A SAD STUDY ON WHAT'S INSIDE NEURAL LANGUAGE MODELS

Irene Dini

SLLD Exam, 17/05/2023

A bit of context: Natural Language processing

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to **process and analyze large amounts of natural language data**. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately **extract information and insights contained in the documents** as well as **categorize** and **organize** the documents themselves.

Neural Language Models

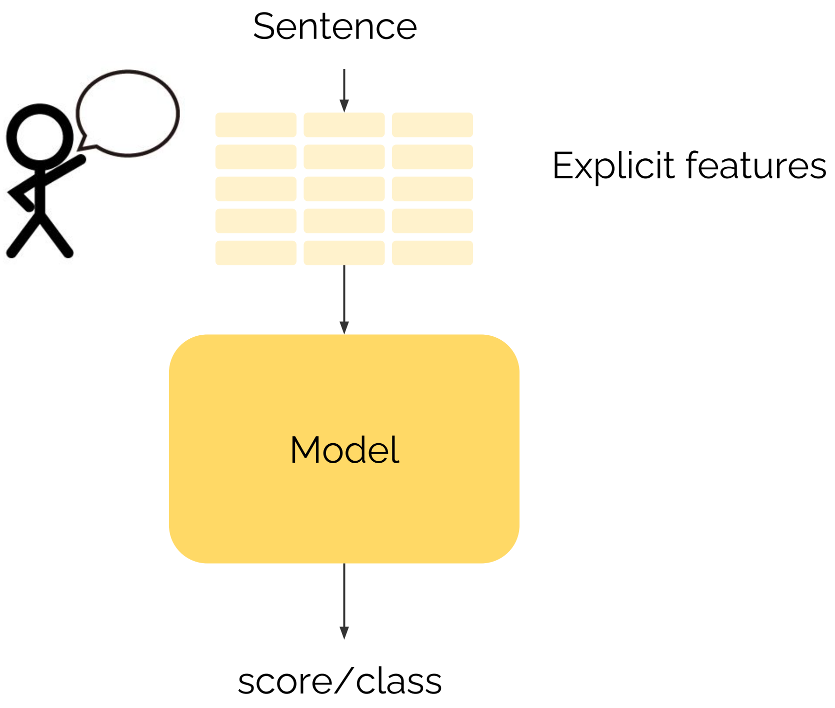
- Neural Language Models (NLMs) have become a central component in NLP systems over the last few years
- **Language modelling**: task of assigning a probability for the likelihood of a given word to follow a sequence of word

Can you please come **<u>here</u>**?

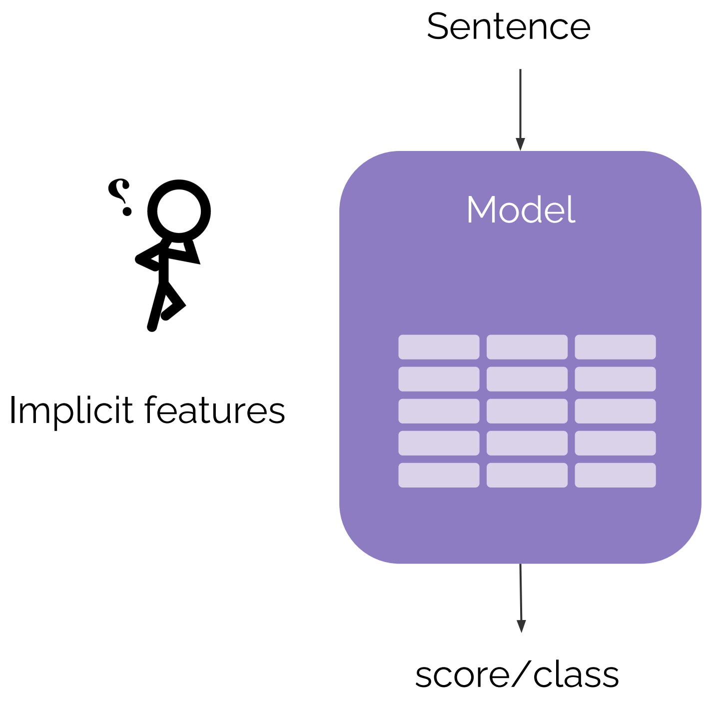
Context

Word being predicted

How it worked



How it works



Project’s goal

Study if and where NLMs encode **linguistic information** about the processed sentence.

Methodology

Leverage on a set of **explicitly encoded** linguistic features to find a set of implicit features that encode the same type of information.

