Social data gathering

Álvaro Domínguez Calvo
Universidad Nacional de Educación a Distancia
adomingue599@alumno.uned.es

Contents

List of Figures

Abstract

The nature of the social data brings forward new tasks related to monitoring and analyzing the opinion of certain topics or companies. This requires collect, process and analyze social data in order to determine what type of information could be a candidate for running supervised or unsupervised Machine Learning algorithms. However, social data contain specific features that are not captured by current trained models, as they are trained on well-written texts. Current Machine Learning models must be adapted to capture these social data nuances. However, it would therefore be useful to conduct an analysis based on the amount of existing data according to these specific characteristics. This could aid to determine which information could be exploited by Machine Learning models and guide future research. In this work it is proposed three multilingual social datasets and a high level analysis to study how much data exist in function of these nuances. The construction process involves compilation, formatting and considering constraints related to the APIs being used.

1 Introduction

Social media has become an interesting source of information for many companies since they could drive marketing plans in function of the online reputation management executed by detecting topics or analyzing the polarity of the posts related to their brand. The first step to take is to gather information related to the brand based on specific keywords that may contain valuable information. This information must be persisted taking into account legal concerns about the source information. Twitter, for example, forces not to publish the textual content of tweets. Typically, this information will be the input to machine learning algorithms since they provided a good framework to infer statistical rules. Depending of the task goal and the nature of the dataset, i.e if it is labelled or unlabelled, supervised or unsupervised algorithms must be applied. However, when dealing with social data the task is not that trivial since there exists specific lingo that complicates classification or clustering tasks. This work will assess the question of what type of information could be useful for further analysis and for these machine learning algorithms by performing an analysis on the amount of data available according to some characteristics of the datasets.

This document is organized as follows. The section 2 is dedicated to explain the gathering and formatting process executed during the building of the three datasets. Section 3 is focused on a high level analysis in function of common and specific variables contained in each dataset. The section 4 is dedicated to highlight the mandatory adaptations that need to be applied to the current Natural Language Processing models if social data is considered. Finally, the sections 5 and 6 are dedicated to conclusions and further work.

2 Data retrieval process: gather and formatting social media data

In this work it is proposed three datasets with social data from three social networks that are Twitter, Reddit and post blogs. Each dataset is build by using three different APIs: the Twitter API for the Twitter dataset, the Twingly API for the blog posts dataset and the Social Searcher API for the Reddit dataset. For each dataset it is considered a set of keywords that will retrieve social information by exact matching within corpus of each document. Since the retrieval must be performed for different languages, the keywords corresponds to named entities or specific lingo used in the social media in question in order to execute retrieval queries whose words do not need to be translated to the target language. As it will be seen, some tweets were retrieved by using the special character *$* which denotes information related to the stock market.

The next sections are intended to explain the gathering process, how the data is formatted in order to persist it according to legal concerns of the social media source information and some limitations related to the use of these type of APIs and the information they provide.

2.1 Gathering process

Typically, in the gathering process of any type of social data is implied the time variable which, for example, can be used to cluster documents by fixed window times. The building process of the different datasets provided comprise periods of time in terms of its retrieval. The Table 1 shows the dates when the retrieval process starts and ends per dataset.

Dataset	**Start date**	**End date**
Twitter	2021-12-04	2021-12-31
Blog posts	2021-11-08	2021-12-06
Reddit	2021-11-08	2021-12-06

Table 1: Start and end dates of the retrieval data

Each dataset is built according to a fixed set of keywords that corresponds to named entities. Each keyword belongs to the domain of music, stock market, news related to a natural disaster, technology and influencers. Each keyword corresponds with a named entity. The Table 2 summarizes

this information.

<table>
<tr><th>Keyword</th><th>Domain</th><th colspan="3">Dataset</th></tr>
<tr><td>paramore</td><td rowspan="3">Music</td><td rowspan="11">Twitter</td><td rowspan="6">Blog posts</td><td rowspan="5">Reddit</td></tr>
<tr><td>my chemical romance</td></tr>
<tr><td>the smashing pumpkins</td></tr>
<tr><td>la palma</td><td>News</td></tr>
<tr><td>dalas</td><td>Influencer</td></tr>
<tr><td>apple</td><td rowspan="2">Technology</td></tr>
<tr><td>microsoft</td><td rowspan="9"></td></tr>
<tr><td>$AMC</td><td rowspan="8">Financial</td><td rowspan="9"></td></tr>
<tr><td>$GME</td></tr>
<tr><td>$AAPL</td></tr>
<tr><td>$HOOD</td></tr>
<tr><td>$MSFT</td></tr>
<tr><td>$NVDA</td></tr>
<tr><td>$TWKS</td></tr>
</table>

Table 2: Keywords with their domain and datasets containing them.

The intention for each dataset is to be multilingual in order to analyze differences between the datasets and the named entities in function of the language. Hence, seven languages are proposed in the retrieval process that are English, Spanish, French, Portuguese, Italian German and Dutch. For each day in the periods of time described in Table 1 a daily retrieval process was executed. For the Twitter dataset, it was retrieved 100,000 tweets per day. The blog posts dataset was built only with a single API call since Twingly indexes articles and stores them in their database and the API call retrieves information persisted from the three last months. However, for the Reddit dataset only 100 posts could be retrieved per day. All the datasets were filtered in order to delete repeated posts.

2.2 Formatting process

The three APIs used to retrieve the social data define a set of properties returned in the response of the API call. However, not all information is

considered in the dataset building process for legal or practical reasons. In the gathering and processing the Twitter information the textual information must be deleted because the Twitter terms and conditions of its API usage. It is important to remark that with the tweet ID or the user ID it can be retrieved more information provided by the Twitter API. In this work, it will be only considered these IDs and the date and language of each tweet. The following schema defines the structure of a tweet in the dataset.

```
// Tweet format
{
        "date": ,
        "id_tweet": ,
        "language": ,
        "user_id":
}
```

Figure 1: Structure of a retrieved tweet.

The Reddit dataset contains information about the sentiment of the text and popularity information related to the number of likes and comments of each post. the Figure 3 shows the schema for a persisted Reddit post in the dataset.

```
{
        // Reddit document format
        "post_id": ,
        "lang": ,
        "date_posted": ,
        "sentiment": ,
        "text": ,
        "ups": ,
        "comments": ,
        "user_name": ,
        "user_id": ,
        "user_url":
}
```

Figure 2: Structure of a Reddit post.

