

FinBert & FinRoberta for Mathematica

by Jürgen Kanz, on January 18th 2023

Introduction

FinBert is an open source Natural Language Processing (NLP) model that has been trained on Financial data and outperforms almost all other NLP techniques for financial sentiment analysis.

Financial data sentiment analysis is more complicated than normal use cases. Let's take a look at an example:

"Although the global market is experiencing a downturn, food companies are doing well thanks to Covid."

Normal NLP techniques will not be able to determine whether the above sentence is good news or has a positive sentiment for food companies. This is because normal NLP techniques, like word2vec, look at each word separately and do not provide context for the words. In order to understand the sentiment of the above sentence, we would need to be aware of the context.

Context awareness is enabled by BERT for sentences. BERT is the acronym for Bidirectional Encoder Representation from Transformer. It is one of the most popular state-of-the-art text embedding models published by Google.

Domain Specific Fine-Tuning of BERT

Many domain specific models have emerged using BERT as the base and are being used for NLP tasks. Some of them are: FinBERT for Finance, BioBERT for Biomedical, VideoBERT for Video captioning categorization, ClinicalBERT for hospitals, and many more continue to evolve.

BERT, RoBERTa and the Wolfram Neural Net Repository

BERT is available in the Wolfram Neural Net Repository [1]. Furthermore, Wolfram provides domain specific fine-tuned versions of BERT like BioBERT, DistilBERT, Pre-trained Distill BERT, and SciBERT [2].

FinBERT is currently not available as a Wolfram Resource.

"The better BERT" is RoBERTa. RoBERTa is a robustly optimized BERT pretraining approach [3]. While BERT is trained on BookCorpus and Wikipedia Data, it is RoBERTa that is already trained on BookCorpus, English Wikipedia, CC-News, OpenWebText and Stories Datasets.

Consequently, the author decided to train BERT and RoBERTa on the enhanced Financial Phrase-bank.

Datasets

Financial PhraseBank

For training of the networks, the sentiment analysis dataset used is the Financial Phrase-bank [4] from Malo et al. (2014) [5]. The “Financial Phrase-bank” consists of, 4845 English sentences selected randomly from financial news found on the LexisNexis database. These sentences then were annotated by 16 people with background in finance and business. The annotators were asked to give labels according to how they think the information in the sentence might affect the mentioned company stock price. The dataset also includes information regarding the agreement levels on sentences among annotators. The distribution of agreement levels and sentiment labels can be seen on the following table.

Agreement level	Positive	Negative	Neutral	Count
100%	25.20%	13.40%	61.40%	2262
75% - 99%	26.60%	9.80%	63.60%	1191
66% - 74%	36.70%	12.30%	50.90%	765
50% - 65%	31.10%	14.40%	54.50%	627
All	28.10%	12.40%	59.40%	4845

Table 1: Distribution of sentiment labels and agreement levels in Financial Phrase-bank

FiQA - Financial Opinion Mining and Question Answering

On top, 997 selected sentences from the “FiQA - Financial Opinion Mining and Question Answering” challenge [6] have been used for training purposes.

Validation Measurements

On the repository webpage [10] in section “Train a classifier model with the subword embeddings”, we can find information about Validation Measurements for different net models.

The following results are based on calculations with “BERT Trained on BookCorpus and Wikipedia Data”, “GloVe 300 - Dimensional Word Vectors Trained on Wikipedia and Gigaword 5 Data”, and finally “RoBERTa Trained on BookCorpus, English Wikipedia, CC-News, OpenWebText and Stories Datasets”. As input the FinancialPhrase-bank was used.

	Loss	ErrorRate
BERT	0.635018	0.293086
GloVE	0.583283	0.236326
RoBERTa	0.364338	0.139319

It becomes clear, that the given BERT net model does not deliver the best results regarding Loss and ErrorRate. Nevertheless, BERT is more or less the current standard transformer, which can be observed in the Python [7] [8] and Matlab community [9].

The best results are provided by RoBERTa regarding Loss and Error Rate. Therefore, I decided to train both BERT versions simultaneously.

Training

The training process itself is not part of this notebook. Nevertheless, some general remarks:

- The Wolfram resource “BERT Trained on BookCorpus and Wikipedia Data” [10] is used to create FinBERT.
- To become FinRoBERTa, the Wolfram resource “RoBERTa Trained on BookCorpus, English Wikipedia, CC-News, OpenWebText and Stories Datasets”[11] is used.
- A balanced training set has been used. Which means that there are the same number of positive, neutral, and negative sentiments applied.
- The training has been completed with sentences with and without sentence punctuation.

