DistilBERT, the alternative to massive models for natural language processing

Master Thesis in Big Data Analytics

Ion Bueno Ulacia

uc3m | Universidad Carlos III de Madrid

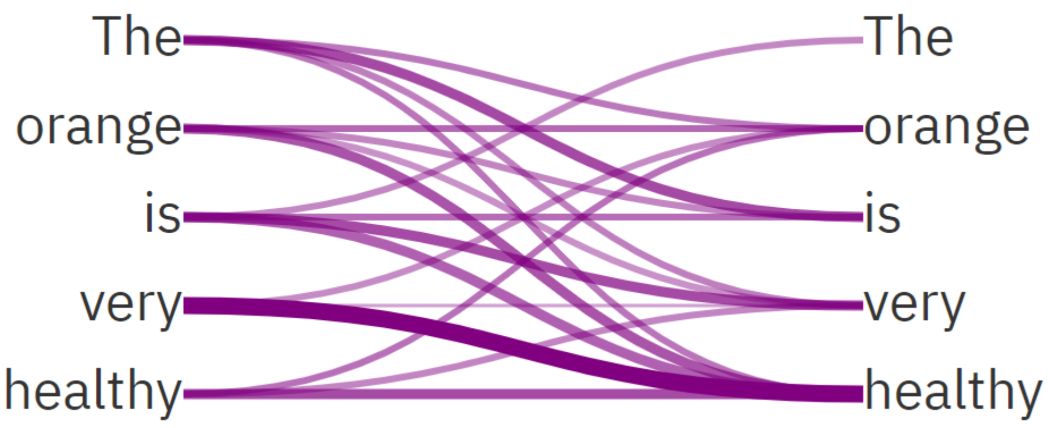
Contents

Natural Language Processing (NLP)

- Area of Artificial Intelligence
- Processing or generation of text
- Multiple applications
- **Used to analyze covid-19 mental health impact**

Self-Attention

- Critical in **BERT**
- Representation of a sentence
- Maps set to set
- Word relations: grammar, semantic…

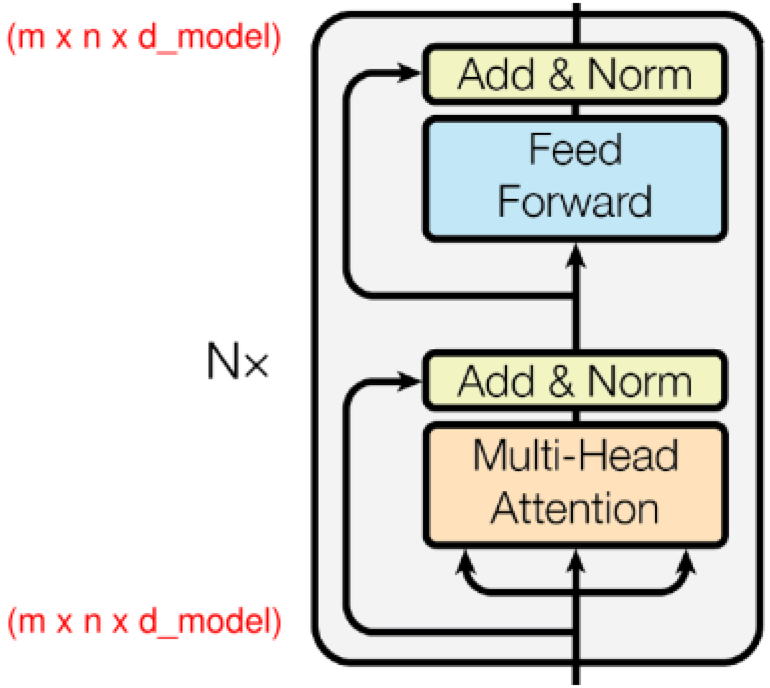


BERT

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, released in 2018 by Google
- Composed by a stack of Transformer encoders
- Uses self-attention to get left and right context