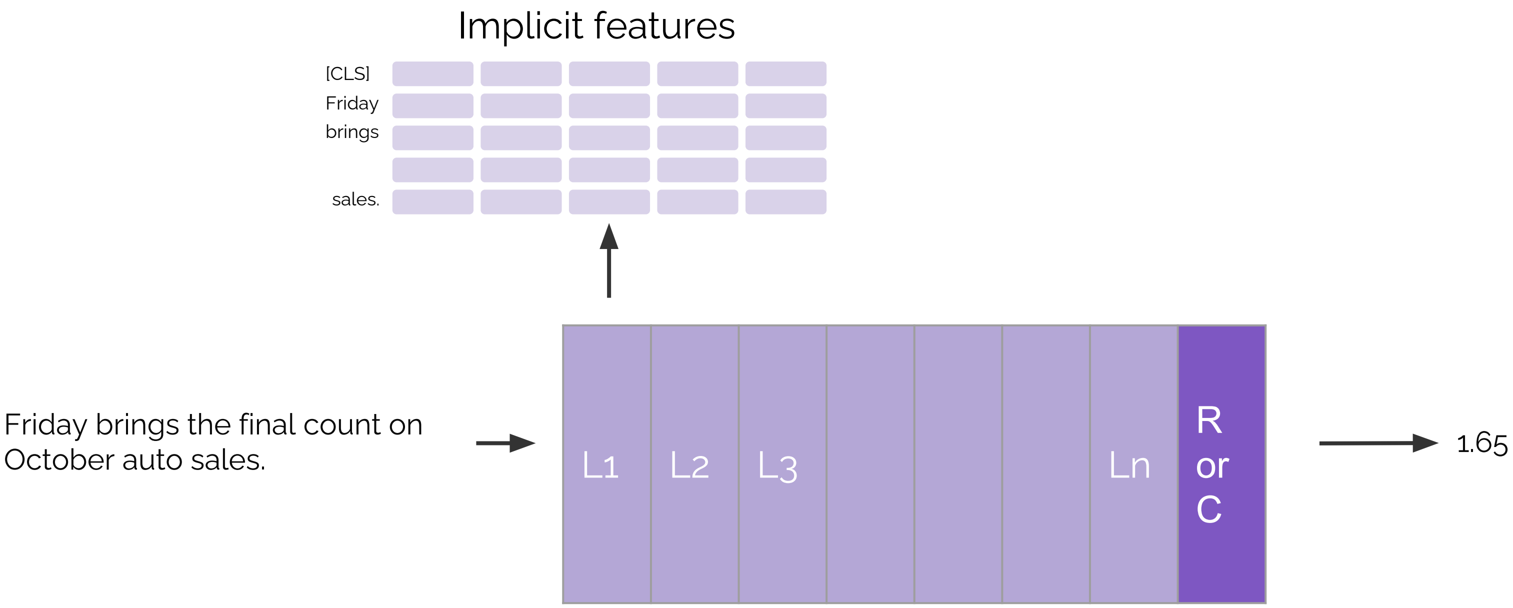
Profiling UD (explicit features)

Profiling-UD is a web-based application devised to carry out linguistic profiling of a text, or a large collection of texts, for multiple languages.

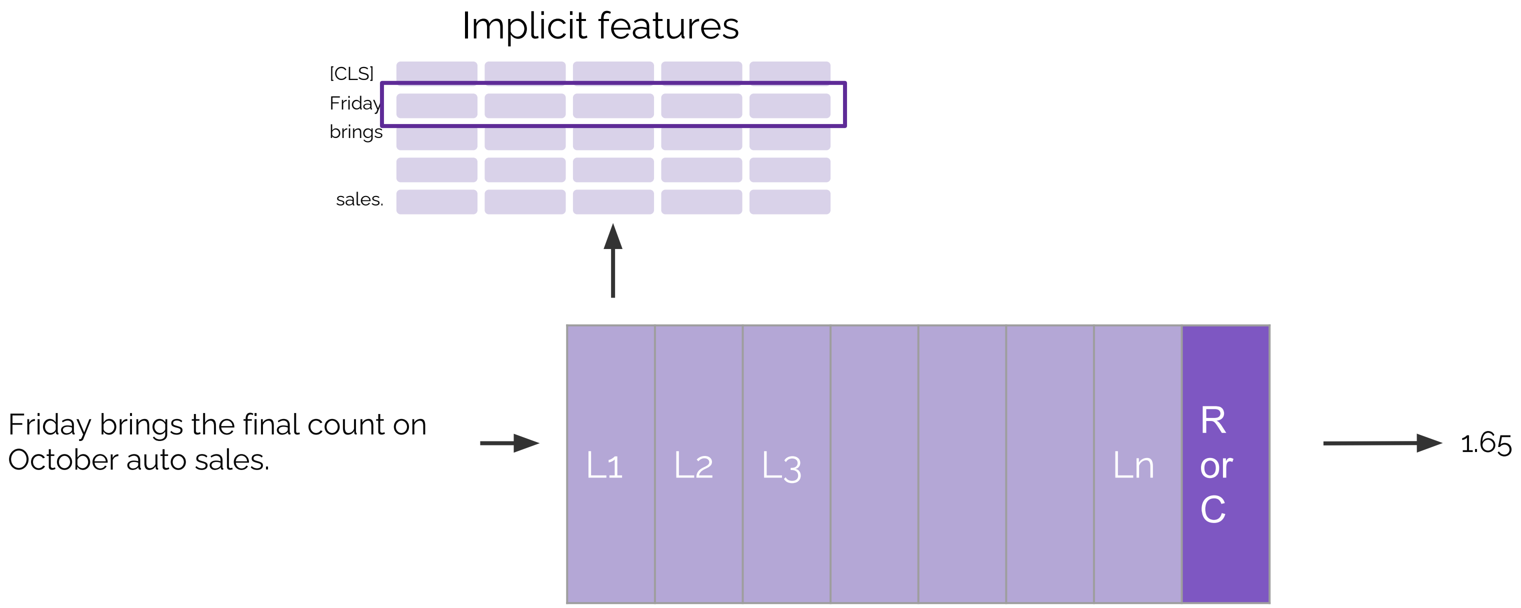
It allows the extraction of more than 120 features, spanning across different levels of linguistic description [...].

Examples of features: number of tokens, average number of character per token, adjectives distribution, nouns distributions, fraction of verbs at present tense, fraction of verb at past tense, average syntactic tree length.

