

#CANTSLEEP Insomnia and Covid-19: an analysis of the causes of sleep disturbances in a pandemic

Data Management and Visualization Project

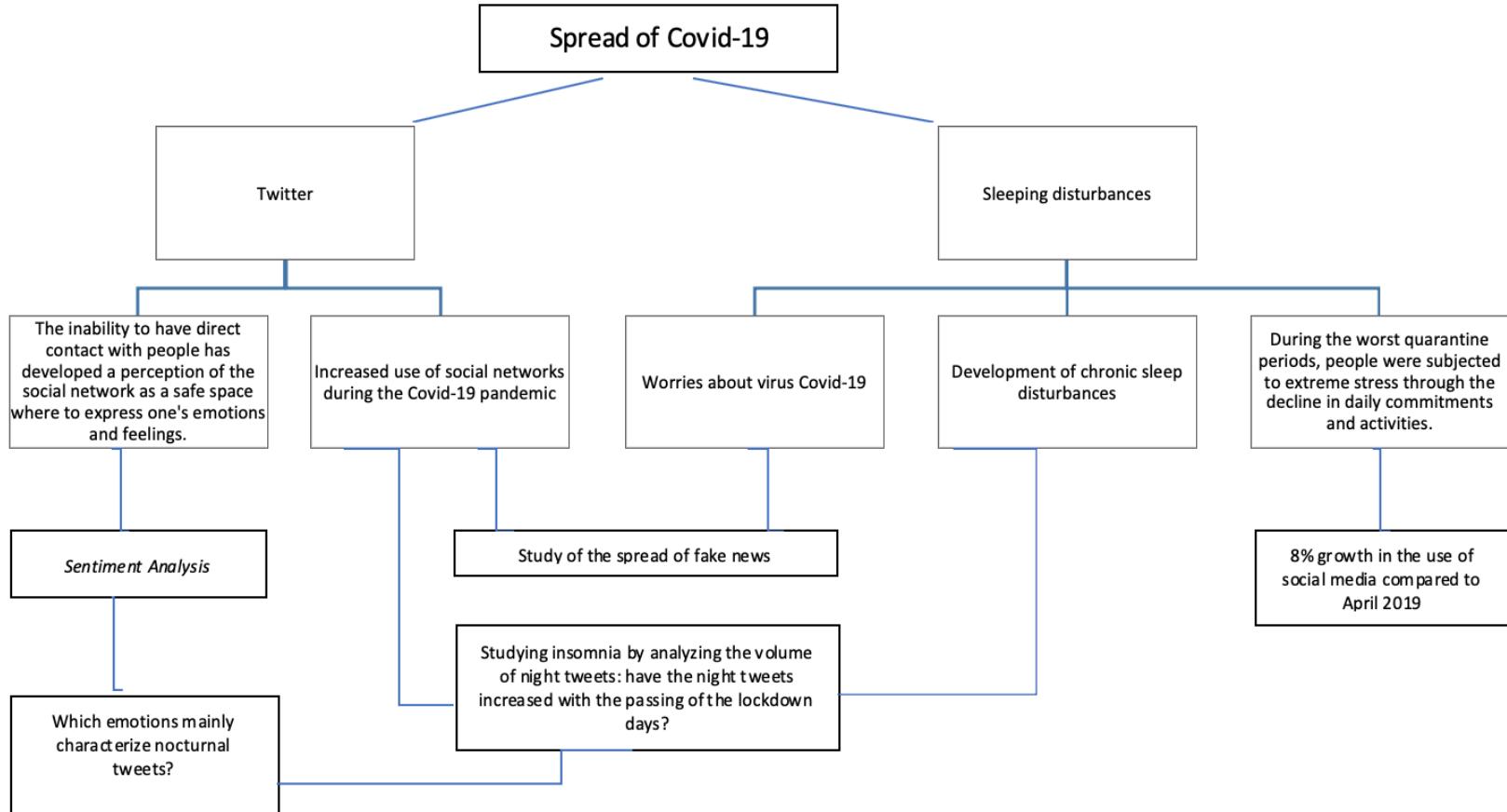
Università degli Studi di Milano Bicocca –CdLM in Data Science

Vittoria Porta, Alessandro Vincenzi, Fabio Pasquale Lombardi

Contents

- Introduction
- Data sources
- Data Extraction and Ingestion
- Real Time Analysis
- Sharded Architecture
- Analysis (data enrichment and transformation)
- Data Quality
- Sentiment Analysis
- Retweet Analysis and Spreading of the Fake News analysis
- Visualization

The main idea:



Introduction

Research questions:

Has the volume of night tweets increased with the passing of the days in the worst months of the pandemic?

