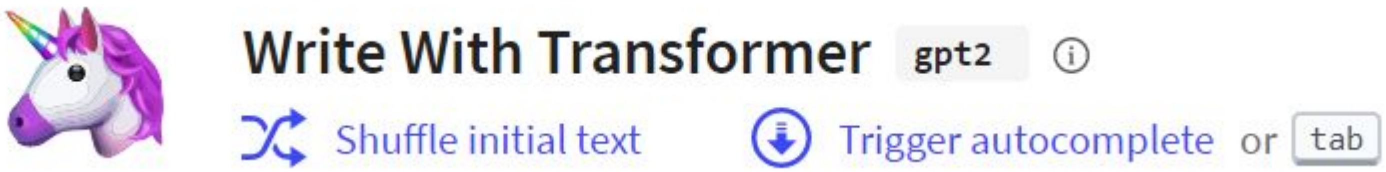
I am going to start my talk at Data Science Portugal Day 2019 with a little help from AI.



https://transformer.huggingface.co/

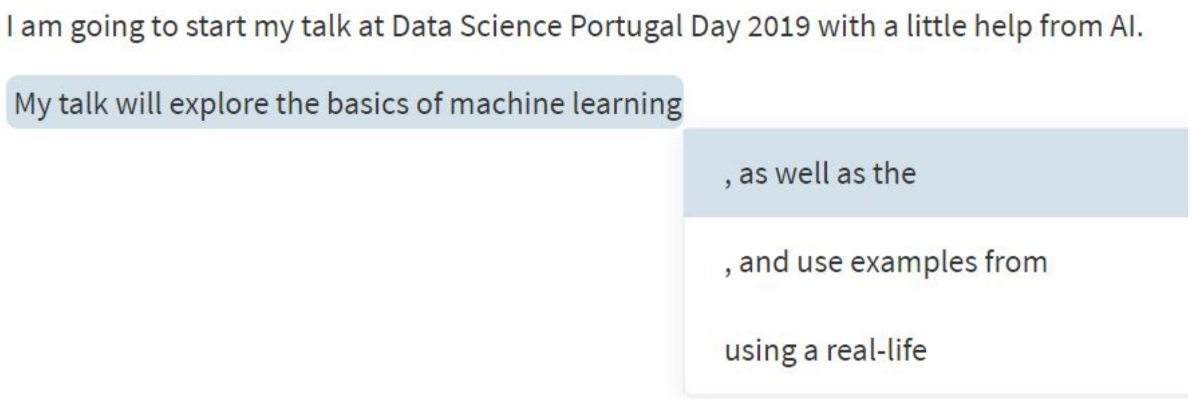
I am going to start my talk at Data Science Portugal Day 2019 with a little help from AI.

I am going to start my talk at Data Science Portugal Day 2019 with a little help from AI.

- My talk will explore
- AI is a lot
- I will try to explain the

https://transformer.huggingface.co/

I am going to start my talk at Data Science Portugal Day 2019 with a little help from AI.



https://transformer.huggingface.co/

I am going to start my talk at Data Science Portugal Day 2019 with a little help from AI.

I am going to start my talk at Data Science Portugal Day 2019 with a little help from AI.

My talk will explore the basics of machine learning , as well as the underlying ideas about how to build models. This is a topic that I have been interested in for a while now, and I 'm going to try to explain it in a fun and interesting way . What I am going to talk about is a "Deep Learning" approach to machine learning which focuses on training the model on large datasets.

