Tweet Classifier for Disasters

Creating a filter to sieve through twitter for disaster-related tweets.

The process, simplified.

Exploring the data

Cleaning the tweets to extract the text information and preparing it for the models. Duplicate tweets are removed

Real-world classification

Using the best model to classify the tweets on the test data and another separate dataset.

Finding the best model

Chosen by evaluation of the model's accuracy score and the specificity and sensitivity

Evaluating the effectiveness

Identify the strengths and weaknesses and how it fits in the process

1.

Data source: Twitter

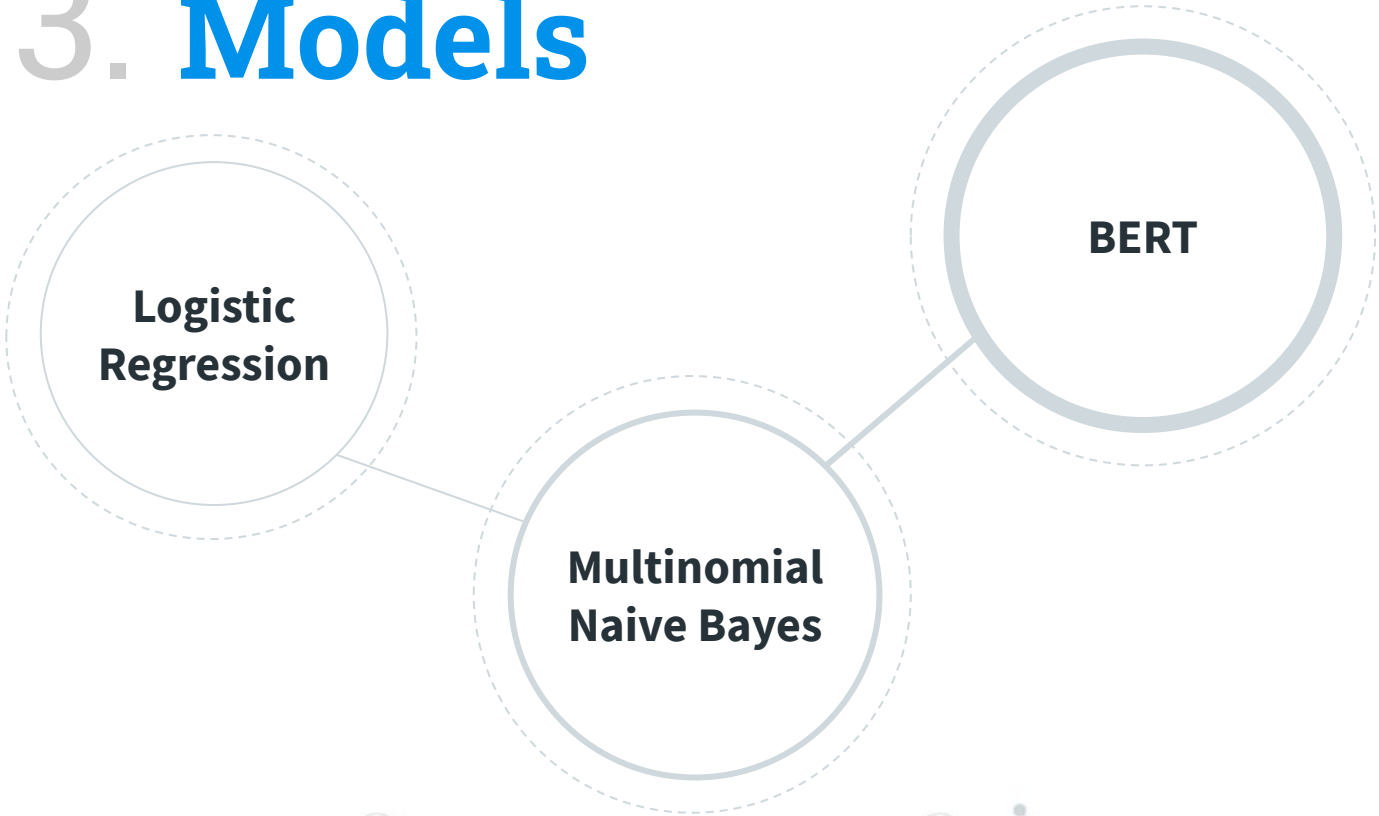
10,000 labelled tweets from FigureEight

2.

