BBCM Public Diplomacy

# DOCUMENT PURPOSE

To record and summarise the data-collection and language-analysis process, and provide a source of answers on methodology and results. For internal use.

# SEED ACCOUNTS: DATA

## Source Data

Two Excel files have been provided:

* **ChinaTwitter20200928.xlsx** containing consolidated Twitter accounts
* **ChinaSocialMedia20200928.xlsx** containing a mix of Twitter and Facebook accounts

Twitter accounts are provided as screen names.

Facebook accounts are provided as a mix of page links and handles.

## Processing

Both of these files have been processed to create a unified social-media account document:

* **lang\_analysis-2020-10-22/master\_file/all\_accounts\_master\_file\_v2.csv**
* The accounts in this file are de-duplicated.
* Along with the account’s handle/link, to assist linking back to the source data each row contains meta-data such as the platform type, and the file+column the data was extracted from.
* We provided account ids retrieved from the APIs for the collectable twitter and facebook accounts to assist linking back to the source data with the collections.
* We created our own structure of the source file which includes generic columns such as country, region, is\_diplomatic\_outlet, is\_state\_media, is\_confucious\_institute and account\_category and the rest of the information such as handles, languages.
* Note that this file include all the accounts collectable/not collectable, there are 12 accounts has no information retrieved from the API 8 of these accounts are not collectable and 4 of them are not retrieved neither by API nor crawler cause they are not available

## Collectability

The collectibility table can be found under **lang\_analysis-2020-10-22/collectibility**

### Twitter:

#### Q: How many accounts in total?

After deduplication based on screen name, there are 230 accounts.

#### Q: How do we determine collectability?

For the 230 accounts, we query the Twitter API to receive the account details for each. Account details are returned for public accounts, with errors returned for removed or private accounts.

#### Q: How many are collectable/uncollectable, summary of why?

225 of the 230 twitter accounts are collectable

5 resulted in errors; .chineseembsa, botschaftchina, ambassadeurlili, chine nouvelle, confuciusgsu all of these accounts are not available.

### Facebook:

#### Q: How do we determine collectability?

A: Scrape accounts ids using web crawler

B: Add the accounts to crowdtangle dashboard

1- Create two lists one for pages and one for groups

2- Create a batch file to add facebook accounts, this file require a specific format (page link and list name)

3- Add the accounts to crowdtangle lists then we query ct API to return the details of the accounts that has been added

4- We use the ids returned by ct to identify the collectable facebook accounts by compare each id returned form ct with the ids we scraped from using the web crawler if the extracted id is found the crowdtangle list then this account is collectable else not collectable

#### Q: How many accounts in total?

A: After deduplication based on handle/link and scraping the accounts ids using web crawler, there are 148 accounts.

#### Q: How many are collectable/uncollectable, summary of why?

A: 145 of the 148 twitter accounts are collectable

3 resulted in errors; Konfuzius-Institut, Институт Конфуция,Чжан Сяо all of these accounts have one thing in common: they are all user profile. Manually accessing the links show that the page is public but somehow crowdtangle couldn't retrieve this information.

# DATA COLLECTION: HISTORIC & ONGOING

## Twitter:

#### Q: What accounts are posts being collected from?

A: The posts is being collected from the accounts in **tw-accounts-v2-collectable.csv**

#### Q: What collection(s) have been set-up, any caveats about the resulting data?

Two collections have been configured: a historic backsearch, and an ongoing collection.

The historic backsearch has been configured to collect all available tweets, as far back as Twitter permits. The Twitter API can return (at most) the most-recent 3600 tweets for each user, these include retweets. Therefore, different timescales for each user will be present, depending on the posting-frequency of a given account.

The ongoing collection is run once per-day, and collects each account’s past two days of tweets, with two days set as a buffer to avoid gaps.

Tweets are always de-duplicated based on their unique tweet id.

#### Q: What is the timescale of the resulting data collection?

For the back search the collection goes from 30 Sep 2020 till 22 Jul 2009

For the ongoing collection from 30 Sep 2020

## Facebook:

#### Q: What accounts are posts being collected from?

A: The posts is being collected from the accounts in **fb-accounts-v2-collectable-20201014.csv**

#### Q: What collection(s) have been set-up, any caveats about the resulting data?

Two collections have been configured: a historic backsearch, and an ongoing collection.

The historic backsearch has been configured to collect posts from the past year (09-10-2019 to 07-10-2020). The data is collected with the CrowdTangle API, and should return all publically available posts in this time-period.

Similarly to the ongoing Twitter collection, the ongoing Facebook collection is run once per-day, and collects each account’s past two days of posts, with two days set as a buffer to avoid gaps.

Posts are always de-duplicated based on their unique Facebook post id.