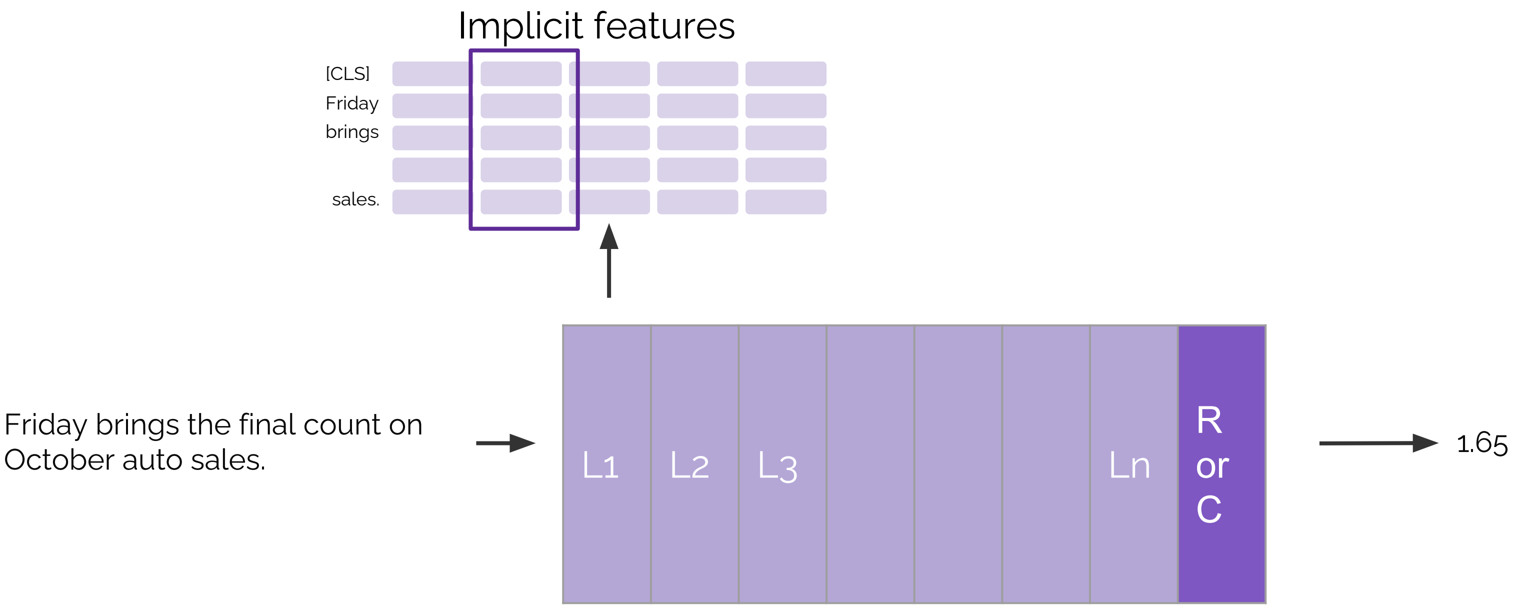
Implicit features extraction



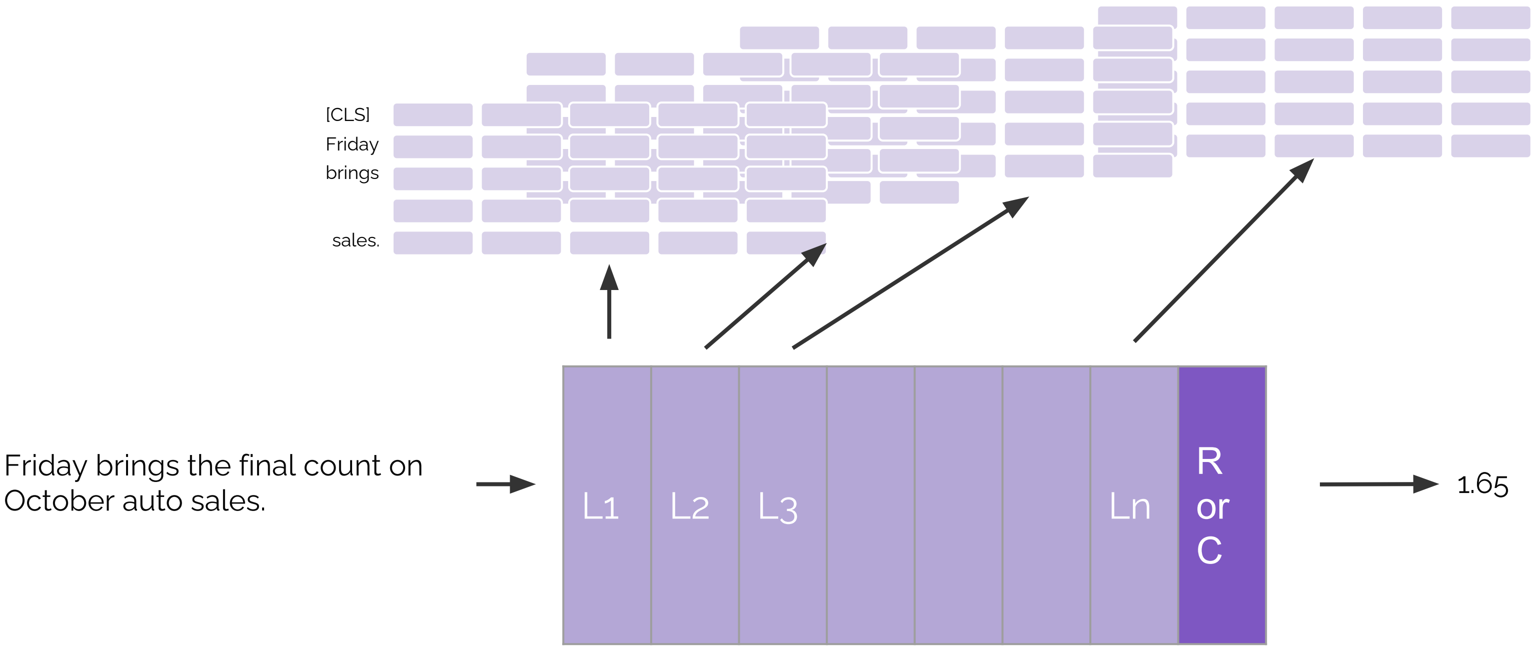
Implicit features extraction



Implicit features extraction



Implicit features extraction



Dataset: sentence complexity

The dataset used to assess the encoding of syntactic information is a dataset of 1200 sentences annotated with their **complexity**.

The task is structured as a **regression** on scores ranging from 1 (simple) to 6 (complex).

