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

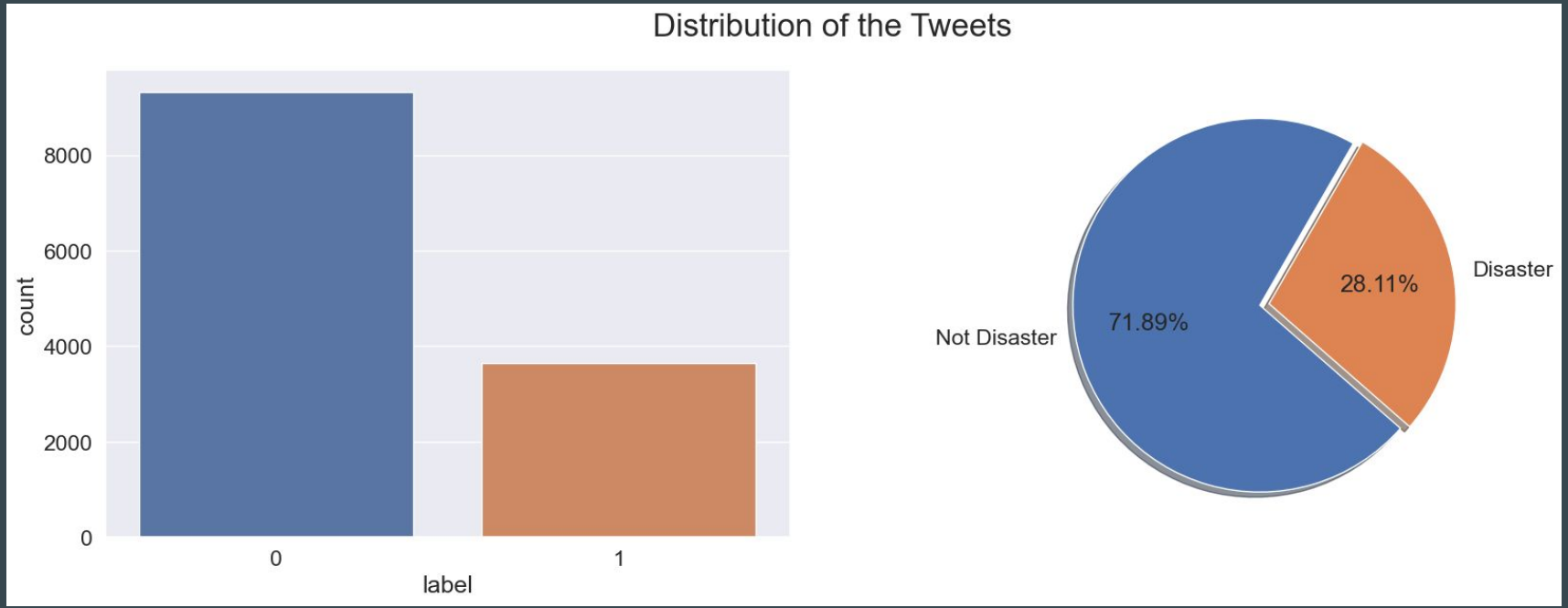
•••

Yanwen Duan
Jiameng Sun

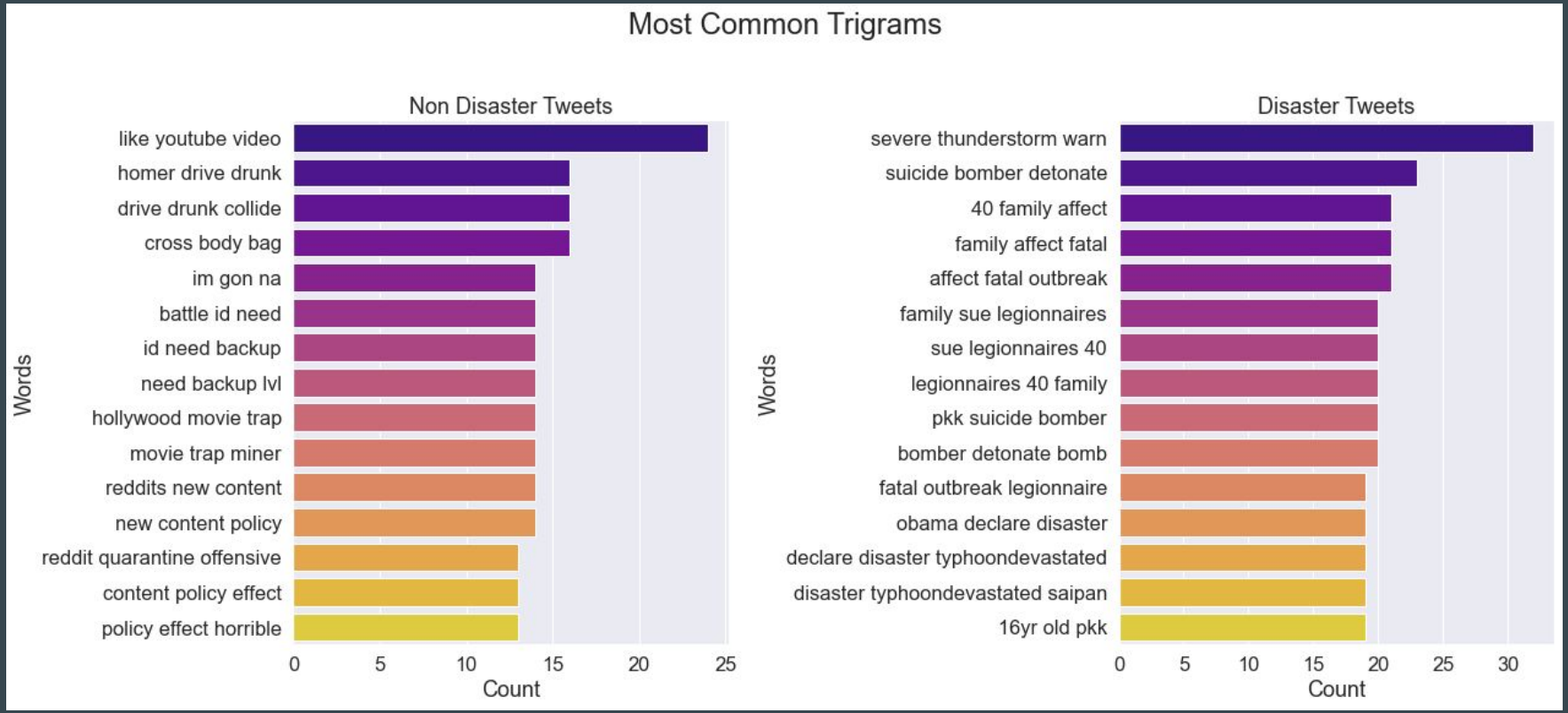
Data Cleaning

	id	keyword	location	text	target	text_clean	tokenized	lower	stopwords_removed	pos_tags	wordnet_pos	lemmatized	lemma_str
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1	Our Deeds are the Reason of this earthquake Ma...	[Our, Deeds, are, the, Reason, of, this, earth...	[our, deeds, are, the, reason, of, this, earth...	[deeds, reason, earthquake, may, allah, forgiv...	[(deeds, NNS), (reason, NN), (earthquake, NN),...	[(deeds, n), (reason, n), (earthquake, n), (ma...	[deed, reason, earthquake, may, allah, forgive...	deed reason earthquake may allah forgive u
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1	Forest fire near La Ronge Sask Canada	[Forest, fire, near, La, Ronge, Sask, Canada]	[forest, fire, near, la, ronge, sask, canada]	[forest, fire, near, la, ronge, sask, canada]	[(forest, JJS), (fire, NN), (near, IN), (la, J...	[(forest, a), (fire, n), (near, n), (la, a), (...	[forest, fire, near, la, ronge, sask, canada]	forest fire near la ronge sask canada
