Multilingual trend analysis. Probabilistic and clustering methods

Álvaro Domínguez Calvo
Universidad Nacional de Educación a Distancia
dominguezcalvoalvaro@icloud.com

Contents

Abstract

In this work we study two methods to detect topics and two clustering algorithms to agglomerate documents semantically related. We propose a bilingual data set to assess if there are topical and clustering differences. The dataset consists of Spanish and English tweets related to COVID and the eruption of the La Palma volcano events. We conclude that there exists topical differences regarding the language because of the lack of data of La Palma tweets. On the other hand, we find that the text normalisation worsen the clustering quality. Lastly, we conclude that it is not possible to stablish differences between the quality of the clusters regarding the language constraint since the proposed corpus is not parallel at language level.

1 Introduction

To detect topics within textual data is useful in the context of social media since it could be possible to track real-time events to monitor, for example, natural disasters or study the reputation of influencers. This task is known as trend analysis and aims to automatize the topic detection within a set of documents.

Supposedly, the results of the application of trend and clustering methods over multilingual data should yield similar results since a topic, semantically, refers to the same concept regardless the language of the documents. Thus, we drive a bilingual trend and cluster analysis in order to determine whether there are differences in the proposed methods.

This work is structured as follows. In the section 2 we describe the methods for the trend analysis. We consider two approaches based on word frequency distributions: The Kullback-Leibler Divergence and a generative model, the Latent Dirichlet Allocation. In section 3 we drive the task of cluster analysis in order to assess the cluster qualities in terms of intrinsic metrics. We consider two clustering algorithms. In the section 4 we propose and describe a bilingual data set of tweets. In the section 5 we propose a set of experiments based on six questions regarding the trend and cluster analysis. In the section 6 we expose the results of the experiments. Finally, we conclude in the section 7.

2 Trend analysis: Description of the methods

Trend analysis aims to uncover topics that are of interest at a given instant of time. In the context of textual data, this analysis can be carried out if we measure the importance of a word regarding of how many times that word appears in a set of documents in a specific moment. There exist methods to determine trending words or latent topics within a collection of documents. In this work, we will use the Kullback-Leibler divergence to measure the importance of a set of words across time and the Twitter Latent Dirichlet Allocation proposed in Zhao et al. (2011) to cluster words that are thematically related.

2.1 Trending words: Kullback-Leibler Divergence

The Kullback-Leibler Divergence (KLD) is defined as a statistical distance that measures the difference between two probability distributions. Formally, let $P(x)$ and $Q(x)$ two probability distributions. The KLD denoted as $KLD(P \mid\mid Q)$ is defined as

$$KLD(P \mid\mid Q) = \sum_{x \in \mathcal{X}} P(x) log(\frac{P(x)}{Q(x)}) \tag{1}$$

In the context of trend analysis, x is a word contained in the vocabulary $V = \{v_1,\ v_2,\ ...,\ v_k\}$ obtained from the set of documents $D = \{d_1,\ d_2,\ ...,\ d_n\}$. However, to compute P and Q given the word x, we need to calculate the probability distributions at different states of the word. These word states corresponds to the probabilities of that word at different instants of time. Hence, $P(x)$ and $Q(x)$ are probability distributions of the word x regarding different dates, hours or the metric time considered.

2.2 Topic modelling: Latent Dirichlet Allocation

The Latent Dirichlet Allocation or LDA, was originally proposed by Blei et al. (2003). LDA is a generative model that aims to compute the probability of a document $d_i \in D$ regarding that there exist n latent topics within the set of documents D. This model assumes that there exists a topic distribution over the words and over the documents. However, it is appropriate to use this model when documents are large and well written. Regarding short texts, such as tweets, misspellings and low word counts should be taken into account. Zhao et al. (2011) propose a LDA model to compute a background word distribution and assume that there exists a topic distribution per Twitter user. The figure 1 summarizes the Twitter LDA model. The parameters, from the left to the right, α_g, β_{word}, β_b and γ are parameters of a Dirichlet prior distribution.

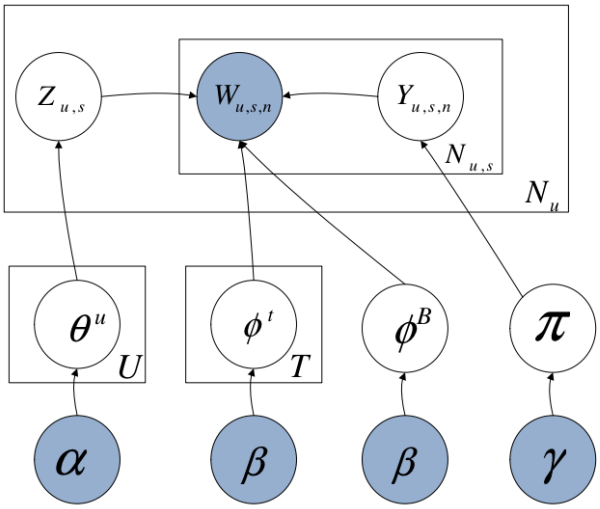

Figure 1: Plate notation of the Twitter LDA model from Zhao et al. (2011).

However, the computation of the document probability is intractable, and sampling methods must be considered. The authors of the Twitter LDA model used the Gibbs sampling to draw instances from the probability function defined in the figure 1.

2.3 High-level comparison of the approaches

LDA clusters words topically without considering the publishing date of the documents. If we group the documents by date, it could be possible to run different independent runs of the model to retrieve topics per date. However, this approach leads to have different vocabularies for each run. In contrast, the KLD computation considers the existence of frequency word distributions over time, allowing to assess the word importance based on word counts in a given date. This importance could be exploited to study tweets related to that word and near the date they are published. To summarize, LDA aims to discover words that are within the same linguistic semantic field, whilst the KLD measures the trending words over time.

LDA yields richer semantic information than KLD since, for example, it allows discovering the number of times a user of Twitter mentions a topic because it assumes that there exists a topic distribution per user. In our work, a user corresponds to a data set type. As we shall see, we propose two types of data sets. Furthermore, LDA allows deleting those words that are not semantically relevant. In turn, the KLD computation does not delete these words and a deletion processing step must be considered.

3 Cluster analysis methods

Cluster analysis aims to group a set of objects within sets that contain similar objects to each other, but are dissimilar to the objects contained in other sets. These type of unsupervised machine learning algorithms compute the clusters considering a similarity measure and a matrix that contains the objects to be clustered. However, the number of clusters, k, must be known beforehand.

There are many types of clustering algorithms like divisive, agglomerative, graph-based etc, that compute the clusters differently. In this work, we consider a divisive algorithm and a graph-based one provided by the CLUTO tool, which is proposed by Zhao and Karypis (2001). This tool is appropriate when working with high dimensional objects. In this case, the objects correspond to documents where the feature space is high, since is defined by the vocabulary of terms retrieved from all the documents.

3.1 Divisive algorithm

This type of clustering computes the groups by dividing the document matrix in two clusters. Then, the algorithm divides these two clusters until k clusters are computed. In each division, the algorithm optimizes a criterion function within each bisection, i.e., locally. When the algorithm finishes, it optimizes the same criterion function but globally, i.e., across all the clusters. The CLUTO tool implements the repeated bisection in its RBR algorithm.