If that were the case, people would tweet more at night since they can't sleep.

Which emotions mainly characterized night tweets?

Where the idea came from?

Generalized anxiety disorder, depressive symptoms and sleep quality during COVID-19 outbreak in China: a web-based cross-sectional survey}, Y. Huang,N. Zhao, June 2020



The technologies we used

Data Sources

- Twitter as primary source

For the **data enrichment** part were used three different data sources:

- The alphabetical list of all countries and capitals of the world
- countries.csv dataset
- List of U.S. state abbreviations

For the **data transformation** part was used the list of tz database time zones.

Finally, it was useful in terms of **visualization**, to focus attention on tweets from the United States of America (a place where Twitter is known to be one of the most used social networks) and for this purpose an additional collection of tweets was integrated.

```
kafka_2.12-2.4.1 — java -Xmx512M -Xms512M -server -XX:+UseG1GC -XX:MaxGCPau...
keeper.server.ZooKeeperServer)
[2020-07-03 11:57:29,470] INFO minSessionTimeout set to 6000 (org.apache.zookeeper.serv...
e.ZooKeeperServer)
[2020-07-03 11:57:29,470] INFO maxSessionTimeout set to 6000 (org.apache.zookeeper.serv...
er.ZooKeeperServer)
[2020-07-03 11:57:29,471] INFO Created server with tickTime 3000 minSessionTimeout 6000 ...
maxSessionTimeout 60000 datadir /tmp/zookeeper/version-2 snapdir /tmp/zookeeper/versio...
n-2 (org.apache.zookeeper.server.ZooKeeperServer)
[2020-07-03 11:57:29,509] INFO Using org.apache.zookeeper.server.NIOServerCnxnFactory a...
s server connection factory (org.apache.zookeeper.server.ServerCnxnFactory)
[2020-07-03 11:57:29,518] INFO Configuring NIO connection handler with 10s sessionless ...
connection timeout, 1 selector thread(s), 8 worker threads, and 64 kB direct buffers. (...
org.apache.zookeeper.server.NIOServerCnxnFactory)
[2020-07-03 11:57:29,532] INFO binding to port 0.0.0.0/0.0.0.0:2181 (org.apache.zookeep...
er.server.NIOServerCnxnFactory)
[2020-07-03 11:57:29,568] INFO zookeeper.snapshotSizeFactor = 0.33 (org.apache.zookeep...
r.server.ZKDatabase)
[2020-07-03 11:57:29,574] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2/snapshot.0 ...
(org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2020-07-03 11:57:29,581] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2/snapshot.0 ...
(org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2020-07-03 11:57:29,622] INFO Using checkIntervalMs=60000 maxPerMinute=10000 (org.apac...
he.zookeeper.server.ContainerManager)
[2020-07-03 11:57:45,185] INFO Creating new log file: log.1 (org.apache.zookeeper.serve...
r.log)