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1	All residents asked to shelter in place are be...	[All, residents, asked, to, shelter, in, place...	[all, residents, asked, to, shelter, in, place...	[residents, asked, shelter, place, notified, o...	[(residents, NNS), (asked, VBD), (shelter, JJ)...	[(residents, n), (asked, v), (shelter, a), (pl...	[resident, ask, shelter, place, notify, office...	resident ask shelter place notify officer evac...
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1	13000 people receive wildfires evacuation orde...	[13000, people, receive, wildfires, evacuation...	[13000, people, receive, wildfires, evacuation...	[13000, people, receive, wildfires, evacuation...	[(13000, CD), (people, NNS), (receive, JJ), (w...	[(13000, n), (people, n), (receive, a), (wildf...	[13000, people, receive, wildfire, evacuation,...	13000 people receive wildfire evacuation order...
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1	Just got sent this photo from Ruby Alaska as s...	[Just, got, sent, this, photo, from, Ruby, Ala...	[just, got, sent, this, photo, from, ruby, ala...	[got, sent, photo, ruby, alaska, smoke, wildfi...	[(got, VBD), (sent, JJ), (photo, NN), (ruby, N...	[(got, v), (sent, a), (photo, n), (ruby, n), (...	[get, sent, photo, ruby, alaska, smoke, wildfi...	get sent photo ruby alaska smoke wildfires pou...