3.2 Graph-based algorithm

Graph based clustering considers a weighted non-directed graph with a set of vertices V and a set of edges E, namely $G = \langle V,\ E \rangle$. The vertices correspond to the objects to be clustered, whilst the edges are the similarities among the objects. The CLUTO algorithm, named as *graph*, computes the graph in a nearest-neighbor fashion. Then, the algorithm divides the graph onto k clusters with a graph partitioning algorithm.

4 Data set description

The data set consists of 281.332 tweets retrieved from the Twitter API. We used the keywords *la palma* and *covid* to retrieve the tweets. We consider two types of data sets named as COVID data set and La Palma data set. For each data set, we retrieved tweets written in Spanish and English. Thus, we built four data sets that are summarized in the table 1.

The tweet crawling consists in storing the tweet identifier. We started the tweet crawling on 23rd November 2021 and finished it on 28th December 2021. However, we retrieved the same tweets by the tweet identifier on 19th March 2022. This involves loss of data due to numerous tweets were deleted.

		Language		
		Spanish	English	Total
Type	COVID	120.109	123.144	243.253
	La Palma	33.868	4.211	38.078
	Total	153.977	127.355	281.332

Table 1: Summary of the available tweets per data set type and language.

5 Experiments

We propose experiments to study differences of the proposed methods and to assess whether there exist variations in function of the language of the data sets. To achieve this, we address the following questions:

1. How the KLD values are distributed across the days regarding the type of data set (COVID or La Palma)? Are there differences?
2. Are there differences in the inferred topics by LDA between the Spanish and English data sets?
3. Can KLD be compared with the LDA approach? Are there KLD trending words contained in some topics discovered by LDA?
4. Are there differences between the graph and divisive clustering methods in terms of intrinsic metrics?
5. Does the performance of the clustering improve when we consider a text normalisation technique?
6. Are there differences in the clustering quality regarding the language of the data sets?

We processed the tweets by removing all the links and numbers. The whitespace is the boundary of a token, and we considered tokens only with letters. On the other hand, the normalisation consists

of three steps. For a tweet: (1) Lowercase all the tokens; (2) Lemmatization; (3) Stem all the tokens. We executed the normalisation step with Spacy[1] in the given order.

On the other hand, we built $TF-IDF$ matrices for each data set for the computation of the KLD and clustering. The $TF-IDF$ scheme weights each word regarding of the term frequency (TF) and inverse document frequency IDF. Formally,

$$TF(w,\ d) = \frac{f_{w,\ d}}{\sum_{w' \in d} f_{w',\ d}} \tag{2}$$

where w is a term and d is the document where t appears. The IDF is given by

$$IDF(w) = log(\frac{|D|}{1 + df(w)}) \tag{3}$$

where D is the set of documents and $df(w)$ is the document frequency of the term w.

To control the corpus-specific stop words, we considered the parameters max_{df} and min_{df} in the construction of the $TF-IDF$ matrices. max_{df} permits to exclude those words when building the vocabulary that have a document frequency strictly higher than its value. On the other hand, min_{df} removes those terms that have a document frequency strictly lower than the given threshold. In this work, we have considered different values for these parameters in function of: (1) The preprocessing applied over the data set; (2) The language of the data set. The table 2 shows these values.

		Spanish		English	
		COVID	La Palma	COVID	La Palma
Raw	max_{df}	0,01	0,01	0,01	0,01
	min_{df}	0,0025	0,0025	0,0025	0,0025
Normalised	max_{df}	0,07	0,07	0,07	0,07
	min_{df}	0,023	0,0002	0,0023	0,0022

Table 2: max_{df} and min_{df} values for raw and normalised data sets.

5.1 KLD experiment

The KLD experiment consist of simply compute the KLD of a set of words across time. However, we must first define the probability distribution of the words over time. Regarding the publication time of the tweets, to compute the distributions, we define a function to retrieve the probability of a word w_i given an instant of time t. That is

$$p(w_i)^t = \frac{count\_words(w_i,\ t)}{\sum_{i=1}^{|W^t|} count\_words(w_i,\ t)} \tag{4}$$

where $count\_word(w_i\ t)$ denotes the number of times the ith word appears in the instant of time t. W^t is the set of words that appear in the instant of time t. Thus, the denominator is the total number of words that appear in the given instant of time.

In this work, the term t refers to a specific hour of a given day. Per hour, we compute the probability p. That is, we retrieve the frequency distribution of the word in a given day. Formally, if we divide a day D_j in a set of timings $D_j = \{t_1,\ t_2,\ ...,\ t_k\}$, the probability distribution of w_i of that day, $P(w_i,\ D_j)$ is computed as

$$P(w_i,\ D_j) := [\ p(w_i)^{t_1},\ p(w_i)^{t_2},\ ...,\ p(w_i)^{t_k}\ ] \tag{5}$$

Thus, given the dates D_k and D_j where $k > j$ and given a word w_i, the KLD is given by

$$KLD(P(w_i,\ D_k)\ ||\ P(w_i,\ D_j)) = \sum_{x \in \mathcal{D}} P(w_i,\ D_k) \cdot log(\frac{P(w_i,\ D_k)}{P(w_i,\ D_j)})$$

where D is the timing division of the days. In this case, the hours of a day. With this KLD definition, it is possible to compute the KLD given a set of words. The considered set of words is the vocabulary inferred in the construction of the $TF-IDF$ matrices over the normalised data sets.

[1] https://spacy.io/

5.2 LDA experiment

The experiment consists in executing the algorithm with a specific parameter configuration over the normalised tweets. The parameter values are $alpha_g = 0.6$, $beta_b = 0.01$, $beta_{word} = 0.01$, $gamma = 20$, $iteration = 20$. The number of topics must be known beforehand. We considered 100 topics.

On the other hand, we remove all tokens whose length is lower than two. We group the tweets by language regardless of the entity, that is, we merge the COVID and La Palma tweets for each considered language and then, we execute the Twitter LDA model over the generated data set.

From the execution result, we assess the word within topics inferred by the model and the topic weights for each word. Besides, we analyse the topic distribution per user, i.e. for the COVID and La Palma data sets.

5.3 Clustering experiments

The experiments consist of executing the two commented clustering algorithms over the raw and normalised tweets for the two considered languages. The clustering is executed with the $TF-IDF$ matrices yielded with the parameter configuration of the table 2. To measure the quality of the clusters, we use intrinsic metrics provided by CLUTO that are: (1) The average similarity within each cluster or internal similarity, $ISim$; (2) The average standard deviation of the average internal similarity, $ISdev$; (3) The average external similarities, $ESim$. On the other hand, we use the cosine similarity as the similarity function, which is given by

$$sim(t_i,\ t_j) = \frac{t_i \cdot t_j}{||t_i|| \cdot ||t_j||} \tag{6}$$

where $t_i \cdot t_j$ denotes the scalar product and $|| \cdot ||$ denotes the module of a vector.

6 Analysis of the results

The analysis will consist in comparing the results obtained for each approach. For the KLD analysis, we rank the words by the summation of the KLD scores across the considered days. We only show the first 85 words from that ranking. On the other hand, the LDA result analysis consists in assessing the first twenty-five topics inferred by the model and evaluating the generated topic distribution per data set. We perform the analysis for the two considered languages.

6.1 Analysis of Kullback-Leibler Divergence results

We drive the KLD analysis by plotting a heatmap for each data set to assess what words show high values of the KLD score across the days.