https://transformer.huggingface.co/

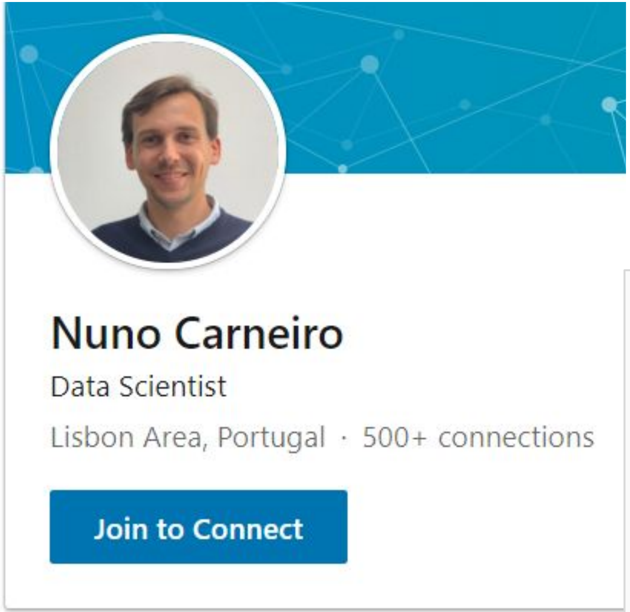
CLEVERLY

Efficient service with a human touch

Text Classification with Deep Learning

Nuno Carneiro
Head of Product and AI, Cleverly

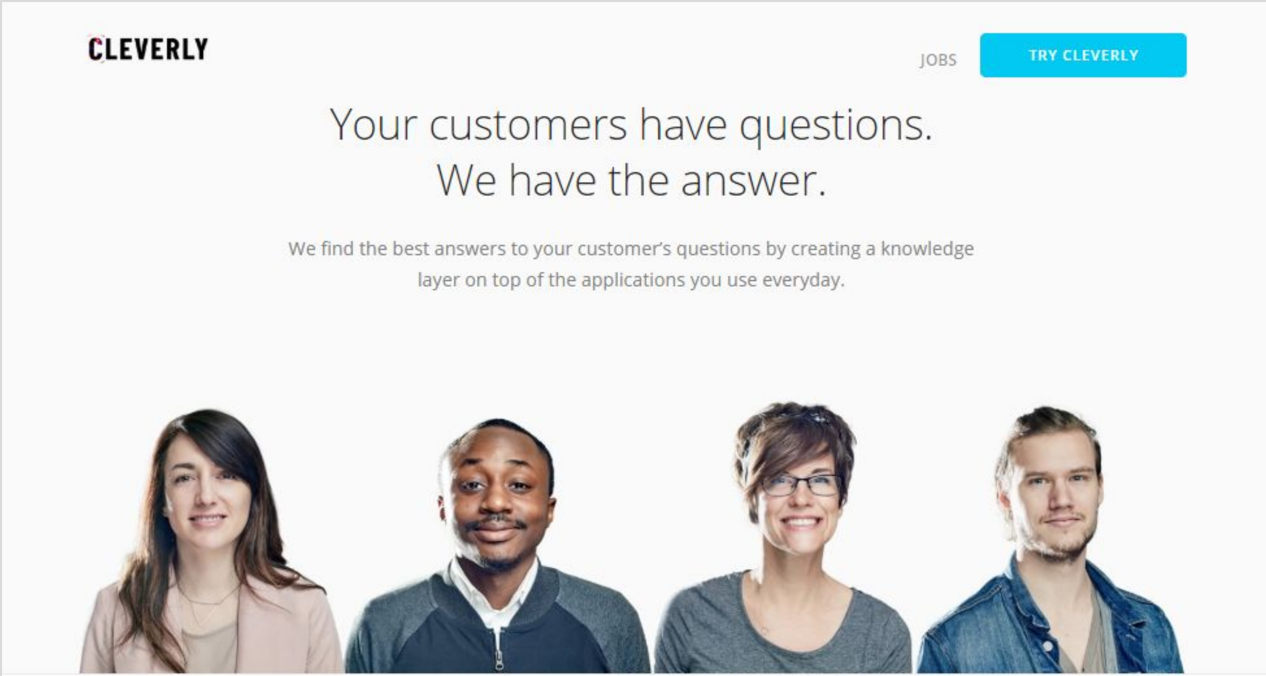
nuno@cleverly.ai
www.linkedin.com/in/nunocarneiro



linkedin.com/in/nunocarneiro

U.PORTO

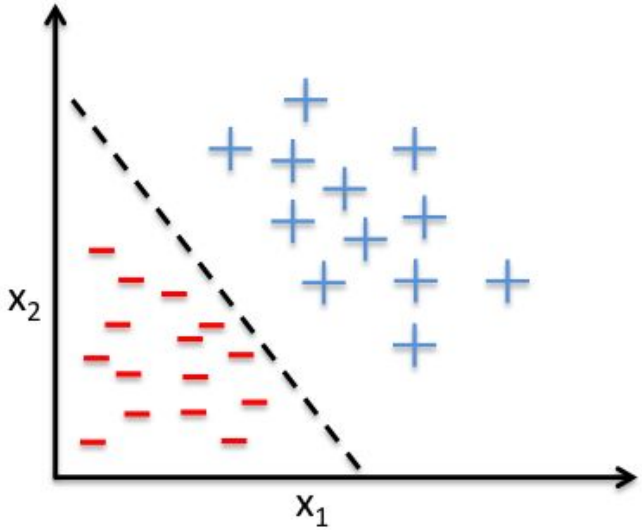
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



Text Classification with Deep Learning

1. What is Text Classification
2. Review of the state of the art
    2.1. Bag of words, TFIDF, Naive Bayes
    2.2. Word embeddings
    2.3. Neural architectures
3. Pre-trained models
4. Case study with BERT

1. Classification is a type of ML problem



Example of a linear decision boundary for binary classification.