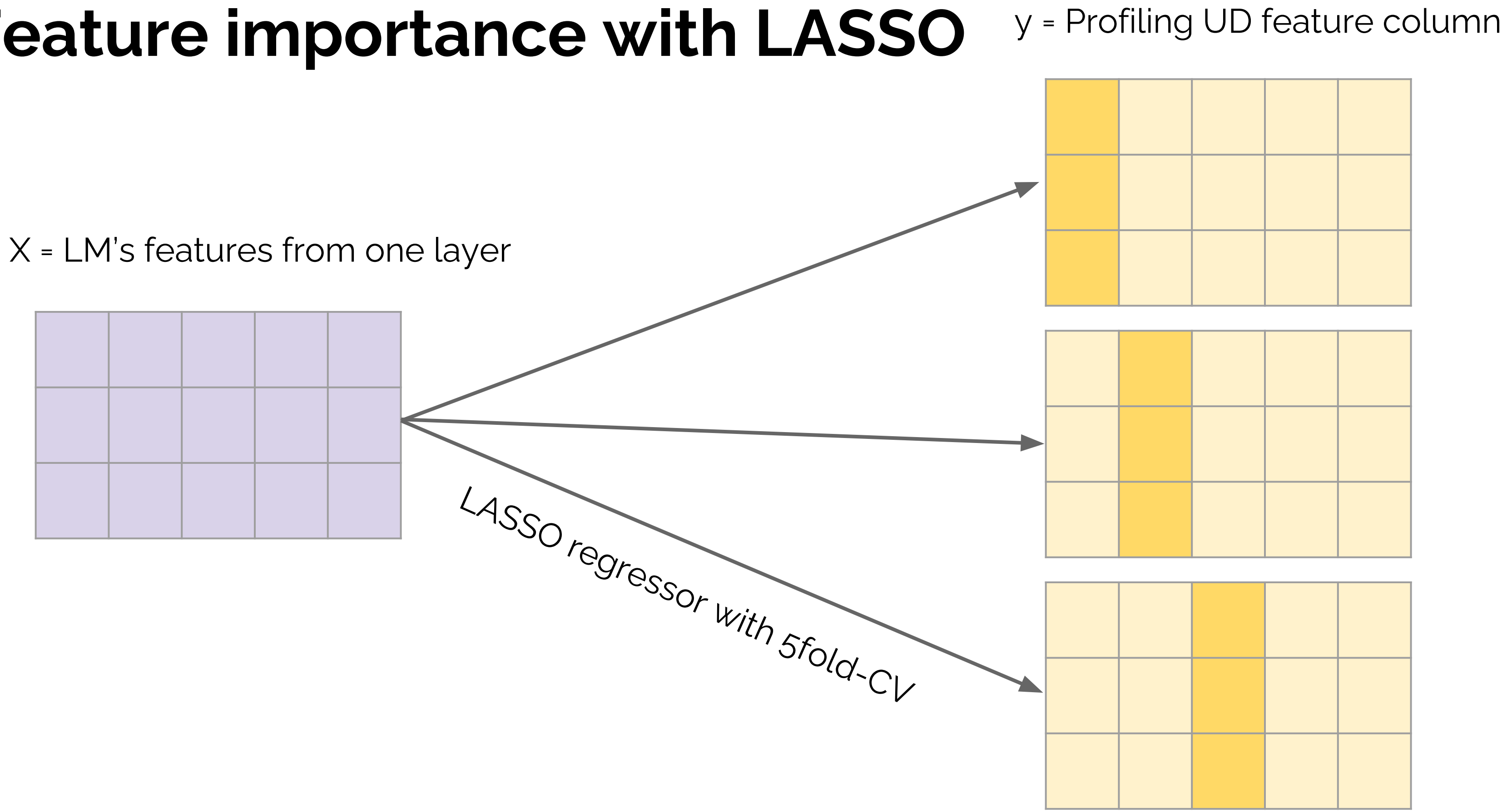
The complexity score is computed as the mean of perceived complexity among 10 human annotators.

TSNE on sentences represented using profiling features.

Starting with a tiny model:

bert-tiny (2 layers, 128 features)

Feature importance with LASSO

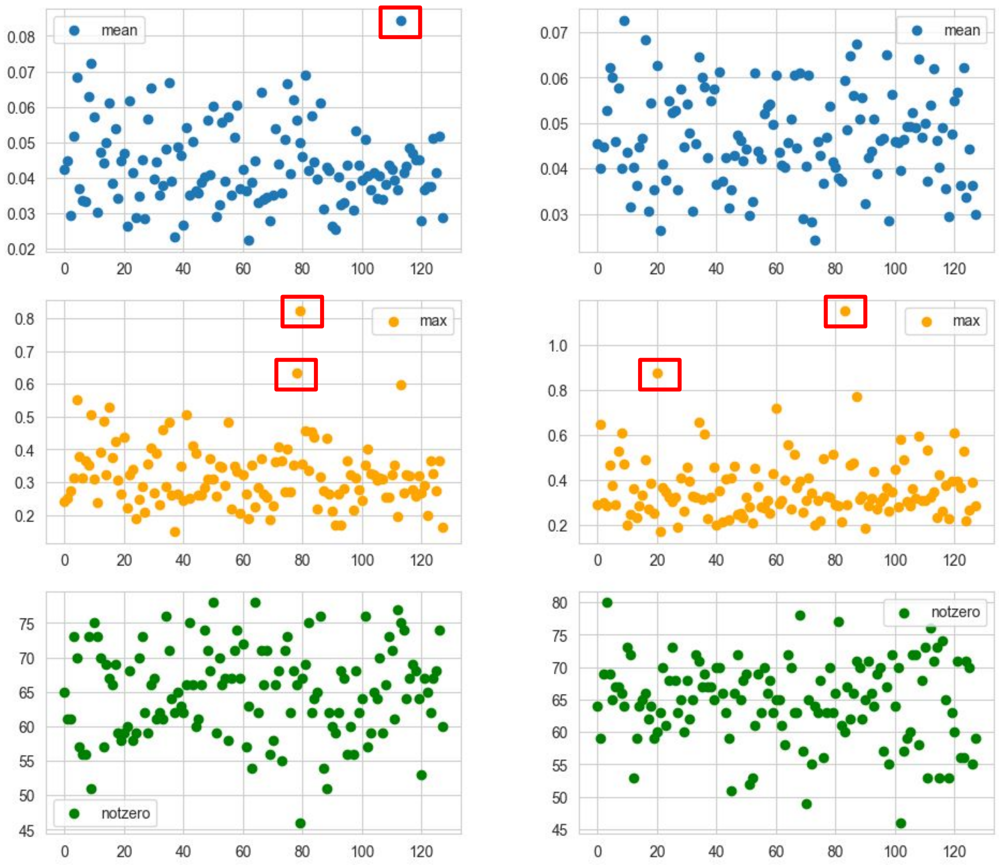


Finding most important features.