Spanish COVID data set. The figure 2 depicts the heatmap for the first 85 words of the ranking words generated by the KLD summation over time. There exist some words related to the linguistic semantic field of the COVID disease. Words like *asintomat*, *brot* which comes from *brote*, meaning outbreak or *antigen* are COVID-related words. Even there are named entities like *sanchez* which refer to the President of the Spanish government. Note that there is a huge word activity in days near to Christmas. If fact, there is no variation of the KLD scores for almost every word in these days.

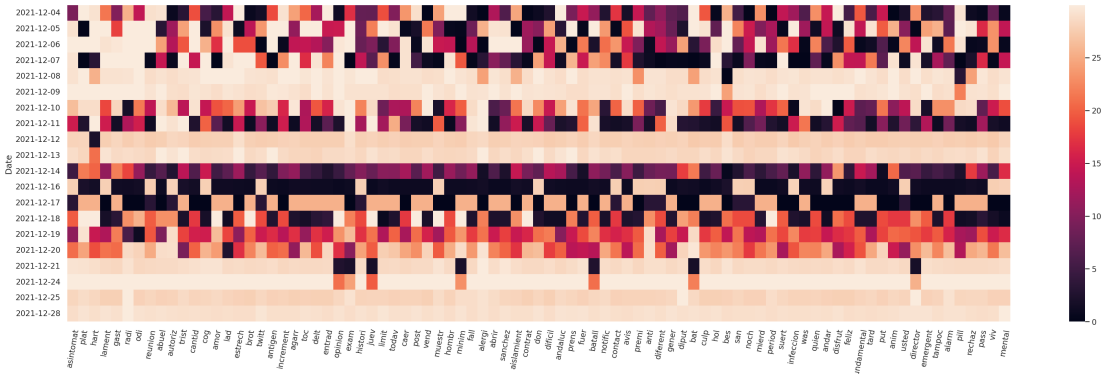

Figure 2: KLD values of the top 85 words from the list of words ordered by the summation of KLD across the dates. On the X, axis we plot the words. The Y axis shows the dates contained in the Spanish COVID data set.

English COVID data set. The figure 3 depicts the heatmap with the same characteristics as the commented above. Note that all the words remain important across the dates except for the 20th of December. Equally, there exist words related to the COVID disease like *infect*, *safety*, *virus* or *coronavirus*.

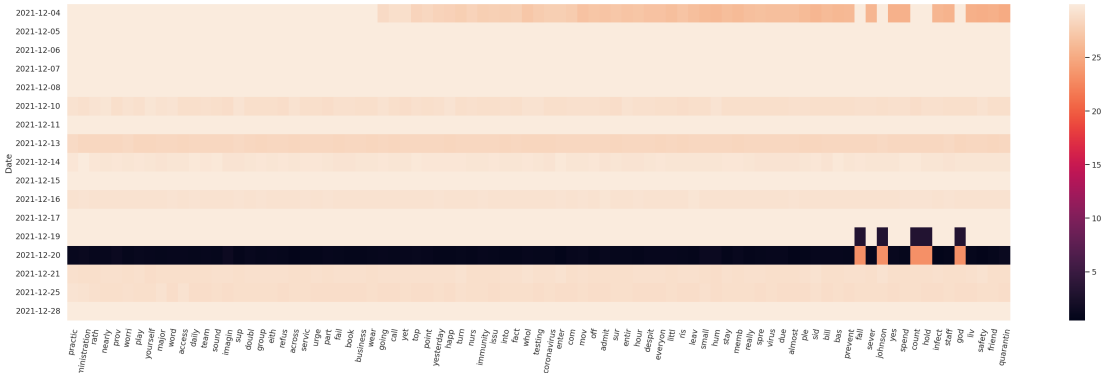

Figure 3: KLD values for the top 85 words from the English COVID data set.

Spanish La Palma data set. The figure 4 depicts the heatmap with the same characteristics as the commented above. There are only two words that relate to an event, e.g. *volcanlapalm* refers to the eruption of the La Palma volcano. The remaining words do not have topical senses.

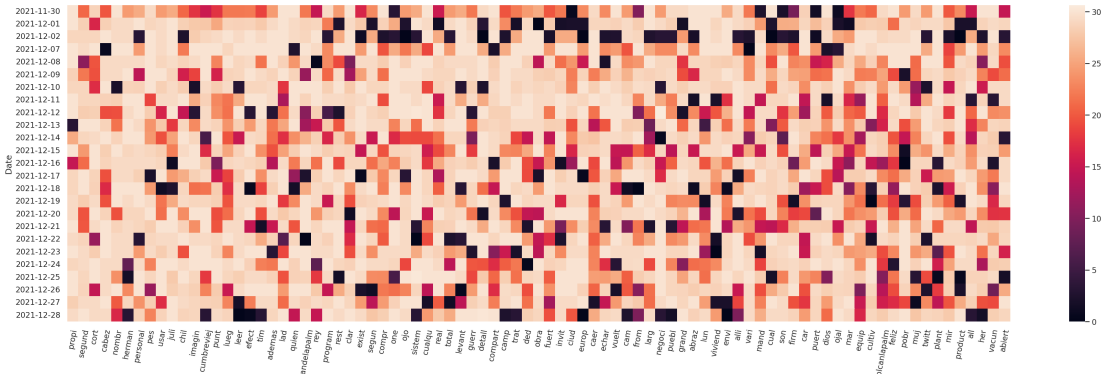

Figure 4: KLD values for the top 85 words from the Spanish La Palma data set.

English La Palma data set. Lastly, the figure 5 depicts the heatmap for La Palma data set. Surprisingly, there are more topical words in this data set in comparison to the Spanish one. Words like *erupt*, *activity*, *destroy* or *hous* (the lemma of house) may refer to the volcano eruption. However, there are still some stop words like *than* or *but*.

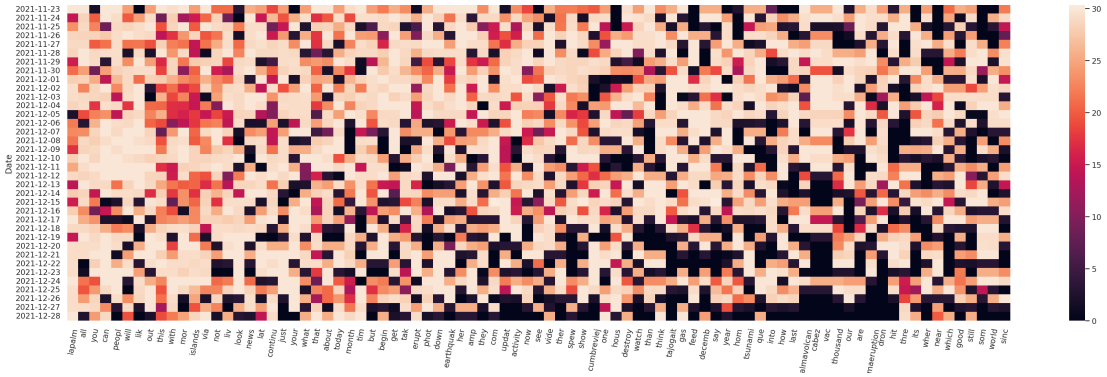

Figure 5: KLD values for the top 85 words from the English La Palma data set.

6.2 Twitter LDA analysis

With this analysis, we address the question 2 of the proposed experiments. The analysis will consist in comparing the generated topics for the proposed languages regardless of the type of the data, i.e. whether it is the COVID or La Palma data set. On the other hand, we analyse if there exist differences between the generated topic distribution per data set given a language. That is, we plot the topic distribution for the Spanish COVID and La Palma data sets and the topic distribution for the English COVID and La Palma data sets.Lastly, due to space limitation and clarity purposes, we only plot the first twenty-five topics per result.