```

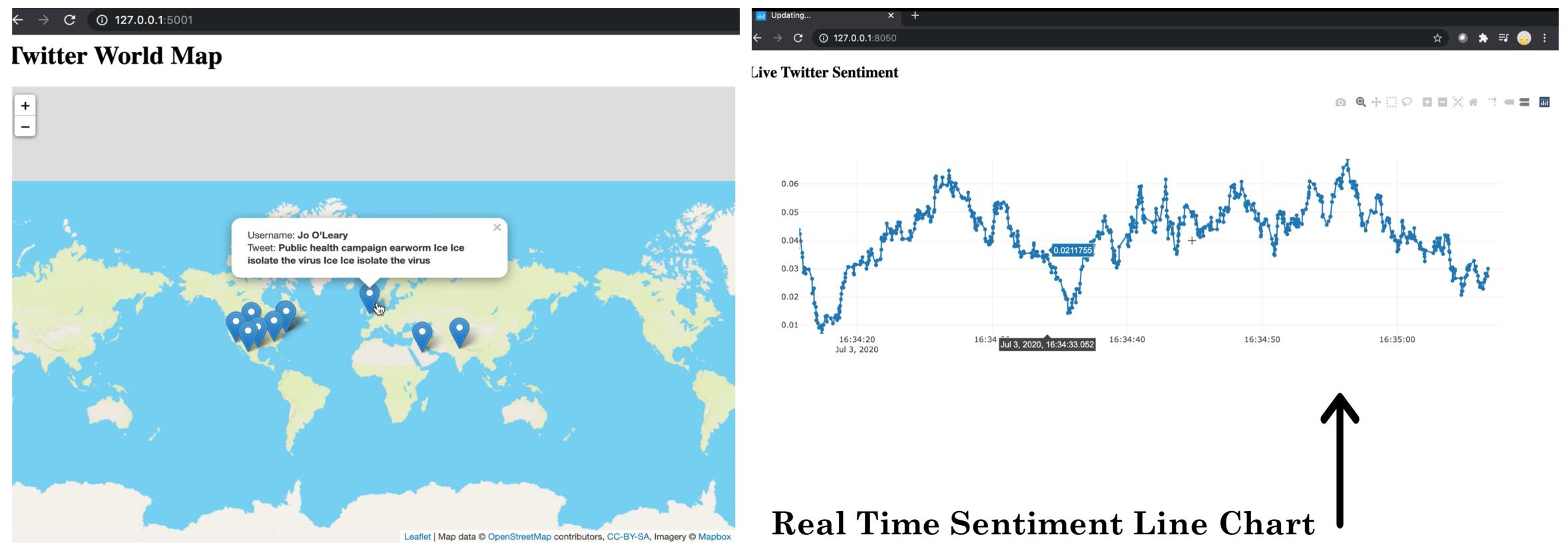
Data extraction: Tweepy

```
Data ingestion: Apache Kafka connection to Python through PyKafka  
Library
```

```
Connection to MongoDB and use of PyMongo
```

```
kafka_2.12-2.4.1 — java -Xmx512M -server -XX:+UseG1GC -XX:MaxGCPauseMillis=2...
,"display_url":"pic.twitter.com/WyUtZFhcXD","expanded_url":"https://twitter.com/Sk...
vNews/status/1278915883411398656/video/1","type":"video","video_info":{"aspect_rat...
io":1, "duration_millis":5980, "variants":[{"bitrate":432000, "content_type": "video...
/mp4", "url": "https://video.twimg.com/amplify_video/1278912360451846144/vid/320x3...
20/KxtR0gVOivki5c1D.mp4?tag=13"}, {"bitrate":1280000, "content_type": "video/mp4", "url...
": "https://video.twimg.com/amplify_video/1278912360451846144/vid/720x720/d_WqvtA...
QmTXnK8TG.mp4?tag=13}], {"content_type": "application/x-mpegURL", "url": "https://video...
.twimg.com/amplify_video/1278912360451846144/pl/NEUQMPb6fdb0500A.m3u8?tag=13"}, {"b...
itrate":832000, "content_type": "video/mp4", "url": "https://video.twimg.com/amplify_v...
ideo/1278912360451846144/vid/480x480/4PpnIh5xjQPRzJSY.mp4?tag=13"}], "sizes": {"thu...
mb": {"w": 150, "h": 150, "resize": "crop"}, "small": {"w": 680, "h": 680, "resize": "fit"}, "medium...
": {"w": 1080, "h": 1080, "resize": "fit"}, "large": {"w": 1080, "h": 1080, "resize": "fit"}}}}}, "...
quote_count": 47, "reply_count": 105, "retweet_count": 171, "favorite_count": 556, "entities": ...
{"hashtags": [], "urls": [{"url": "https://t.co/cPiwL2FAot", "expanded_url": "https://twi...
itter.com/i/web/status/1278915883411398656", "display_url": "twitter.com/i/web/s...
tatus/1\ud83d\udc06", "indices": [117, 140]}], "user_mentions": [{"screen_name": "BethRigby", "name...
": "Beth Rigby", "id": 19902709, "id_str": "19902709", "indices": [103, 113]}], "symbols": [], ...
, "favorited": false, "retweeted": false, "possibly_sensitive": false, "filter_level": "low", "l...
ang": "en"}, "is_quote_status": false, "quote_count": 0, "reply_count": 0, "retweet_count": 0, ...
favorite_count": 0, "entities": {"hashtags": [], "urls": [], "user_mentions": [{"screen_name": ...
"SkyNews", "name": "SkyNews", "id": 7587032, "id_str": "7587032", "indices": [3, 11]}, {"scr...
een_name": "BethRigby", "name": "Beth Rigby", "id": 19902709, "id_str": "19902709", "indices": [116, ...
126]}], "symbols": [], "favorited": false, "retweeted": false, "filter_level": "low", "lang": ...
"en", "timestamp_ms": "1593770316957"}
```

```
twitter_copy.py — 84x24
Last login: Fri Jul 3 11:56:06 on ttys003
(base) vitti@MacBook-di-Vittoria ~ % cd Desktop/DATAMANPROJ/code/twitterlive
(base) vitti@MacBook-di-Vittoria twitterlive % python twitter_copy.py
```