Method52 collection jobs can be found at:

* <https://method52.casmconsulting.co.uk#!/job/bbc-collections:fb:fb-backsearch/bbc-monitoring>
* <https://method52.casmconsulting.co.uk#!/job/bbc-collections:fb:fb-ongoing-collection/bbc-monitoring>

#### Q: What is the timescale of the resulting data collection?

A: For the back search the collection goes from 30 Sep 2020 till 22 Jul 2009

B: For the ongoing collection from 6 Oct 2020

# Volume Over Time

* We have created pipelines to extract volume over time plots for the collections for both fb and tw
  + For the overall activity
    - <https://method52.casmconsulting.co.uk/index.html#!/job/vol-analysis:tw-volumes/bbc-monitoring#data-vol-analysis-tw-volumes>
    - <https://method52.casmconsulting.co.uk/index.html#!/job/vol-analysis:fb-volumes/bbc-monitoring#data-vol-analysis-tw-volumes>
  + For the timeframe activity
    - <https://method52.casmconsulting.co.uk/index.html#!/job/vol-analysis:fb-backsearch-volume-within-timeframe-copy/bbc-monitoring#visualisation-vol-analysis-fb-backsearch-volume-within-timeframe-copy>
    - <https://method52.casmconsulting.co.uk/index.html#!/job/vol-analysis:tw-backsearch-volume-within-timeframe/bbc-monitoring#visualisation-vol-analysis-fb-backsearch-volume-within-timeframe-copy>

# Choosing TimeFrame

Q: How did we choose the time frame:

A: Created a graph to show the year where the most users are active. It is obvious that this time should be year 2020 because we set our facebook back search over that year so any time period before 2020 will have users less than 2020

B: Created a graph to show the month where the most users are active. It turns out that Sep is the month with the highest number of active users users

Both graphs in A and B can be found in here **lang\_analysis-2020-10-22/timeframe selection /time-frame-with-the-highest-number-of-active-users.png**

C: The previous two graphs prompted us to choose a timeframe which is Sep-2020.

We created two graphs to show some statistics about the number of users in this time frame and number of posts. The first graph was the number of active users per day in Sep which shows the number of distinct users in each day of Sep-2020. This graph can be found under **lang\_analysis-2020-10-22/timeframe selection /The-number-of-active-users-per-day-in-September.png**

D: The final graph we have created was to show the number of posts/tweets in each day of Sep-2020 and if you compared this graph with the previous one you will see a correlation between the number of posts/tweets and the number of users. This graph can be found under **lang\_analysis-2020-10-22/timeframe selection /The-number-of-activities-per-day-in-September.png**

Q: Generate rate calculation for accounts within this timeframe (posts/Tweets a day)

A: This file is calculated by dividing the number of posts/tweets for each user by 30 and the results can be found in file **lang\_analysis-2020-10-22/timeframe selection /rate\_calculation\_per\_account\_within\_time\_frame.csv**

This file has column called **rate\_of\_user\_activity\_per\_day** which shows the rate of the posting/tweeting for each user per day (e.g. if the user produce 15 tweets/posts in Sep then the rate over the month will be 0.5 tweet/post per day)

# LANGUAGE ANALYSIS

Note that all the following language calculations are calculated within a chosen timeframe which is Sep 2020 as described in the previous section The overall number of tweets/posts within the timeframe is : 54328

#### Q: How is the language of a post being determined (using which field, and which process)?

A: For tw :

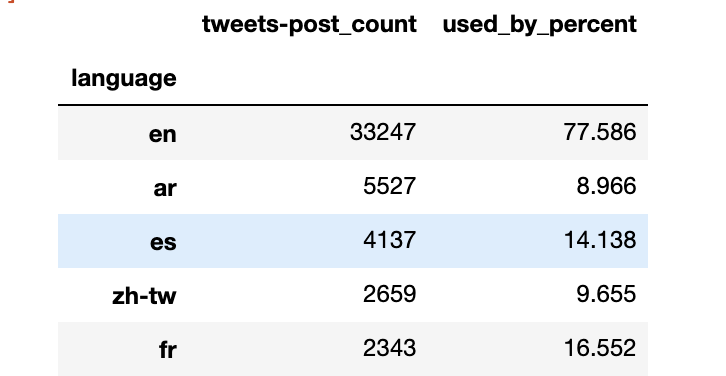
1. Read ongoing and back search collection tables
2. Use language annotator to annotate the tweets based on the text in the tweet (we annotate the tweets based on the actual language being used)

B: For fb :

1. Read ongoing and back search collection tables
2. Use language annotator to annotate the post based on the message column which represents the actual text in the post (similar to tw we annotate the posts based on the actual language being used)

Q: How is the language break down created?