6.2.1 Words per topic results

Each topic has a set of words that represents the topic, with weights associated to each word. For each language, we plot the word distribution per topic for the first twenty-five topics.

Spanish topics. The figure 6 depicts the generated Spanish topics by the model. The first topic is meaningless since it contains English words. The topic #5 mentions COVID-related words as *cov*, *variant* and *omicron*. Another interesting topic is the #23. It contains a named entity and words

surrounding it, like *libert*, which comes from *libertad* and *ataq*, from the Spanish word *ataque*. Note that there are only five topics that contains words related to the volcano eruption of La Palma.

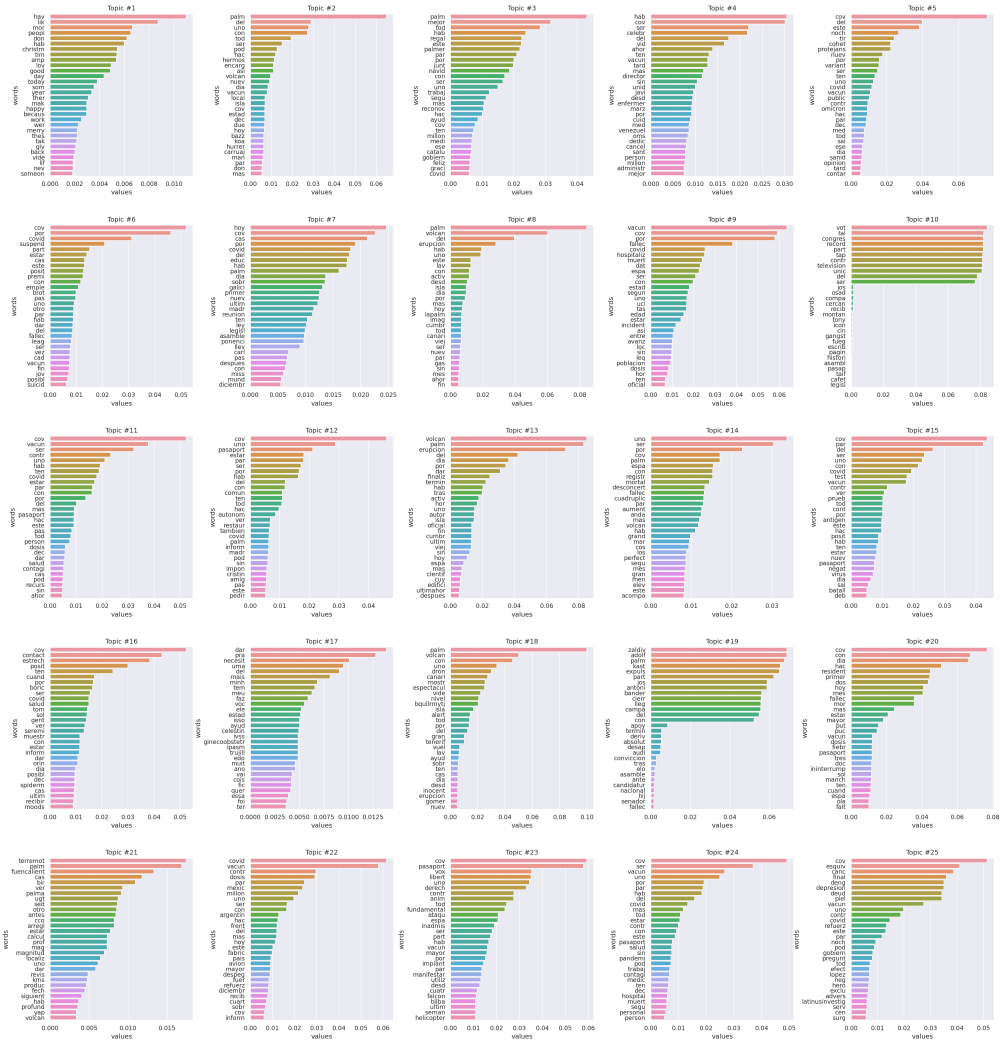

Figure 6: Topics for the Spanish tweets. Each plot represents the weights for each word within the topics, i.e. how representative a word is within the topic.

English topics. The figure 7 depicts the generated English topics by the model. The topic #20 is the only one that contains words related to the volcano eruption. The remaining topics are COVID-related with some exceptions. For example, the topic #14 contains named entities like *boris*, *johnson* or *covid*.

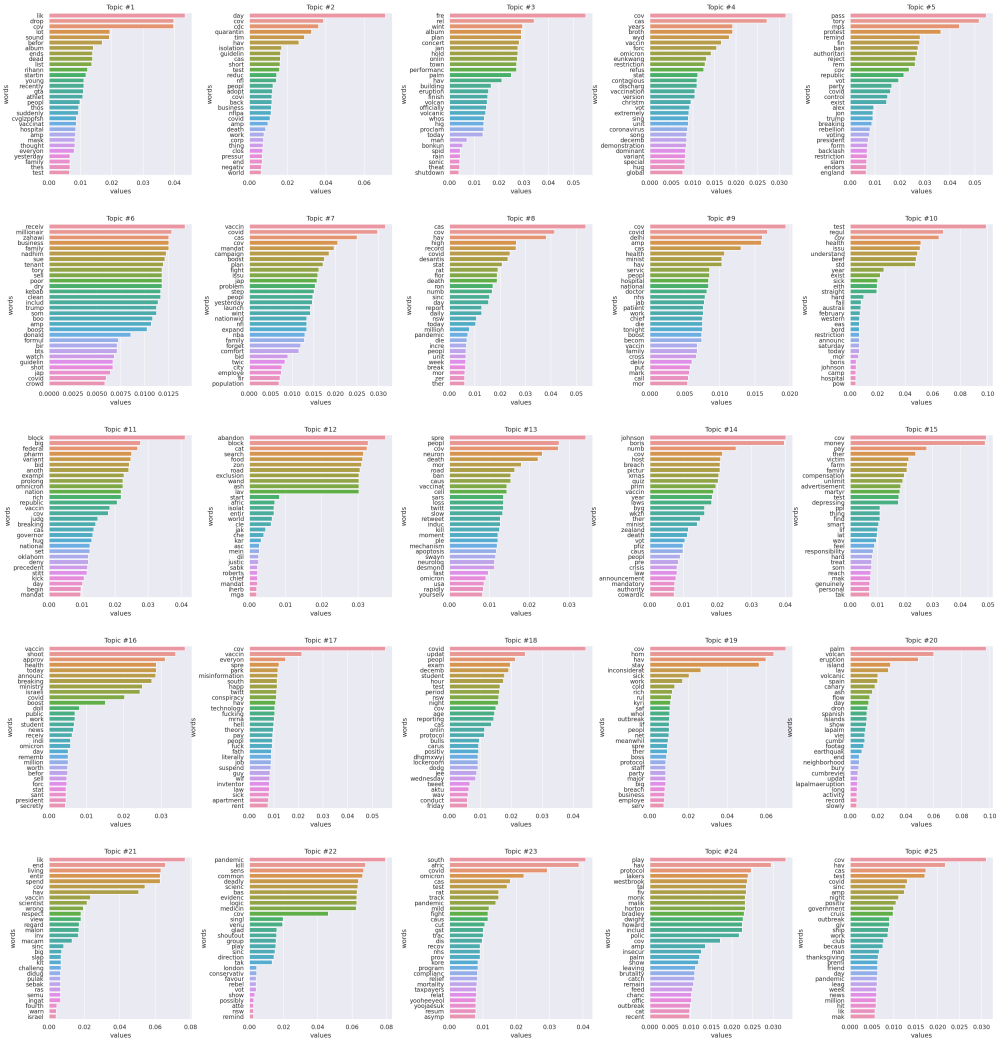

Figure 7: Topics for the English tweets.

