IEEE CS Global Student Challenge (GSC-21)

Challenge 2: Multi-label Sentiment Analysis for Covid-19

Competition Kickoff Webinar

May 10, 2021

Xiangliang Zhang, Qiang Yang

King Abdullah University of Science and Technology (KAUST)，Saudi Arabia --- The sponsor of GSC-21

Multi-label Sentiment Analysis Challenge

- Covid-19 Tweets Dataset Collection
- Dataset Annotation
- Dataset Information
- Baseline Model Performance
- Challenge Website and Submission

Covid-19 Tweets Dataset Collection

- The dataset was collected by, an open source Twitter crawler, called Twint[1].
- The query includes *covid-19, coronavirus, covid, corona*, etc.
- The matched tweets were saved as JSON documents and pre-processed, by noise removing.
- The final dataset is presented as CSV document.

[1] https://github.com/twintproject/twint

Dataset annotation

- The sentiment categories were determined by domain experts after reviewing a subset of the collected tweets and discussing for several rounds.
- The final determined set of labels reflect the complicated sentiments in pandemic.

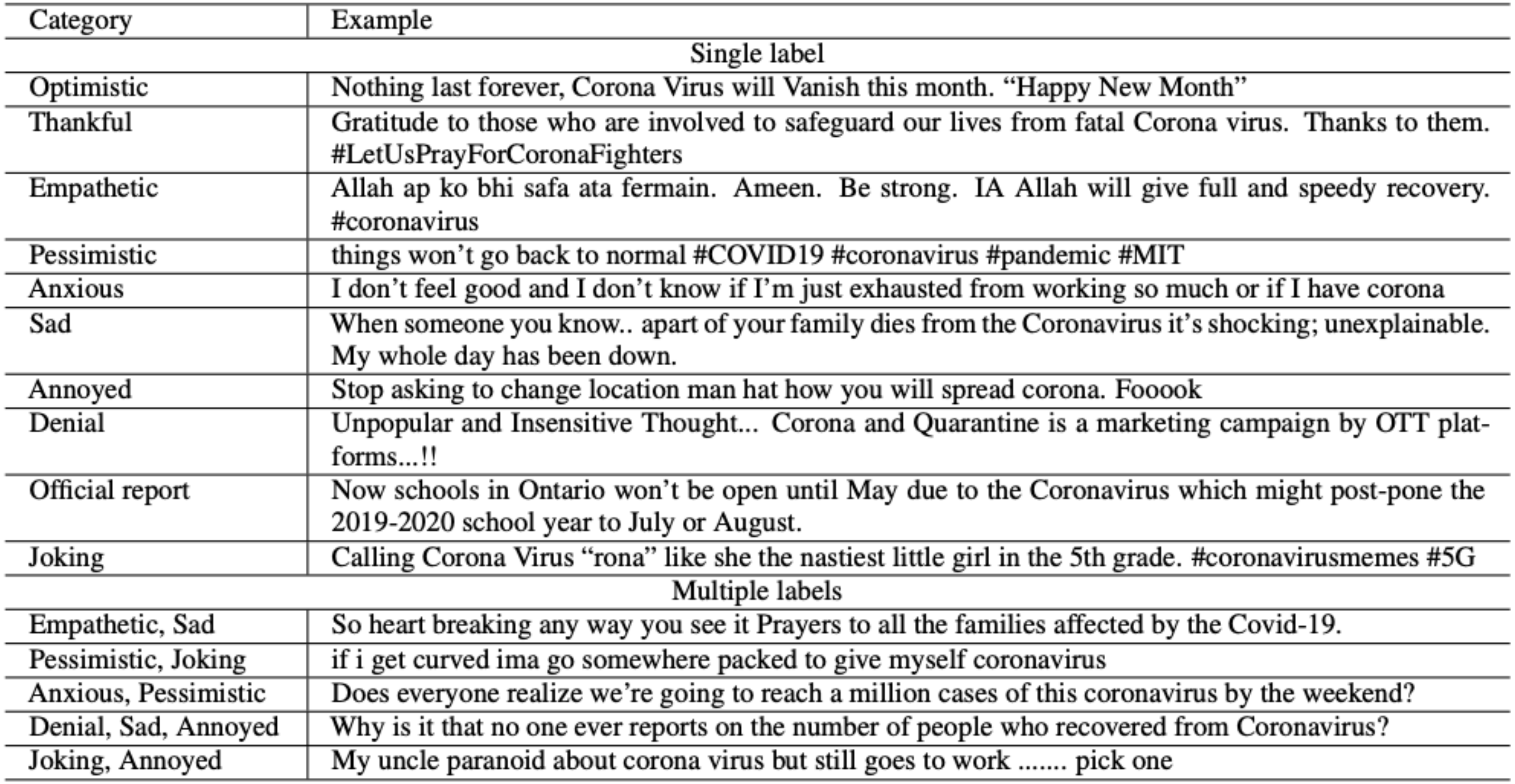
  1. ***optimistic*** (representing *hopeful, proud, trusting*),
  2. ***thankful*** for the efforts to combat the virus,
  3. ***empathetic*** (including *praying*),
  4. ***pessimistic*** (*hopeless*),
  5. ***anxious*** (*scared, fearful*),
  6. ***sad***,
  7. ***annoyed*** (*angry*),
  8. ***denial*** towards conspiracy theories,
  9. ***surprise*** (unprecedented),
  10. ***official report***,
  11. ***joking*** (*ironical*).