Preliminary Analysis: Real Time Interactive Dashboards

Since the data extraction process would take a long time, we wanted to have a tool that could visualize the data we were streaming in real time, allowing us to constantly check the quality of the data visually.

Real Time Sentiment Line Chart
technologies used: Python, Flask, vaderSentiment Python library (Valence Aware Dictionary and sEntiment Reasoner)

Live Twitter Map technologies used
Python, Flask, Leaflet

Demo available at
https://drive.google.com/drive/folders/1T8rJVkNA05zm8P_yzRKrGoVWSA0IR4nG?usp=sharing

Sharded Architecture

- The architecture that has been deployed has:
- 2 Shard clusters in replica set (3 nodes for each cluster)
 - mongoshard11, mongoshard12, mongoshard13, mongoshard21, mongoshard22, mongoshard23
- Config cluster in replica set (3 nodes)
 - mongocfg1, mongocfg2, mongocfg3
- 1 Router mongos instance
 - mongos1
- First, the collection was saved in MongoDB, after which, using the --mongorestore command, it has been reinserted into a new sharded collection.

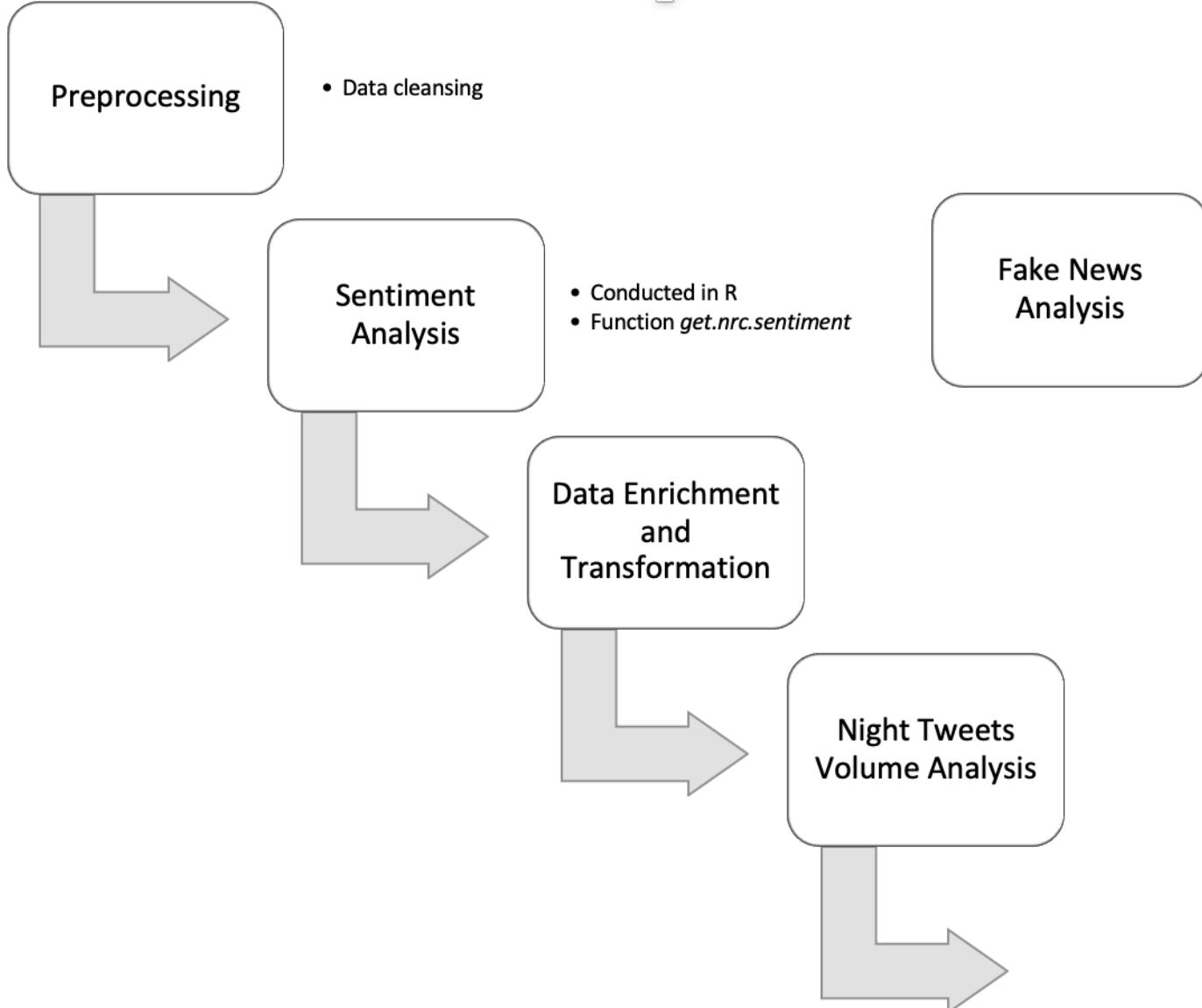
```
> db.TwitterCollection.getShardDistribution()

shard1 at shard1/mongoshard11:27017,mongoshard12:2701
: 1GiB docs : 174078 chunks : 31