6.2.2 Topic distribution of COVID and La Palma data sets

Lastly, we analyse if there exist differences between in the topic distributions of the COVID and La Palma data sets, regardless of the language. That is, for the COVID and La Palma data sets, we analyse how much these data sets mention the inferred topics. This can be evaluated regarding the topic distribution per data set, since it shows how a data set mentions each generated topic.

COVID data sets. The figure 8 depicts the topic distribution for the COVID data sets given the two considered languages. In general, it appears that the Spanish topics are more mentioned than the English ones. However, recall that the English topics are more informative, in a topical sense, than the Spanish ones. Note that, regarding the figure 8, all topics are mentioned for both languages.

On the other hand, there are some Spanish topics that are less mentioned. For example, the topics #13 and #18 presents this situation (blue bars). Note that these topics correspond to events related to the volcano eruption regarding the figure 6.

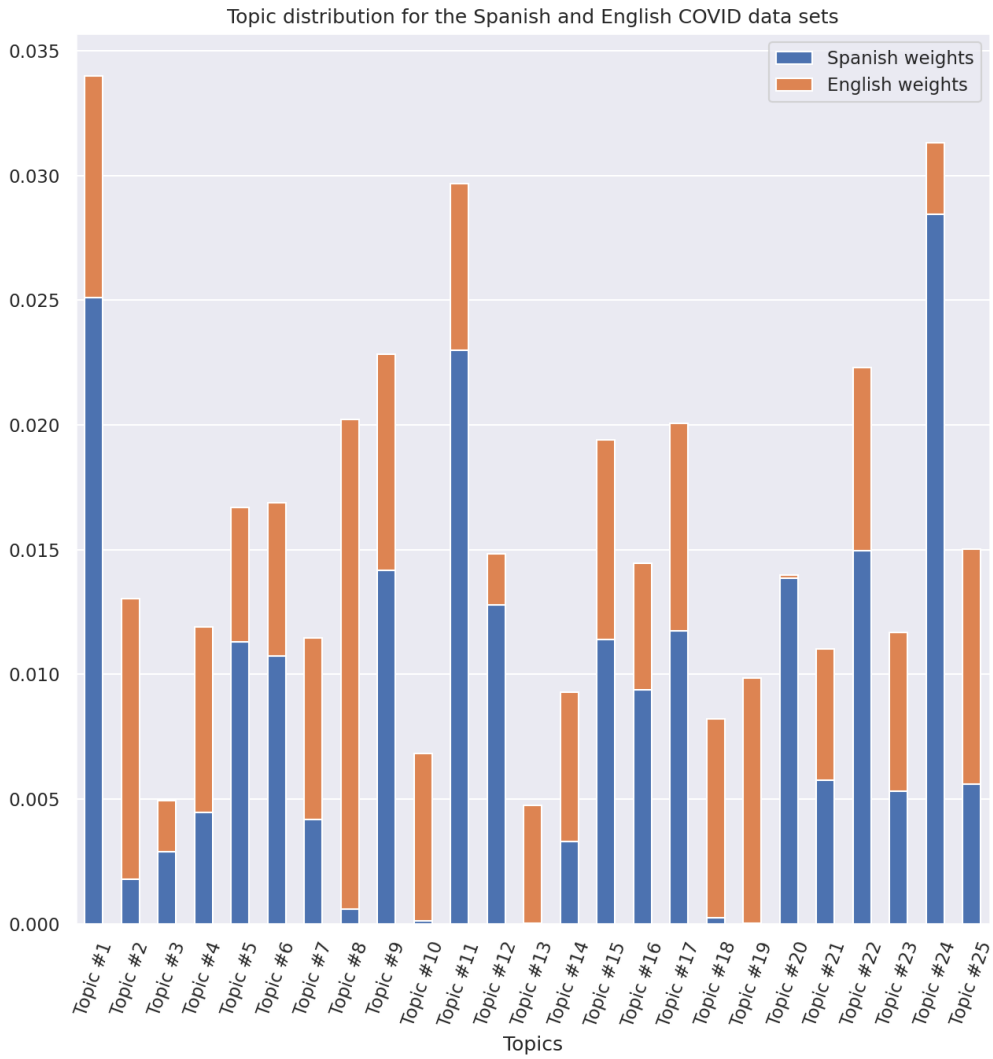

Figure 8: Topic distributions for the COVID data sets. The X axis denotes the considered topics. The Y axis measures weights that indicate how much a topic is mentioned regarding the language variable.

La Palma data sets. The figure 9 depicts the topic distribution for La Palma data sets given the two considered languages. In general, in comparison to the figure 8, the topics are not mentioned as much as they are in the COVID data sets. The Spanish topics #13 and #18 are mentioned relatively often. On the other hand, the English topic #20 is highly mentioned. This topic is related to the event of the volcano eruption, regarding the words within the topic #20 of the figure 7.