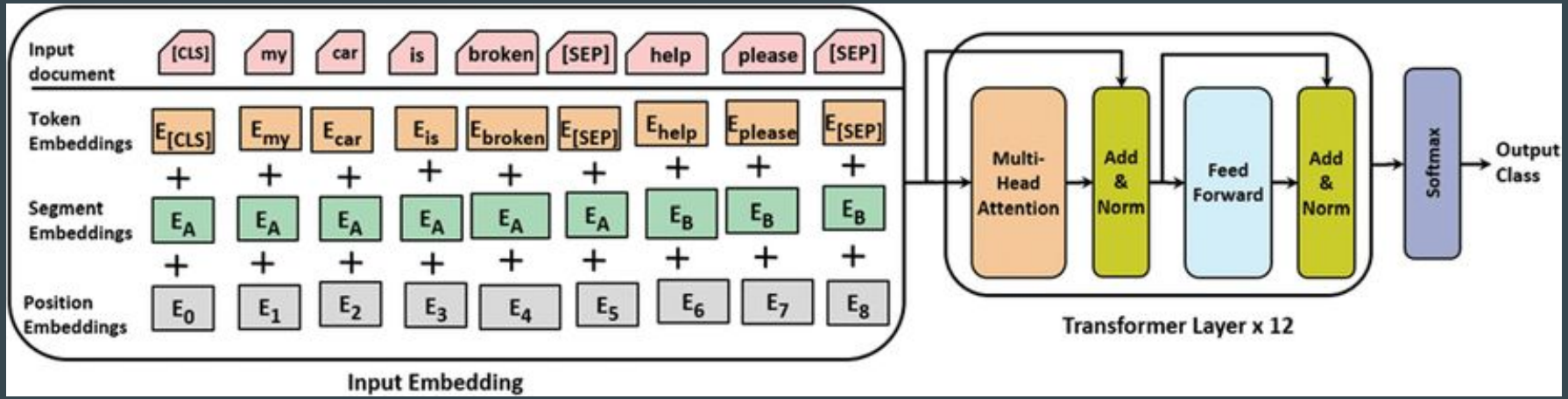
Data Visualization I



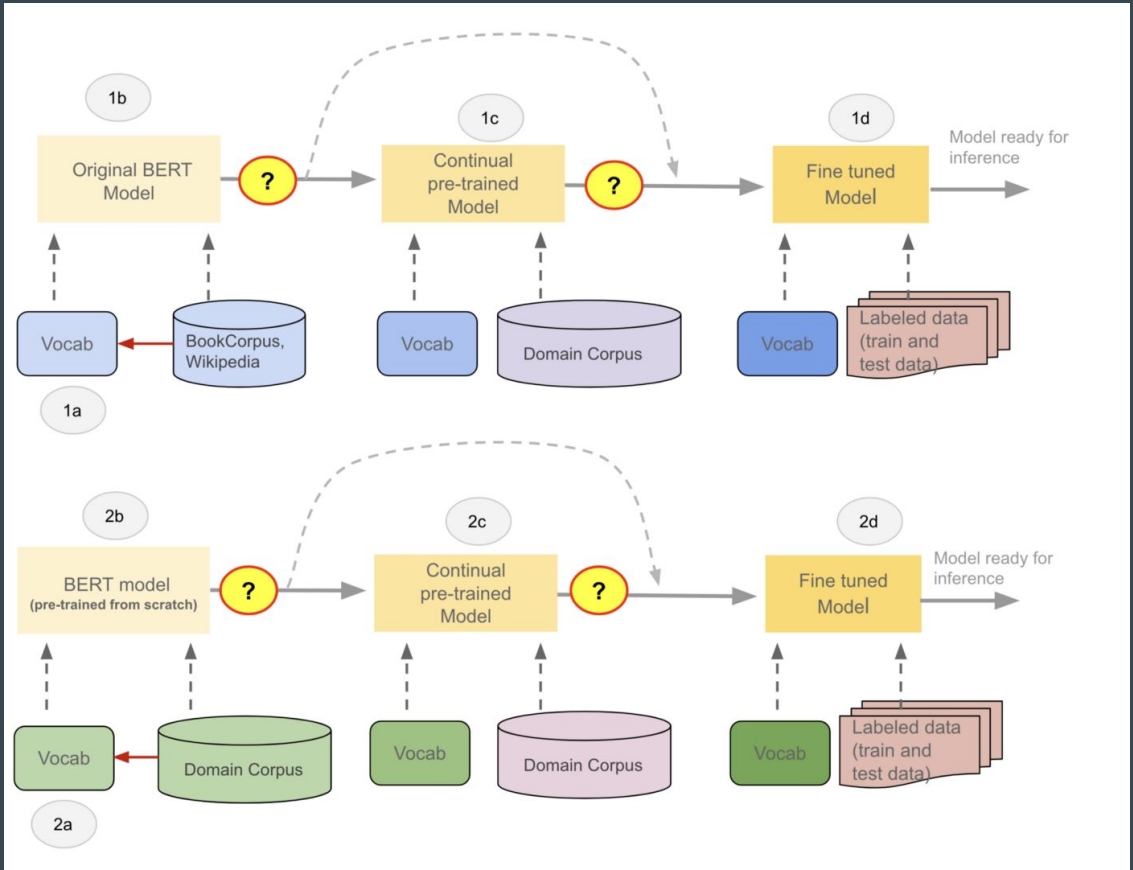
Data Visualization II



Model I - BERT


Input document
[CLS]
my
car
is
broken
[SEP]
help
please
[SEP]
Token Embeddings
Segment Embeddings
Position Embeddings
Input Embedding
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Transformer Layer x 12
Softmax
Output Class

Model II - BERTweet



Fine tune pre-trained models from Hugging Face

- Model is trained based on text classification tasks
- Train on a 16GB GPU from Google Colab
- Since each text in our dataset is relatively short between 10 to 20 tokens, we used smart batch padding on the fly instead of pre-padding to max length 512. This helps speed up the training.

```
kwargs = {"finetuned_from": model_args.model_name_or_path, "tasks": "text-classification"}
if data_args.task_name is not None:
    kwargs["language"] = "en"
    kwargs["dataset_tags"] = "glue"
    kwargs["dataset_args"] = data_args.task_name
    kwargs["dataset"] = f"GLUE {data_args.task_name.upper()}"

if training_args.push_to_hub:
    trainer.push_to_hub(**kwargs)
else:
    trainer.create_model_card(**kwargs)
```

BERT vs BERTweet? Old Data vs New Data?

BERT-Base trained with the train set of the first dataset (7613 samples)

Epoch/Metric	Training Loss	Test Loss	Test Accuracy	Test F1
1	0.50	0.45	0.81	0.77
2	0.38	0.42	0.83	0.78
3	0.33	0.43	0.83	0.78