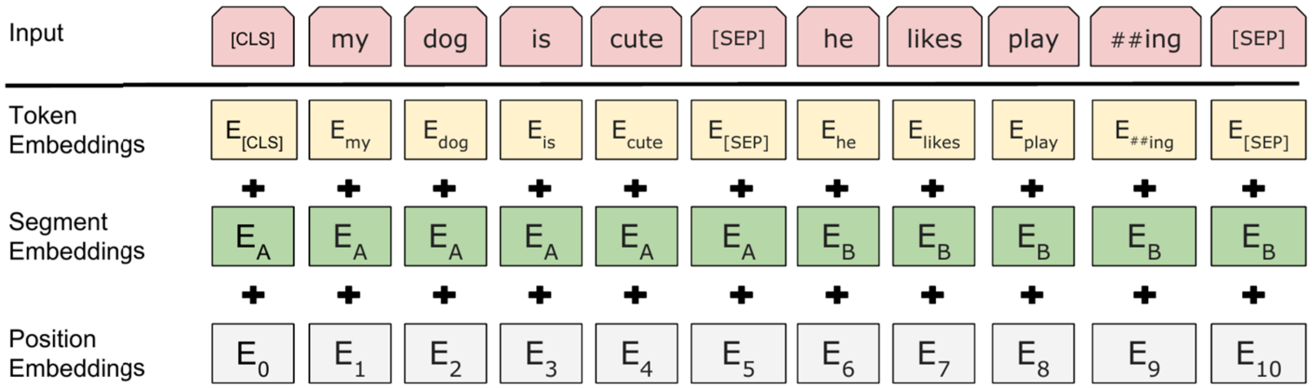
BERT Architecture

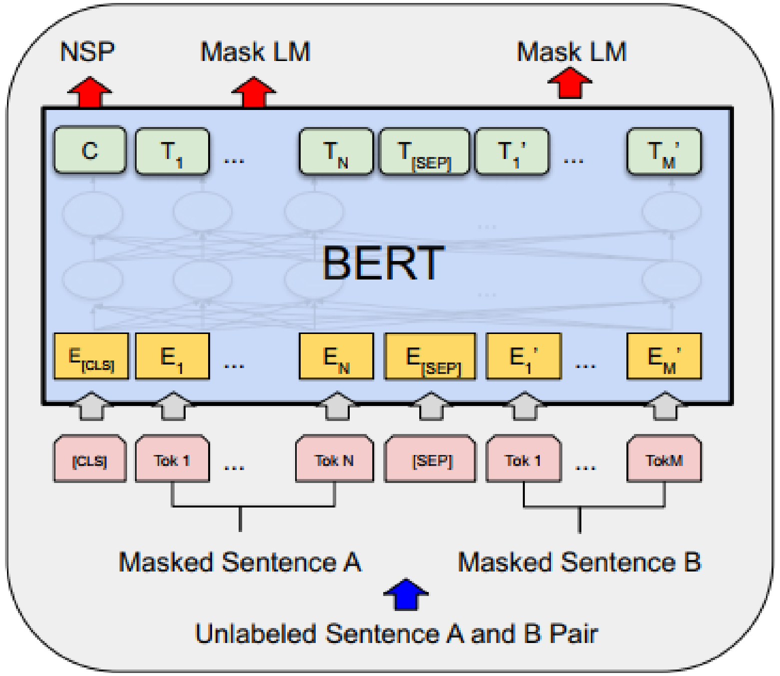
- Multi-Head attention block
- Input and output with same dimensionality
  - *m*: number of sentences
  - *n*: number of words
  - *d_model*: embedding dimension



Input-Output BERT

- Input: addition of 3 embeddings
- Output: classification token and attention vectors