nated data per chunk : 33.11MiB
nated docs per chunk : 5615

shard2 at shard2/mongoshard21:27017,mongoshard22:2701
: 1.02GiB docs : 178055 chunks : 31
nated data per chunk : 33.87MiB
nated docs per chunk : 5743

5
: 2.02GiB docs : 352133 chunks : 62
d shard1 contains 49.43% data, 49.43% docs in cluster,
d shard2 contains 50.56% data, 50.56% docs in cluster,
```

```
:\\Users\\User\\Desktop\\cluster\\mongodb-
tarting mongocfg3 ... done
tarting mongoshard22 ... done
tarting mongoshard23 ... done
tarting mongoshard13 ... done
tarting mongoshard12 ... done
tarting mongocfg2 ... done
tarting mongocfg1 ... done
tarting mongoshard21 ... done
tarting mongoshard11 ... done
tarting mongos1 ... done
```



Analysis

- **Conducted in R**
- **Pre-processing and missing values treatment and data filtering**
- **Data cleaning**
- Data Enrichment and Transformation on field ‘user.location’ and ‘created_at’

Data Quality



To analyse the locations, we assessed the presence of **homonyms** (same name for different concepts) that could give rise to any conflicts, both between states, between cities and between state-code.



Refer to secondary sources.



Constantly monitor the stream through real time dashes.