Load FinBERT and FinRoBERTa

Download FinBERT and FinRoBERTa from the web

Download FinBERT





Download FinRoBERTa

Store the files in the same folder as this notebook.

C4tndwV051E07tLg0Y1rQeizWqKGxJnbSeBd11oc7FRAWvavVxMAWRcGlh8Ddwk6dq64BhFfgyY+we :
 KGs1UH0Wkd5QfjWSQprst0CMypVqxHtkrDhQ7TceZrMEJ6P25EnZkLn39LqZrIH0WhcM0dC7y7dWI :
 01LjjRa / K1a4gFaTWKsrzpVCqRoupRStf472rp9hT3 / dyUDrCYntiZ15YAwrbwfOyMwNa4 /
 VSA3GiK50BLybaBuIVXKU10bfyWoU12UzvUcXsIOyF20bdrXTYB1g8cm6EqDvkvCPB4hu / Gtio /
 RrjVdbRQ53v+Ar0jrcSsK+x2LEpZMHcQjts / TBAw7kAvZcsrIStDvtjqfowZdhG2dx / vru1tPIQh /
 uRFu1drntc24qMjWub6kyAvsXd3oDoG46+eETP7c / k3S1toU2mBsyhJg5M2hjrohStt64dlfDhRa+
 dqBF4hch7nbcZXLcytanaShrVrUm1K3eDQ+
 K5jfIILLv56IJ8tLwvSjSAlDBSWBYmCG09525RqD4EOYFmqFUS+3r2nznujz8Ak7555nMP2Lp1LvqQ3 :
 H22e5Lfk21zzcngSXfMyZTvJQhKfKHhsBHqRinCoz5i7fYDhkBnvC5ialhxeS+
 VHFTdsLXh35L6QQ7wE64Lh8T3Y4 / YzRMGqMlGGemTqayCTLz /
 TWMNiw0KxHDqoFlhQkeBQYJCZbtZ0VKjjlGVUyo813v0wRGjoa8ISl+yprBFs0Yu8XwqYEnJ / 3
 XoHsp0wukEdF671SxwUSZmi38zw8Bu5aBKsXoes1eCUs0D4AbrqIa30udI5iZr++
 ewrpBjQxQ0deR0VhB1+mpAiGsNfViqoLZBqMso2EIC4nZR9hLZpoLlXgPn+J94YKN+nzPcCi+3
 jDDya8jSYGUBBgBuDTQ3sbw / 6r1KLsmHc06izc1E2W6+fqyxJRvAIwiUcIZghZhcShzJRWbqs9Dw0 :
 8N5M3x7x3lq4QNd6WPI6QCfhi4h3abUDGDv6WOTuZQ0j7wqoL+9URZxoGaxGZEAWUp8LSgfet2AVZf :
 T5T8QRZSve8snBT3oPxhdpEdHqGbAmXVDKsRUKARPGx6ixzfVnd8rMm70CxUdJT08eA5mj2Nm1amo :
 k8d4aQpeaVjrgyuX2ja9x5cA8VZCd0DTY0m19d2dIHT1m5UwLtfomggNraqhXei0I1weAa5vEwUwi :
 gWaQ4TQ9KLXkImFWKcuXBKeL4075CT4b1js16Inmnw7BGq+
 bWLWLfY1hCL4CiK6dHDj1IrMmGbo6jbosHYtILtk1vgnXzUTZeGPPjLJiKyd0lchBx9pX6maNjnISme /
 EwrF1021kCEDvnCyuBk+u7Mwa3RopL2u6iCoM6ZoZ1Sbgm2PKN4gDC7ChAEVnVN5X3HjIGG3Jrt700 :
 zoiuKiiaWx6arnhm8di23mHt+5WsMnFRILKSxsrk855jhJg+AnaqoMGHdRtr3J8AGsRauz9MYvpdF7 :
 yXUquuymK3UNTzxwpo73V21ZiS1eABU1RfQrmGUKkmByktTf7Mwo1YX31W5EqLuPxw4tz6JCLqEk1 :
 0N2zkI0L0prM2H6GoDc / MqN46ZGogEUiW+znRuTIQ7N17sh29zQqVvKn7TVk03Jhue+
 W0p68ZPqZKR9Br67v1P70vfRcrbVFs+1f1AKDTXoEp1BYGOVWV2x+
 ieQe7sIduf1afuJKHGSgEysYqQ1a7u61gszKhAMgYBnzHMFwi1qsfpUBmaUa /
 eK0eCW0jgrIWjVmkUQ1vVswvBxk+iWgBqkfemS1VFRns6s7DLa0IV12x8pIpnG4NF2UPqoGh4qra6g /