1. Aggregate the LASSO best coefficient among all profiling-features regressors.

2. Find outliers: points at least 3 std away from the mean

Layer 1 Layer 2



Performances removing / using only outliers

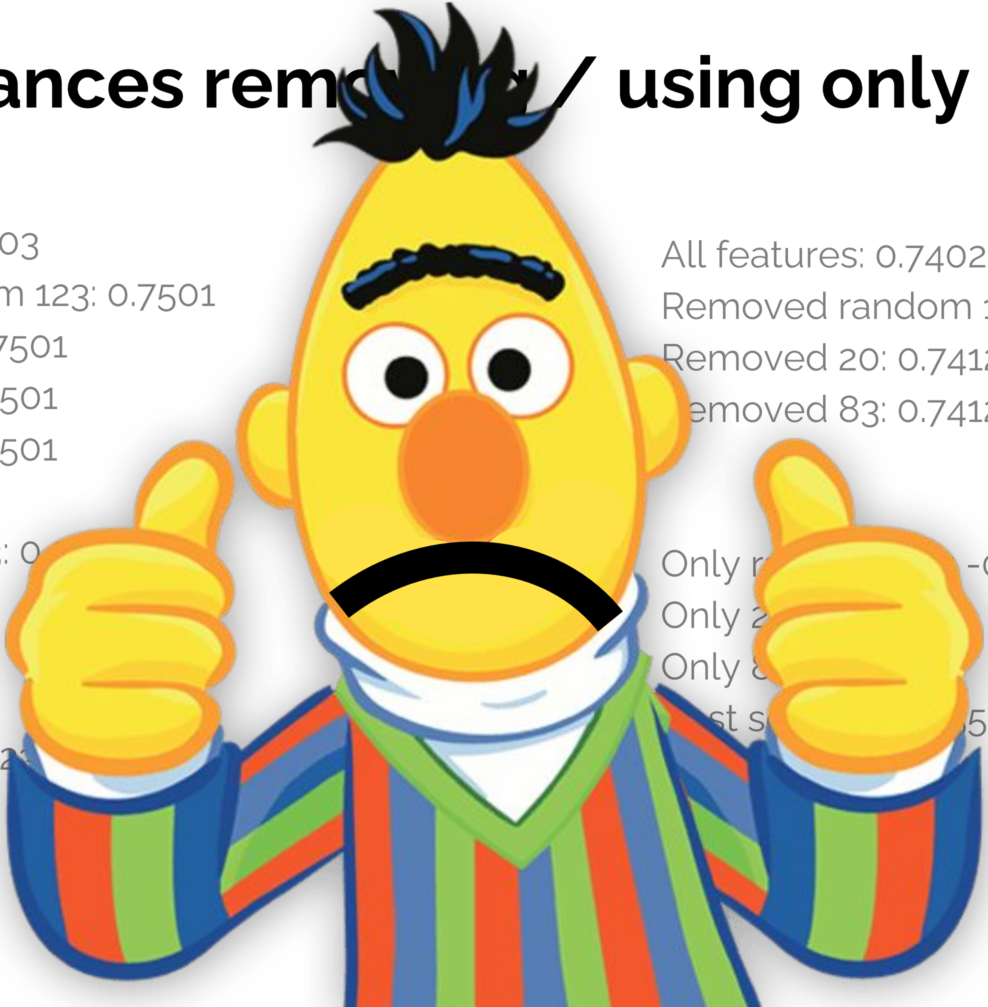
All features: 0.7503
Removed random 123: 0.7501
Removed 113: 0.7501
Removed 78: 0.7501
Removed 79: 0.7501

Only random 123: 0.0152
Only 113: 0.0440
Only 78: 0.0523
Only 79: -0.0113
Best score (only 23) = 0.4512

All features: 0.7402
Removed random 118: 0.7412
Removed 20: 0.7412
Removed 83: 0.7412

Only random 118: -0.0147
Only 20: 0.0201
Only 83: 0.0291
Best score (only 55) = 0.3338

Performances rem / using only outliers



All features: 0.7503
Removed random 123: 0.7501
Removed 113: 0.7501
Removed 78: 0.7501
Removed 79: 0.7501

Only random 123: 0
Only 113: 0.0440
Only 78: 0.0523
Only 79: -0.0113
Best score (only 2