Lastly, the Twingly API was used to retrieve the blog posts information regarding the keywords and languages considered. In comparison with the

later structures, Twingly provides more information like the images attached for each post or the links contained in it. The blog post structure is depicted in the Figure 3.

```
// Blog post document format
{
        "author": ,
        "authority": ,
        "blog_id": ,
        "blog_name": ,
        "blog_rank": ,
        "blog_url": ,
        "coordinates": ,
        "id": ,
        "images": ,
        "indexed_at": ,
        "inlinks_count": ,
        "language_code": ,
        "latitude": ,
        "links": ,
        "location_code": ,
        "published_at": ,
        "reindexed_at": ,
        "tags": ,
        "text": ,
        "title": ,
        "url"
}
```

Figure 3: Structure of a blog post

2.3 Limitations

The Twitter API limits the number of API calls per day. This led to the retrieval process to take several hours because wait clauses must be coded to avoid the rejection of API calls by the Twitter side. Furthermore, the textual information can not be shared publicly due to legal Twitter concerns. However, with the tweet ID it is possible to retrieve the metadata that defines a tweet such that the geographical position where the tweet was posted, the number of retweets etc.

Twingly provides a great API to retrieve blog posts. However, the supported format is XML which is not the trend nowadays. Nonetheless, there exists packages that allow to translate files from XML to JSON format.

Lastly, the Social Searcher API is the most restrictive. Only 100 posts per API call can be retrieved and the language of the posts is not guaranteed to match with the one specified in the API call. However, the API response contains attributes that can be exploited in the sentiment analysis context like the sentiment property.

3 Data analysis

Mining information from social data is not a trivial task since there are some aspects that must be taken into account. First the noise data will difficult the execution of Machine Learning algorithms, noise that is mostly related to misspellings. Secondly the lack of textual data in the microblogging sites obstructs the application of these algorithms. This motivates to enrich the dataset with external sources like Wikipedia (Tang et al., 2014) or with the links that can potentially contain.

However, in this work it will be assessed the question of what type of data is likely to be used in further analyses and what tools covers the variables considered in this study. For example, exclude those languages that do not contain enough information to train Machine Learning models or exploit determined attributes to enrich the textual information. For this, two analyses are proposed at different levels of granularity: (1) At dataset level where it is depicted some statistics related to the languages and the dates when the social data was posted; (2) At named entity level where it is studied whether some specific lingo for a given dataset could be used in these potential analyses.

3.1 Analysis at dataset level

For the analysis at dataset level it will be considered two common variables across the three datasets: the language and the number of posts per date.

3.1.1 Twitter dataset analysis

The figure 4 shows top 100 dates ordered by the number of tweets published. This suggests an analysis of the tweets posted those days in order to detect

sudden events. However, the Twitter dataset is the largest one and cannot be compared with the other at first glance.

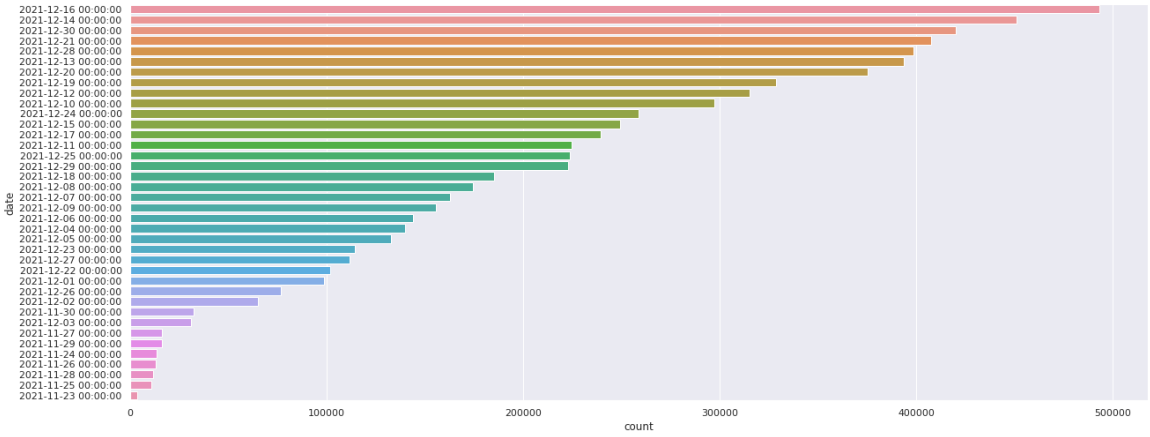

Figure 4: Number of tweets published per date. Only the dates with higher values are shown.

For the language categorical variable, the count of tweets of each value is shown in the Figure 5. Since there is more available data for the English language, it is feasible to study a specific task with a monolingual approach an then tackle the problem but with a multilingual perspective. However, when the task becomes multilingual it must be taken into account that the NLP tools are limited for certain languages. CoreNLP provides a pipeline to tokenize, perform a PoS tagging, lemmatize and detect named entities given a document for multiples languages (Manning et al., 2014).

In the case of Twitter the challenges not only appears from the multilingual side. Deal with short texts with misspellings makes tougher Machine Learning tasks like clustering or classification since the lack of data and the noise will affect in the performance of the trained model in terms of their prediction capability. An approach to avoid this issue could be to retrain the current models in order to adapt them to the nuances of the tweet data (Farzindar and Inkpen, 2020). However, this probably will lead to models trained over these specific features and the desired generalization of the retrained models will not be guaranteed. This fact motivates the design of systems that normalize the input text in order to reuse the trained models for the NLP tasks.

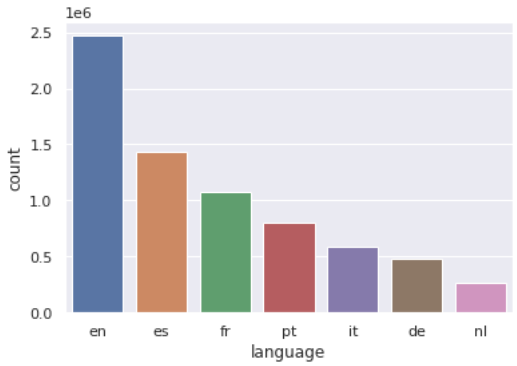

Figure 5: Tweet posts count per language.

3.1.2 Blog post dataset analysis

The distribution of blog posts per language is quite similar to the Twitter dataset as shown in the Figure 6. Also, for the amount of posts published per date, there exists some days the number of posts soars.