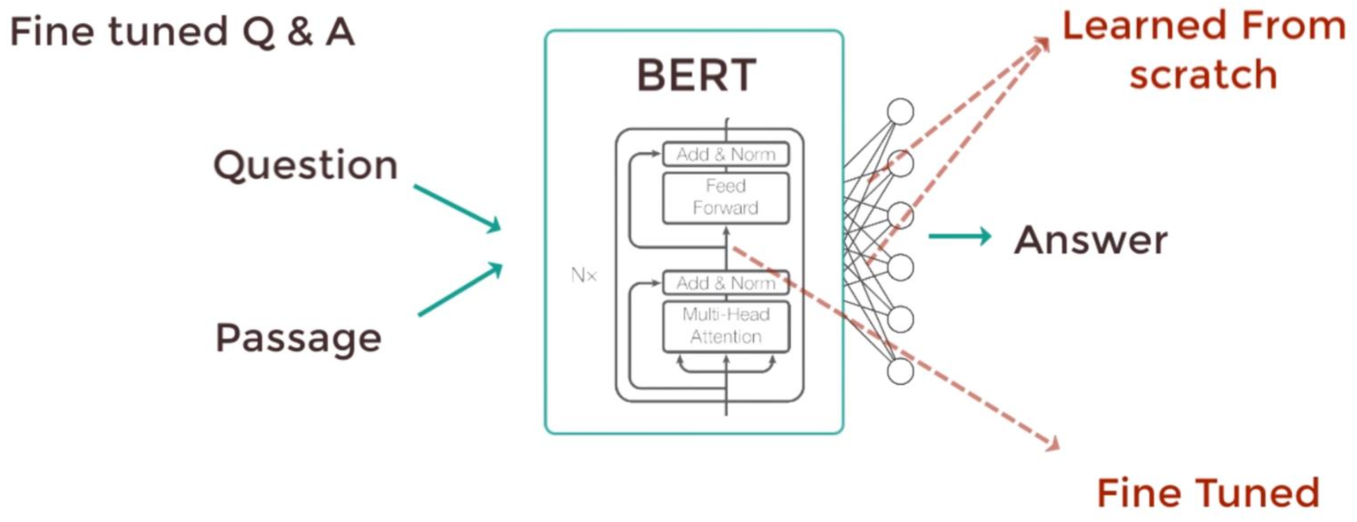


Pre-Training BERT

- 16 TPUs for 4 days, data from Wikipedia and BookCorpus
- Masked language modelling
- Next Sentence Prediction

Fine-Tune BERT

- Specialization in a task
- BERT + output layer and specific dataset
- Computationally inexpensive compared with pre-training



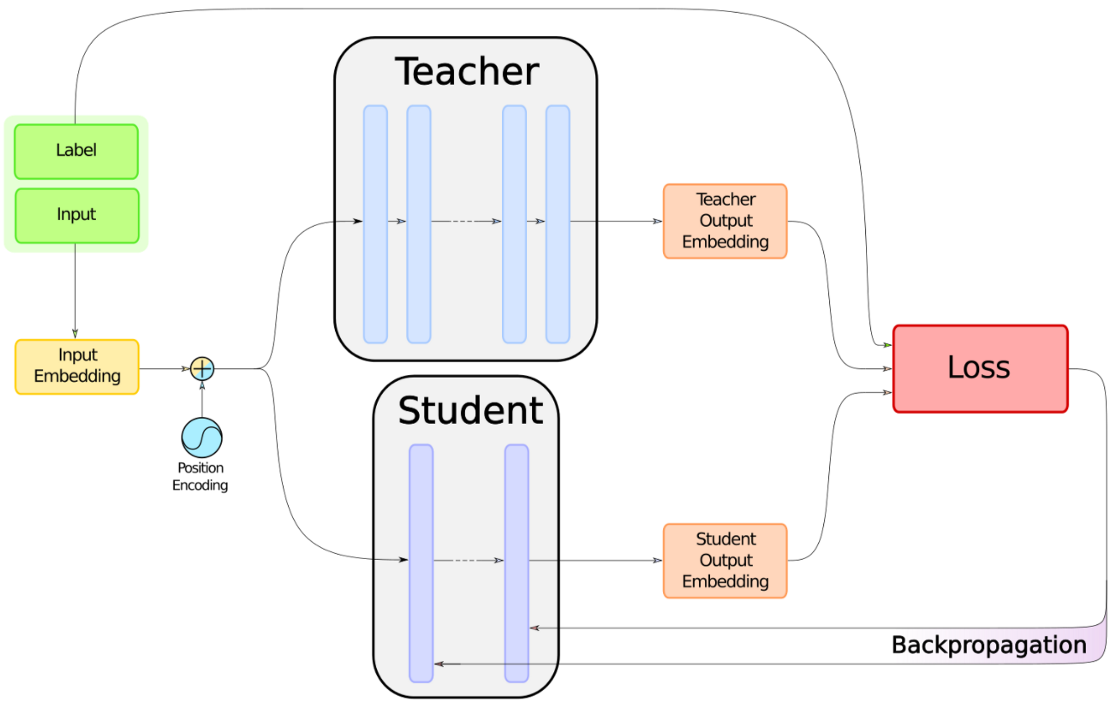
Evolution of NLP models

DistilBERT

- **40%** smaller and **60%** faster, while retaining **97%** of BERT’s language understanding capabilities
- Knowledge distillation
- Same architecture than BERT