The following models were trained with the new train set (13289 samples)
BERT-Base (Total time: 744.282s)

Epoch/Metric	Training Loss	Test Loss	Test Accuracy	Test F1
1	0.371	0.350	0.857	0.713
2	0.276	0.343	0.863	0.744
3	0.196	0.383	0.856	0.742

BERTweet-Base (Total time: 770.702s)

Epoch/Metric	Training Loss	Test Loss	Test Accuracy	Test F1
1	0.386	0.358	0.861	0.710
2	0.331	0.341	0.864	0.740
3	0.191	0.363	0.869	0.745

- We have fine tuned both BERT and BERTweet with old or new train set
- The models with new data have higher Accuracy.

  One of the potential reason: Data makes more impacts on performance than Model itself does.

- BERTweet can achieve similar accuracy as traditional BERT with the same train set

  One of the possible reason we can think of is: Both models have large amount of parameters (100M +), while our dataset size is relatively small (10K +).

Web Application Based on BERTweet

- We have built a web application with Streamlit
- By typing in context of a tweet and hitting the Detect button, the app can predict whether the tweet is about real disaster or not with corresponding probability.

Summary

- Data vs Model ?
  - Our projects show that more data brings better results. The benefit of data even outweighs the improved model in our case.
  - We believe that with more data, the BERT and BERTweet models can keep improving. After all, the number of our samples(13289) is 1/10000 of the number of BERT's parameters (110M).