Sentiment Analysis

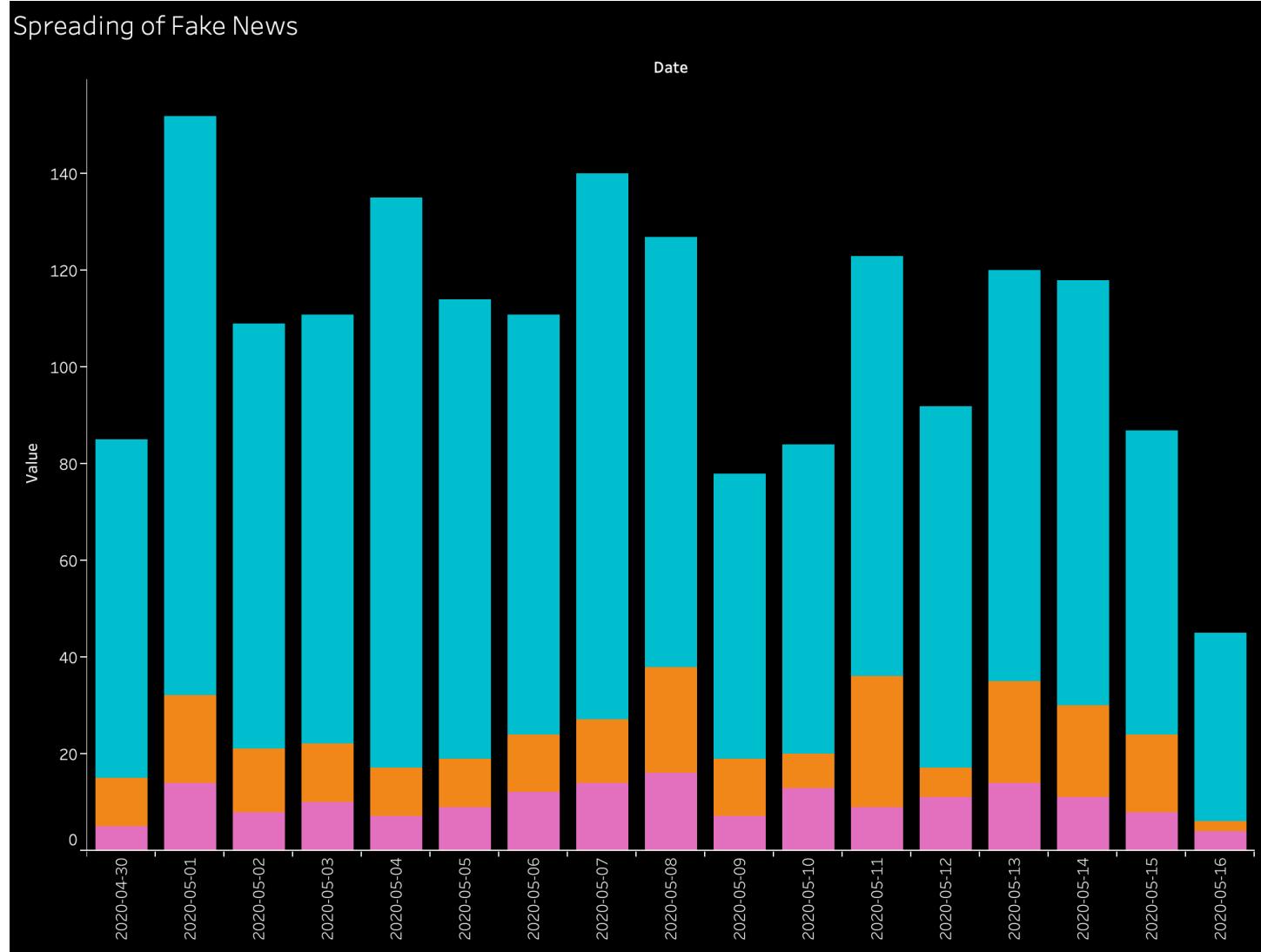
Python:
vaderSentiment
Python library
(Valence Aware
Dictionary and
sEntiment Reasoner)

R :get_nrc_sentiment'
which calls the NRC
sentiment dictionary to
calculate the presence of
eight different emotions

