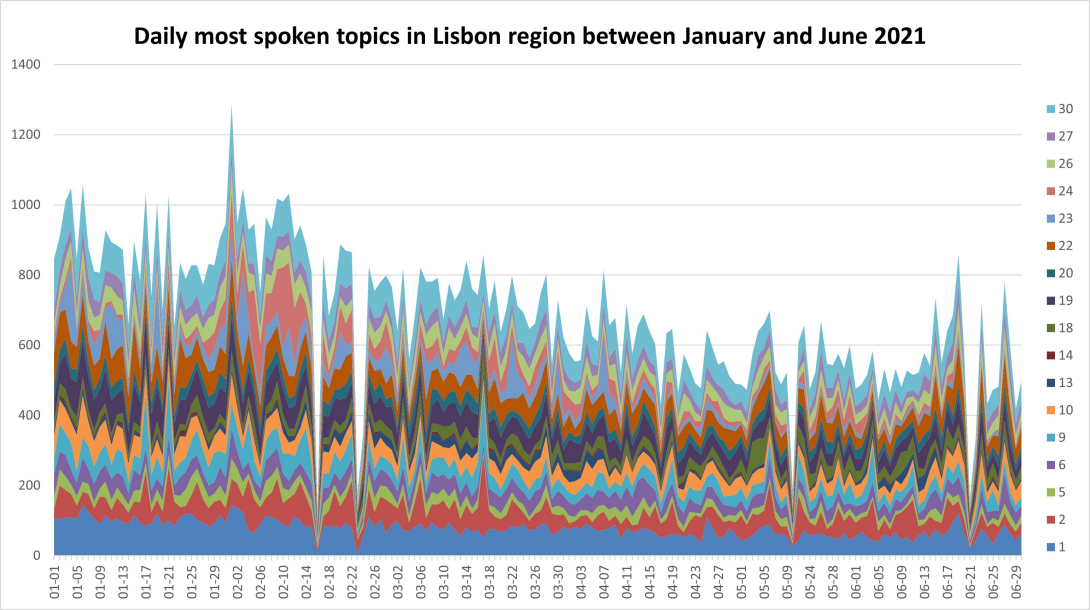

Figure 6.3: Daily topics evolution in Lisbon region during the first 6 months of 2021

Topic 8, one of the most talked about in Algarve and Lisbon, it's related with food and cuisine, and with some photos or videos probably related with the subject publish at the social networks. Texts about training's and online courses on various topics were also found in several texts, including technology, therapy and massage, we can find all of these services both in Lisbon and in Algarve with the frequency, including at the hotels when talking in massages for instance, so this subject being associated with these two regions makes sense and it's expected.

To better characterize each topic, Table 6.2 presents the top terms and it is count for each of these topics that has not been yet characterized in Section 6.1.2.

6.2.3 What is the daily topic evolution and trend in each region?

Lisbon

Figure 6.3 shows us that topic 30, related with faith, has a well define increase of tweets about it at the begging and at the end of January and also in the middle of June. We can also verify that topic 24, related to electoral voting, gets more relevance between the 3rd and the 15th February.

Alentejo

Table 6.2: The most representative terms for the most representative topics found by the model and the topic name chosen according to these terms and to the texts found in each of the topics IDs.