Dataset annotation(Cont.)

- We recruited over 50 experienced annotators to make every tweet labeled by **at least three** annotators.
- Example tweets were provided in advance to annotators with suggested categories.
- Each tweet was allowed to be assigned to **multiple labels.**
- IRA and Kappa coefficient: 0.904 and 0.381.

Dataset information

Examples



Category	Example
	Single label
Optimistic	Nothing last forever, Corona Virus will Vanish this month. “Happy New Month”
Thankful	Gratitude to those who are involved to safeguard our lives from fatal Corona virus. Thanks to them. #LetUsPrayForCoronaFighters
Empathetic	Allah ap ko bhi safa ata fermain. Ameen. Be strong. IA Allah will give full and speedy recovery. #coronavirus
Pessimistic	things won’t go back to normal #COVID19 #coronavirus #pandemic #MIT
Anxious	I don’t feel good and I don’t know if I’m just exhausted from working so much or if I have corona
Sad	When someone you know.. apart of your family dies from the Coronavirus it’s shocking; unexplainable. My whole day has been down.
Annoyed	Stop asking to change location man hat how you will spread corona. Fooook
Denial	Unpopular and Insensitive Thought... Corona and Quarantine is a marketing campaign by OTT platforms...!!
Official report	Now schools in Ontario won’t be open until May due to the Coronavirus which might post-pone the 2019-2020 school year to July or August.
Joking	Calling Corona Virus “rona” like she the nastiest little girl in the 5th grade. #coronavirusmemes #5G
	Multiple labels
Empathetic, Sad	So heart breaking any way you see it Prayers to all the families affected by the Covid-19.
Pessimistic, Joking	if i get curved ima go somewhere packed to give myself coronavirus
Anxious, Pessimistic	Does everyone realize we’re going to reach a million cases of this coronavirus by the weekend?
Denial, Sad, Annoyed	Why is it that no one ever reports on the number of people who recovered from Coronavirus?
Joking, Annoyed	My uncle paranoid about corona virus but still goes to work .... pick one

