Predict which Tweets are about real disasters and which ones are not

...

Yanwen Duan
Jiameng Sun

Are these tweets about real disasters?

Tweets about real disasters can save life

We would like create a web application to tell you which tweets are about about real disasters.

Input:

Output:

Input:

Output:

Existing Methods

- Data
- Model

Existing Method -- Data (2015)

```
train.head(100)
```

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1
...	...	...	...	...	...
95	137	accident	Charlotte	9 Mile backup on I-77 South...accident blockin...	1
96	138	accident	Baton Rouge, LA	Has an accident changed your life? We will hel...	0
97	139	accident	Hagerstown, MD	#BREAKING: there was a deadly motorcycle car a...	1
98	141	accident	Gloucestershire , UK	@flowri were you marinading it or was it an ac...	0
99	143	accident	NaN	only had a car for not even a week and got in ...	1

```
test.head(100)
```

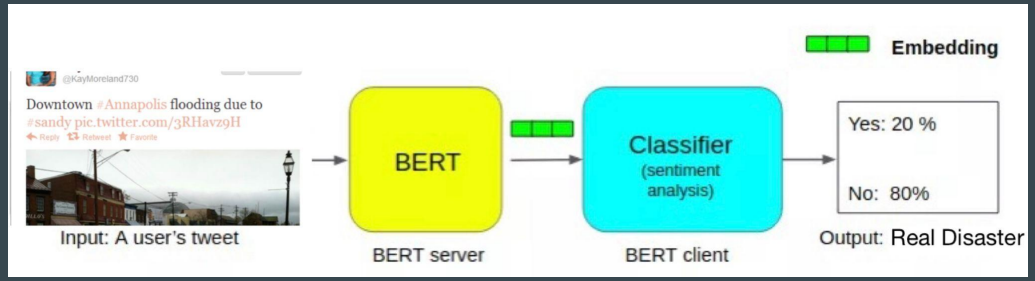
	id	keyword	location	text
0	0	NaN	NaN	Just happened a terrible car crash
1	2	NaN	NaN	Heard about #earthquake is different cities, s...
2	3	NaN	NaN	there is a forest fire at spot pond, geese are...
3	9	NaN	NaN	Apocalypse lighting. #Spokane #wildfires
4	11	NaN	NaN	Typhoon Soudelor kills 28 in China and Taiwan
...	...	...	...	...
95	323	annihilated	NaN	'If your nature appropriates it love will burn...
96	324	annihilated	NaN	@NinaHoag - 'if you shred my Psych work our fr...
97	325	annihilated	upstate NY	@thehill this is 1 example of y the Conservati...
98	326	annihilated	NaN	Aug 3 1915 ÛÓKAISERJAEGERS WIPED OUT.; Francis...
99	333	annihilated	NaN	They should all die! All of them! Everything a...

Data Resources:
https://appen.com/open-source-datasets/
https://www.kaggle.com/c/nlp-getting-started/

Existing Method -- Model

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google

- Deep bidirectional:
  - BERT learns information from both the left and the right side of a token's context during the training phase
- Pre-training:
  - Masked Language Model
  - Next Sentence Prediction
- Fine-tuning:
  - Classification tasks (used in our case)
  - Question Answering tasks
  - Named Entity Recognition



References:
https://www.kaggle.com/ratan123/in-depth-guide-to-google-s-bert
https://en.wikipedia.org/wiki/BERT_(language_model)

Limitation and proposed improvement

Limitation : Source data may not be correctly labeled

```
In [21]: train_dupli = train.groupby("text").agg({'target':lambda x: list(x)})

In [34]: train_dupli = train_dupli[train_dupli.target.map(len)>1]
         train_dupli.head(100)
```

Out[34]:

text	target
#Allah describes piling up #wealth thinking it would last #forever as the description of the people of #Hellfire in Surah Humaza. #Reflect	[0, 0, 1]
#Bestnaijamade: 16yr old PKK suicide bomber who detonated bomb in ... http://t.co/KSAwlYuX02 bestnaijamade bestnaijamade bestnaijamade be ÛÒ	[1, 1, 1, 1, 1, 1]
#KCA #VoteJKT48ID 12News: UPDATE: A family of 3 has been displaced after fired damaged housed near 90th and Osborn. Fire extinguished no i ÛÒ	[1, 1]
#Myanmar Displaced #Rohingya at #Sittwe point of no return http://t.co/cgf61fPmR0 #Prison like conditions #genocide IHHen MSF Refugees	[1, 1]
#Newswatch: 2 vehicles collided at Lock and Lansdowne Sts in #Ptbo. Emerg crews on their way	[1, 1]
...	...
incident with injury:I-495 inner loop Exit 31 - MD 97/Georgia Ave Silver Spring	[1, 1]
like for the music video I want some real action shit like burning buildings and police chases not some weak ben winston shit	[1, 0]
that exploded & brought about the\nbeginning of universe matches what's\nmentioned in the versethe heaven and Earth\n(thus the universe)	[0, 0]
that horrible sinking feeling when you ÛªÛªve been at home on your phone for a while and you realise its been on 3G this whole time	[1, 0, 0, 1]
wowo--=== 12000 Nigerian refugees repatriated from Cameroon	[1, 0]

- In train.csv, some duplicate tweets are labeled incorrectly, which may affect the final classification results.
- To alleviate potential effects, we are considering relabeling the existing dataset or utilizing a new dataset

Improvement 1: new data

The new dataset contains over 11,000 tweets associated with disaster keywords like “crash”, “quarantine”, and “bush fires” as well as the location and keyword itself.