Topic ID	**(Word, Count)**	**Topic Name**
13	[(’és’, 2122), (’oh’, 1846), (’juro’, 1820), (’falar’, 1807), (’olha’, 1576), (’tens’, 1548), (’gajo’, 1427), (’aí’, 1398), (’então’, 1385), (’lá’, 1373), (’tá’, 1353), (’tas’, 1344), (’ahahah’, 1260), (’bro’, 1258), (’tão’, 1218), (’ti’, 1213), (’bom’, 1150), (’mal’, 1149), (’vais’, 1129), (’sabes’, 1105)]	Random messages
2	[(’pessoas’, 6726), (’vida’, 3800), (’pessoa’, 3149), (’coisas’, 2964), (’tão’, 2253), (’coisa’, 2074), (’vezes’, 1915), (’mal’, 1890), (’falar’, 1765), (’alguém’, 1731), (’ninguém’, 1677), (’melhor’, 1643), (’outros’, 1630), (’têm’, 1486), (’cada’, 1440), (’saber’, 1399), (’vez’, 1383), (’tempo’, 1372), (’tipo’, 1287), (’vcs’, 1275)]	Daily life on Covid time
20	[(’dormir’, 5101), (’casa’, 2330), (’tão’, 2282), (’amanhã’, 2177), (’bom’, 1812), (’acordar’, 1776), (’manhã’, 1648), (’queria’, 1628), (’sono’, 1622), (’ficar’, 1608), (’noite’, 1591), (’acordei’, 1553), (’dias’, 1443), (’cedo’, 1417), (’mal’, 1366), (’tempo’, 1345), (’vontade’, 1342), (’cabeça’, 1334), (’consigo’, 1288), (’cama’, 1221)]	New Year messages
6	[(’jogo’, 3096), (’golo’, 2359), (’puta’, 2123), (’porto’, 1769), (’benfica’, 1437), (’bola’, 1415), (’sporting’, 1096), (’falta’, 1068), (’jogar’, 960), (’crl’, 934), (’amarelo’, 833), (’vamos’, 761), (’árbitro’, 759), (’lá’, 758), (’gajo’, 746), (’marcar’, 711), (’contra’, 710), (’penalti’, 700), (’foda’, 685), (’jogador’, 645)]	Futebol games arbitration
5	[(’casa’, 2476), (’saudades’, 2308), (’lá’, 1547), (’vamos’, 1148), (’alguém’, 1062), (’praia’, 1005), (’onde’, 790), (’aí’, 780), (’queria’, 767), (’beber’, 764), (’amanhã’, 750), (’bom’, 739), (’sol’, 714), (’semana’, 712), (’sair’, 690), (’vir’, 674), (’vem’, 657), (’lisboa’, 638), (’cá’, 612), (’noite’, 611)]	Going out to visit Lisbon
14	[(’linda’, 3314), (’bom’, 2714), (’tão’, 2205), (’lindo’, 1845), (’amo’, 1713), (’obrigado’, 1544), (’és’, 1522), (’obrigada’, 1340), (’coisa’, 1177), (’deus’, 935), (’parabéns’, 892), (’amor’, 860), (’juro’, 669), (’saudades’, 599), (’lt’, 595), (’melhor’, 577), (’mulher’, 551), (’adoro’, 547), (’amiga’, 509), (’gosto’, 497)]	Love and friendship
19	[(’casa’, 1811), (’vez’, 1548), (’anos’, 1350), (’lá’, 1243), (’mãe’, 1194), (’carro’, 1002), (’ia’, 943), (’disse’, 856), (’tempo’, 770), (’pai’, 761), (’quase’, 749), (’vi’, 671), (’hora’, 660), (’ontem’, 660), (’outra’, 641), (’coisa’, 623), (’tava’, 610), (’fiquei’, 600), (’vida’, 585), (’andar’, 544)]	Family memories
26	[(’jogo’, 2384), (’melhor’, 1919), (’jogar’, 1333), (’time’, 1242), (’jogador’, 999), (’tá’, 862), (’bom’, 860), (’joga’, 835), (’bola’, 827), (’contra’, 806), (’campo’, 782), (’santos’, 780), (’pode’, 753), (’ronaldo’, 706), (’gol’, 674), (’futebol’, 671), (’flamengo’, 665), (’seleção’, 664), (’pro’, 642), (’palmeiras’, 635)]	Futebol players