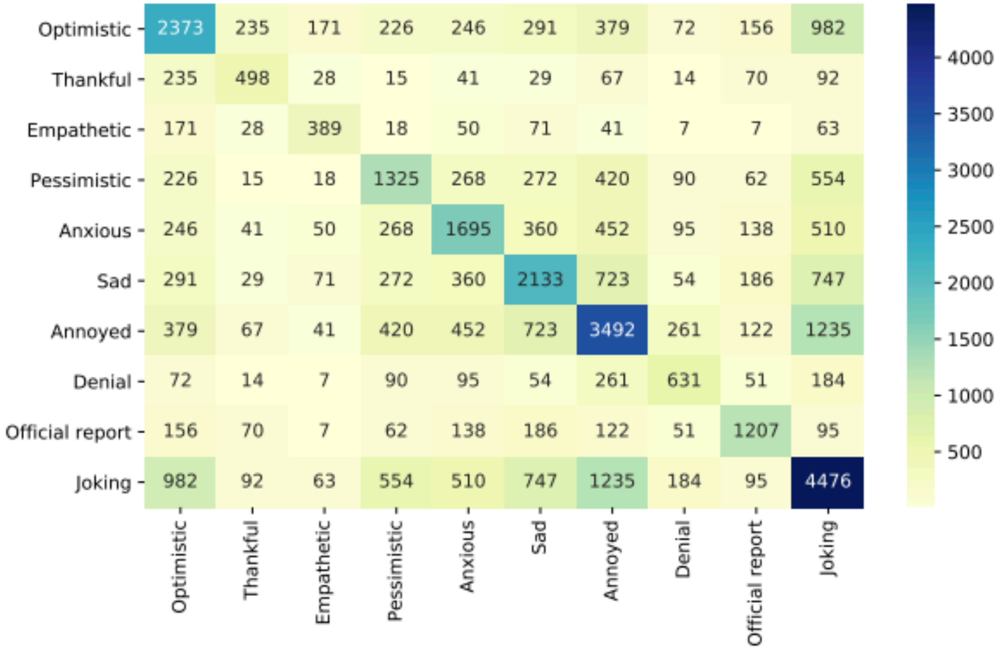
Dataset information (Cont.)

- # of tweet samples: 10K

- Label distribution

Opti.	Than.	Empa.	Pess.	Anxi.	Sad	Anno.	Deni.	Offi.	Surp.	Joki.
0.2373	0.0498	0.0389	0.1325	0.1695	0.2133	0.3492	0.0631	0.1207	0.1820	0.4476

- Heatmap of label co-occurrence



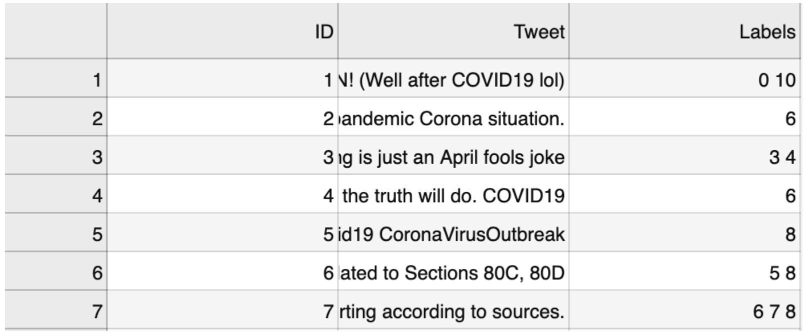
	Optimistic	Thankful	Empathetic	Pessimistic	Anxious	Sad	Annoyed	Denial	Official report	Joking
Optimistic	2373	235	171	226	246	291	379	72	156	982
Thankful	235	498	28	15	41	29	67	14	70	92
Empathetic	171	28	389	18	50	71	41	7	7	63
Pessimistic	226	15	18	1325	268	272	420	90	62	554
Anxious	246	41	50	268	1695	360	452	95	138	510
Sad	291	29	71	272	360	2133	723	54	186	747
Annoyed	379	67	41	420	452	723	3492	261	122	1235
Denial	72	14	7	90	95	54	261	631	51	184
Official report	156	70	7	62	138	186	122	51	1207	95
Joking	982	92	63	554	510	747	1235	184	95	4476

Dataset information (Cont.)