The tweets were collected on Jan 14th, 2020.

Some of the topics people were tweeting:

The eruption of Taal Volcano in Philippines

Coronavirus

Bushfires in Australia

Iran downing of the airplane flight PS752

```
In [11]: new_data.head(200)
```

Out[11]:

	id	keyword	location	text	target
0	0	ablaze	NaN	Communal violence in Bhainsa, Telangana. "Ston...	1
1	1	ablaze	NaN	Telangana: Section 144 has been imposed in Bha...	1
2	2	ablaze	New York City	Arsonist sets cars ablaze at dealership https:...	1
3	3	ablaze	Morgantown, WV	Arsonist sets cars ablaze at dealership https:...	1
4	4	ablaze	NaN	"Lord Jesus, your love brings freedom and pard...	0
...	...	...	...	...	...
195	195	ambulance	Hyeongjun world	The Untamed boys singing “Love Siren” from My ...	0
196	196	ambulance	Theale, England	...or pulling over for an ambulance	0
197	197	ambulance	London innit	I hope there’s going to be an ambulance waitin...	0
198	198	ambulance	NaN	รักติดไซเรน (My Ambulance) .Cover is coming ! ...	0
199	199	ambulance	Sicily (❤️) Mielno 🇵🇱	When driving drunk you can choose your way to ...	0

Improvement 2: pretrain BERT

BERT can be improved and optimized.

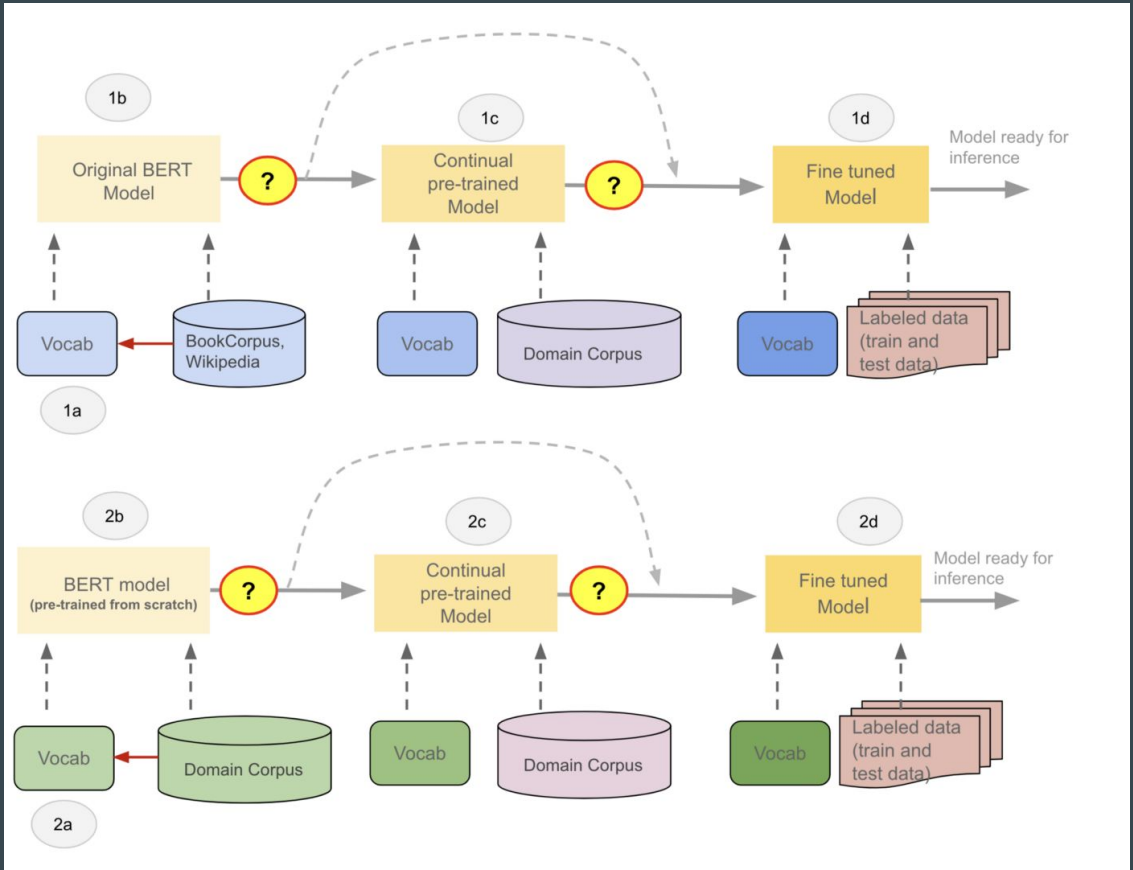
Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

One area we would like to focus is to pretrain BERT with domain-specific text data.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36, no. 4 (2020): 1234-1240.

Canete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. "Spanish pre-trained bert model and evaluation data." Pml4dc at iclr 2020 (2020): 2020.

Improvement 2: pretrain BERT



Improvement 3:

They leverage knowledge distillation during the pre-training phase and show that it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

Datasets, experiment, coding,

To pretrain BERT, we will find a large database of tweets and follow procedures in Jinhyuk Lee et al (2020).

To fine-tune BERT, there are good examples in previous solutions.

We will use both the 2015 data and 2020 data to fine-tune BERT. We will split both datasets into train set and test set.

Finally, we are wondering if we should focus on pretraining BERT or reducing size of BERT given constraints in resources?

Thank you!