Data Mining of COVID-19 Project

**1. Data Collection**

***< Part 1>***

***Collect user Account ID who tweeted from Italy (IT), Poland (PL), and Japan (JP).***

Dataset 1 was collected from May 31, 2020 to June 2, 2020. Concluding unique user id (N = 100,125). Using Twitter streaming API to filter tweets meet the following 3 conditions.

(1) Tweets concluding more than one keywords.

(2) Tweets written in Italian or Polish or Japanese.

(3) Tweets posted from Italy or Poland or Japan.

|  |
| --- |
| **keywords** |
| ['corona', '#coronavirus', '#COVID19Italia', '#Koronawirus', '#COVID19Pandemic', '新型コロナウィルス', 'コロナ','新型肺炎', 'covid', 'covid19', 'sarscov2', '#corona virus', '#Coronavirus', 'SARS-CoV-2','covid-19', 'corona virus', '#2019nCoV', '#codvid\_19', '#codvid19', '#conronaviruspandemic', '#coronaflu', '#coronaoutbreak', '#coronapandemic', '#Coronapanik', '#coronavid19'] |
| *References of keywords choosing* |
| (1) IEEE Dataport <https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset>  (2) Andrzej Jarynowski, Wojta-Kempa, Belik (2020) Trends in Perception of COVID-19 in Polish Internet. doi: <https://doi.org/10.1101/2020.05.04.20090993>  (3) http://twita.di.unito.it/dataset/40wita  (4) Twitter COVID-19 Stream. <https://developer.twitter.com/en/docs/labs/covid19-stream/filtering-rules> |

**Languages**

[Italian (it), Polish (pl), Japanese (ja)]

**Geo-locations**

[7.05809, 36.71703, 18.37819, 46.99623] #IT

[14.24712, 49.29899, 23.89251, 23.89251] #PL

[124.15717, 24.34478, 145.575, 45.40944] #JP

(<https://www.geodatos.net/en/coordinates>)

As shown in Table 1, there was a deviation in counting the number of users who meet the country and language conditions. For some uncertain reason, some users are repeatedly marked as tweeted form more than one country (e. p. one user exits both in IT and JA country data). However, we cannot further determine in which country this user tweeted from, and similar users account for a small part. Therefore, we will filter out the repeatedly marked users in user posts collection part.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Filter by Language** | **Filter by Country** | | | | |
| Italy | Poland | Japan | Total | Actual Total |
| Italian user | 17261 | 14935 | 5099 | 37295 | 30777 |
| Polish user | 2324 | 2635 | 685 | 5644 | 5045 |
| Japanese user | 14544 | 9843 | 43934 | 68321 | 64368 |
| Total | 34129 | 27413 | 49718 | 111260 | 110190 |
| Actual Total | 34103 | 27428 | 49701 | 111232 | **100125** |

Table1. Number of user accounts who tweet concluding one of the keywords

|  |  |
| --- | --- |
| Data Name | Description |
| id | Tweet id |
| retweeted | Indicates whether this tweet has been retweeted by the authenticating user. (True/ False) |
| created at | Date when this Tweet was created. |
| favorite count | Number of this tweet was “likes”. |
| retweet count | Number of this tweet was retweeted. |
| lang | Tweet language. |
| coordinates | Represents the geographic location of this Tweet as reported by the user or client application. |
| text | Tweet text. |
| account id | User account id. |
| user followers count | Number of followers this account currently has. |
| description | The user-defined location for this account’s profile. |
| country | The Tweet including COVID-19 keywords was sent from which country, identified by latitude and longitude of the country. Tagged as IT, PL, and JP. |

Table2. Data description of Dateset1

(“account id” and “country” were what we need most)

***<Part 2>***

***Collect tweets and user account information from user id in dataset1.***

Using Twitter search API, we could get approximately 3,200 tweets data from each user (N = 100,125) in total. Although, we can’t be able to collect data from some of users due to their privacy settings or the possibility of deleting accounts. Data in the dataset2 was described as table3.

|  |  |
| --- | --- |
| Data Name | Description |
| id | Tweet id |
| text | Tweet text |
| created at | Time when this Tweet was created. |
| lang | Tweet language |
| retweeted | Indicates whether this tweet has been retweeted by the authenticating user. (True/ False) |
| retweet count | Number of this tweet was retweeted. |
| favorite count | Number of this tweet was “likes”. |
| possibly sensitive | This field only surfaces when a Tweet contains a link. The meaning of the field doesn’t pertain to the Tweet content itself, but instead it is an indicator that the URL contained in the Tweet may contain content or media identified as sensitive content. |
| coordinates | Represents the geographic location of this Tweet as reported by the user or client application. |
| in reply to status id | If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet’s ID. |
| quoted status id | This field only surfaces when the Tweet is a quote Tweet. This field contains the integer value Tweet ID of the quoted Tweet. |
| user id | User account id |
| user screen name | User screen name |
| user followers count | Number of followers this account currently has. |
| user statuses count | number of Tweets (including retweets) issued by the user. |
| user created at | Datetime that the user account was created on Twitter. |
| user description | The user-defined UTF-8 string describing their account. |
| user location | The user-defined location for this account’s profile. |

Table 3. Data Description of Dataset2

A sample of dataset2, concluding 378,192 tweets (114,369 in Italian; 1,539 in Polish; 214,494 in Japanese) were collected by 125 unique users from Italy. Due to 3 languages tweets in dataset2, we need to use different tools to do the pre-process and analysis.