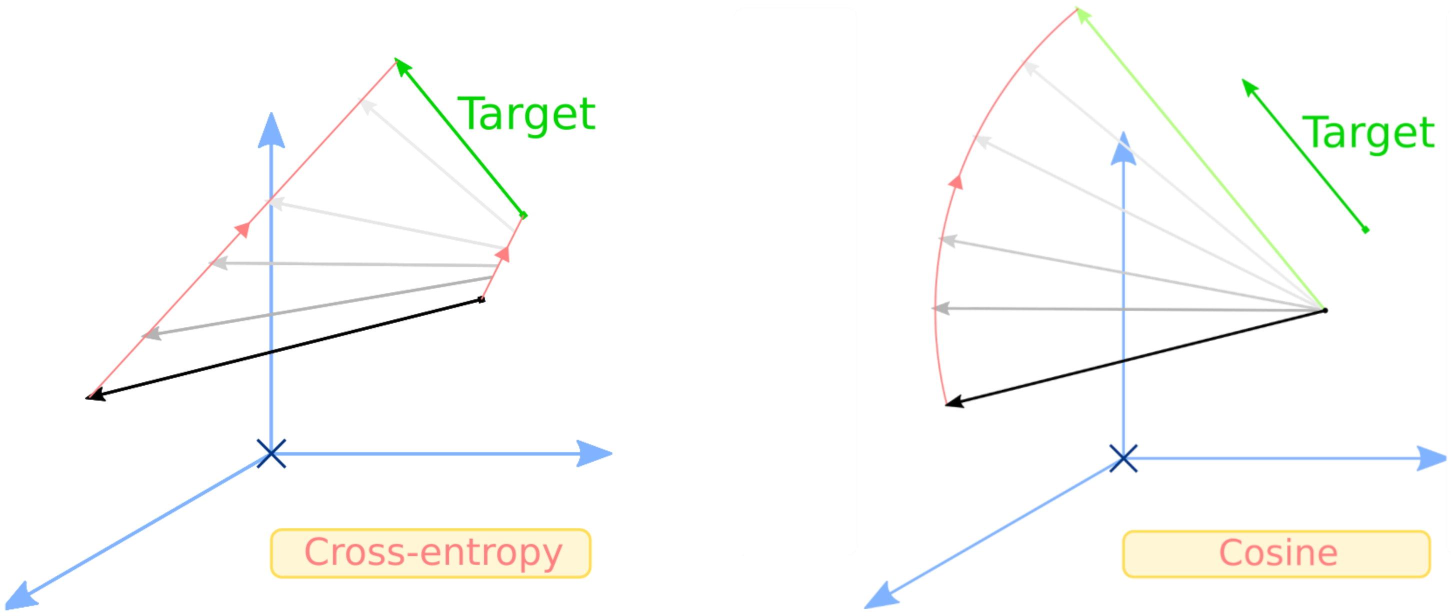
Knowledge Distillation (I)

- Smaller model (student) is trained to **mimic** a larger model (teacher)
- Training objective is a linear combination of **3 losses**



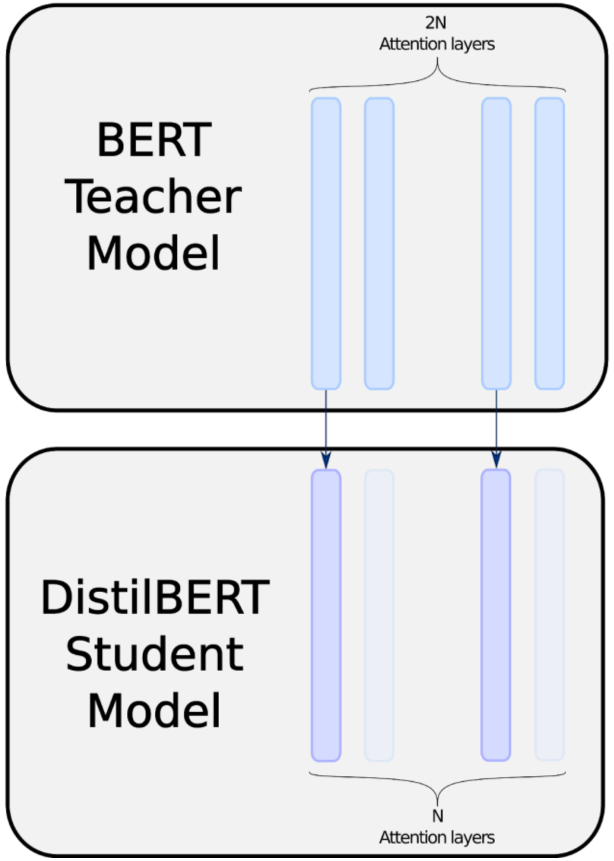
Knowledge Distillation (II)

1. **Original loss**: masked language modelling, next sentence prediction is removed
2. **Teacher-student cross entropy loss**: aim to mimic the teacher
3. **Teacher-student cosine loss:** align vector to the target



Student Architecture and Initialization

- Token-type embeddings removed
- **Layers** reduced by a factor of 2
- Optimized using modern linear algebra frameworks
- Initialize one layer out of two
- Trained on very large batches with dynamic masking



Experiments

- **Multiclass** and **multilabel** classification
- Employing BERT and DistilBERT

Multiclass Dataset

- News Aggregator by UCI Machine Learning Repository
- **4 categories:** business, science, entertainment, health

text	category	labels
'Field of Dreams' Anniversary Dream Come True for Fans	entertainment	2
Facebook CEO sees telemedicine opportunity with $2B Oculus acquisition	science	1
Google unveils self-driving cars that don't need steering wheels or brake pedals	science	1
Coleman: Casey Kasem: Pop's 'gateway drug'	entertainment	2
Homeland Security warns against using Internet Explorer until Microsoft fixes ...	science	1

