Group 3: Disaster Tweets Classification

Anupam Kumar, Changchang Liu, Xuefang Hu

Content

- Description
- Dataset
- Methodology
- Data Analysis and Preprocessing
- Building the Classifier Model
- BERT and its variants
- Experiments and Results
- Conclusion
- Further topics

Description

- Tweeter: one of the most important communication channel in times of emergency

  But, tweets are not always clear

- **Our goal: Build a model to accurately predicts whether a tweet is about real disaster**

Dataset

https://www.kaggle.com/c/nlp-getting-started/data

Data format

- The text of a tweet
- A keyword from that tweet (although this may be blank!)
- The location the tweet was sent from (may also be blank)
- ~11k rows in both train and test dataset

Columns

- id - a unique identifier for each tweet
- text - the text of the tweet
- location - the location the tweet was sent from (may be blank)
- keyword - a particular keyword from the tweet (may be blank)
- target - in train.csv only, this denotes whether a tweet is about a real disaster (1) or not (0)

Dataset: train

train

id	keyword	location	text	target
48	ablaze	Birmingham	@bbcmtd Wholesale Markets ablaze http://t.co/lHYXEOHY6C	1
49	ablaze	Est. September 2012 - Bristol	We always try to bring the heavy. #metal #RT http://t.co/YAo1e0xngw	0
50	ablaze	AFRICA	#AFRICANBAZE: Breaking news:Nigeria flag set ablaze in Aba. http://t.co/2nndBGwyEi	1
52	ablaze	Philadelphia, PA	Crying out for more! Set me ablaze	0
53	ablaze	London, UK	On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE http://t.co/qqsmshaJ3N	0
54	ablaze	Pretoria	@PhDSquares #mufc they've built so much hype around new acquisitions but I doubt they will set the EPL ablaze this season.	0
55	ablaze	World Wide!!	INEC Office in Abia Set Ablaze - http://t.co/3lmaomknnA	1
56	ablaze		Barbados #Bridgetown JAMAICA ÛÒ Two cars set ablaze: SANTA CRUZ ÛÓ Head of the St Elizabeth Police Superintende... http://t.co/wDUEaj8Q4J	1
57	ablaze	Paranaque City	Ablaze for you Lord :D	0
59	ablaze	Live On Webcam	Check these out: http://t.co/rOI2NSmEJJ http://t.co/3Tj8ZjiN21 http://t.co/YDUiXEfIpE http://t.co/LxTjc87KLS #nsfw	0
61	ablaze		on the outside you're ablaze and alive but you're dead inside	0
62	ablaze	milky way	Had an awesome time visiting the CFC head office the ancop site and ablaze. Thanks to Tita Vida for taking care of us ??	0
63	ablaze		SOOOO PUMPED FOR ABLAZE ???? @southridgelife	0
64	ablaze		I wanted to set Chicago ablaze with my preaching... But not my hotel! http://t.co/o9qknbfOFX	0
65	ablaze		I gained 3 followers in the last week. You? Know your stats and grow with http://t.co/TlyUliF5c6	0
66	ablaze	GREENSBORO,NORTH CAROLINA	How the West was burned: Thousands of wildfires ablaze in California alone http://t.co/vl5TBR3wbr	1
67	ablaze		Building the perfect tracklist to life leave the streets ablaze	0
68	ablaze	Live On Webcam	Check these out: http://t.co/rOI2NSmEJJ http://t.co/3Tj8ZjiN21 http://t.co/YDUiXEfIpE http://t.co/LxTjc87KLS #nsfw	0
71	ablaze	England.	First night with retainers in. It's quite weird. Better get used to it; I have to wear them every single night for the next year at least.	0
73	ablaze	Sheffield Township, Ohio	Deputies: Man shot before Brighton home set ablaze http://t.co/gWNRhMSO8k	1
74	ablaze	India	Man wife get six years jail for setting ablaze niece http://t.co/eV1ahOUCZA	1
76	ablaze	Barbados	SANTA CRUZ ÛÓ Head of the St Elizabeth Police Superintendent Lanford Salmon has r ... - http://t.co/vpIR5Hka2u http://t.co/SxHW2TNNLf	0
77	ablaze	Anaheim	Police: Arsonist Deliberately Set Black Church In North CarolinaåÊAblaze http://t.co/pcXarbH9An	1
78	ablaze	Abuja	Noches El-Bestia '@Alexis_Sanchez: happy to see my teammates and training hard ?? goodnight gunners.?????? http://t.co/uc4j4jHvGR'	0