**Pre-processing text data**

According to Kaggle(https://www.kaggle.com/ragnisah/ ), most text data are cleaned by following below steps.

(1) Remove punctuations.

(2) Tokenization, converting a sentence into list of words

(3) Remove stopwords.

(4) Lammetization/stemming, tranforming any form of a word to its root word

**Sentiment analysis** discloses the overall feelings inside tweet text data. The main approach of sentiment analysis could be divided as learning-based and lexicon-based methods (Mahmud et al., 2018). In addition, I’m considering the following methods.

Table 4. Analytical Tools for Sentiment Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Languages | Tools | Description | References |
| Italian | *Spark* | Label the tweet with positive, negative or neutral sentiment | [1] |
|  | *Sentix*  *(Sentiment Italian Lexicon)* | Consists in an italian lemma, part-of-speech (noun, verb, adjective, adverb), WordNet synset ID, a positive and a negative score from SentiWordNet, a polarity score ranging from -1 to 1, and an intensity score ranging from 0 to 1. | [2] |
|  | *Sentita* | “Results” is a list of strings with the sentence, the positive polarity score and the negative polarity scores. “polarities” is a list of lists with the positive and negative polarity score for each sentence | [3] |
|  | *LIWC*  (Linguistic Inquiry and Word Count) | Identifies emotion-related words as positive, negative or neutral.  <https://liwc.wpengine.com/> | [4], [5] |
|  | *SentiStrength* | Report two sentiment strengths: -1 (not negative) to -5 (extremely negative), 1 (not positive) to 5 (extremely positive). Estimate the strength of positive and negative sentiment in short texts, even for informal language. | [5], [6] |
| Polish | *Sentiment-analysis-on-Twitter-data* | Predict positive / negative sentiment and determine the best model for the tweets with unknown sentiment. | [7] |
|  | *Using Keras and Word2vec* | Predict the sentiment of Polish language texts as either positive, neutral or negative with the use of Python and Keras Deep Learning library. | [8] |
|  | *SentiStrength* | Same as above. | [5] |
| Japanese | *Asari* | Classify text into positive/negative class, without the need for training, and could get the confidence of positive and negative emotions | [9] |
|  | *PN Table* | Classify text into positive/negative class, and confidence (-1~1) | [10] |
|  | *Sentimentja* | Classify emotions as happy, sad, disgust, angry, fear and surprise.  Could choose to use stable version or train our own model. | [11] |

References

[1]. Italian-Sentiment-Analysis-with-Spark,

<https://github.com/giuseppegambino/Italian-Sentiment-Analysis-with-Spark/blob/master/tweetSentimentRadici.py>

[2]. Sentix, Twita, <http://valeriobasile.github.io/twita/downloads.html>

[3]. Sentita, a sentiment analysis tool for Italian,

<https://nicgian.github.io/Sentita/>

[4]. Salas-Zárate MP, Paredes-Valverde MA, Rodríguez-García MÁ, Valencia-García R, Alor-Hernández G (2017). Sentiment Analysis Based on Psychological and Linguistic Features for Spanish Language. Intelligent Systems Reference Library: 73–92. Available: <http://dx.doi.org/10.1007/978-3-319-51905-0_4>.

[5]. Analytical Text Programs. <https://quanttext.com/analytical-text-programs/>

[6]. SentiStrength, <http://sentistrength.wlv.ac.uk/>

[7]. <https://github.com/AleksandraWozniak/Sentiment-analysis-on-Twitter-data>

[8]. POLISH SENTIMENT ANALYSIS USING KERAS AND WORD2VEC <https://ermlab.com/en/blog/nlp/polish-sentiment-analysis-using-keras-and-word2vec/>

[9]. <https://github.com/Hironsan/asari>

[10]. 単語感情極性対応表　高村大也, 乾孝司, 奥村学（2006スピンモデルによる単語の感情極性抽出, 情報処理学会論文誌ジャーナル,Vol.47 No.02 pp. 627—637. <http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html>

<https://blog.aidemy.net/entry/2017/08/10/170715>

[11]. PythonとKerasによるディープラーニング Francois Chollet, 巣籠悠輔, 株式会社クイープ.　<https://github.com/sugiyamath/sentiment_ja>

[12]. Awesome Sentiment Analysis. <https://awesomeopensource.com/project/laugustyniak/awesome-sentiment-analysis>

[13]. Mahmud M, Kaiser MS, Hussain A et al. Applications of deep learning and reinforcement learning to biological data. *IEEE T Neur Net Learn Syst* 2018; 29(6): 2063–2079.

7/15/20

Sample selection

Poland & Polish 2635

user followers count < 10000 2522

Filter text concluding

'corona|Koronawirus|koronawirus|coronavirus|covid|pandemic|covid19|sarscov2|Coronavirus|SARS-CoV-2|covid-19|corona virus|2019nCoV|codvid19|coronaoutbreak|coronaflu|Coronapanik'

482 / No 2→ 478

random 400

follower\_count > 20000 15 Check them Official

>10000 23 official

Italy

Italian 242

follower<10000 209

~~33622 in total → text containing keywords 6817 → first account id 4997~~

33622→ first id 17261 → keywords 3847 → followers<10000 3693

17261 keep account id first

415/33622 already collected