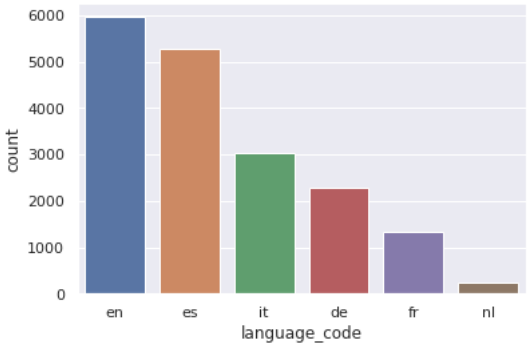

Figure 6: Blog posts count posted per language.

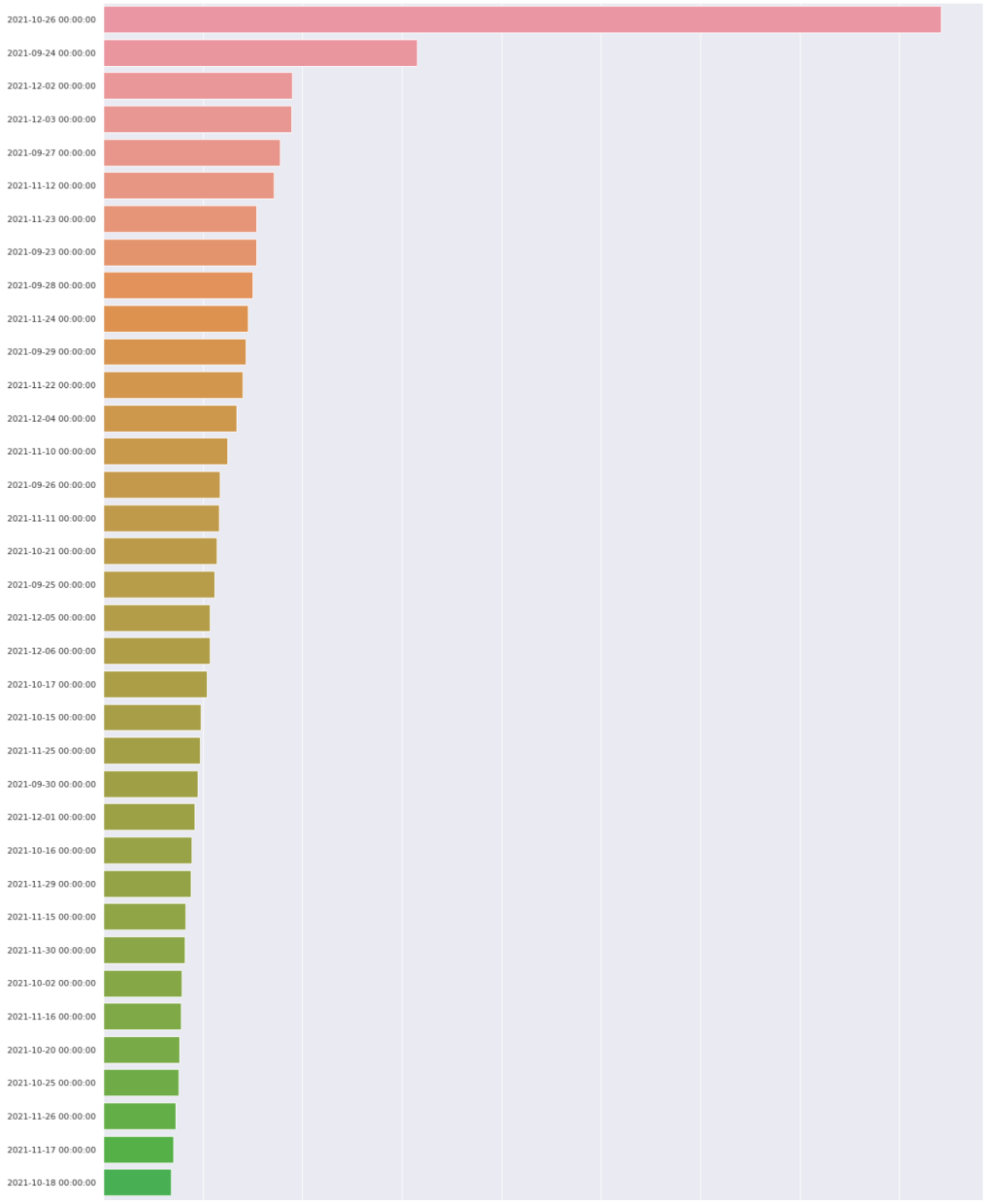

Figure 7: Blog posts count posted per date. Only the dates with higher values are shown.

3.1.3 Reddit dataset analysis

Something quite different occurs with the Reddit dataset. First of all, despite the predominant language is the English, there are fewer posts for the

remaining languages. Furthermore, there are languages not reflected in the proposed ones. This is due issues of the Social Media Searcher API. Note that the total number of posts are fewer in comparison to the later datasets. The Figure 8 depicts the count of Reddit posts per language.

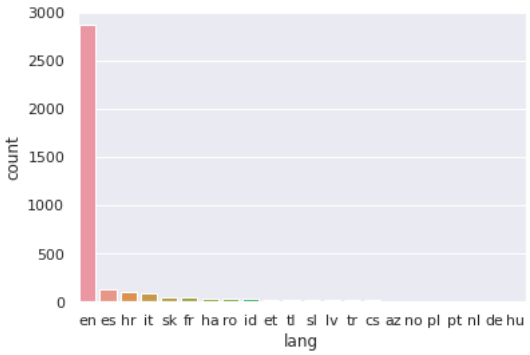

Figure 8: Reddit posts count posted per language.

From the Figure 9 it can be stated that, regardless the social data, the time variable could be used to mine information related to events by considering, for example, the hypothesis that posts with common words and published in similar dates will form clusters of related posts.