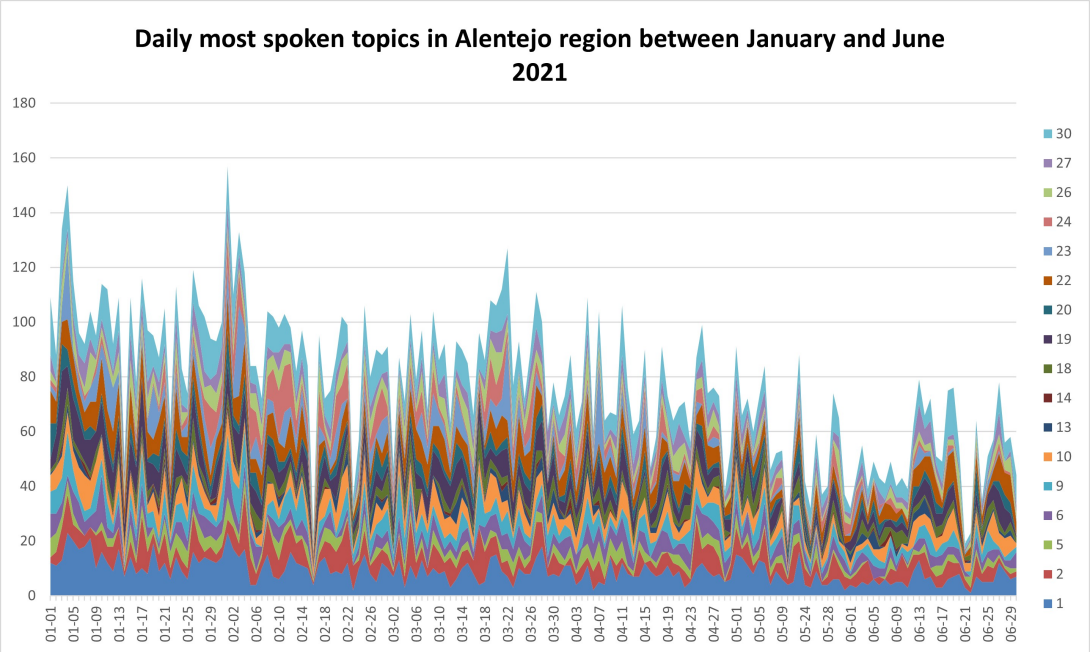

Figure 6.4: Daily topics evolution in Alentejo region during the first 6 months of 2021

Figure 6.4shows us that we have a lot of mentioned topics changes, within each day, in Alentejo. We see that topic 24, related to electoral voting, only has relevance in the region between 25th January and 22th March, with the focus between 25th January and 22th February. We also see four higher existing pics for topic 30, about faith, in specific dates as the 3rd January, 30th January, 2nd February and 24th March, day's where is likely to exist religious events. We notice clearly a pic for topic 10, about the presidents in Brazil and Portugal, in the 3rd January, and of topic 26, about Futebol players , in the same day and at the 2nd February.

Algarve

Figure 6.5show us that topic 30, has more pics in Algarve that as in the other regions, showing us that is likely that Algarve has more religious events or more religious people comparing to the other regions. Besides that we found a similar behavior of Lisbon and Alentejo, with topic 24, related with electoral voting, having the most relevance between the end of January and the middle of February.

North

Figure 6.6reveals some similarity in the daily topics distribution comparing to the other regions, but we see a difference with the timing for the topic 30, that is having mostly is pic between January and February for all other regions, and is happening in the 8th March in the North region.

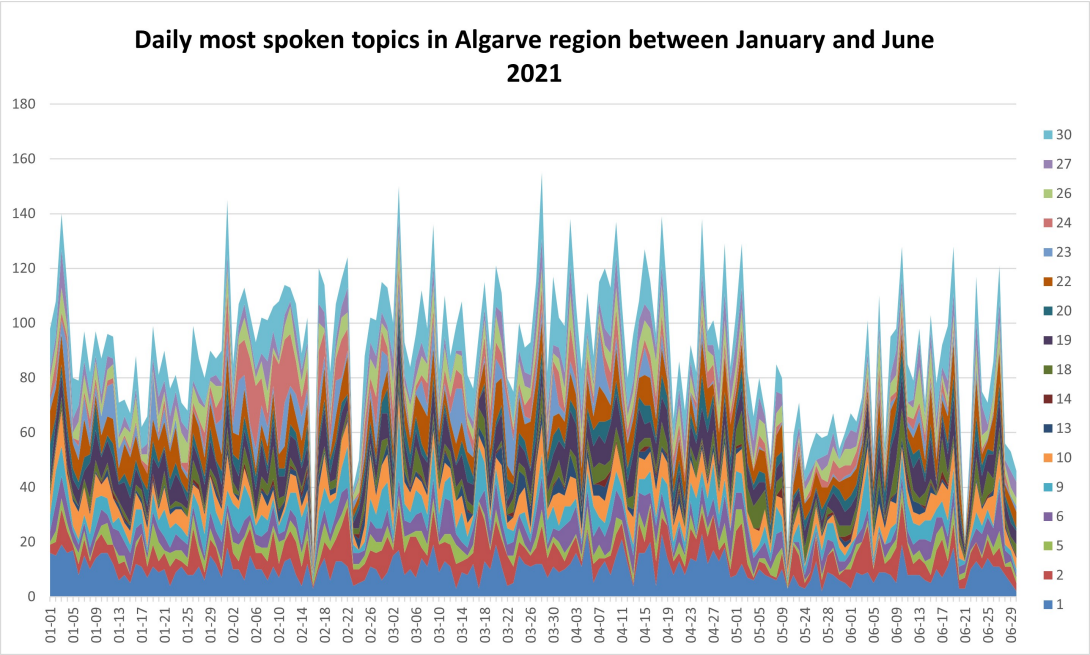

Figure 6.5: Daily topics evolution in Algarve region during the first 6 months of 2021