- Data splits for the challenge
  - Training data (5K) : validation data (2.5K) : testing data (2.5K) = 0.5: 0.25: 0.25

Type	Opti.	Than.	Empa.	Pess.	Anxi.	Sad	Anno.	Deni.	Offi.	Surp.	Joki.
Training	0.2360	0.0488	0.0372	0.1240	0.1684	0.2176	0.3450	0.0628	0.1208	0.1828	0.4514
Validation	0.2524	0.0560	0.0396	0.1444	0.1744	0.2020	0.3456	0.0604	0.1232	0.1840	0.4420
Testing	0.2248	0.0456	0.0416	0.1376	0.1668	0.2160	0.3612	0.0664	0.1180	0.1784	0.4456

  - Quick glance of the data



	ID	Tweet	Labels
1	1	N! (Well after COVID19 lol)	0 10
2	2	andemic Corona situation.	6
3	3	g is just an April fools joke	3 4
4	4	the truth will do. COVID19	6
5	5	id19 CoronaVirusOutbreak	8
6	6	ated to Sections 80C, 80D	5 8
7	7	rting according to sources.	6 7 8

Label index varying from 0 to 10:

Optimistic (0), Thankful (1), Empathetic (2), Pessimistic (3),

Anxious (4), Sad (5), Annoyed (6),

Denial (7), Surprise (8), Official report (9), Joking (10)

Baseline model performance

- Baseline model
  - The data were pre-processed, e.g., removing user information, interactions(e.g., retweet, like), emojis and emoticons, and filtering out noisy symbols and texts, such as retweet symbol “RT” and hyperlinks and some special symbols including line break, tabs and redundant blank characters, and word tokenization, steaming and tagging with the NLTK tool[2].
  - The model use the **pretrained XLNet[3] for multi-label text classification** where a fully connected network with the sigmoid activation function was added to finetune the embeddings.
  - Performance
    - **F1-macro: 0.50**
    - **F1-micro: 0.56**

[2] https://www.nltk.org/

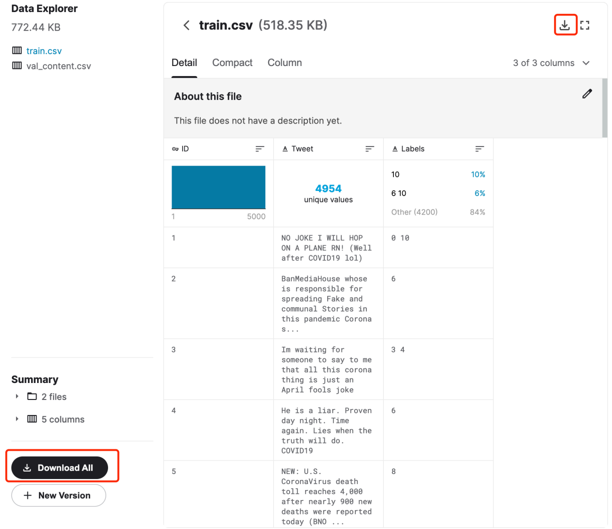
[3] XLNet: Generalized Autoregressive Pretraining for Language Understanding. Zhilin Yang et al. NeurIPS 2019

Challenge Website

- Competition website:

https://www.kaggle.com/c/sentiment-analysis-of-covid-19-related-tweets

- Data acquisition:
  - Using Kaggle API to download data with the terminal command (kaggle competitions download -c sentiment-analysis-of-covid-19-related-tweets)
  - Download data directly from the web page