In the example: finding the right parameters of the linear function allows us to find a separation to correctly classify Pluses and Minuses.

Image: https://sebastianraschka.com/Articles/2015_singlelayer_neurons.html

1. In Text Classification, we predict a label(s) for an observation of text.

Commerce

I would like to order a new computer.

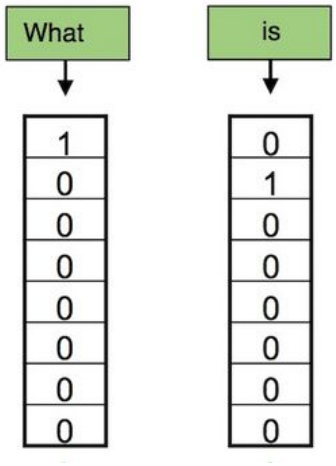
Refund request

The previous order never arrived, please also issue that refund ASAP!

Refund request (85%)
Order issue (5%)
New order (1%)
...

How to represent text numerically so that it be used by machine learning models?

2.1 Classical representations of text are based on the bag of words model.



Each word is represented by a sparse vector.

Source: Malinowski, Mateusz & Fritz, Mario. (2016). Tutorial on Answering Questions about Images with Deep Learning.

2.1 A sentence is thus represented by a vector of ones and zeros.

#	Example	Features											
		the	cat	sat	on	the	mat	.	what	is	behind	table	coffee
1	The cat sat on the mat.	1	1	1	1	1	1	1	0	0	0	0	0
2	What is behind the table?	0	0	0	0	1	0	0	1	1	1	1	0
...	...	0	0	0	0	0	0	0	0	0	0	0	0
n	The coffee is on the table.	1	0	0	1	1	0	0	0	1	0	1	1

Issue: Sparse representation leads to huge number of features.

2.1 The Naive Bayes model is the standard baseline for text classification.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Simple and efficient. The Naive Bayes model achieves decent results based on assumptions of independence of word occurrence.

Image: https://becominghuman.ai

Word embeddings

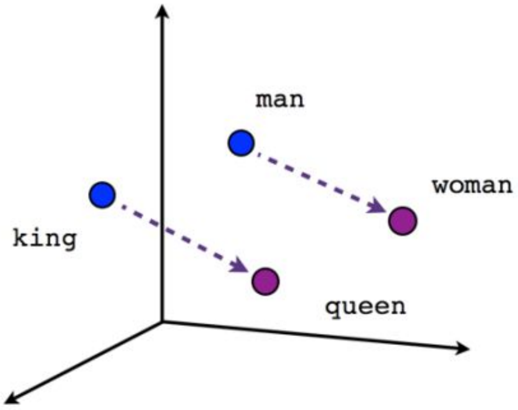
A new way to represent text through dense numeric vectors

2.2 Dense word vectors became the norm after 2013, when the *word2vec* paper was published.

#	vocabulary	embedding					
		d1	d2	d3	d4	..	dM
1	cat	0.55	-0.28	-0.96	0.84	..	0.87
2	sat	-0.86	0.00	0.76	-0.96	..	0.00
...	...	0.13	0.00	0.19	0.00	..	0.00
n	mat	0.59	-0.09	-0.52	0.00	..	0.26

Each word is represented by a dense vector which contains its semantic meaning.