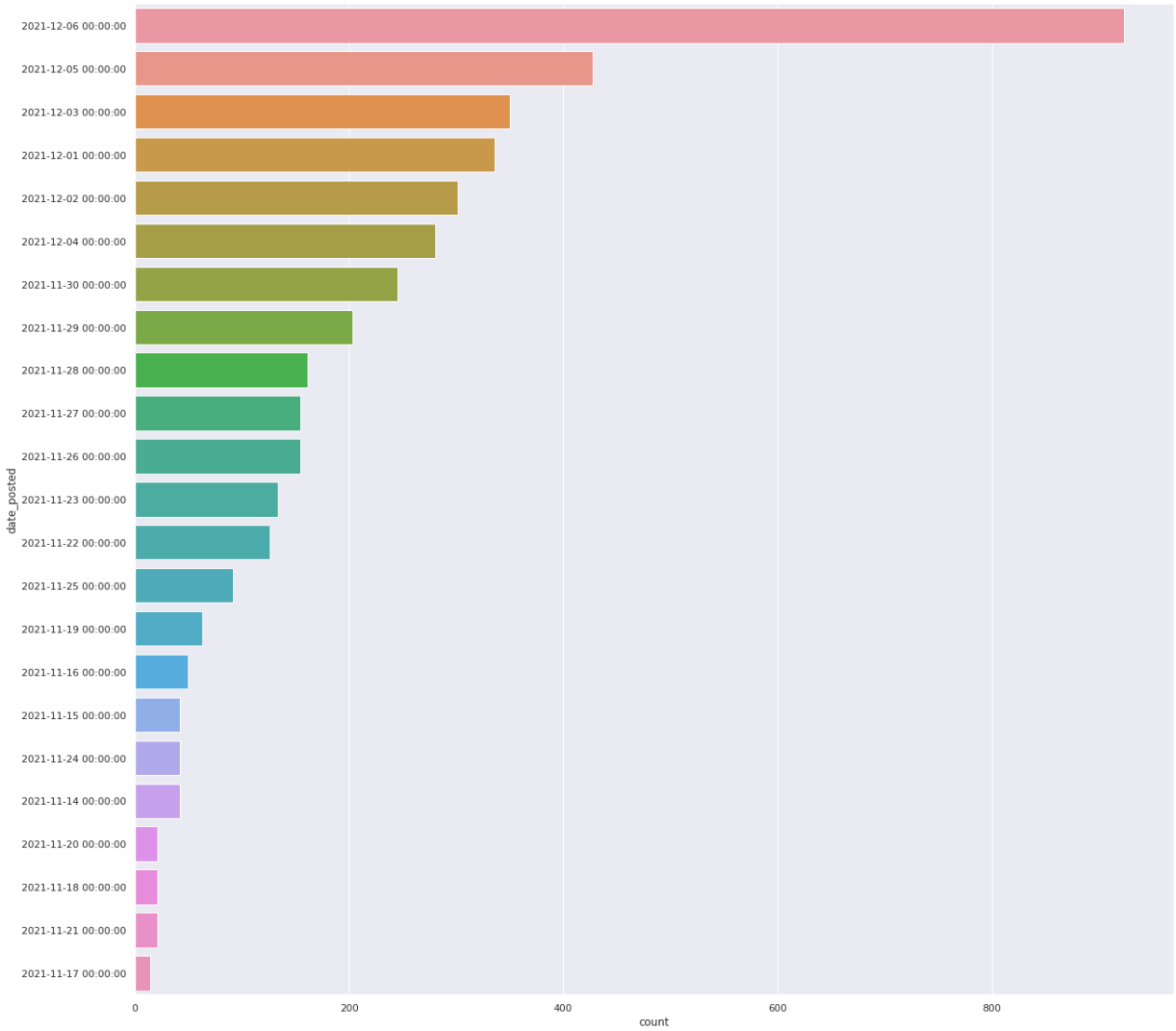

Figure 9: Reddit posts count posted per date. Only the dates with higher values are shown.

3.2 Analysis at named entity level.

This analysis is intended to elucidate variables that could be candidates to be considered in further analyses. This analysis will focus on specific features of the social media source for the different named entities considered.

3.2.1 Twitter named entity analysis

The Figure 10 depicts the count of tweets per language and named entity. The Twitter dataset contains higher number of tweets for the *covid* keyword and for each language considered. This fact could explain the uncertainty related to the average number of links per named entity as is shown in the Figure 11.

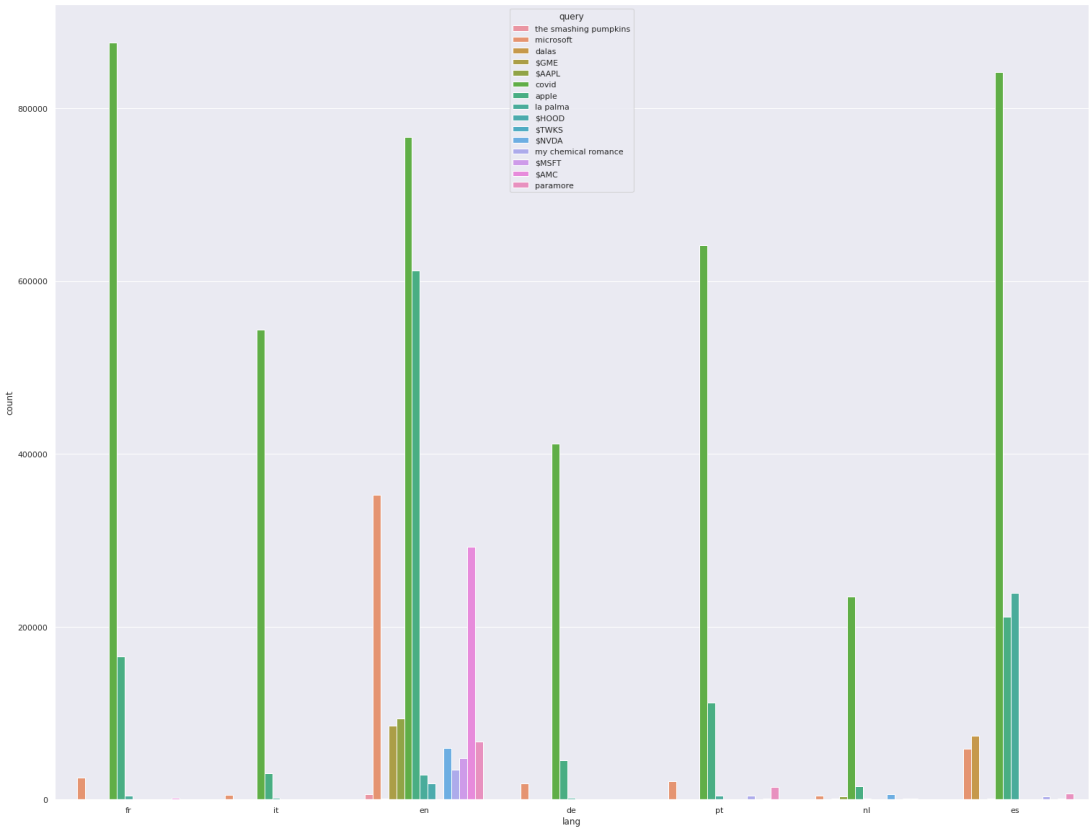

Figure 10: Tweet posts count posted per language and named entity.