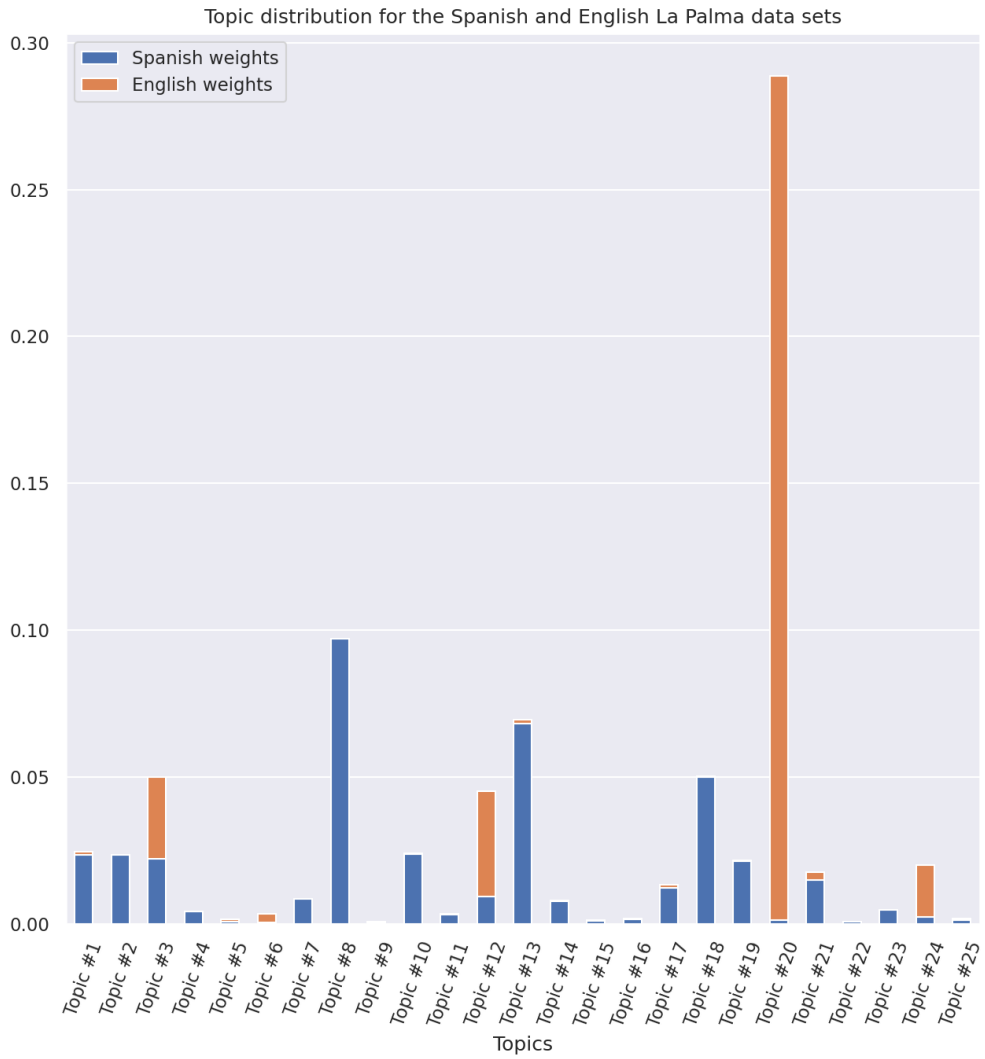

Figure 9: Topic distributions for La Palma data sets.

6.2.3 Topic comparison

For both languages, the generated topics are similar concerning what type of topics are discovered. There are two main sets: (1) Topics related to the virus, the vaccine and the Omicron COVID variant; (2) Topics related to the volcano eruption of La Palma. However, the model generates only a few topics related to the volcano eruption regardless of the language. This limitation could be due to the lack of tweets within La Palma data set. Nonetheless, it appears that, for the English tweets, the model generates a well-defined topic related to the volcano eruption, the topic #20 of the figure 7. Overall, the English topics contain fewer meaningless words than the Spanish topics.

In conclusion, and regarding the question 2, there exist differences in the first twenty-five topics generated for each language. The English topics are more meaningful than the Spanish ones. Despite there is only one topic related to La Palma data set, the words contained within it are well-clustered semantically since they refer to the volcano event. However, the absence of topics related to La Palma data set is due to the lack of tweets related to this data set.

6.2.4 Topic comparison with KLD

As we mention in the section 2.3, we expected the results to be different due to the assumptions that each approach consider. At a topical level, LDA yields more information than KLD. However, LDA cannot track how the words evolve across time. On the other hand, LDA is able to remove stop words. Recall that, for the KLD computation, we used the words contained in the generated vocabulary from the construction of the $TF - IDF$ matrix over the normalised tweets. This involved to tweak the parameters max_{df} and min_{df} in order to remove corpus-specific stop words.

In conclusion, LDA can be used to retrieve topics regardless of the time constraint, whilst KLD is proper if we aim to study how the words evolves across time. Thus, we cannot compare the results regarding the words generated by both approaches. Nonetheless, both methods assume that the word importance is based, in the end, on word frequency distributions and the results are different because of the usage of these word frequency distributions.

6.3 Analysis of clustering results

The cluster analysis consists in studying the internal similarities ($ISim$), their standard deviations ($ISDev$) and the internal similarities ($ESim$) for each cluster yielded by the considered clustering algorithms. Given the data sets, we study these metric values and discuss their variations regarding the data set language and the applied tweet normalisation. Note that CLUTO orders the clusters in descendant order of the $ISim$ and $ESim$ metrics. Thus, we analyse how these metrics decrease across the considered clusters. this analysis addresses the questions four, five and six.

6.3.1 Analysis of the results of the COVID data set

Average internal similarities, $ISim$. The figure 10 depicts the average internal similarities, $ISim$ for the English and Spanish tweets from COVID data sets. In general, the graph method yields more clusters that contain similar objects than the divisive algorithm. The $ISim$ metric for the divisive algorithm decreases faster than the graph method. Surprisingly, if we consider the graph method and the raw tweets, the $ISim$ is, in general, higher compared to the $ISim$ in the case of processed tweets (red and green dotted lines). In the plot, the red dotted line is, in general, over the green dotted one. This situation is more accentuated in the case of the Spanish tweets. However, for the Spanish tweets, the $ISim$ starts to decrease faster from $Cluster_{id} = 22$ onwards in comparison to the English tweets.

Regarding the divisive algorithm, RBR, note that the $ISim$ remains equal overall, but the metric is slightly better concerning raw tweets.

In conclusion, there exists variabilities in terms of the data set language. The average similarities per cluster are better regarding English tweets. The processing does not seem to improve the clustering $ISim$ metrics.

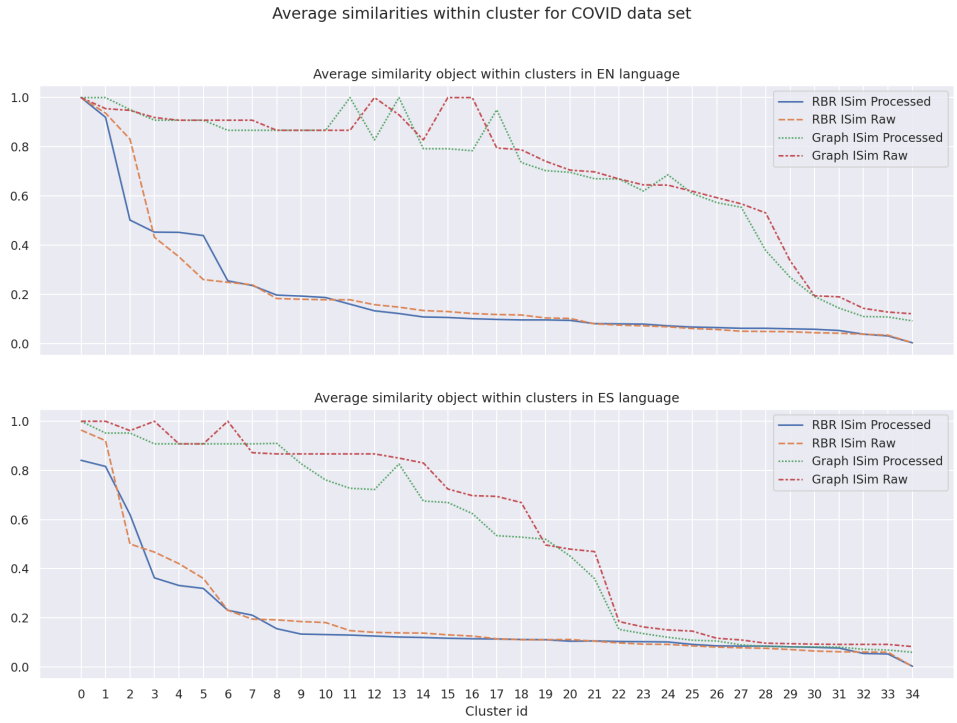

Figure 10: Average internal similarities ($ISim$) per cluster for English (EN) and Spanish (ES) COVID data sets, respectively. The X axis denotes the cluster identifier and the Y axis measures the $ISim$. The decreasing tendency is due to CLUTO orders the clusters in decreasing order of the internal and external similarities, i.e. $ISim$ and $ESim$.