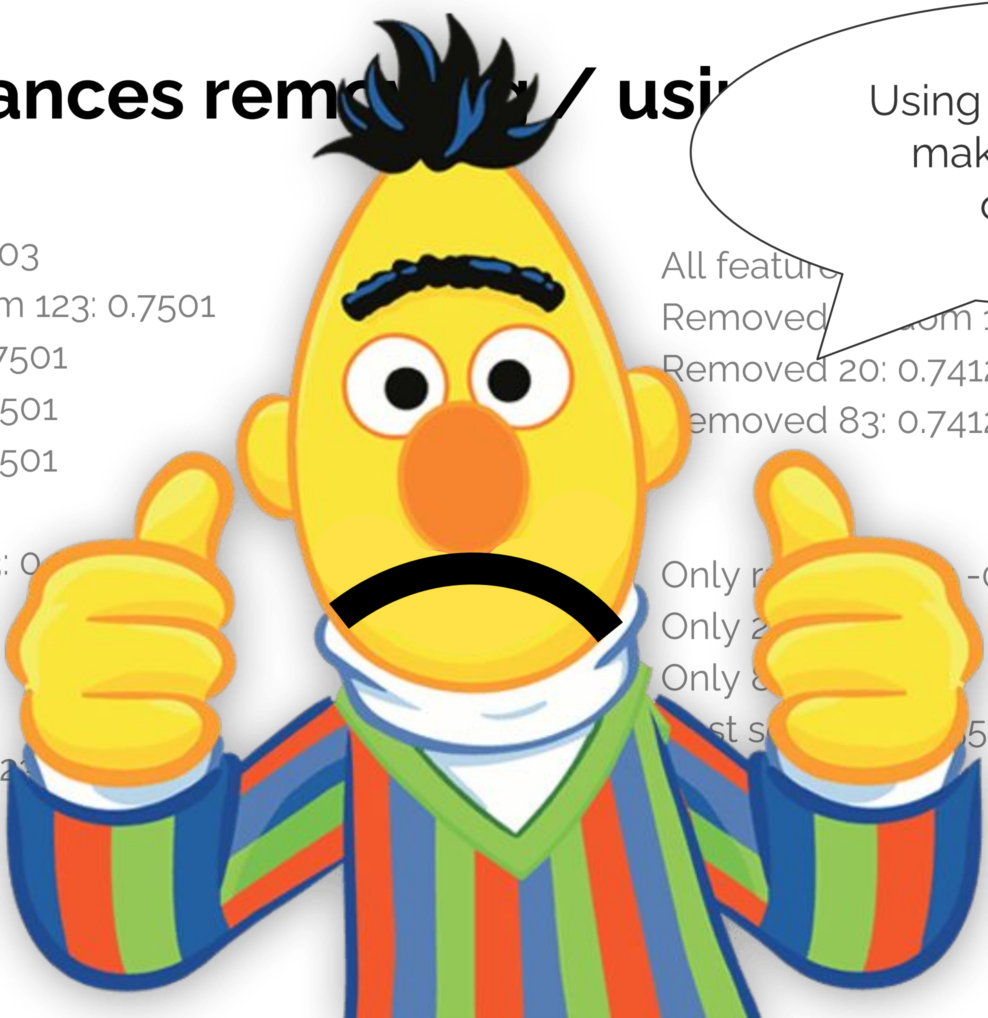
All features: 0.7402
Removed random 118: 0.7412
Removed 20: 0.7412
Removed 83: 0.7412

Only r -0.0147
Only 2
Only 8
st s 5) = 0.3338

Performances rem / usi

All features: 0.7503
Removed random 123: 0.7501
Removed 113: 0.7501
Removed 78: 0.7501
Removed 79: 0.7501

Only random 123: 0
Only 113: 0.0440
Only 78: 0.0523
Only 79: -0.0113
Best score (only 2

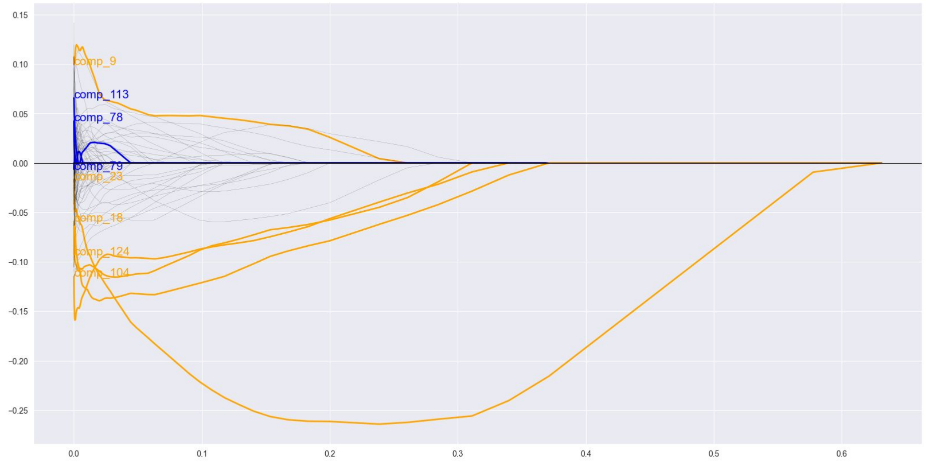


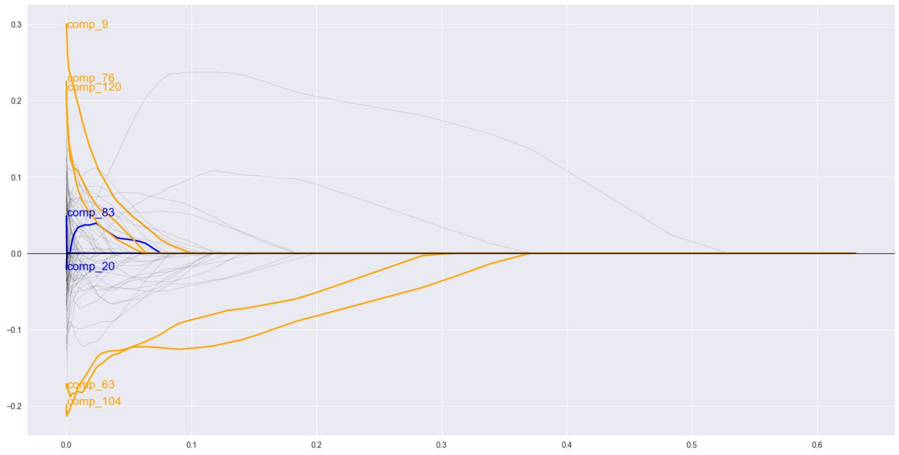
Using BSS algorithms makes little or no difference