2.2 The *word2vec* algorithm created word vectors with meaningful real-word relationships.

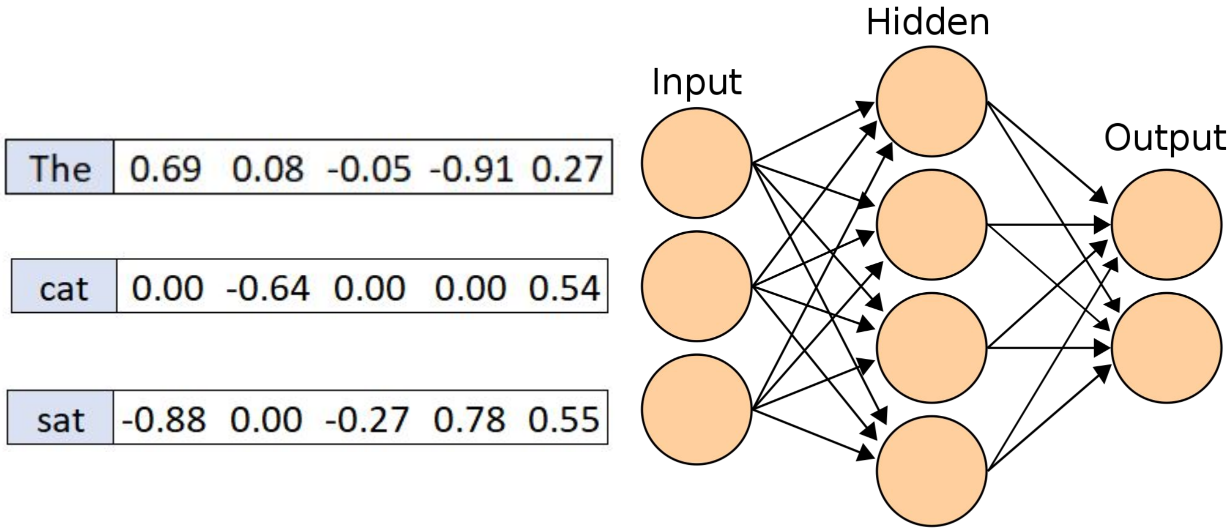


Male-Female

People started using word vectors which were pre-trained by other researchers.

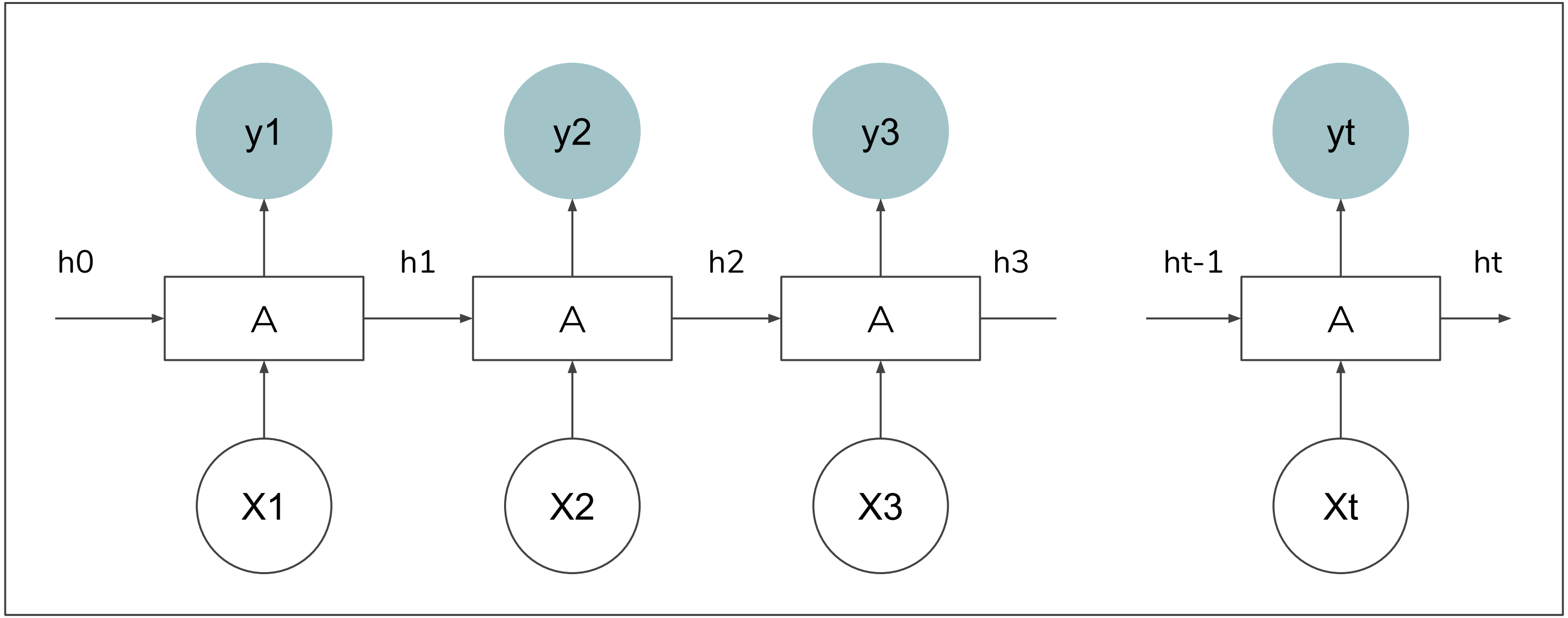
You can easily go online and download word vectors for most words in the English language trained with *word2vec* or *GloVe* on datasets such as Wikipedia or Google News..

2.2 Dense word vectors allow the construction of neural architectures where the input size is now limited to the sentence length.