 htdMYUFkbHeSB57nJVkImPkjjePdhfA300PrvwYsqNoIr9hV01qaOyg0mlbKKf+
 T1urP0sgvo0veea17N7qYmZJg5MrWpKwUzdtQ+RCemAe5maUZAGZ8qd5QimRqye0L1NLV2x1o3EH8 :
 dYVMFawaFmpbmyXLgraSNbVpgyixM91IMzw00epHzBPDoeVsJlqYeg51PGiKb2 /
 u8mYsuqNyIOGA5d1uF7jHIIvXAAyt6fe+7VBB1NeVJXZich700t7m / L4kscGWZuLMovp+
 cLy598Hb3dNT7crCxicxMp2nRhr6IA5mejawbKiv629akieEqNINGJinY / TS0cT08luHuixUDINuAX :
 JGQXDPVixn7TJ2xygf7g5YmbxQcWzFgjeS3Zy0HfIDNCwL36sijqpXt0Axt0TbQIAeTpzow5vgOK+x :
 qzJmhgKYCKIJ1dYYjroHLXoHJPT7ttAdIW21qL43DpaVucwp8K1PghkxrtHepCGCk4GqEgmHsN12 :
 1rb4NR1g0TEGD7HYRXzrq00w0jOTcMBvvY0QdJz / kWE3y3BX11dXro8+
 vnSeYKenrwyL4rC2g7Kb1YK14Y6NzD1L / AJcWFcBtJrDT / sebWhcTJN / t5IGcN0Pm7RP3HeuU6nynho
 +8y7sTAynHi5WJ+3bq4fu0m1uc4FzF17nqQ7wSgmsK4fWs5Vt0Q /
 peHBjS0cF6aXt2kcbT7NfOX0egXRbZWL6E2jZWoQ3U7ahYWzt89Ne /
 o3KHvWquacNE1zxuR9nMofpc783RgGTh2CGS1DJR155rQZdaFWbeidJZBjblCP+tdR5 /
 GqAw89ZN7mjTWmpnI9U / SILv+ku4eFqmRM2othwICUi8b5c1ge1EOdwFqXraf /
 FzsGzxpB9XVx7YrcFH0njzpipj2XrgJ81LJ0bfjZx3q1Jv1RWB7Rtv8B / umXm6puqF++z7c8HSBs+8
 QH1mpZcMBNU1AcBOaw1um4nFKX51Dygg22Fr /
 mjW81gZCC1qdsueNbGweB68FiJdcNGHFaUuZihM1rNMx55 / vwECam1MYc0Kvn08BStoksP2dL812K6fa
 ++3Mw / 6FxVFJcjANBG / aLerFshbCqW2kDnViD853cyCAOHLlitfkjo10xU+o5tQ745v4Abwaac4VcN :
 vkmXYXfGiZxn66rFfpS74c5yY8EMRt81kT572eZKhpfAV2euCJuznmGmsnsnpfpq6Qbx+I / 5
 lZZuYqy0BQhQ8417IoKw+cepYc1NQmv00cxvUffsXP0FWsrD9wN72s8MHn8vdT6WSZ4NPZGX1Xnqe :
 9kTjvYcrcncVa8uiTXdrW2876dV2PDbyyLu2kUzZjEoo2S /
 Ah94pFQtcgAmwyadVPf8pZt0PcNg3IPXbid9 /
 sdhQYQqpLZv0NwiGvQCKzPf0cVaqGtHRA5sqqvutiBuJ7q3I6vLKMp1+5
 f011c9kfgn0ZTH2wiy5e5CqvJdWiQiDuup0+
 ZR157gSp6L55xL44eaXSBvrCm2py2S3njh007A8VM3uzNk2PeG2sE1X1xRkE254H2g /

```
MWwFsJD3Y593K5MFSrj9RvM3Zrbeb+
bxTLb6gCOfwTvmEUZW7z6QV7zFsSsmcwmCeLW6vPzgdvZbOQEb004yKA/
FWVcfKSIASeqZSTWkxmedUQxhr/i0IrGjWqgbyFx1qbsQo4jIW/BsZ2JDhHjtPeTYDKXULGza+OmJoD:
RKIHC59063rLyfWxXiYu3YY59sk4IMnEw9Xb0WVxvxd8S1n91K4HCU9cq15gVUMtJZf8TDUdfoQa:
rn//s5iM351741iL7NZn+Sxc/iX2adBuo1KLzmu77pG735tk9/ZvEzrn5/+ubWQdXh1wt9NRTB7k3H:
9H8B9vHpN7Zr/s1AsoNTCL36s5wG/P+Xf/P1Df8WrVh4jZP9D+GMYRo="]
```