About this file

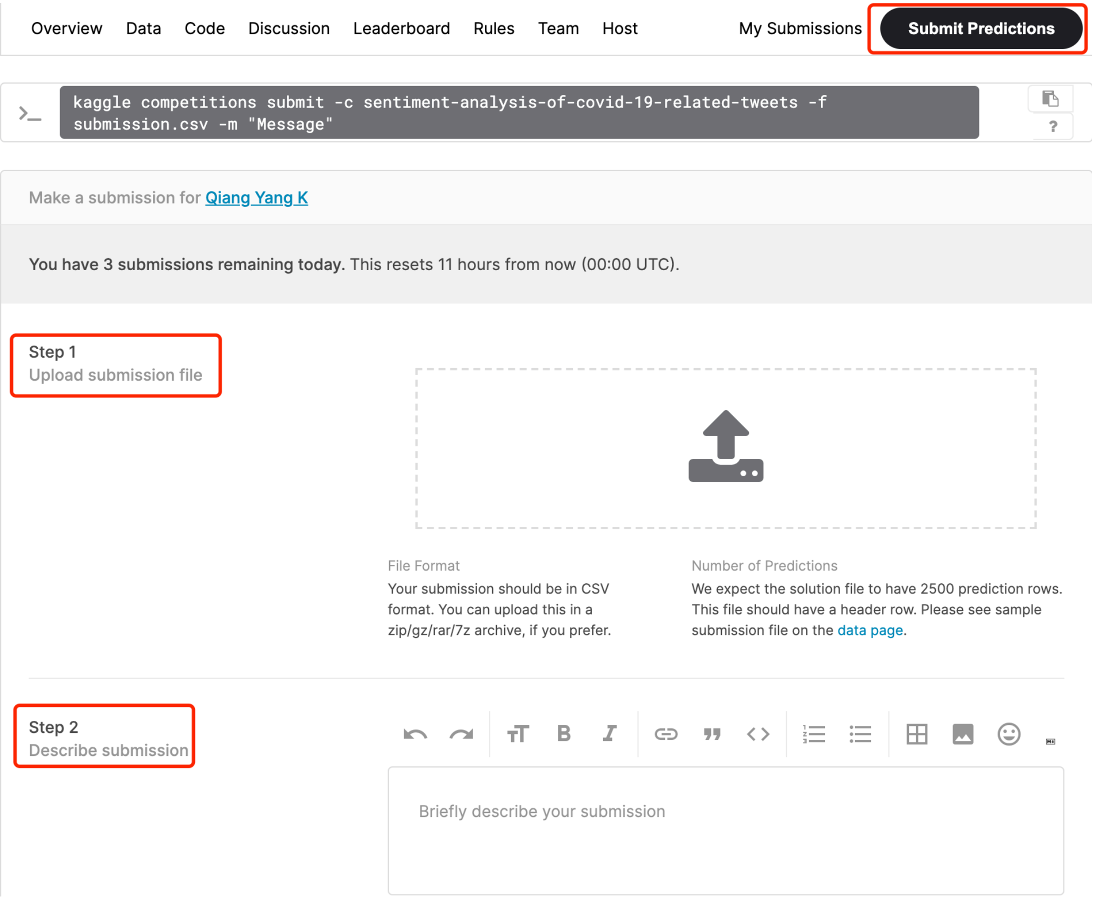
This file does not have a description yet.

ID	Tweet	Labels
		10 10% 6 10 6% Other (4200) 84%
1	NO JOKE I WILL HOP ON A PLANE RN! (Well after COVID19 lol)	0 10
2	BanMediaHouse whose is responsible for spreading Fake and communal Stories in this pandemic Corona s...	6
3	Im waiting for someone to say to me that all this corona thing is just an April fools joke	3 4
4	He is a liar. Proven day night. Time again. Lies when the truth will do. COVID19	6
5	NEW: U.S. CoronaVirus death toll reaches 4,000 after nearly 900 new deaths were reported today (BNO ...	8

Note: KEEP AN EYE ON THE OVERVIEW WEBPAGE

Challenge Submission

- Result submission:



```
kaggle competitions submit -c sentiment-analysis-of-covid-19-related-tweets -f
submission.csv -m "Message"
```

Make a submission for Qiang Yang K

You have 3 submissions remaining today. This resets 11 hours from now (00:00 UTC).

Step 1
Upload submission file

File Format
Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions
We expect the solution file to have 2500 prediction rows. This file should have a header row. Please see sample submission file on the data page.

Step 2
Describe submission

Briefly describe your submission

Notes:

1) Scoring metric: Macro F1-score

2) Maximum daily submissions: 3

3) File type: CSV

4) File content: 2500 rows with header, referring to the Evaluation page