The average number of links per named entity and the relatively low variance demonstrate that links could be a proper candidate to enrich, somehow, the tweet data with the content of the links and cover their inherent lack of information.

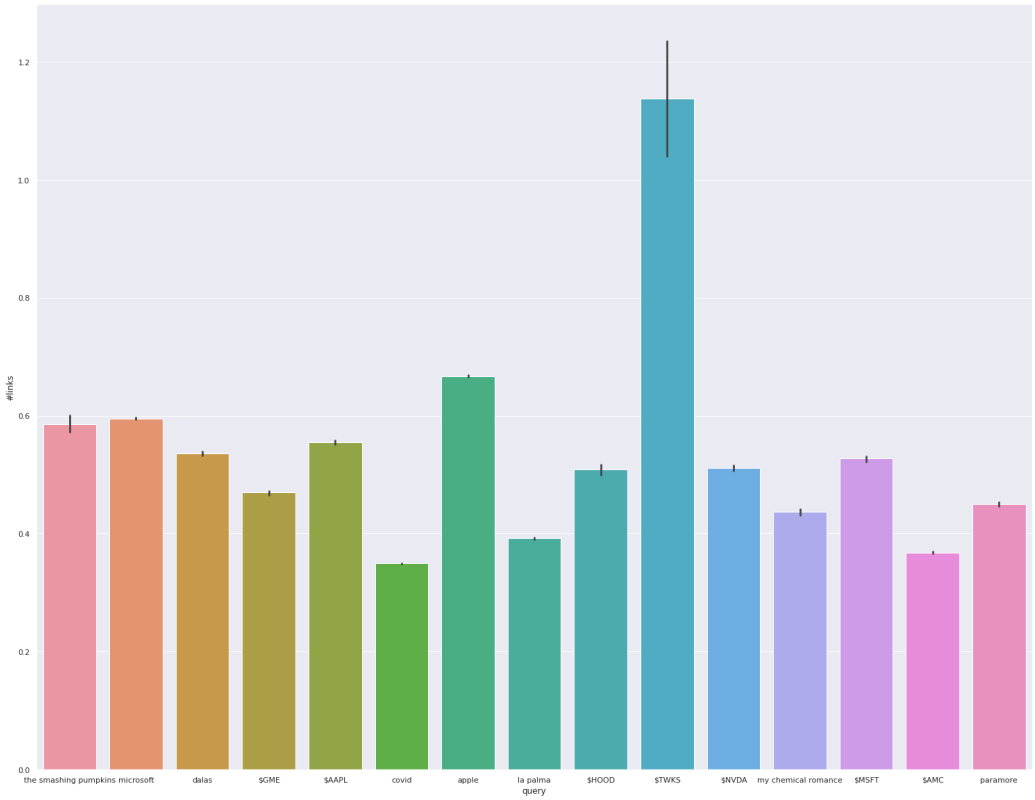

Figure 11: Average number of links per named entity.

A hashtag is a specific Twitter feature. Hashtags are intended to act as a summarization of the tweet in terms of the topic they cover. However, exploiting information from hashtags is a difficult task. First of all not all tweets will contain hashtags. The Figure 12 plots the average number of hashtags per named entity regardless the language. There are hardly any hashtags for the keyword *dalas* in contrast with the keyword *$AMC* which reaches the maximum average amongst the remaining keywords. However, it is important to highlight that the maximum average is 0.75 approximately and the average number of hashtags per named entity is, in general, small. Therefore, it is expected that there will be tweets that do not contain a hashtag.

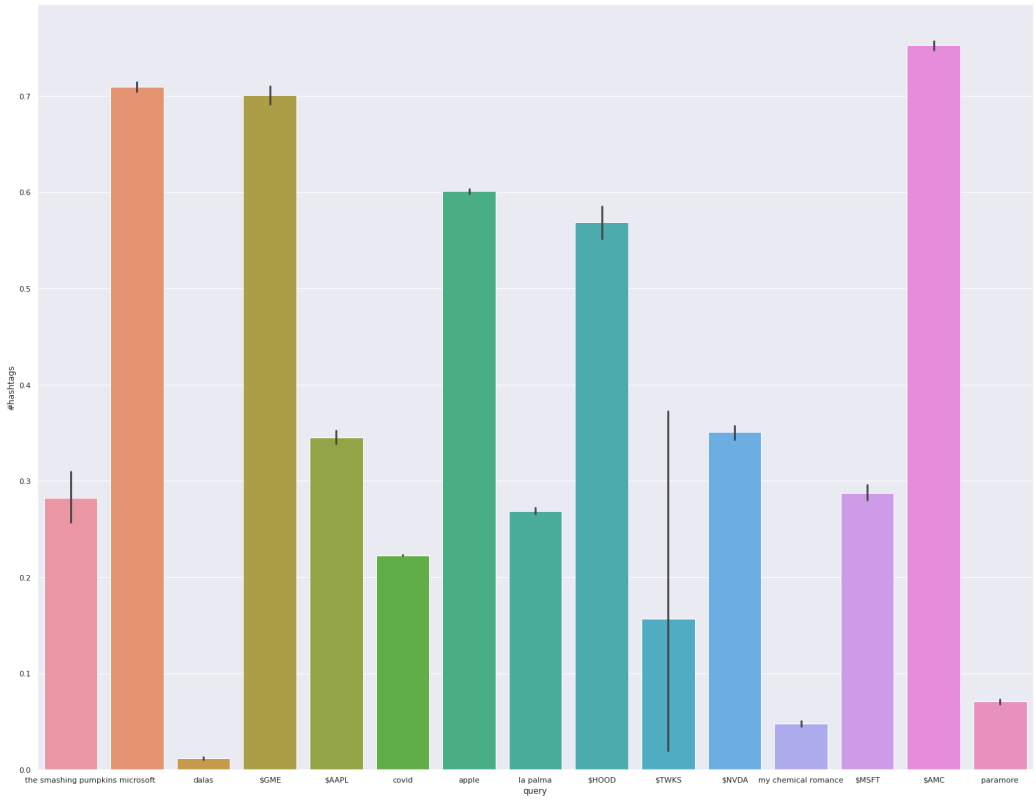

Figure 12: Average number of hashtags per named entity.

Lastly, a tweet may be a repost of other tweet. This is known as retweet and the number of retweets of a tweet could be a good indicator of its popularity. The Figure 13 shows the number of tweets that are a retweet of other for each named entity. For the *covid* keyword there are a large amount of retweets. However, for the remaining named entities, the number of tweets and retweets are similar.