Multilabel Dataset

- Toxic Comment Classification Challenge by Kaggle
- **6 categories:** toxic, severe toxic, obscene, threat, insult, identity hate

text	toxic	severe_toxic	obscene	threat	insult	identity_hate	labels
Check out the history ! — | |The WelshBuzzard| —	0	0	0	0	0	0	[0, 0, 0, 0, 0, 0]
Keep this under your hat but i heard he was gay dude.	1	0	0	0	0	0	[1, 0, 0, 0, 0, 0]
support ship Ships of this class would effectively be destroyers, or big frigates?	0	0	0	0	0	0	[0, 0, 0, 0, 0, 0]
Brilliant. Thanks so much.	0	0	0	0	0	0	[0, 0, 0, 0, 0, 0]
. All of them are confirmed by officials that their death are related to the operation	0	0	0	0	0	0	[0, 0, 0, 0, 0, 0]

Preprocessing

- Cleaning and encoding
- Splitting: 80% training (10% validation) and 20% test
- **Tokenization**
- Truncate and padding

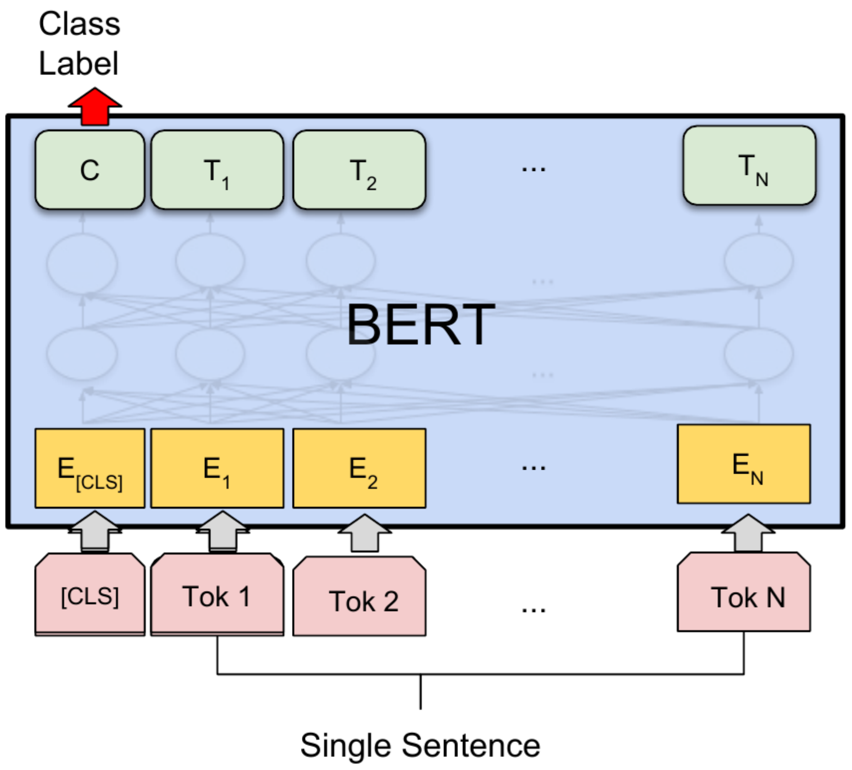
```
Original sentence:
one but two flagrantly fake thunderstorms

Token IDs:
[  101  2028  2021  2048  5210 17884  2135  8275  8505 19718  2015   102
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0]

After tokenization:
['[CLS]' 'one' 'but' 'two' 'flag' '##rant' '##ly' 'fake' 'thunder'
 '##storm' '##s' '[SEP]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]'
 '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]'
 '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]'
 '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]'
 '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]'
 '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]' '[PAD]'
 '[PAD]' '[PAD]' '[PAD]']
```