2.3 Recurrent neural networks

2.3 Recurrent neural network architectures model an inherent characteristic of text: it is sequential.



2.3 A Long-Short-Term-Memory network is a type of recurrent neural network used in many modern models.

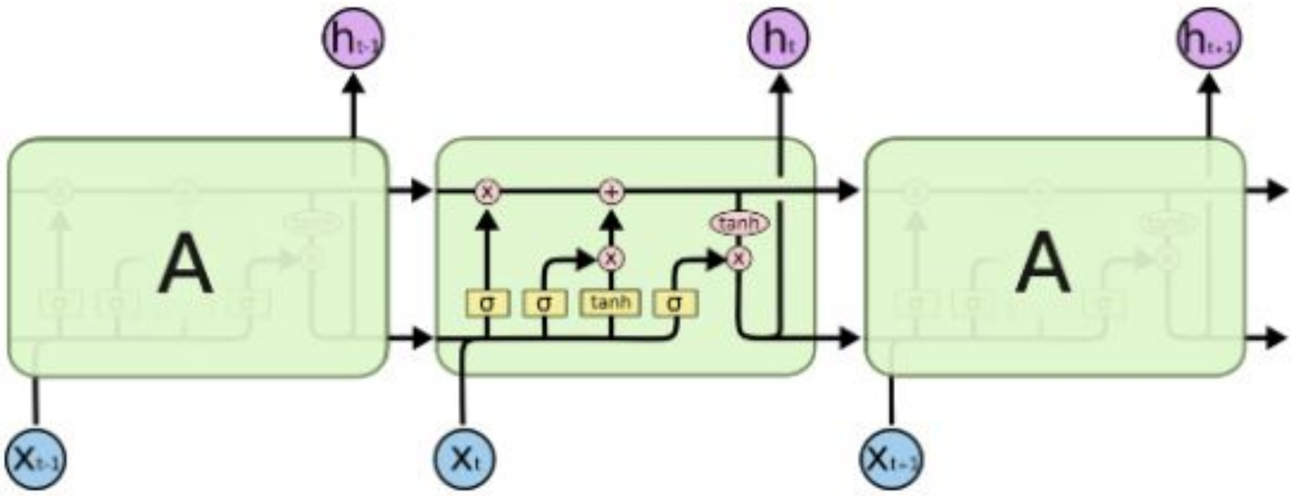


Image: https://colah.github.io/

3. Pre-trained models

3. ELMo introduced the concept of contextual word embeddings.

Image: http://jalammar.github.io/illustrated-bert/

3. The Transformer architecture proposed that “Attention is all you need”.

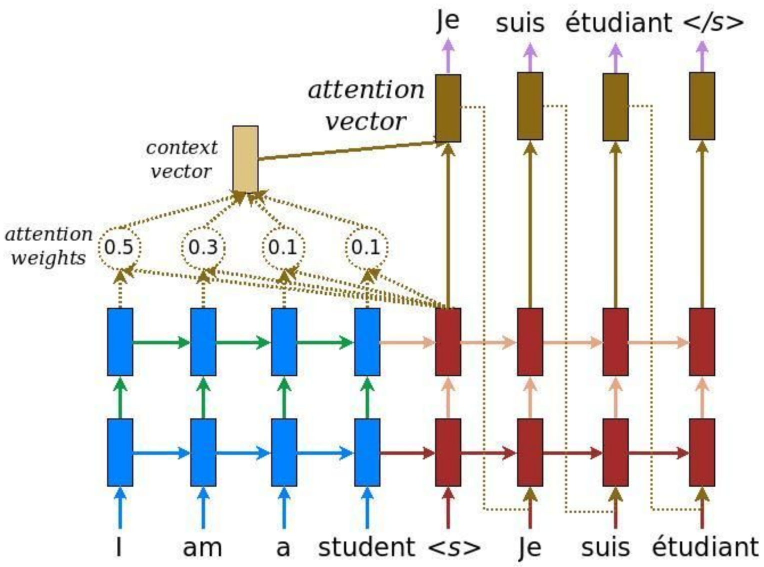
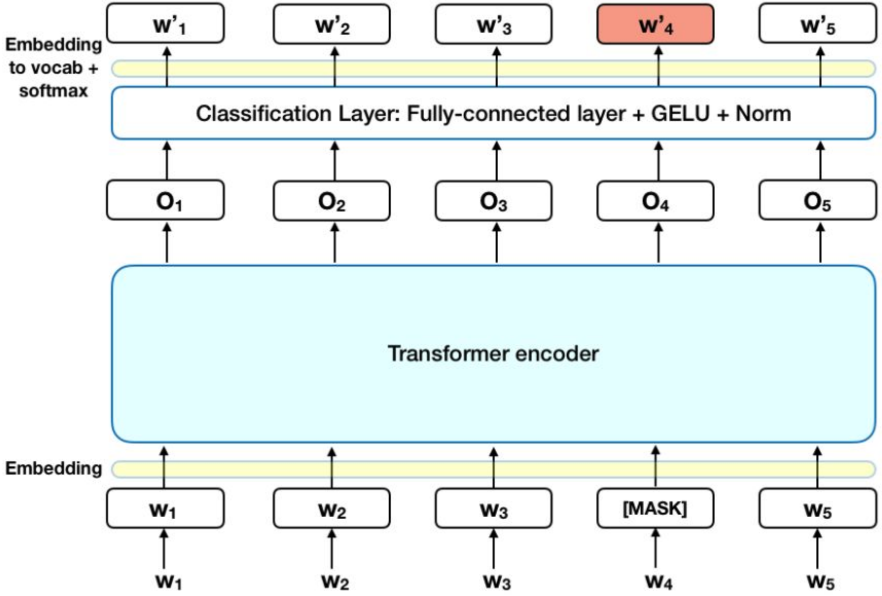