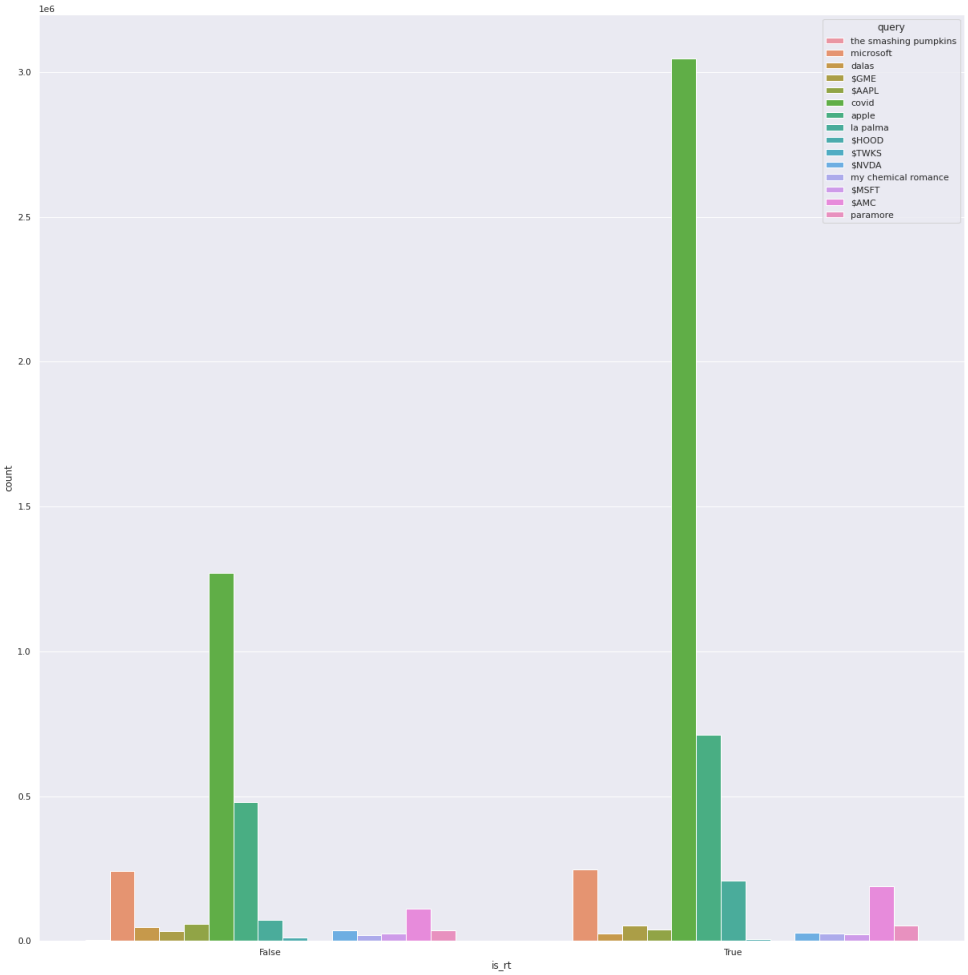

Figure 13: Count of tweets that are retweet per named entity.

3.2.2 Blog posts named entity analysis

For the blog posts dataset the average number of links per named entity is higher that the ones computed for the Twitter dataset. This reinforces the fact of using these links to enrich the textual data.

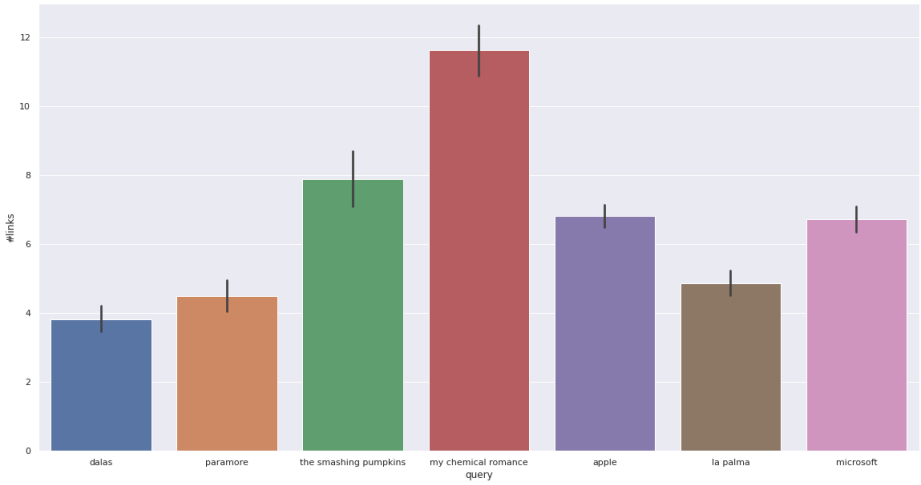

Figure 14: Average number of links per named entity.

The Figure 15 plots the average number of tags per named entity. These values vary drastically compared to the averages computed for the Twitter dataset.

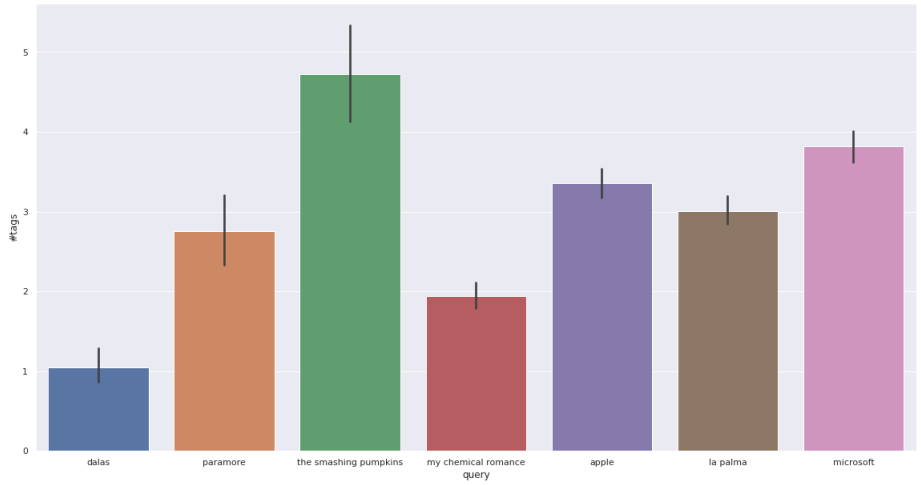

Figure 15: Average number of tags per named entity.