Natural Language Processing

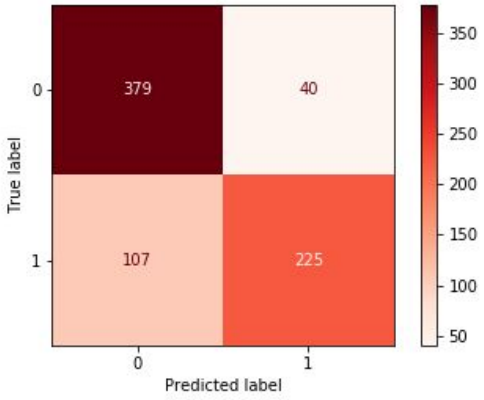
to allow computers to process and analyze large amounts of natural (human) language data.

3. Models



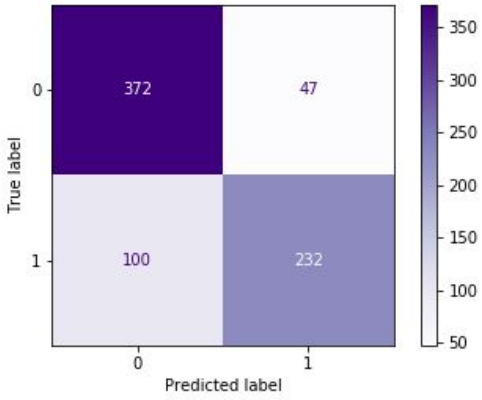
Logistic Regression

1. Removing of usernames and links.
2. Converting text to vectors. (*TF-IDF Vectorizer*)
3. Finding the best parameters for the model. (*GRIDSEARCH*)



Multinomial Naive Bayes

1. Removing of usernames and links.
2. Converting text to vectors. (*Count Vectorizer*)
3. Finding the best parameters for the model. (*GRIDSEARCH*)



Bidirectional Encoder Representations from Transformer

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
Oct 2018

- Process words in relation to all the other words in a sentence, rather than one-by-one in order.
- Considers the full context of a word by looking at the words that come before and after it.
- State-of-the-art scores from benchmarks such as SQuAD, GLUE, SWAG and NER
- Comes in two sizes (Base, Large)

Task 1: Masked Language Modelling

Predicting on randomly masked words

Task 2: Next Sentence Prediction

Predicting if a second sentence is the subsequent of the first sentence.

WordPiece Tokenizer

- Used to segment words into subwords
- Includes individual characters and most frequent/likely combinations

Original: New crime: knowing your rights Punishable by death

Tokenized: 'new', 'crime', ':', 'knowing', 'your', 'rights', 'punish', '##able', 'by', 'death'

- Helps with dealing with out-of-vocabulary words
- Deals with words and their different adjectives/tenses

Original: ducks duckling ducklings ducking

Tokenized: 'ducks', 'duck', '##ling', 'duck', '##lings', 'duck', '##ing'