Dataset: test

test

id	keyword	location	text
43			What if?!
45			Awesome!
46	ablaze	London	Birmingham Wholesale Market is ablaze BBC News - Fire breaks out at Birmingham's Wholesale Market http://t.co/irWqCEZWEU
47	ablaze	Niall's place | SAF 12 SQUAD |	@sunkxssedharry will you wear shorts for race ablaze ?
51	ablaze	NIGERIA	#PreviouslyOnDoyinTv: Toke MakinwaÛªs marriage crisis sets Nigerian Twitter ablaze... http://t.co/CMghxBa2XI
58	ablaze	Live On Webcam	Check these out: http://t.co/rOI2NSmEJJ http://t.co/3Tj8ZjiN21 http://t.co/YDUiXEfIpE http://t.co/LxTjc87KLS #nsfw
60	ablaze	Los Angeles, Califnordia	PSA: IÛªm splitting my personalities. ?? techies follow @ablaze_co ?? Burners follow @ablaze
69	ablaze	threeonefive.	beware world ablaze sierra leone & guap.
70	ablaze	Washington State	Burning Man Ablaze! by Turban Diva http://t.co/hodWosAmWS via @Etsy
72	ablaze	Whoop Ass, Georgia	Not a diss song. People will take 1 thing and run with it. Smh it's an eye opener though. He is about 2 set the game ablaze @CyhiThePrynce
75	ablaze	India	Rape victim dies as she sets herself ablaze: A 16-year-old girl died of burn injuries as she set herself ablazeÛ_ http://t.co/UK8hNrbOob
84	ablaze		SETTING MYSELF ABLAZE http://t.co/6vMe7P5XhC
87	ablaze	scarborough, ontario	@CTVToronto the bins in front of the field by my house wer set ablaze the other day flames went rite up the hydro pole wonder if it was him
88	ablaze		#nowplaying Alfons - Ablaze 2015 on Puls Radio #pulsradio http://t.co/aA5BJgWfDv
90	ablaze	121 N La Salle St, Suite 500	'Burning Rahm': Let's hope City Hall builds a giant wooden mayoral effigy 100 feet tall & sets it ablaze. http://t.co/kFo2mksn6Y @John_Kass
94	ablaze	Wandering	@PhilippaEilhart @DhuBlath hurt but her eyes ablaze with insulted anger.
99	accident	Homewood, PA	Accident cleared in #PaTurnpike on PATP EB between PA-18 and Cranberry slow back to #traffic http://t.co/SL0Oqn0Vyr
101	accident		Just got to love burning your self on a damn curling wand... I swear someone needs to take it away from me cuase I'm just accident prone.
103	accident		I hate badging shit in accident
106	accident	USA	#3: Car Recorder ZeroEdgeâ¨ Dual-lens Car Camera Vehicle Traffic/Driving History/Accident Camcorder Large Re... http://t.co/kKFaSJv6Cj
108	accident	Massachusetts	Coincidence Or #Curse? Still #Unresolved Secrets From Past http://t.co/7VG8Df9pLE #accident
111	accident	Bexhill	@Traffic_SouthE @roadpol_east Accident on A27 near Lewes is it Kingston Roundabout rather than A283
115	accident	Anime World	@sakuma_en If you pretend to feel a certain way the feeling can become genuine all by accident. -Hei (Darker than Black) #manga #anime
116	accident		For Legal and Medical Referral Service @1800_Injured Call us at: 1-800-465-87332 #accident #slipandfall #dogbite
122	accident	Cowtown, Caliii !!	There's a construction guy working on the Disney store and he has huge gauges in his ears ?? ...that is a bloody accident waiting to happen

