Censorship of Twitter - Unsupervised Topic Modeling

Mathieu Desponds - 283229 Robin Jaccard - 310682 Eva Luvison - 286547

Abstract – **Twitter censorship covers governmental notice and take-down requests to Twitter, which are enforced as long as in accordance with Twitter's Terms of Service [1]. Many countries have laws that may apply to Tweets and Twitter account content. Governments or authorities can submit a removal request to Twitter indicating that specific content is illegal in their jurisdiction. In order to gain insight into those jurisdictions that are not always very transparent, one can analyse the topics of censure per country. In this paper, we present our approach to the problem.**

I. Introduction

The limit of free speech is not always very clear and differs in function of the origin and point of view. In this project, we analyse a sample of tweets that have been censored in different countries. These tweets were censored between January 2021 and March 2021 and are not publicly available but have been collected by scanning the internet archive.

Topic modeling is a natural language processing technique that has for goal to find abstract topics in a collection of documents. Many algorithms have emerged in the past 20 years in order to tackle this task. In this paper, we focus on four of those algorithms and we compare them using a coherence score. First, we use Latent Dirichlet allocation (LDA) and then two algorithms that are more short-text-oriented and fit more tweets: Gibbs Sampling Dirichlet Mixture Model (GSDMM) and Biterm Topic Model (BTM). We also use BERTopic which makes use of a state-of-the-art sentence transformer. Results will allow us to gain insights into the different topics and better understand the censure targets of different countries.

II. Dataset

A. Presentation

The given dataset is composed of 41'727 tweets from the 1st January 2021 to the 31st March 2021. 23'081 of those tweets have a unique text. Each tweet has 39 features but only the text of the tweet, the language and the country in which it has been withheld are relevant to our work.

We have tweets that are withheld only from 7 countries that are Brazil, France, Germany, Israel, India, Russian Federation and Turkey. The goal is to find the different topics of censure for each country so we split the dataset by country and tweets that are censured in multiple countries are assigned to the dataset of each country where they are censored. The texts are in their original language and there are fifty of them Figure 1.

B. Data processing

The processing of the tweets has an effect on the performance of our models. The following pipeline of processing is applied except for BERTopic. Thanks to the sentence transformer, we only translate, treat the hashtags and remove the URLs.

Removing numbers, URLs, Stop words and Punctuation: The numbers and URLs are removed from the text as it plays no role in topic modeling.

Emojis: Most of the emojis are removed from the text. However, we transform some sequences of emojis into text when it is useful for classification. For example, the sequence of emojis that writes HOT is changed by hot written with normal characters.

Traduction: As the dataset contains various languages even inside each country, we translate all tweets into English.[2]

Hashtags: After translation, we split the CamelCase words and other combinations of words. Note that the translation library could properly translate the CamelCase words. We also remove the hashtag character.

Contraction: Contractions are extended in order to ease the deletion of the punctuation without losing the meaning. Otherwise some words like "ll", "re" would appear. [3]

Lemmatization: As the dataset is not very large, lemmatization is particularly important since it moves sparsity.

III. Methods

We implemented the following methods: Latent Dirichlet allocation (LDA), Gibbs Sampling Dirichlet Mixture Model (GSDMM), Biterm Topic Model (BTM), BERTopic.

A. Latent Dirichlet Allocation

LDA [4] is a probabilistic generative model for collections of discrete data, such as textual data. It is a hierarchical Bayesian model with three levels in which each item in a collection is described as a finite mixture over an underlying set of topics. [4] Each word in a document is assigned to a separate subject and the documents contain a distribution of topics.

However, LDA do not perform very well on short texts such as tweets. The reason behind this is that it captures the document-level word co-occurrence patterns to reveal topics. Hence, it suffers from severe sparsity in short texts.

B. Gibbs Sampling Dirichlet Mixture Model

GSDMM [5] is essentially a modified LDA which assumes that a tweet contains only one topic

GSDMM principle can be explained using the analogy "Movie Group Approach". For instance, think of a student's group (tweets) all having a list of their favourite movies (words). These students are further indiscriminately allocated a 'K' number of tables. On the professor's directions these students must mix up tables with the following 2 rules:

- Completeness: Choose a table with more students
- Homogeneity: Choose a table where students share similar movie interests

This shuffling within the tables continues until a point reaches a plateau where the number of clusters does not move. This algorithm has the advantage that it doesn't need to know the number of clusters in advance and find the optimal number by itself.

C. Biterm Topic Model

Biterm Topic Model [6] uses a different approach from LDA in order to treat the sparsity issue. In BTM we learn the topics by directly modeling the generation of word co-occurrence patterns (i.e. biterms) in the whole corpus. The major advantages of BTM are that

- BTM explicitly models the word co-occurrence patterns to enhance the topic learning
- BTM uses the aggregated patterns in the whole corpus for learning topics to solve the problem of sparse word co-occurrence patterns at document-level.