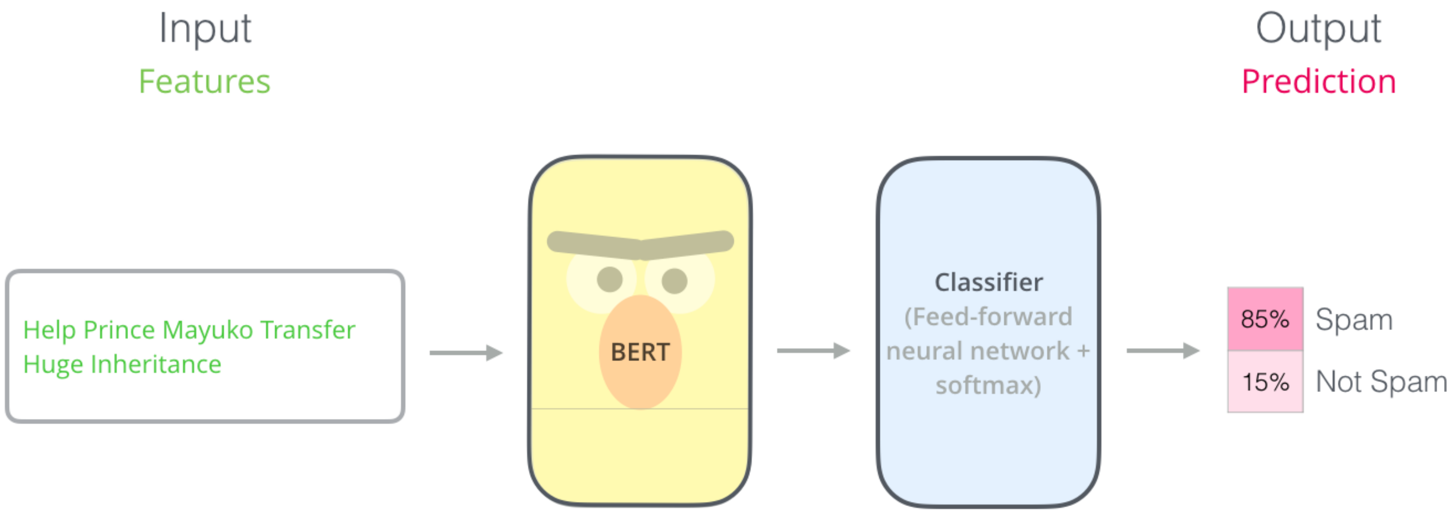
Model Input-Output

- Input: single sentence
- Output: classification token C



Model Architecture

- BERT/DistilBERT + MLP
- New parameters to train: MLP layer weights
- Softmax (multiclass)/Sigmoid (multilabel) to calculate probabilities



Multiclass Training Results

- BERT

Epoch	train loss	val loss	acc	train time	val time	total time
1	0.191402	0.140939	0.953955	0:23:45	0:00:50	0:24:35
2	0.101084	0.128902	0.960376	0:23:43	0:00:50	0:49:08
3	0.064532	0.143761	0.962921	0:23:35	0:00:50	1:13:33
4	0.040264	0.171136	0.962803	0:23:34	0:00:50	1:37:57

- DistilBERT: **50%** faster

Epoch	train loss	val loss	acc	train time	val time	total time
1	0.192959	0.134109	0.953807	0:11:56	0:00:25	0:12:22
2	0.099991	0.124602	0.961353	0:11:56	0:00:25	0:24:44
3	0.064531	0.142409	0.962300	0:11:56	0:00:25	0:37:05
4	0.042051	0.158912	0.962063	0:11:56	0:00:25	0:49:27

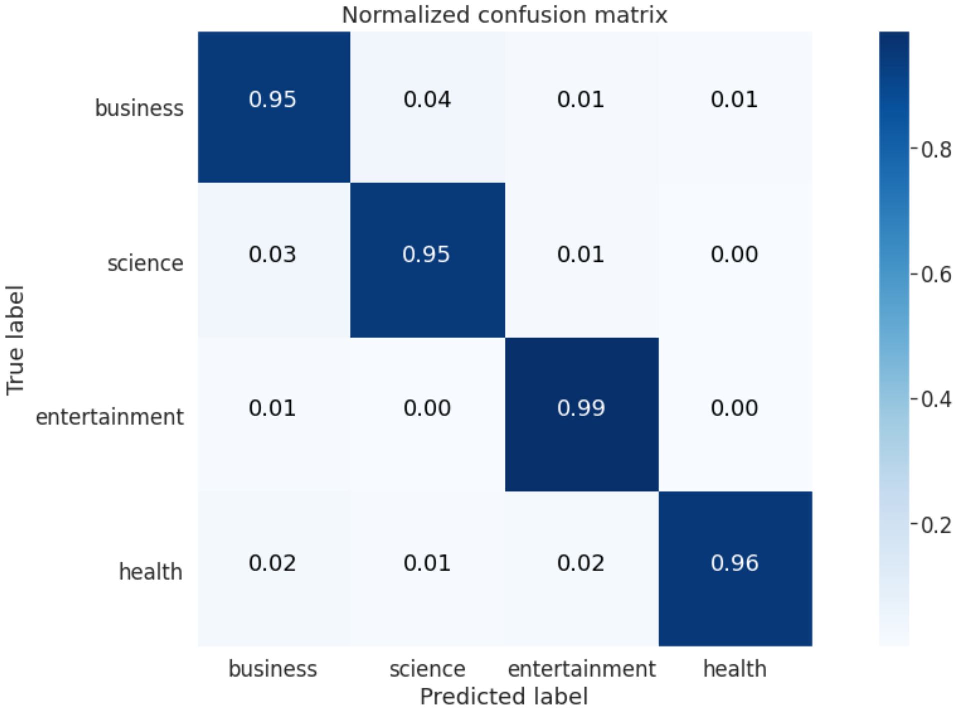
Multiclass Evaluation Results (I)