Finally, it seems that there are two predominant languages in this dataset since the number of blog posts written in English and Spanish for each named entity are greater than the number of posts for the remaining languages. The Figure 16 depicts these statistics.

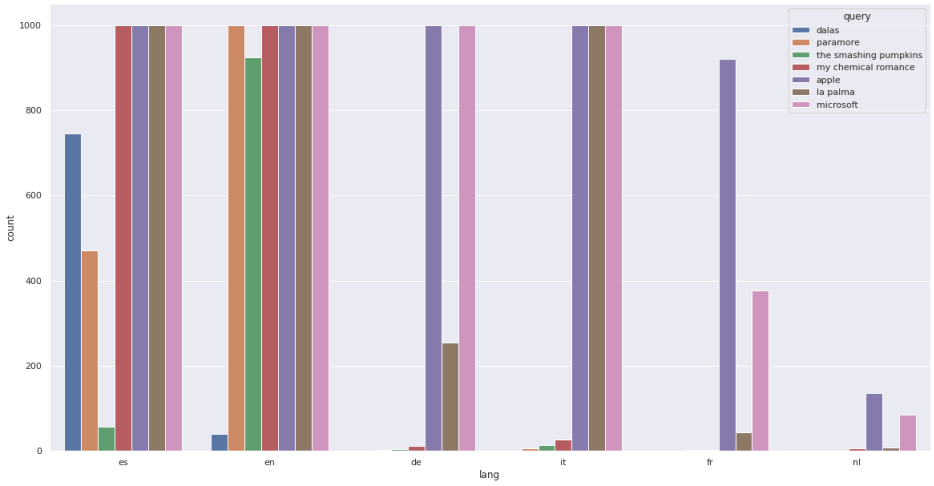

Figure 16: Blog posts count per language and named entity.

3.2.3 Reddit named entity analysis

Similar to the retweets, the number of comments and likes of a Reddit post could be a good indicator of popularity. The Figures 17 and 18 shows the average number of comments and likes per named entity. In general, the number of comments are similar except for the *apple* and *la palma* named entities.

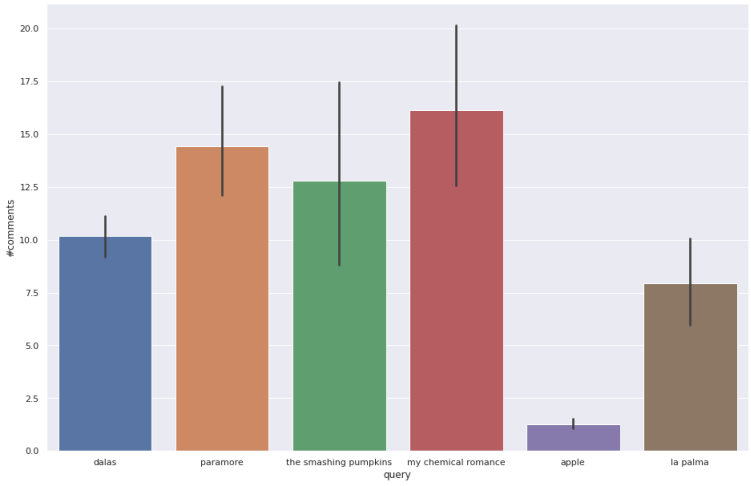

Figure 17: Average number of comments per named entity.

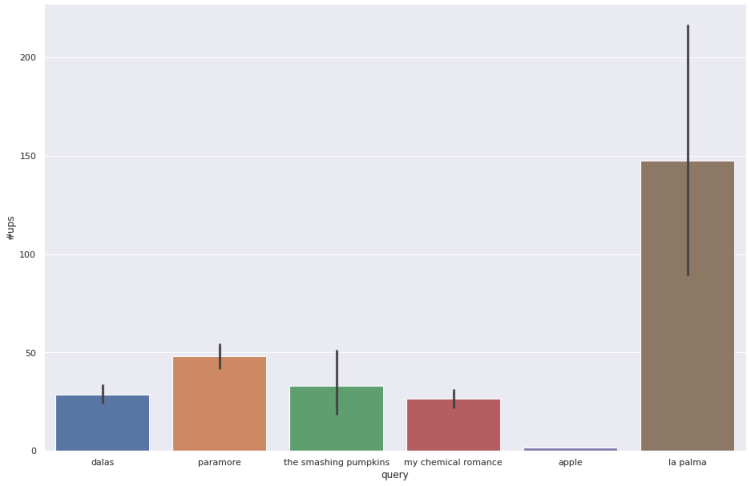

Figure 18: Reddit posts count per named entity.

As expected, the number of Reddit posts in English is the biggest. However, it is interesting to remark that *la palma* is a Spanish named entity. Despite of this, the number of posts in English containing this named entity are still greater than the number of posts for the Spanish language.

Although in this work only five languages are considered, the Figure 19 shows other languages. This is due to issues of the Social Searcher API.

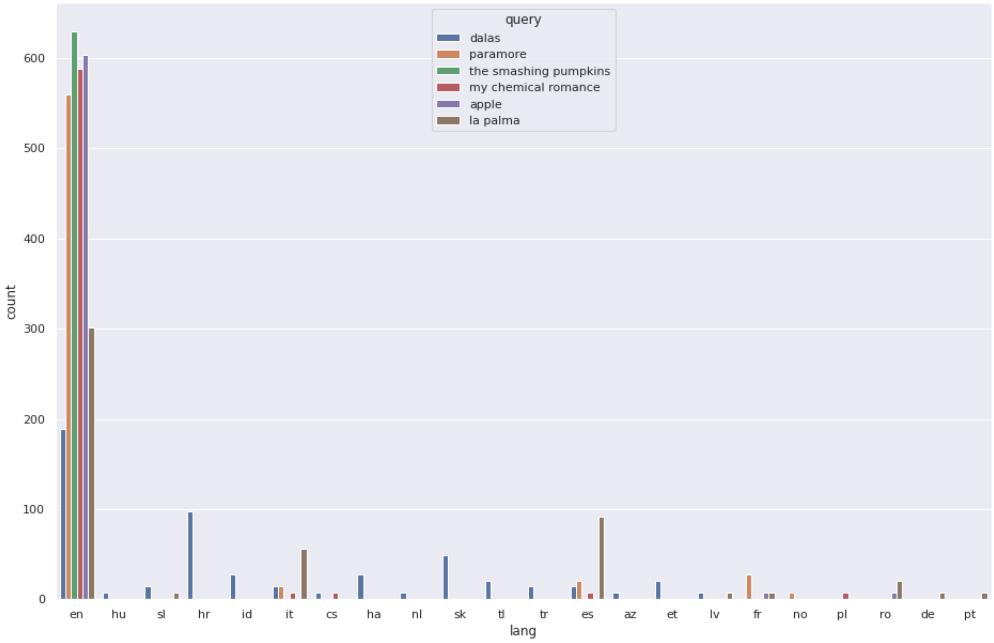

Figure 19: Blog posts count per language and named entity.