D. BERTopic

BERTopic [7] takes advantage of the superior language capabilities of state-of-the-art transformer models. It also uses other powerful ML algorithms such as UMAP and HDBSCAN in order to reduce the dimensionality of the embeddings and cluster them. Finally, it make use of c-TF-IDF to give meaningful names to the different clusters.

It also has the advantage that the outliers are grouped in a "outliers" cluster. However, due to the high flexibility of the library. Many parameters/algorithms have to be chosen which is not always easy is unsupervised learning.

IV. Results

A. Comparing the different methods

The different algorithms are now tested on the tweets censored by France. We made this choice because we have a deeper understanding of the topics censored in France, hence it is easier to label them. It was also helpful to understand the initial text of the tweet in order to judge the quality of the pre-processing (including the translation).

1. Topic coherence

When testing different models of topic modeling, it is important to choose an appropriate metric to evaluate the performance of each model. Even if evaluation can be challenging due to the unsupervised training process, there are several metrics that can be used for this purpose. Each metric has its own strengths and limitations. One of the common metric used for topic modeling is topic coherence.

Coherence mesures the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference [8].

The two coherence measures most commonly used for topic modeling are the CV coherence and the UMass coherence.

CV coherence creates content vectors of words using their co-occurrences and, after that, calculates the score using normalized pointwise mutual information and the cosine similarity [9]. On the other hand, UMass coherence calculates how often two words, w_i and w_j appear together in the corpus and it's defined as

$$C_{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

where $D(w_i, w_j)$ indicates how many times words w_i and w_j appear together in documents, and $D(w_i)$ is how many time word w_i appeared alone [9].

For both coherences, the greater the number, the better is coherence score. A comparison of the different models can be found in Table 1. For this comparison, all models were run with 10 topics on the tweets censored in France. The reported scores are the ones found by doing a grid search over the possible parameters in order to have the best coherence score possible for each algorithm.

Method	CV	Umass	Label accur.
LDA	0.337	-16.439	-
GSDMM	0.406	-9.63	-
BTM	0.504	-7.377	-
BERTopic	0.793	-0.477	0.576

Table 1: Results for each method

2. Labelling

However, like any other tool, topic coherence has certain limitations and potential downsides. One of those downsides is the limited ability to capture semantics. Coherence measures typically focus on the co-occurrence of words in the corpus, rather than on the meanings of those words [10]. This can make it difficult for coherence measures to capture the full semantics of a topic and may result in topics that are technically coherent but not semantically meaningful. In order to give more meaning to our results, we decided to manually label a part of the data.

We labelled 125 French tweets and compared our results to the ones found by the algorithms. To label the tweets, we first decided on the different labels based on the French government policies, the subject we saw during the exploratory data analysis and the actuality in French society. Then, we used crowd-sourcing methods where all of us (three) labelled independently the tweets and we then aggregated the results using majority vote.

To compare the labelling with the results of the algorithms, we assign to the topics of the algorithms the same labels that we use. Finally, we compare the label of the 125 tweets to the label assigned by the algorithm in order to get the accuracy.

Due to truncated tweets and difficulty in assigning label to the clusters given by the algorithm, this metric is not adapted for poor algorithm like LDA, BTM and GSDMM. Therefore this method is only relevent for BERT where cluster labelling was easier.

B. Best results

Our results show that BERTopic algorithm consistently outperforms the other algorithms, producing the highest scores on all of the evaluation metrics. This suggests that BERTopic is a strong choice for discovering coherent and diverse topics in this type of data. We will further explore the results that were obtained on the French dataset using BERTopic and how it generalizes to other countries.

1. Best model for tweets censored in France

Our analysis of censored tweets from France, using BERTopic, identifies several underlying themes. The topics are various and related to hate speech and politics. Figure 2 presents an overview of the different topics found with the most significant words for each topic.

Further analysis of these topics could provide insights into the concerns and priorities of French Twitter users.

2. Generalization to other countries

We wanted to see how the parameters found for France ($params_F$) would generalize to the other countries. In fact, finding good parameters that could be applied to all countries would be great and would spear time of tuning parameters for a new country.

When applying $params_F$ to India, Germany and Turkey we saw the same phenomena. $params_F$ were not adapted because the clustering resulted in two or three general topics whereas we wanted to go deeper and find more particular topics to really understand what government censors. To tackle this problem, we reduced the $\epsilon_{clusterselection}$ of HDBSCAN and the $dist_{min}$ in UMAP, and adapted the number of neighbors and components with the number of tweets censored for the country so that tweets in the same cluster have to be "closer" than with $params_F$. What is interesting is that the best parameters used for India ($params_I$) were really close to the ones of Turkey and Germany.

