Social data gathering

Álvaro Domínguez Calvo
adomingue599@alumno.uned.es

Abstract

1 Introduction

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 in the 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. 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. The table 2.1 summarizes this information.

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

<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="10">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>$AMC</td><td rowspan="8">Financial</td><td rowspan="8"></td></tr>
<tr><td>$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.

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 lasts months. However, for the Reddit dataset only 100 posts could be retrieve without taking into account of repeated posts.

2.2 Formatting process

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

```
        "user_id": ,
        "user_url":
}

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

// 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"
}
```

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 can be retrieved 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

3.1 Corpora analysis at dataset level

3.2 Corpora analysis at named entity level

4 Experiments

5 Conclusions