Feeding Bert

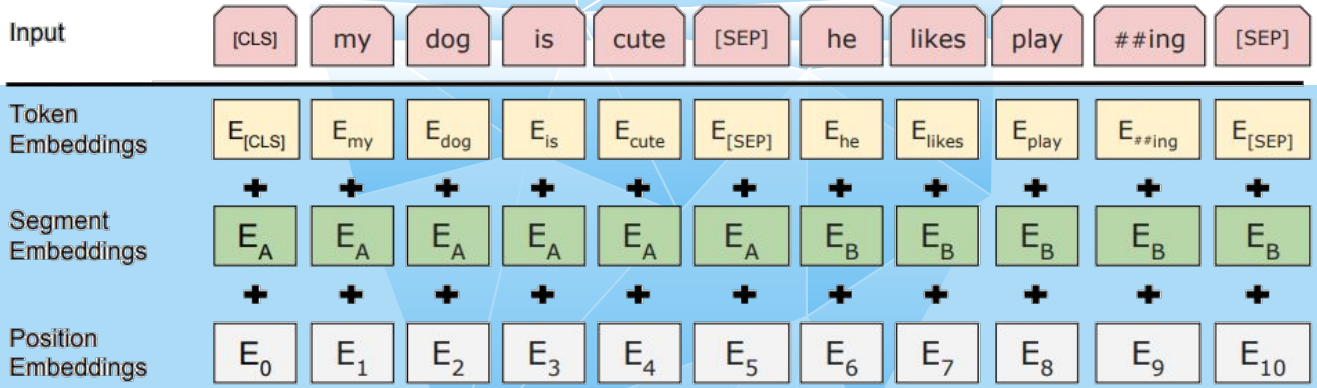
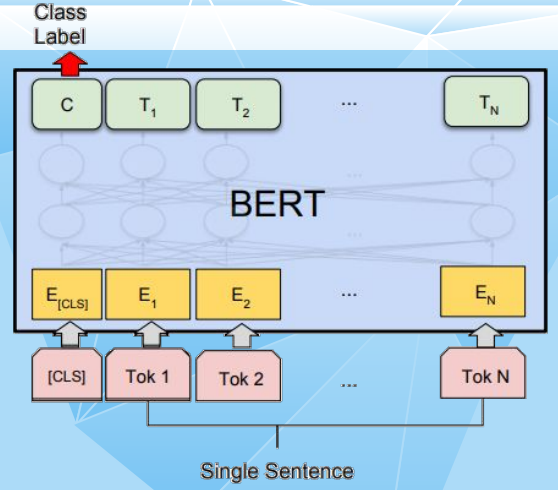


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Inner workings: Bert base classifier

Pre-trained by Google.

- Book Corpus (800,000,000 words)
- Wikipedia articles (2,500,000,000 words)
- WordPiece Tokenizer (30,522 tokens)



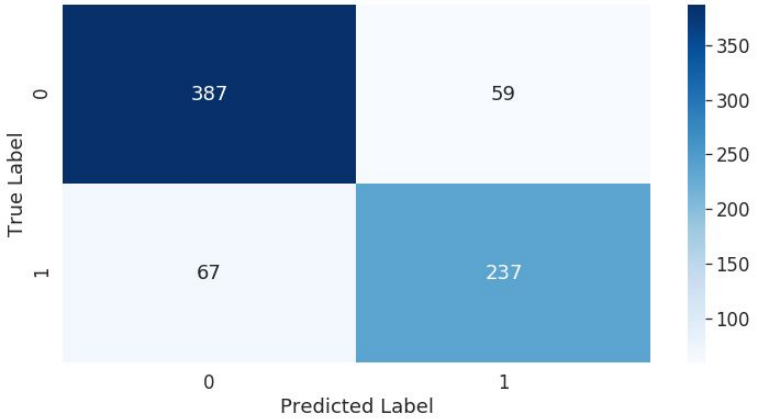
(b) Single Sentence Classification Tasks: SST-2, CoLA

Features

- 12 Transformer encoders
- 12 Multihead attention layers
- Hidden size of 768
- Maximum of 512 tokens per sentence
- Inclusive of one classifier token.

BERT

1. Removing of usernames and links.
2. Tokenizing using BERT's tokenizer.
3. Prepared the data into BERT's required format.
4. Fine-tune the model's output to match the number of labels (2)



Findings

Model Scores

	Accuracy	Specificity	Sensitivity
Logistic Regression	**80.43%**	**90.45%**	**67.77%**
Multinomial Naive Bayes	**80.43%**	**88.78%**	**69.88%**
BERT	**83.60%**	**87.00%**	**78.62%**