Methodology

- **Libraries: Scikit-learn** and **Tensorflow**
- **matplotlib** for some visualization
- **BERT** model and its variants **ALBERT, RoBERTa**

Data Analysis and Preprocessing

Raw Data Analysis:

Data Remove duplicate data

Examine the target data balance

Check for null values

Preprocessing:

Build a text standardization function that is specific to the content found in tweets: build a lookup dictionary with common twitter phrase abbreviations, tweet terms that match keys in the lookup dictionary will be expanded to their non-abbreviated form.

Before taking the next step we should explore it with some visualizations: A comparison between disaster and non-disaster word clouds shows how often different words appear in the tweets. We can see that the word new appears quite frequently in both classes, which may add noise to the training process → Remove the Word *new* from Text

Figure 1: Non-disaster word cloud

Figure 2: Disaster word cloud

Fill Missing Keywords

There are fifty-six missing keywords in the training data set and twenty-six in the test set. We will use spaCy's part of speech tagging feature to fill the missing words

Create Data Splits

A training split is produced leaving 20% of the data for validation purposes.

form these data splits into TensorFlow Datasets and configure them before building our model

Building the Classifier Model

Traditional framework for baseline and use BERT variants to check the accuracy gain over the baseline

- LogisticRegression using TfIdf vectorization and Word2Vec transformation using gensim library.

- Methods: BERT, ELECTRA, RoBERTa, and ALBERT.

BERT and its variants

Why BERT?

BERT is fully bidirectional reading both directions at the same time, with interactions from both directions. It gives incredible accuracy and performance on small datasets.

How it works?

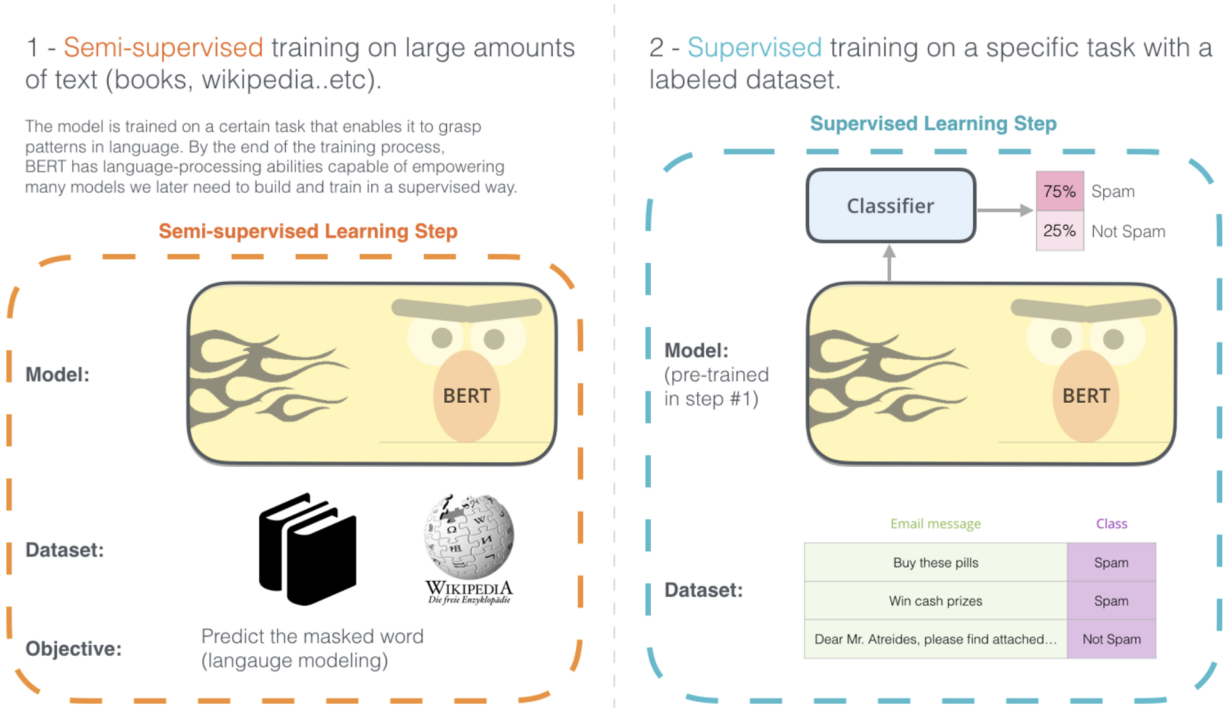
It is a language representation model, and only provides the encoder part. It takes input text data and creates fixed shape segments or sentence embeddings.