Image: https://www.tensorflow.org/tutorials/text/nmt_with_attention

Great introduction to the Transformer: http://jalammar.github.io/illustrated-transformer/

3. Since 2018, most breakthroughs in NLP come from the training of huge architectures based on the Transformer.



BERT architecture

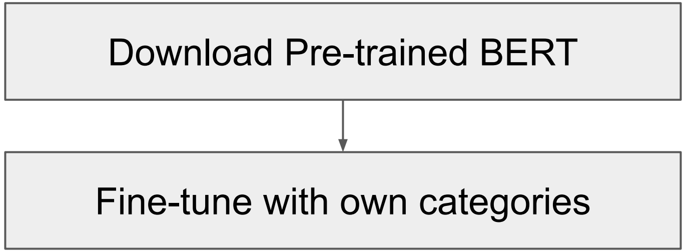
Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B
1	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0
2	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0	69.2	97.1	93.6/91.5	92.7/92.3
3	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2
4	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9
5	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1
6	Microsoft D365 AI & MSR AI	MT-DNN-ensemb[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	91.1/90.7
7	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6
8	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7
9	XLM Systems	XLM (English only)		[illegible]	[illegible]	[illegible]	90.7/87.1	88.8/88.2

Most of the models in the GLUE Benchmark leaderboard are based on the Transformer architecture.

Click on a submission to see more information

4. Case study: using BERT and transfer learning for text classification

4. Fine-tuning BERT for text classification.



There are different ways to fine-tune BERT, such as:

- Train last layer with same task
- Train whole network with same task
- Train whole network with adjacent task