- DistilBERT is **50%** faster and **0.06%** less accurate

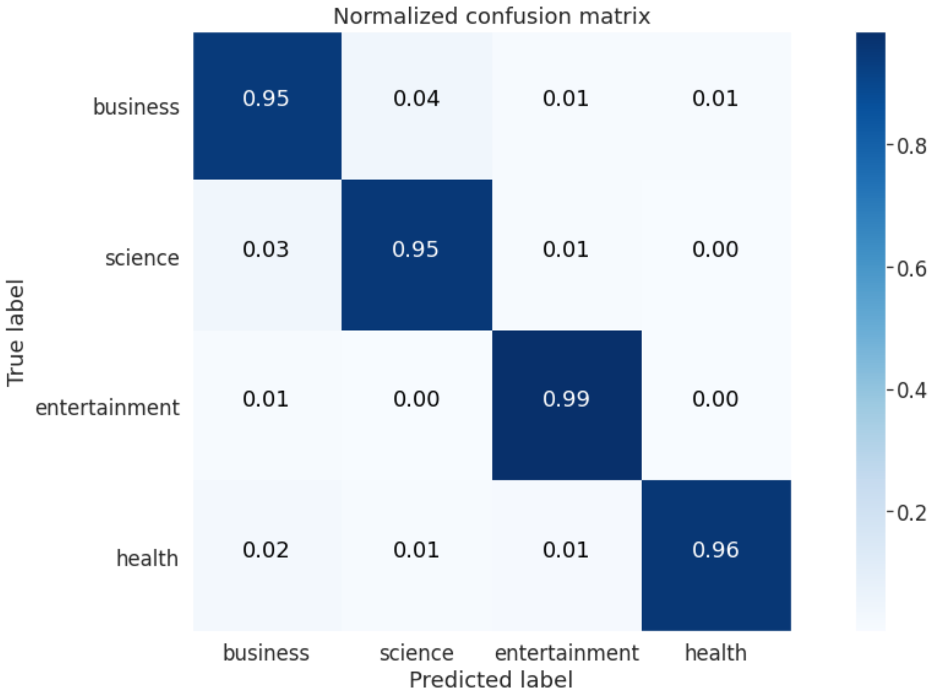
Model	loss	acc	time
BERT	0.132820	0.964205	0:02:05
DistilBERT	0.135481	0.963566	0:01:04

Multiclass Evaluation Results (II)

BERT



DistilBERT



Multilabel Training Results

- BERT

Epoch	train loss	val loss	hamm	train time	val time	total time
1	0.094784	0.048438	0.945196	0:11:46	0:00:25	0:12:11
2	0.045753	0.045861	0.934125	0:11:45	0:00:25	0:24:20
3	0.033957	0.046269	0.939297	0:11:44	0:00:25	0:36:29
4	0.026291	0.047548	0.942118	0:11:44	0:00:25	0:48:38

- DistilBERT: **50%** faster

Epoch	train loss	val loss	hamm	train time	val time	total time
1	0.114079	0.047593	0.944439	0:05:56	0:00:13	0:06:09
2	0.049229	0.045350	0.937831	0:05:56	0:00:13	0:12:18
3	0.036902	0.045316	0.941386	0:05:56	0:00:13	0:18:27
4	0.029288	0.047954	0.941972	0:05:56	0:00:13	0:24:36

Multilabel Evaluation Results (I)

- DistilBERT is **50%** faster and **0.0007%** more accurate

Model	loss	hamm	time
BERT	0.054059	0.939548	0:01:02
DistilBERT	0.052840	0.939555	0:00:32

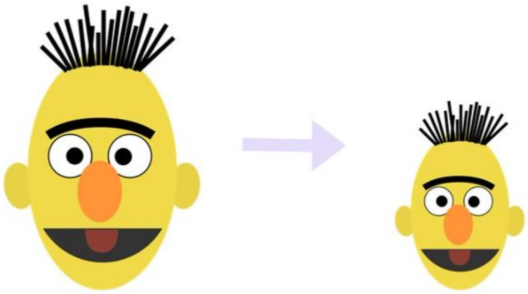
Multilabel Evaluation Results (II)