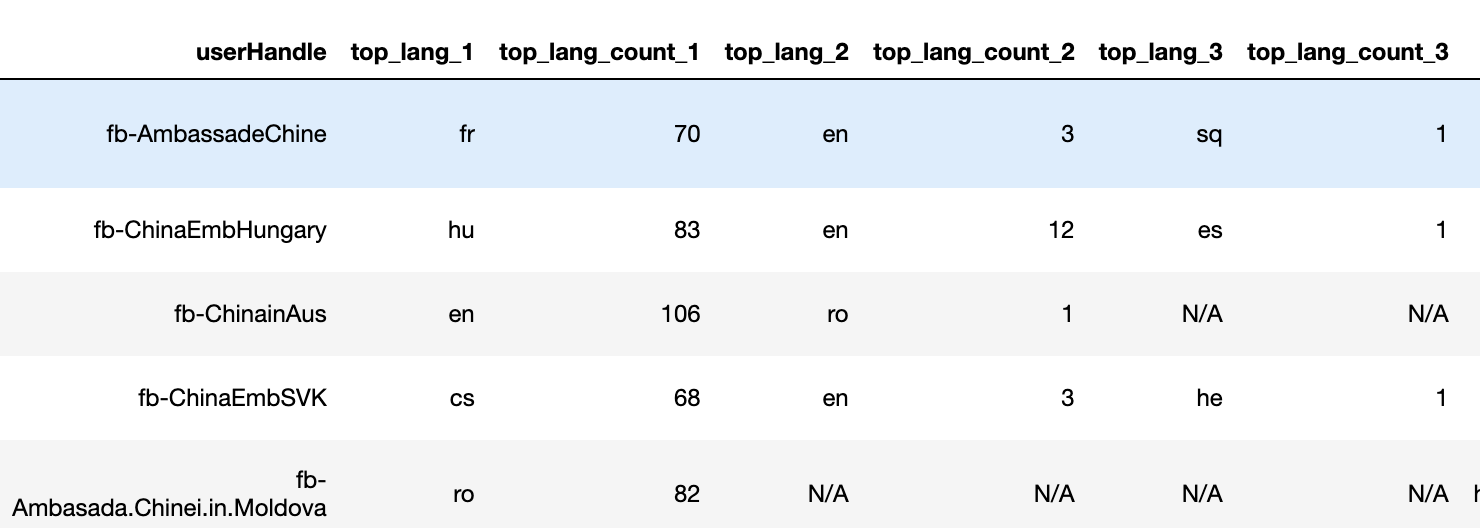
A: Created a summary statistics about the number of users used each language and for each language the number of posts/tweets in file called **language\_summary\_statistics.csv**

****

This is a screenshot of the file which can be found under **lang\_analysis-2020-10-22/language break down /language\_summary\_statistics.csv** folder**.** This file has a language column which represents each language , tweets-post-count which represents for each language the number of posts and tweets and the used\_by\_user which represents the percentage of users that used each language

B: Created a language break down based on the user in file called **user\_language\_break\_down.csv**

1. For each user we identify the number of languages used in his/her posts/tweets. The number of languages differ but we chose up to the top 3 languages.
2. Count for each user the number of times he used each language to post/tweet



This is a screenshot of the file which can be found under

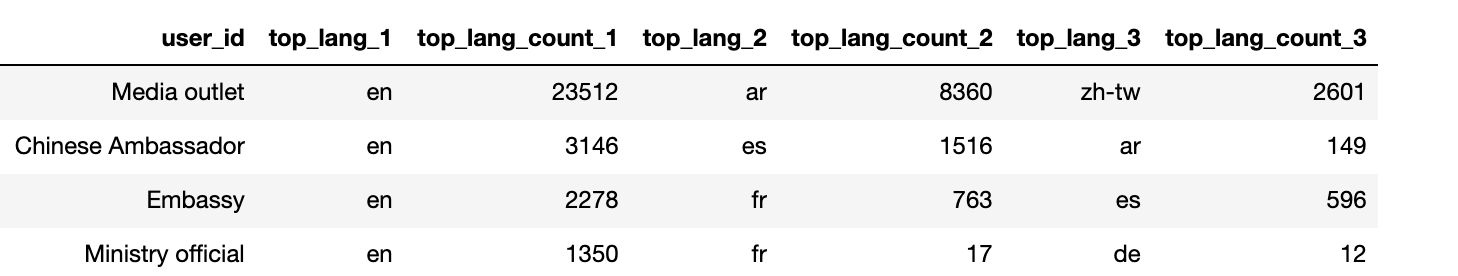
**lang\_analysis-2020-10-22/language break down /user\_language\_break\_down.csv**

Notice that this file has the full information about each user. For each user e.g. AmbassadeChina you can find the top 3 languages. Note that not all the users post in three languages. For each user we provide the **ISO code** of the top language and the number of times the user used it to post/tweet.

C: Finally we also created language breakdown based on the user category in called

**category\_language\_break\_down.csv**

1. For each category we identified the number of languages used in posts/tweets similar to users. We also chose up to 3 languages.
2. Count for each category the number of times a language used to post/tweet



This is a screenshot of the file which can be found under

**lang\_analysis-2020-10-22/language break down /category\_language\_break\_down.csv**

For each category e.g. media outlet you can find the top 3 languages. Note that not all the categories post in three languages (e.g Deputy chief of mission has only es). For each category we provide the **ISO code** of the top language and the number of times the category used it to post/tweet.