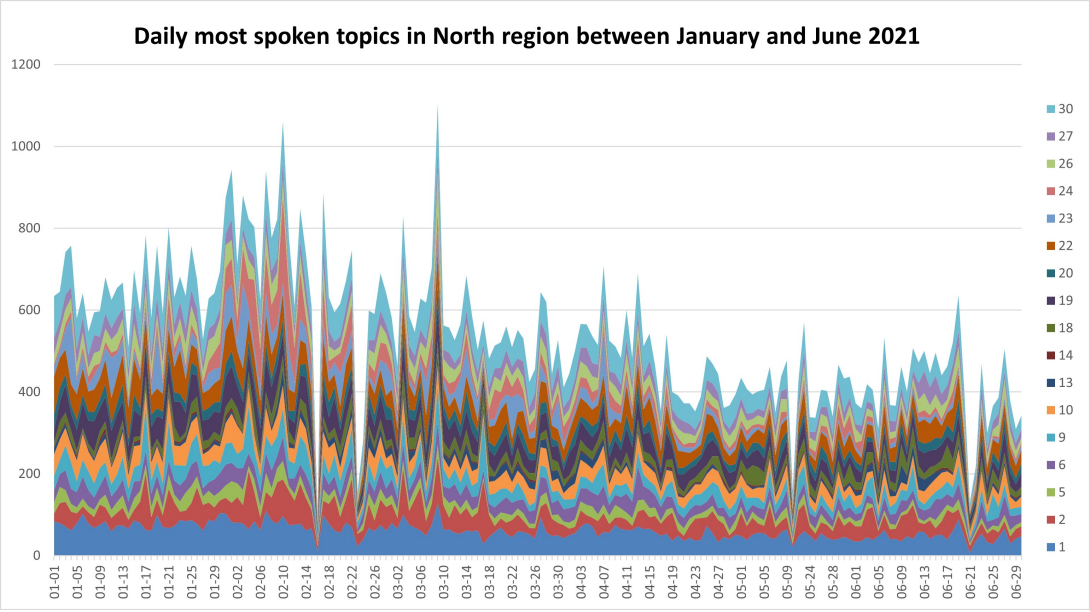

Figure 6.6: Daily topics evolution in North region during the first 6 months of 2021

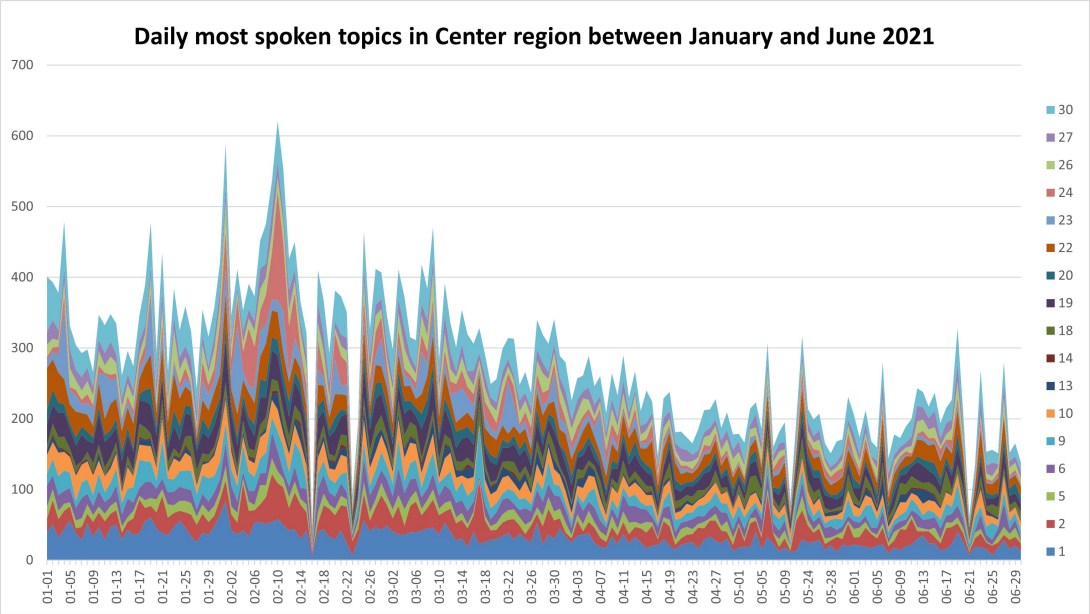

Figure 6.7: Daily topics evolution in Center region during the first 6 months of 2021

Center

Figure 6.7 shows us that the Center's region daily topics distribution has a huge similarity with the Lisbon region, being the users behavior apparently similar in both regions, what is comprehensive since both regions are closest to each other geographically talking, comparing to all other regions.