Out[]=

Sentence
TikTok considers London and other locations for headquarters
Disney cuts ad spending on Facebook amid growing boycott: WSJ
Trail of missing Wirecard executive leads to Belarus, Der Spiegel reports
Twitter says attackers downloaded data from up to eight non-verified accounts
U.S. Republicans seek liability protections as coronavirus aid battle looms
Wall Street Week Ahead: Fund managers navigate 'Night of the Living Dea
Take Five: Hoping for that V-shape in earnings
Evictions nearly back to pre-pandemic levels in some U.S. areas: Cleveland Fed study
Google bans ads on coronavirus conspiracy theory content
Flight to suburbs boosts U.S. homebuilding, but COVID-19 surge erodes consumer sentiment
Delta will avoid pilot furloughs if they agree to reduced minimum pay: memo
S&P 500 ends higher as traders weigh stimulus and virus worries
FTC considering deposing top Facebook executives in antitrust probe: WSJ
Where did all the bulls go? Rallies stall as EU summit begins
BlackRock profit beats Street estimates as markets rebound
Dollar falls, euro hovers at four-month high as EU negotiates rescue fund
Boohoo tells UK Home Secretary it is probing charges of malpractice in supply chain
Exclusive: EU in talks with Moderna, BioNtech, CureVac to secure possible COVID vaccines
Former Pemex boss arrives back in Mexico for graft trial, enters hospital
Twitter says about 130 accounts were targeted in cyber attack this week
  rows 1-20 of 60  

The column “Sentence” contains 60 sentences from across the internet. The “Sentence” sentiments are listed in the columns “Mathematica FinBERT” and “Mathematica FinRoBERTa.” It is recognizable, that few sentiments differ among these two networks.

Sentiment analysis has been done with the “Huggingface FinBERT solution for Python” [7][8], and with the Matlab FinBERT solution [9]. A comparison of all the sentiments shows that there are differences between the sentiments.

The person who wrote the notebook is not a native English speaker, so some mistakes might happen during sentence interpretation. The sentiments calculated by Mathematica FinRoBERTa fit the expectations of most of the author. In general, the differences might be caused by a lack of

training settings that are standard.

Therefore, the user is free to choose which solution he/she prefers.

References

- [1] Wolfram Neural Net Repository, <https://resources.wolframcloud.com/NeuralNetRepository>
- [2] Wolfram Neural Net Repository, Text Domain, <https://resources.wolframcloud.com/NeuralNetRepository/inputdomain/Text/>
- [3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, 2019, RoBERTa: A Robustly Optimized BERT Pretraining Approach, <https://arxiv.org/abs/1907.11692>, arXiv:1907.11692v1
- [4] The dataset of the Financial Phrase-Bank can be found here: https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10
- [5] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala, 2014, Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65, 4 (2014), 782–796. <https://doi.org/10.1002/asi.23062> arXiv:arXiv:1307.5336v2
- [6] selection of FiQA-2018 data, <https://sites.google.com/view/fiqa/home>
- [7] <https://github.com/yya518/FinBERT>
- [8] <https://finbert.ai/sentiment>
- [9] <https://github.com/matlab-deep-learning/transformer-models>
- [10] BERT Trained on BookCorpus and Wikipedia Data, <https://resources.wolframcloud.com/NeuralNetRepository/resources/BERT-Trained-on-BookCorpus-and-Wikipedia-Data/>
- [11] RoBERTa Trained on BookCorpus, English Wikipedia, CC-News, OpenWebText and Stories Datasets, <https://resources.wolframcloud.com/NeuralNetRepository/resources/RoBERTa-Trained-on-BookCorpus-English-Wikipedia-CC-News-OpenWebText-and-Stories-Datasets/>