All featu
Removed dom 118: 0.7412
Removed 20: 0.7412
emoved 83: 0.7412

Only r -0.0147
Only 2
Only 8
st s 5) = 0.3338

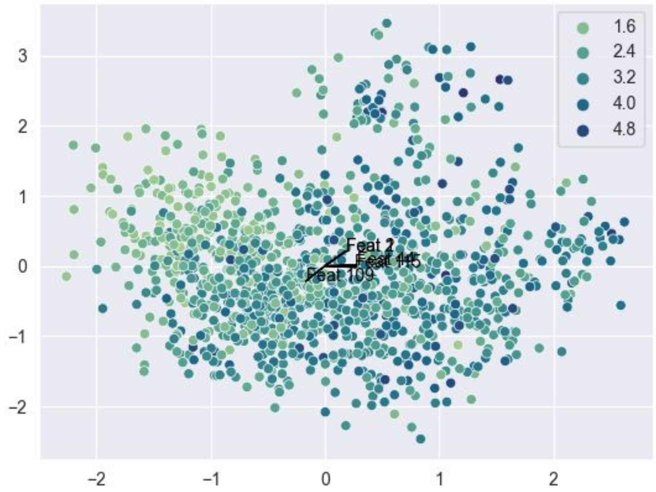
Comparison with Lasso coefficient calculated directly on the target task.





Overcoming collinearity

PCA with whitening



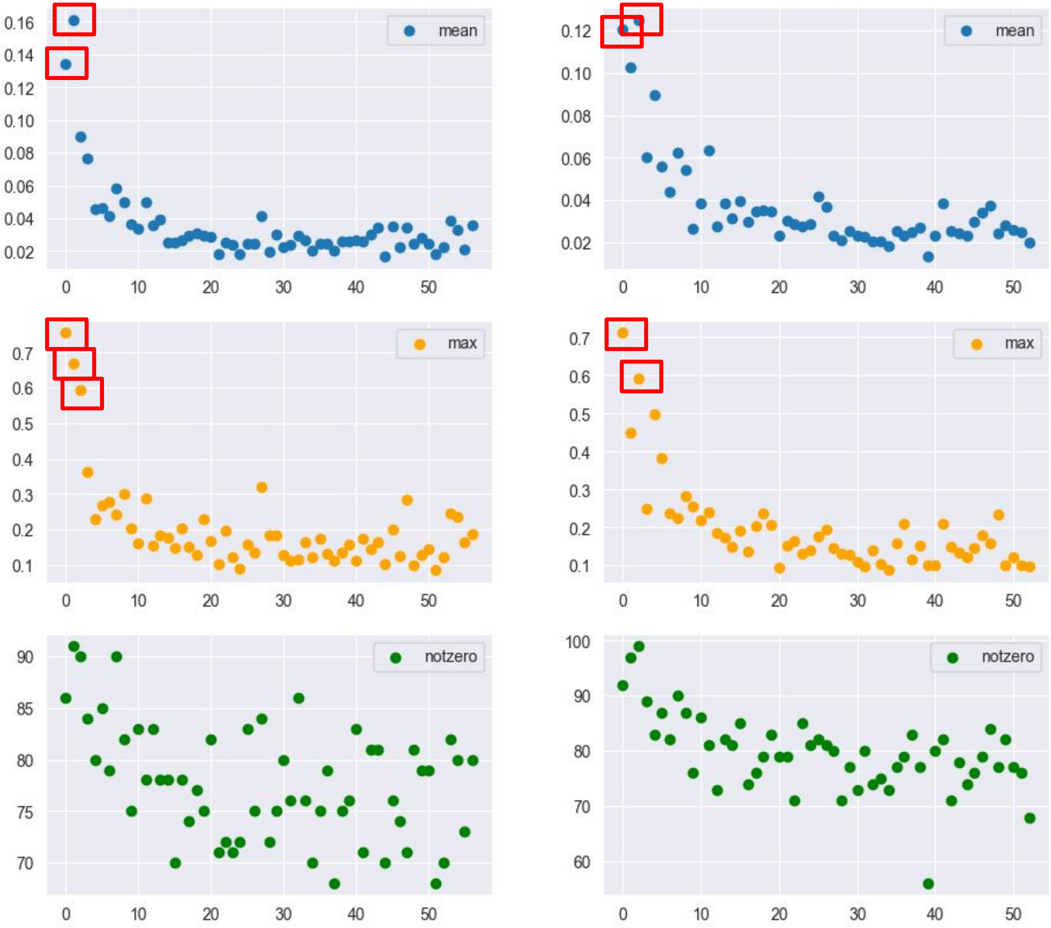
lm_feature_24: 0.1926
lm_feature_10: -0.1919
lm_feature_82: -0.1871
lm_feature_64: 0.1799
lm_feature_14: 0.1750

lm_feature_54: 0.2199
lm_feature_6: 0.2058
lm_feature_80: 0.1939
lm_feature_56: -0.1859
lm_feature_9: 0.1816

lm_feature_45: 0.2773
lm_feature_113: 0.2627
lm_feature_116: 0.2624
lm_feature_114: 0.2118
lm_feature_20: 0.1957

lm_feature_110: -0.217
lm_feature_61: 0.2141
lm_feature_3: 0.2129
lm_feature_2: 0.2129
lm_feature_8: -0.2085

Using LASSO again to find "outlier" features.