7 Conclusions

This final chapter presents both our final conclusions and insights about what Portuguese people are talking about, in each region of the country, during the first semester of 2021, and about topic modeling applied to short text corpora.

7.1 Topic modeling application insights

As was mentioned in Section 1.2, one of this research's objective was to answer the following question:

Will the MM model actually work better with our short text data, comparing to the LDA model, since the first one is known by is better performance with short text since it assumes that each document is part of one only specific topic?

After applying both models to our data, and measuring the coherence of the results obtained from both models, we could then conclude, by analyzing the coherence values and also looking at the representative words in each cluster, for both models, that the performance of the MM model was better than the LDA model, with a much higher topic quality. These results made us assume that the Multinomial Mixture model is a better model for short text than the LDA. However, considering the unsupervised nature of the topic models, this evaluation remains remains a major challenge. Nevertheless, based on our tests, the results of our research reveals that the MM model is a good option when talking about short text analysis.

Nonetheless, we remains have into consideration, that other evaluation methods should also be took into account to fully assess all the aspects of the MM model, such as it is classification and generalization capabilities, comparing to those of the LDA.

7.2 Topics among the country: some marketing insights

The application of information retrieval tools on tweets is nowadays a subject of very interest, as it can provide the markets insights about current trends, public interests or the public reaction to breaking news for instance [14]. As an example, we have the business

managers of companies, that may use sentiment analysis for quality control of the clients comments in the internet, using that information to match if it is services or products are addressed to the consumer needs, as well as knowing the current consumers opinions about it. From a project like ours, we can also figure out the topic trends by Portuguese region, and that would be interesting has a business insight for the companies to know what are the best products to sell in each regions in Portugal, based in the region's population interests revealed in it is written tweets.

Having that said, and considering the results found in Sections 6.2.2 and 6.2.3, we can write down the principal conclusions and possible business purposes for the Portuguese companies:

1. Politics, Religion and Futebol seem to be the most talked about topics in all the country regions, being good options of business industries to invest on, considering the most talked about topics on Twitter;

2. Lisbon and Center reveal similar daily topics distribution behavior, being both regions where probably the marketing companies for the general subjects could be together, being the specific geologically targeting campaigns more interesting to do at regions like Alentejo, Algarve or the North;

3. Algarve seems to have more pics about the religious subject, so companies related to the area should be attempt to this situation when doing it is business strategies;

4. Food and cuisine is also one topic that is frequently addressed in the general country, presenting an higher importance in Algarve, Lisbon and in the North region, being the best targeting for this subject;

5. Politics discussions about Brazilian and Portuguese presidents present more relevance in the Alentejo region, being likely to be a good place to invest in electoral events.

Bibliography

[1] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improvinglda topic models for microblogs via tweet pooling andautomatic labeling," *In Proceedings of the 36th international ACM SIGIR conference on Research anddevelopment in information retrieval*, p. 889–892, 2014.

[2] C. Lin and Y. He, "Joint topic model for sentiment analysis," *Computer Science Proceedings of the 18th ACM conference on Information and knowledge management*, 2009.

[3] B. J., M. H., and Z. X., "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, p. 1–8, 2011.

[4] B. Xiang and L. Zhou, "Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training," *ACL 1 Computer Science*, 2014.

[5] A. De Waal, "Topic models with structured features," *Ph.D. thesis, North-West University, Potchefstroom Campus*, 2010.

[6] A. Huang, D. Milne, E. Frank, and I. Witten, "Clustering documents using a wikipedia-based concept representation," *Advances in Knowledge Discovery and Data Mining*, p. 628–636, 2009.

[7] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries," *Proceedings of the 10th annual joint conference on Digital libraries, pp. 215–224. ACM*, 2010.

[8] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research, vol. 3, pp. 993–1022.*, 2003.

[9] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "In-dexing by latent semantic analysis," *Journal of the American Society of Information Science, vol. 41, no. 6, pp. 391–407*, 1990.

[10] T. Hofmann, "Probabilistic latent semantic analysis," *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, p. 289–296, 1999.

[11] ——, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1-2, p. 177–196, 2001.

[12] J. Canny, “Gap: a factor model for discrete data,” *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 122–129, 2004.

[13] L. Hong and B. Davison, “Empirical study of topic modeling in twitter,” *Proceedings of the First Workshop on Social Media Analytics, pp. 80–88. ACM*, 2010.

[14] D. Khartabil, “Data mining and visualisation of twitter using topic modelling,” 2013.

[15] A. Bermingham and A. Smeaton, “On using twitter to monitor political sentiment and predict election results,” 2011.