Finally, the number of Reddit posts per sentiment and named entity is shown in the Figure 20. The unbalancing of data with respect to this variable is not appropriate to train classifier models.

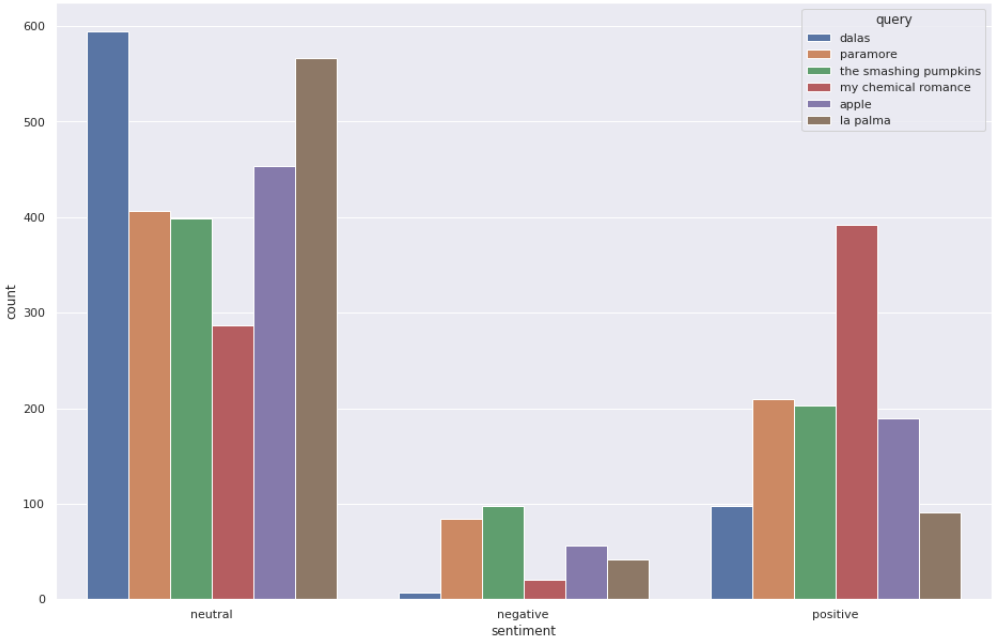

Figure 20: Reddit posts count per sentiment and named entity.

4 Adaptations of NLP tools

The specific lingo used in social networks requires to readapt the current NLP tools in order to capture these nuances. Hashtags, abbreviations, misspellings, emoticons or inconsistent capitalization or punctuation must be taken into account by the already trained models. Current tokenizers must cover these patterns. (Nikfarjam et al., 2015) coded regular expressions for tweets and reached a F-measure value of 0.96. However, to code manual rules implies that they probably will need to be reconsidered if the domain application is changed since they focused on tweets that mention drugs.

Part of speech taggers must face the out of vocabulary words (OOV) since the misspellings or short forms will be incorrectly classified by the current taggers. Hence, re-training the current models and extend the tag set is mandatory (Farzindar and Inkpen, 2020). (Toutanova et al., 2003) proposes a system based con Conditional Random Fields and clustering OOV words to reach an accuracy value of 0.85 but they did not consider a including new tags in the tag set. (Owoputi et al., 2013) extended the tag set for Twitter to cover hashtags, links, etc and obtained an accuracy value of 0.9. Likewise, chunkers and parsers must be reconditioned. (Khan et al., 2013) proposes a

dependency parser but was designed over specific Twitter dataset.

5 Conclusions

The analysis at corpora level arouses that there exists a dominant natural language across all the datasets. The amount of retrieved data in English surpasses the amount for the remaining considered languages. This fact must be a trigger for developing multilingual tools for Natural Language downstream tasks like PoS tagging, tokenizers or stemmers but designed to cover the nuances of the social data. For the Twitter dataset and the keyword *covid* there is a huge stream of tweets for any language considered regarding the Figure 10. On the other hand, the small amount of hashtags per named entity could reflect that exploiting semantic information from this Twitter specific lingo would likely be tough. For the blog posts dataset derived from Twingly there exist more posts related to certain named entities as shown in the Figure 16. This is probably due to the geographic location where the real object represented by the named entity belongs to. This fact could affect to the popularity of a given named entity in a specific country. Lastly, the unbalanced nature of the Reddit dataset could be a good starting point to build a balanced dataset at the level of the target variable to be predicted.

6 Further work

Deeper analysis must be conducted in order to determine relative importances of the words for a given dataset and named entity in order to exploit them in downstream tasks like topic modelling and event detection.

Since the Twitter and blog posts dataset are not annotated, unsupervised approaches could be executed in order to detect events or discover topics. Clustering algorithms and generative models like Latent Dirichlet Allocation (Blei et al., 2003) could be a relatively good starting point. However, to evaluate the quality of the clusters, intrinsic metrics must be considered since the dataset is unlabelled. Two metrics could be the average similarity within a cluster objects or the average similarity between the objects of one cluster with the objects of other clusters. However, if the dataset were annotated, extrinsic metrics like F-value or Normalized Mutual Information (NMI) could be used to evaluate the clusters quality (Curiskis et al., 2020).

References

Blei, D. M., Ng, A. Y., and Edu, J. B. (2003). Latent dirichlet allocation michael i. jordan.

Curiskis, S. A., Drake, B., Osborn, T. R., and Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034.

Farzindar, A. and Inkpen, D. (2020). Natural language processing for social media, third edition. *Synthesis Lectures on Human Language Technologies*, 13:1–219.

Khan, M., Dickinson, M., and Kuebler, S. (2013). Does size matter? text and grammar revision for parsing social media data. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 1–10, Atlanta, Georgia. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. pages 55–60.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22:671–681.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.

Tang, G., Xia, Y., Wang, W., Lau, R., and Zheng, F. (2014). Clustering tweets usingwikipedia concepts. In *LREC*, pages 2262–2267. Citeseer.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, page 173–180, USA. Association for Computational Linguistics.