Models for sentiment analysis

SURFMETRICS

Considered Models

- BERT
- RoBERTa (Facebook)
- DistilBERT
- XLNet

- Key Takeaway:

  If we're looking for the best prediction metrics Roberta is a good choice.

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base:** 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

GPT-2

- No references to use GPT in text classification on Kaggle
- GPT-2 in its nature is a generative model while BERT isn't
- Found one reference to implement it in text classification

Database?

- Bought online
- Scrapped and labeled manually :
  - Scrap tweets related to E,S or G using ESGBERT or FinBERT. Then do the classification if positive or negative by hand to finetune models
  - Use the finetuned model to complete the database.
- Try unsupervised finetuning (continue the unsupervised training with a costum scrapped database)

FineTuning Process for RoBERTa

- First Layers capture linguistic synthax
- Only ¼ of the layers need to be fine-tuned to obtain 90% of the original quality.
- Batch_size = 16 on 10 epochs and small lr ($<10^{e}$-4)
- Almost the same approach could be applied to XLNet and BERT
- Idea: training the model to analyse the sentiment on each of the three segments (E, S or G)

Possible Approach: Ordinal Classification

- Objective: We want to have a certain proximity in classes:
  - If we want to scale our sentiment analysis from 0 to k the model should understand that it's an ordinal classification (0 negative to k positive)

- Instead of having a k-classifier we use k binary classifiers where the ith classifier calculates P(y>= i-1)

- We finally use the rule: P(y=x) = P(y>=x) - P(y>=x+1)

Topic Modeling

- Could be used to determine the most relevant topics on Twitter about a certain firm

Possible Value Chain