Standard deviations of $ISim$. The figure 11 depicts the standard deviations of $ISim$ for the English and Spanish tweets and given the COVID data set. For the processed English tweets and the graph method, the $ISDev$ values (green dotted line) are, overall, higher compared to the obtained ones for the raw tweets. This means that there exists less $ISim$ variability in the case of non-processed tweets. However, in the bottom plot, the contrary happens if we consider the Spanish tweets. There exists more $ISim$ variability when the tweets are raw.

On the other hand, the divisive algorithm yields clusters whose $ISDev$ are smaller if we consider the raw tweets regardless of the language (orange dotted line).

In conclusion, if we consider the graph method, there exist variabilities in function of the language in terms of the $ISim$ metric. For the divisive algorithm, the variabilities are smaller for the raw tweets, but recall that the average similarities per cluster are smaller in comparison to the obtained ones with the graph method.

Figure 11: Standard deviations of $ISim$ per cluster for English and Spanish COVID data sets, respectively.

Average external similarities, $ESim$. The figure 12 depicts the average external similarities for the English and Spanish tweets given the COVID data set. For both languages, if we consider the graph method, the $ESim$ is practically zero. Thus, and recalling the results for the $ISIm$, the graph method yields clusters that are supposedly, in general, tighter than the ones obtained with the divisive algorithm.

Regarding the RBR algorithm, the $ESim$ is accentuated yet small (0.004 is the maximum value). Note that, the $ESim$ is higher overall in the case of processed tweets regardless the language (blue line).

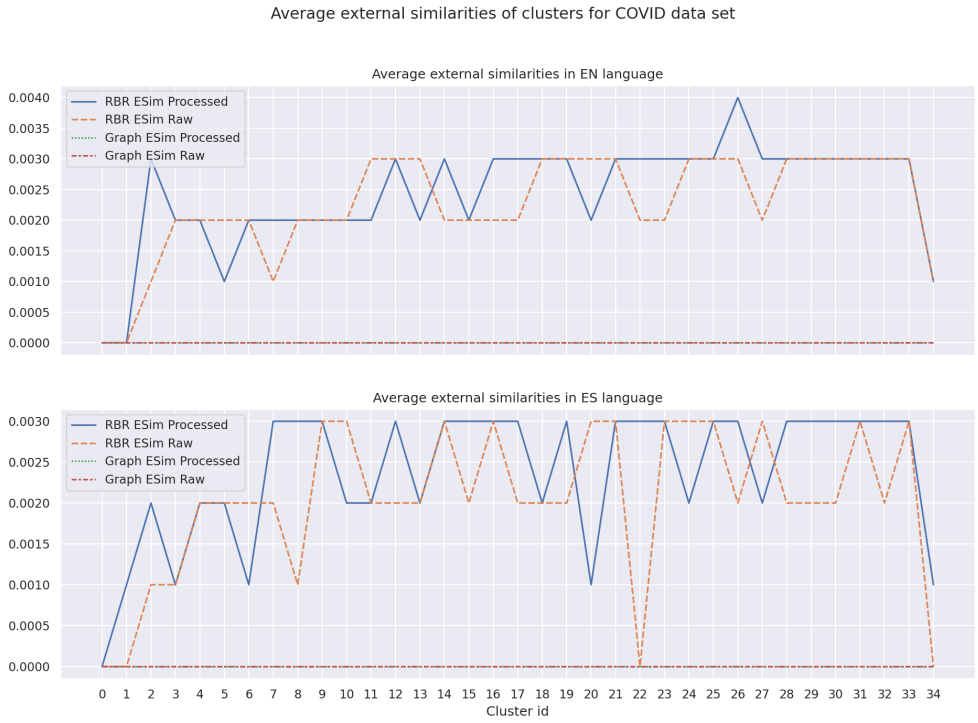

Figure 12: Average external similarities, $ESim$, for the English and Spanish COVID data sets, respectively.

6.3.2 Analysis of the results of the La Palma data set

Average internal similarities, $ISim$. The figure 13 depicts the average internal similarities, $ISim$, for the English and Spanish tweets from La Palma data sets. Overall, as the COVID $ISim$ results, the graph method builds clusters with more similar objects than the divisive algorithm. For the English and processed tweets (green dotted line for the top plot), the $ISim$ values are, in general, higher in comparison to the obtained ones for the raw tweets. With respect to the divisive algorithm, the $ISim$ is, in general, equal when $Cluster_{id} \geq 12$. Note that for smaller values of $Cluster_{id}$, the $ISim$ values present more variability.

On the other hand, considering the Spanish tweets, the graph method builds, in general, clusters with higher $ISim$ values when tweets are raw too if we compare them to the obtained ones from the processed tweets. In addition and regarding the divisive algorithm, the same situation occurs w.r.t. the English tweets. Note that the $ISim$ values are, in general, equal concerning the raw tweets and when the $Cluster_{id} \geq 17$. For lower values of $Cluster_{id}$, the built clusters over the processed tweets present higher values of $ISim$ if we compare them with the obtained values for the non-processed ones.

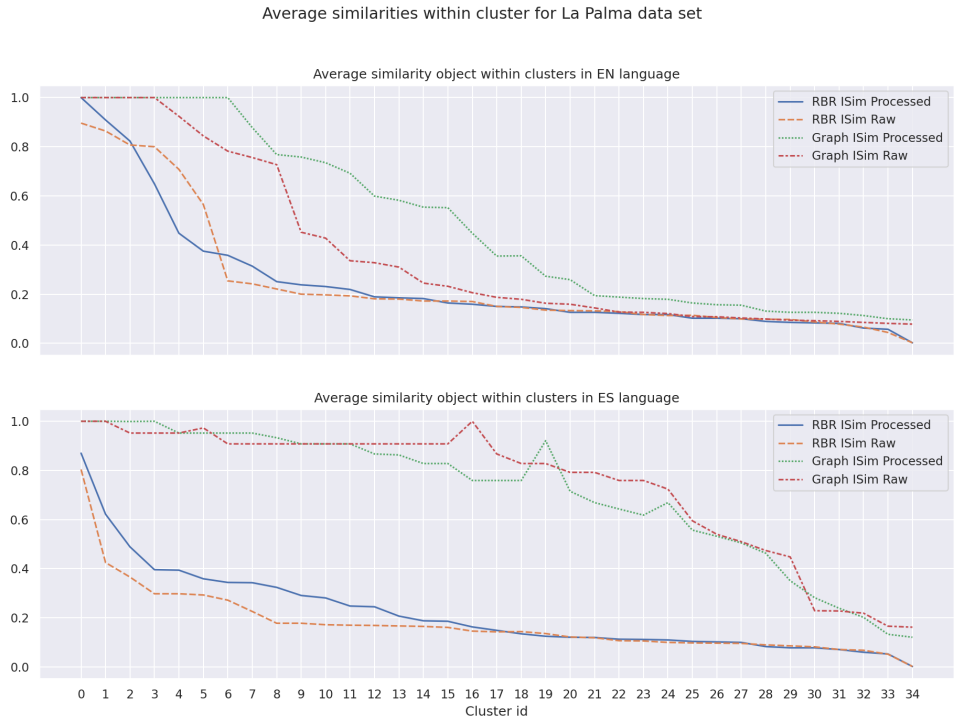

Figure 13: Average internal similarities ($ISim$) per cluster for English and Spanish La Palma data sets, respectively.