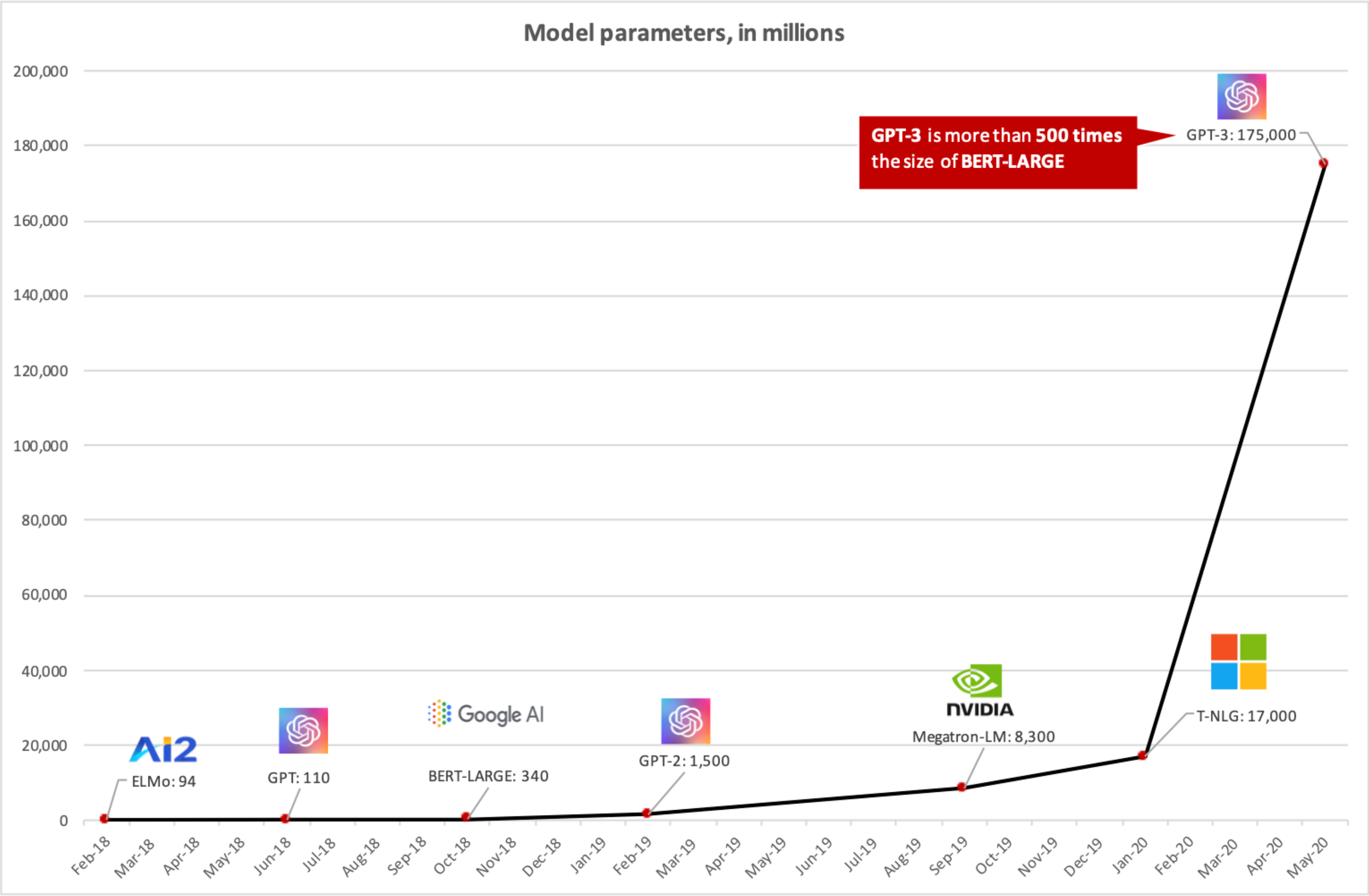
```
, darn, darn.  keep this here and OFF my page or i'll file another
complaint.  IM DONE!!!!  HE GONE!!!!  KA-BOOOOOOOM!!!!
```

	labels	BERT preds	BERT probs	DistilBERT preds	DistilBERT probs
toxic	1	1	0.959324	1	0.828968
severe_toxic	0	0	0.003004	0	0.002743
obscene	0	0	0.182350	0	0.025668
threat	0	0	0.004620	0	0.002254
insult	0	0	0.015846	0	0.025421
identity_hate	0	0	0.003184	0	0.002350

Conclusion

- Increasing parameters trend
- Computational cost, inference time and scaling problems
- Knowledge distillation to compress





Thank you for your attention