4. Using BERT for text classification in customer service.

Data: Customer service emails from one specific account

Dataset 1

7.000 emails with labels

40+
categories

Challenge: BERT often fails to converge/train for small number of observations

4. But we have a second dataset with different labels.

Dataset 2

Dataset 1

7.000 emails with labels

40+ categories

140.000 emails with labels

200+ categories

4. Solution: Train BERT on the big dataset which has same data distribution but different labels

Pretrained BERT-base

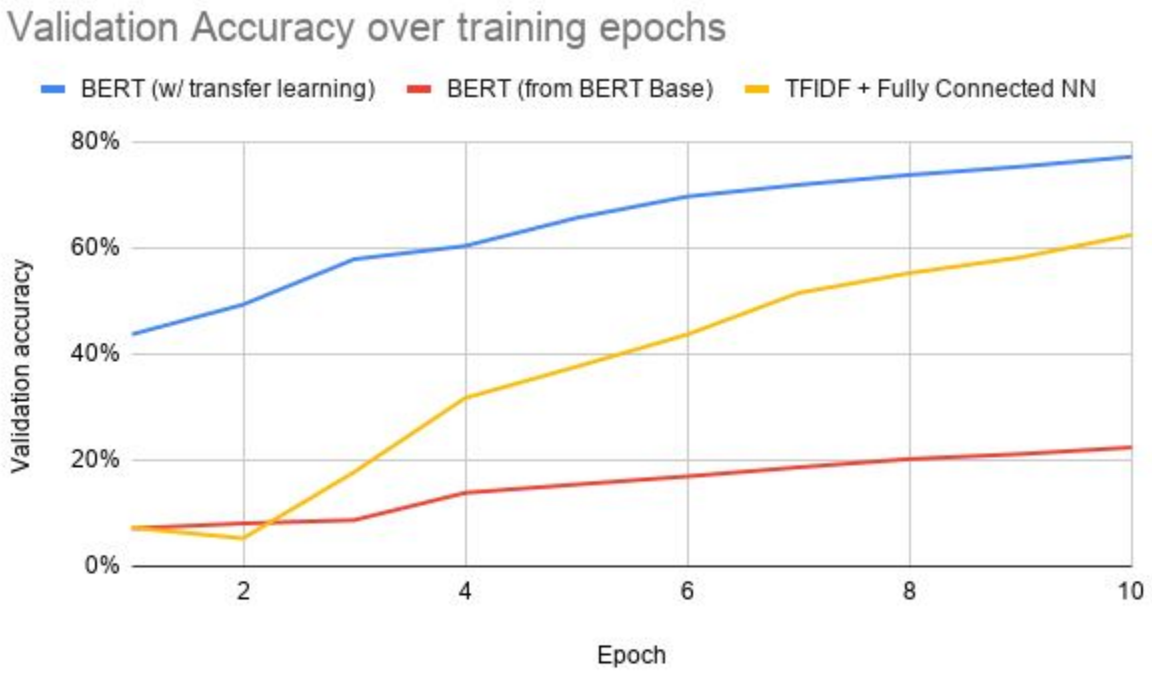
↓

Fine-tune on adjacent task (Dataset 2)

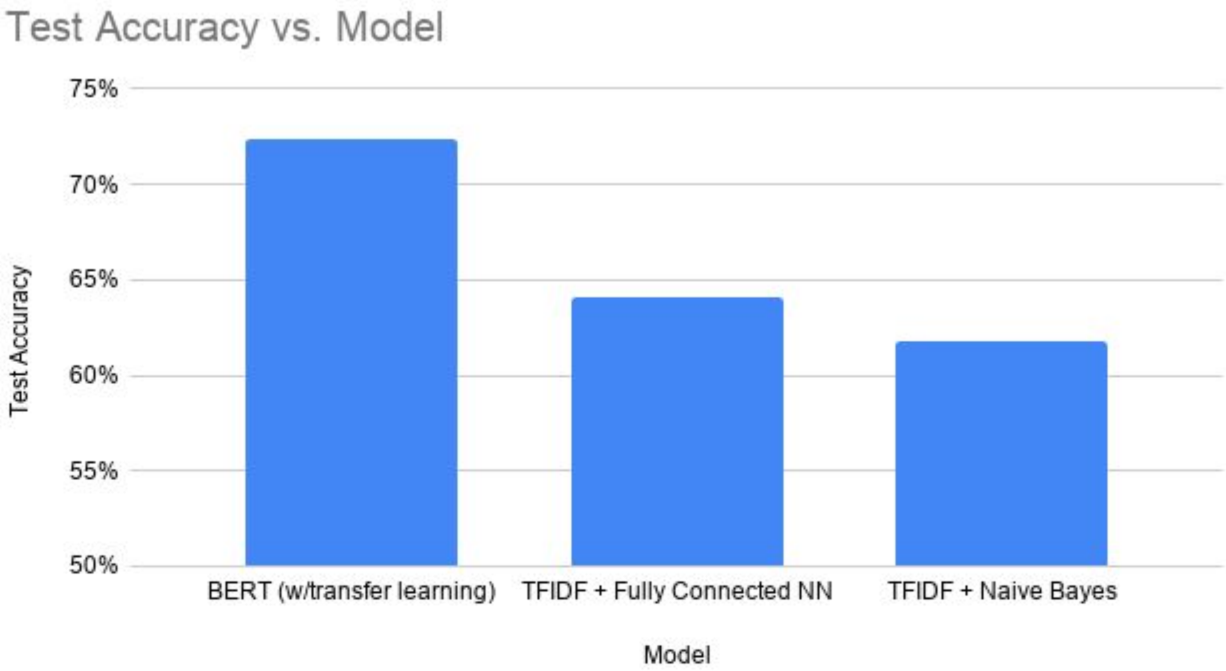
↓

Fine-tune on our own task (Dataset 1)