BERT and its variants

BERT

We used BERT model architecture with 12 hidden layers, 768 hidden units, and 12 attention heads. This architecture has 110 million parameters.

BERT: Bidirectional Encoder Representations from Transformers



The two steps of how BERT is developed. You can download the model pre-trained in step 1 (trained on un-annotated data), and only worry about fine-tuning it for step 2. [Source for book icon].

BERT and its variants

ELECTRA

It provides a more efficient alternative using replaced token detection as pre-training task instead of replacing tokens with mask. ELECTRA small takes less time to train due to less number of parameters to train but at a cost of overall performance.

BERT and its variants

ALBERT

‘A Lite BERT’. Increasing model size and increased parameters often results in better performance on downstream tasks in natural language representations, but can often be memory and compute limitations, especially on mobile devices.

ALBERT is designed to address those issues. It incorporate two parameter reduction techniques: factorized embedding parameterization and cross-layer parameter sharing.

BERT and its variants

RoBERTa

‘Robustly Optimized BERT Pre-training Approach

After building a model, a lot of effort is invested in hyperparameter tuning to get the optimal performance from the network. This process can be computationally very expensive. The authors found that BERT was quite undertrained. They proposed modifications to the BERT pre-training procedure improving the overall robustness of the model and performance.

Experiments and Results

We used TensorFlow hub for implementation of BERT, Electra-small, and ALBERT-base while trans- formers module for the implementation of Roberta-base. The text preprocessing before implementation of Roberta-base was less heavy than for implementations of other BERT variants.

- around 7k tweets to classify into disastrous and non-disastrous tweets
- 80% data for training and 20% for validation
- trained each BERT variant model for 3 epochs
- took 5e-5 as learning rate for BERT, ALBERT0base, 1e-5 for ELECTRA-small and 2e-5 for RoBERTa-base model
- KFold with 6 splits for RoBERTa-base model but not for other architectures
- used GPU on kaggle to run all of the models and compare the training time

BERT variants	training-time (sec.)	val-accuracy	F1-score	# parameters
BERT	1156.578	0.812	0.768	110M
Electra-small	422.668	0.766	0.725	14M
ALBERT-base	1125.614	0.812	0.761	31M
RoBERTa-base	426.732	0.837	**0.812**	123M
Distilbert		0.841	0.806	66M

Conclusion

With more parameters and dynamic masking, RoBERTa-base out- performed the other models in training time, val-accuracy as well as F1-score. Though ELECTRA-small has the lowest val-accuracy and F1-score, it has nearly

ALBERT-base and BERT-base perform similarly in all of those three metrics. ALBERT will perform better when the structure becomes larger. So ALBERT-base does not have significantly good performance since it has much fewer parameters and do not have large enough structure.

Further Topics

- Due to compute constraints as we were using Google Colab and Kaggle GPU, we did not perform much hyperparameter tuning on BERT variants model
- For future work and given more time, we would like to use KFold split on dataset for all architecture and do a more elaborate EDA of the text data and implement several baseline models like Random Forest, SVM and LightGBM using gensim and other Word2Vec transformation.

Thank you!