Standard deviations of $ISim$. The figure 14 depicts the standard deviations of $ISim$ for the English and Spanish tweets of La Palma data set. Comparing both clustering algorithms, the graph method yields clusters with higher standard deviations regardless of the language. For the English data (top plot of the figure) and considering the graph method, the $ISDev$ is overall higher regarding the processed tweets, i.e. the green dotted line is over the red one. The same situation occurs if we consider the Spanish tweets.

In relation to the divisive algorithm and regardless of the language, if we consider the processed tweets, the $ISDev$ is, in general, higher in comparison to the $ISDev$ computed for the raw tweets.

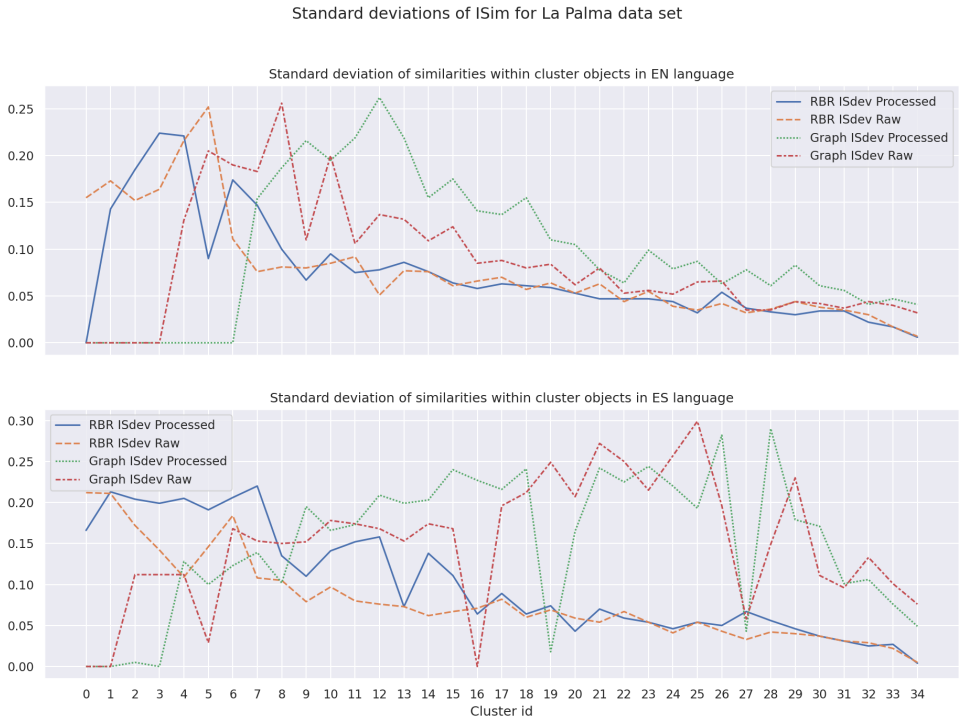

Figure 14: Standard deviations of $ISim$ per cluster for the English and Spanish La Palma data sets, respectively.

Average external similarities, $ESim$. The figure 15 depicts the average external similarities for the English and Spanish tweets given the La Palma data set. Overall, the graph method yield tighter clusters since the external similarities are smaller in comparison to the divisive algorithm. Note that, despite that, the $ESim$ is relatively small (0.003 is the maximum value). On the other hand, for the English data and considering the graph method, it seems there are no substantial differences in terms of the application of the normalisation. There are oscillations, but none of the $ISim$ curves for the raw and normalised are above another. However, for the divisive algorithm, the $ESim$ is overall higher regarding the processed tweets (blue curve).

In conclusion, the graph method generates tight clusters for the Spanish tweets since the average external similarities are zero. This is depicted at the bottom plot of the figure. However, there exist external similarities regarding the Spanish tweets.

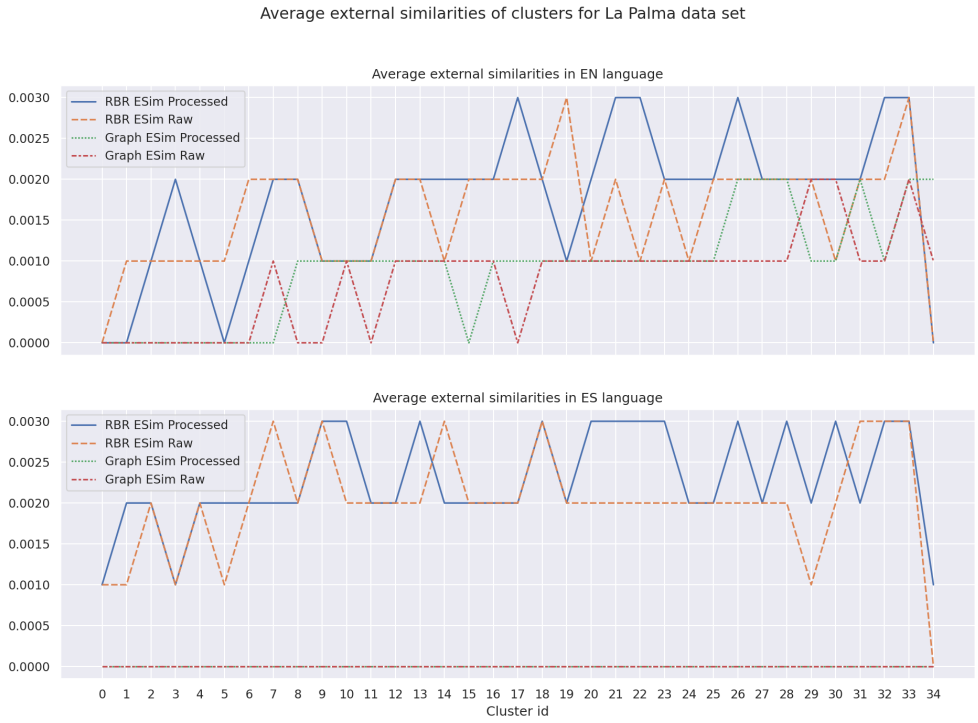

Figure 15: Average external similarities ($ESim$) for the English and Spanish La Palma data sets, respectively.

7 Conclusions

Related to the trend analysis task, the KLD experiments address the first question and show that there is a notable difference between the COVID and La Palma data sets regardless of the language. With respect to the COVID data set, it appears that most of the words remain important across the considered dates. In contrast, the KLD values for the La Palma data set are less uniform across time. However, there are fewer tweets for the La Palma data compared to the COVID one. Moreover, there are differences if we consider the language variable. There exist some topical words in the presented heatmaps, except for the Spanish La Palma data set.

On the other hand and regarding the second question, the LDA model generates more understandable topics when executed on the English data compared to the generated ones with the Spanish data. However, there are only a few topics related to La Palma events for both languages. This limitation is due to the lack of La Palma tweets. With respect to the third question, we cannot compare the results obtained with the KLD and LDA methods since they use word frequency distributions in different ways.

Regarding the clustering results, there exist differences between the divisive and graph-based clustering algorithms. The graph-based method yield tighter clusters in comparison to the divisive algorithm. This addresses the fourth question proposed in the experiments. Interestingly and addressing the fifth question, the average internal similarity results are better if we consider the raw tweets, except for the English tweets of the La Palma data set. Regardint the sixth question, we cannot state that there exists differences in the clustering quality regarding the language, since the proposed corpora are not parallel at language level. With a parallel corpus, we could investigate whether the quality of clustering varies from one language to another.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

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 *ECIR*.

Zhao, Y. and Karypis, G. (2001). Criterion functions for document clustering. experiments and analysis.