This led us to the idea that $params_I$ were better general parameters. To confirm this hypothesis, we applied $params_I$ to France and although the results were worse than with $params_F$, it was still good.

To summarize, we found two types of censorship policy. On one hand, countries like France and Russian Federation where the topics are more heterogenous. therefore we can allow for a bigger distance inter-topic. On the other hand, countries like Germany and India have a main topic censored (porn and topics related to Pakistan respectively). In the latter, if we want to divide the main topic into smaller ones we have to make our algorithm group only tweets that are really close.

Our proposed method is to start with $params_I$ that seems to apply to all countries and then tune the parameters specifically to countries depending on the number of topics the governments censors and how precise we want the topics to be.

More details about the parameters and topics can be found here.

V. Conclusion

We have analyzed multiple approaches of topic modeling in order to identify the underlying topics present in the censored tweets. After evaluating the results, we found that the BERTopic algorithm was the best performer among the algorithms tested. The use of transformer-based models like BERTopic allows us to effectively capture the semantics within the data, which in turn enables us to more accurately identify and extract relevant themes and topics from the censored tweets. Moreover the $params_I$ can be used on a new country to have a first insight of what countries censor.

This work allowed us to gain insight into the different topics that are censored by the government in different countries and tools that might be applied to new data. The different clusters for each country can be found here.

VI. Appendix

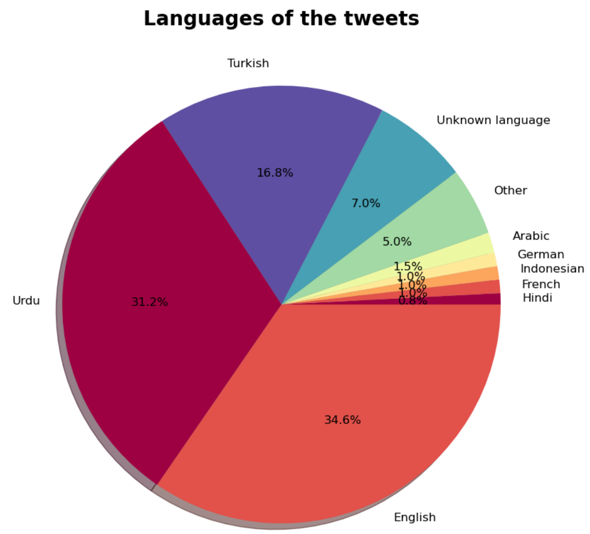

Figure 1: Distribution of the languages

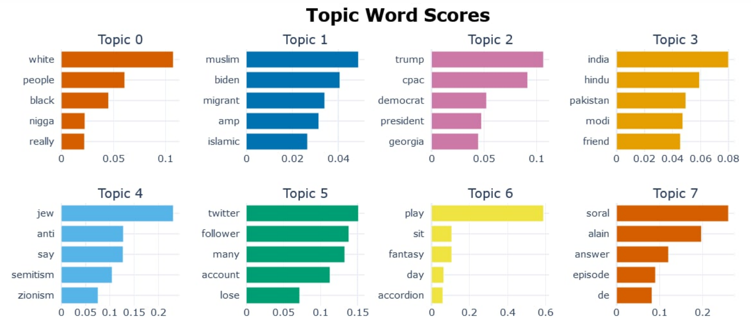

Figure 2: Topics censored in France

References

[1] Wikipedia, "Censorship of twitter." `https://en.wikipedia.org/wiki/Censorship_of_Twitter`. Last accessed 21 December 2022.

[2] N. Baccouri, "deep-translator." `https://pypi.org/project/deep-translator/`. Last accessed 10 December 2022.

[3] Kootenpv, "contractions." `https://github.com/kootenpv/contractions`. Last accessed 10 December 2022.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research 3 993-1022*, 2003.

[5] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," 2014.

[6] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," pp. 1445–1456, 2013.

[7] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[8] S. Kapadia, "Evaluate topic models: Latent dirichlet allocation (lda)." `https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0#:~:text=But%20before%20that%E2%80%A6-,What%20is%20topic%20coherence%3F,are%20artifacts%20of%20statistical%20inference.`, 2019. Last accessed 21 December 2022.

[9] E. Zvornicanin, "When coherence score is good or bad in topic modeling?." `https://www.baeldung.com/cs/topic-modeling-coherence-score`, 2022. Last accessed 21 December 2022.

[10] F. Rosner, A. Hinneburg, M. Roder, M. Nettling, and A. Both, "Evaluating topic coherence measures," 2014.