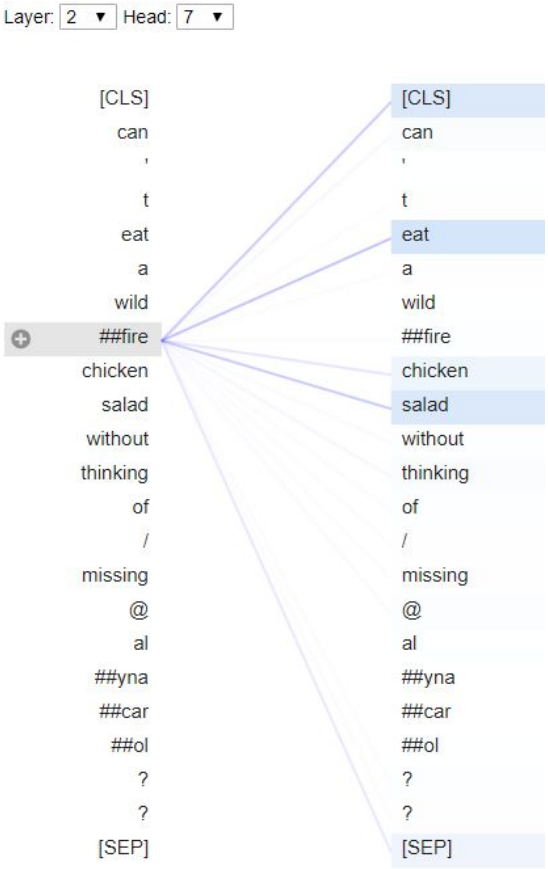
Accuracy scores for two Disaster datasets

	Figure Eight	CrisisLex 6
Logistic Regression	**80.43%**	**74.31%**
Multinomial Naive Bayes	**80.43%**	**72.99%**
BERT	**83.60%**	**85.45%**

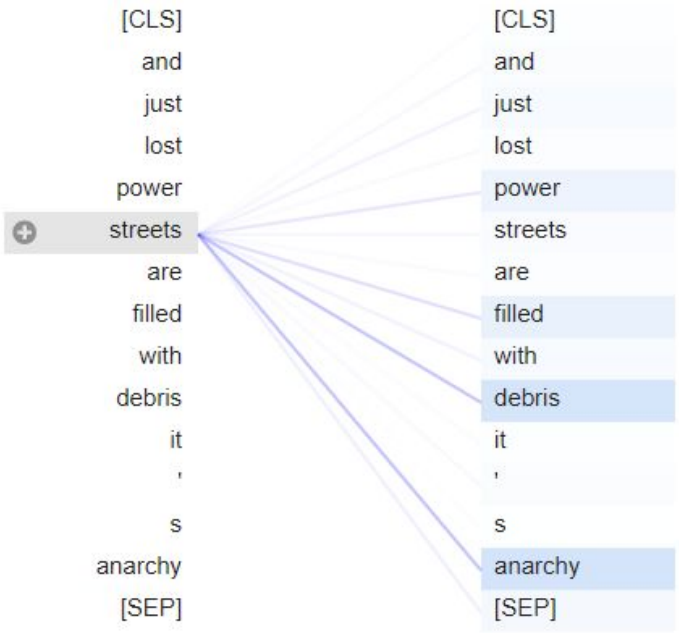
BERT Triumphs

Taken to identify that the word wildfire is relating to a salad and not a disaster. Other models picked it up as a disaster.

“can't eat a wildfire chicken salad without thinking of / missing @Alynacarol ??”



BERT Triumphs



Able to correctly identify tweets with incorrect labels based on the description mentioned in the tweet.

"and just lost power Streets are filled with debris It's anarchy"

BERT Challenges:

Struggles with fiction/lyrics

“Is this the real life? Is this just fantasy? Caught in a landslide. No escape from reality.”

Unable to understand sarcasm/metaphors

“My brain feels like it's vibrating inside my skull. The MRI is going to look like a seismic readout of a faultline during earthquake season.”

Some requires external context

“There you go folks here's you update of Sandy”

Advantages of BERT

- Performs well on unseen data
- Can be fine-tuned for specific tasks
- Able to take context into account.
- Able to handle new words.
- Fast to set up for such a complex model

Disadvantages of BERT

- Complex model to understand
- Large model size (400mb for standard BERT)
- Hard to trouble shoot
- Computationally expensive to train

BERT has a more widespread application and is more capable of handling unseen data.

“All models are wrong, but some are useful”

- **George E. P. Box**

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

- https://arxiv.org/abs/1810.04805
- https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
- https://mccormickml.com/
- https://huggingface.co/transformers/v2.2.0/model_doc/bert.html#bertforsequenceclassification
- http://jalammar.github.io/illustrated-bert/
- https://github.com/jessevig/bertviz

Sources