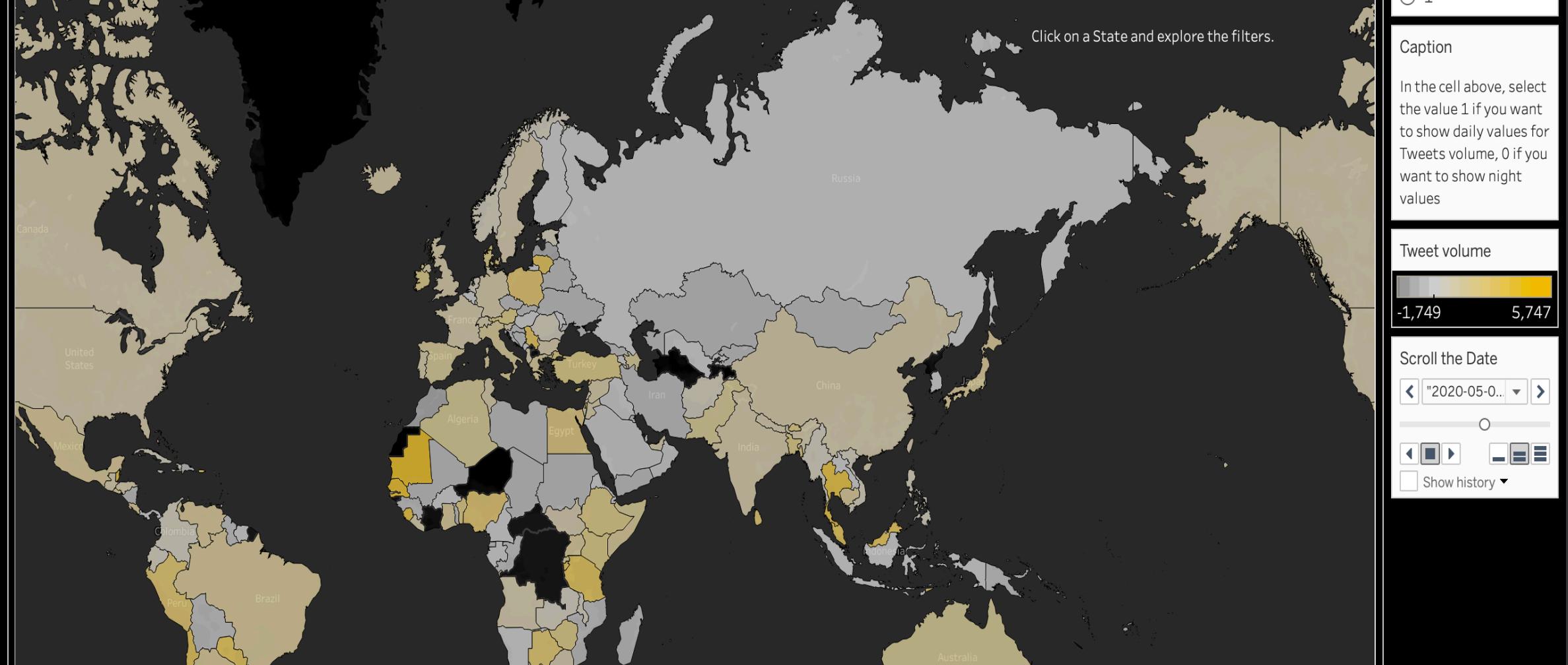


- anger
- anticipation
- disgust,
- fear
- joy
- sadness
- surprise
- trust

Retweet Analysis and Spreading of the Fake News analysis



- **Retweet Analysis:** the importance of creating a dataset *ad hoc*. Identify who has been retweeted and who retweeted.
- **Fake News Analysis:** The fake news concerning the Covid-19 virus can refer to three main categories: the conspiracy (or the belief that the virus was somehow the result of the failure of diplomacy between states); false homemade treatments and transmission. The analysis was conducted in order to search the tweet texts of the keywords that referred to one or more categories



Data Visualization

- Tableau
- Gephy
- Real Time Dashboards