4. A model trained with transfer learning achieves high accuracy scores with few epochs.



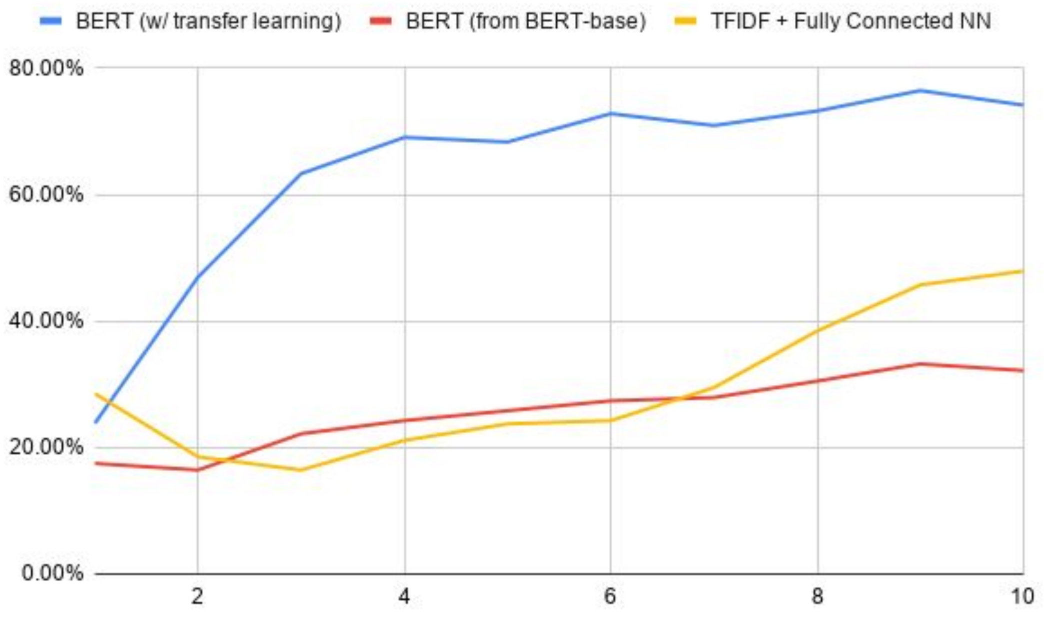
4. The BERT model achieves higher test accuracy than other architectures.



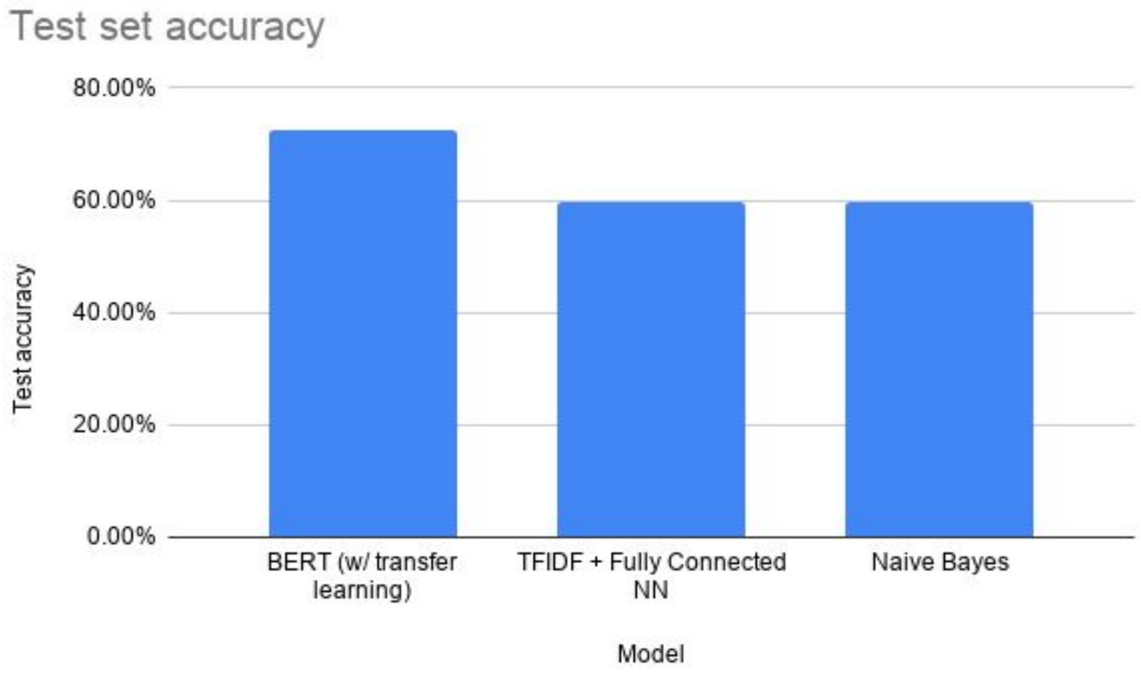
4. The same results are observable when training with different datasets and transfering that knowledge to a new dataset.

Datasets

- Customer service emails
- Different accounts in same industry
- Fine-tune for one specific account



4. The same results are observable when training with different datasets and transfering that knowledge to a new dataset.



Summary

- It is an exciting time to work on Natural Language Understanding
- Transformer based models achieve great results in text classification problems
- A lot of the potential comes from being able to transfer knowledge across different domains
- Go and try it out yourself!

References

- http://ruder.io/
- https://colah.github.io/
- https://sebastianraschka.com
- https://mccormickml.com
- http://jalammar.github.io/
- https://github.com/huggingface/transformers

Thank you!

nuno@cleverly.ai
www.linkedin.com/in/nunocarneiro