Performances removing top components

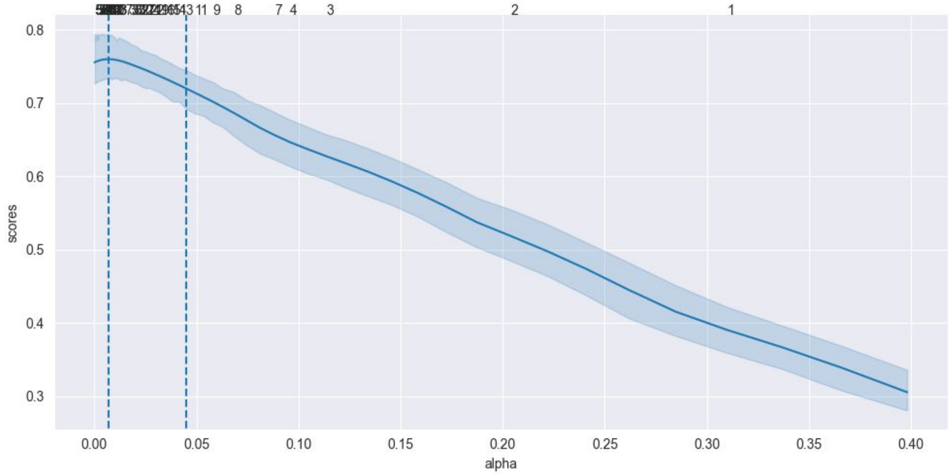
All features: 0.7552
Removed random 19: 0.7547
Removed 0: 0.1424
Removed 1: 0.7001
Removed 2: 0.7365

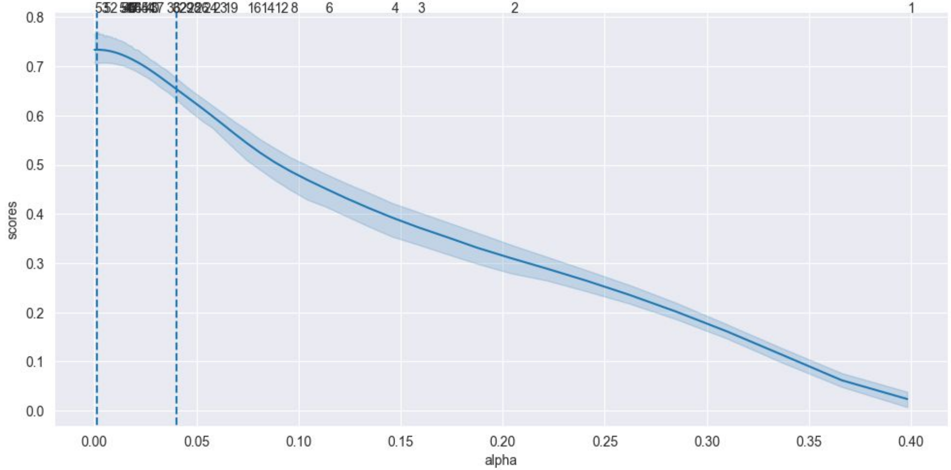
Only random 19: -0.0153
Only 0: 0.5203
Only 1: 0.0285
Only 2: 0.0004
Best score (only 0): 0.5203

All features: 0.7335
Removed random 46: 0.7334
Removed 0: 0.5284
Removed 2: 0.7193

Only random 46: -0.0150
Only 0: 0.1694
Only 2: -0.0025
Best score (only 4): 0.2367

Lasso regressor performances and number of features, varying alpha.





Studying bigger models

Bert-mini (4 layers, 256 features)

Bert-medium (8 layers, 512 feature

What happens with more layers (Bert-mini)

Layer 1

All features: 0.7195
Best score (only 172): 0.5328

PCA (114)
All features: 0.7607
Removed random 22: 0.7593
Removed 0: 0.2567
Removed 1: 0.3649
Removed 2: 0.7590

Only random 22: -0.0155
Only 0: 0.3846
Only 1: 0.2854
Only 2: -0.0155

Layer 2

All features: 0.7310
Best score (only 190): 0.5118

PCA (114)
All features: 0.7698
Removed random 9: 0.7705
Removed 0: 0.3287
Removed 1: 0.3455
Removed 2: 0.7704

Only random 9: -0.0157
Only 0: 0.3348
Only 1: 0.3061
Only 2: -0.0158

Layer 3

All features: 0.7321
Best score (only 190): 0.4429

PCA (111)
All features: 0.7715
Removed random 56: 0.7718
Removed 0: 0.5845
Removed 1: 0.5097
Removed 2: 0.5276

Only random 56: -0.0152
Only 0: 0.1387
Only 1: 0.1859
Only 2: 0.1692

Layer 4

All features: 0.7355
Best score (only 139): 0.1841

PCA (103)
All features: 0.7538
Removed random 38: 0.7545
Removed 0: 0.6148
Removed 1: 0.6939
Removed 2: 0.7106
Removed 4: 0.6086

Only random 38: -0.0185
Only 0: 0.1000
Only 1: 0.0301
Only 2: 0.0133
Only 4: 0.0921

Correlation between the two set of features drastically decreases!

And even more layers (Bert-medium)

Layer 1

All features: 0.5305
Best score (only 400):0.4450

PCA (228)
All features: 0.7212
Removed random 209: 0.7211
Removed 0: 0.2632
Removed 1: 0.0070
Removed 2: 0.7221

Only random 209: -0.0172
Only 0: 0.2607
Only 1: 0.3568
Only 2: -0.0171

Layer 2

All features: 0.5435
Best score (only 270): 0.3740

PCA (221)
All features: 0.7281
Removed random 15: 0.7142
Removed 0: 0.4323
Removed 1: -0.0647
Removed 2: 0.7300

Only random 15: -0.0091
Only 0: 0.1848
Only 1: 0.4165
Only 2: -0.017

Layer 3

All features: 0.5565
Best score (only 270): 0.3969

PCA (220)
All features: 0.7422
Removed random 102: 0.7411
Removed 0: 0.3985
Removed 1: 0.0664
Removed 2: 0.6983

Only random 102: -0.0154
Only 0: 0.2049
Only 1: 0.3507
Only 2: 0.0127

Layer 4

All features: 0.5540
Best score (only 270): 0.4108

PCA (222)
All features: 0.7394
Removed random 197: 0.7398
Removed 0: 0.3955
Removed 1: 0.0843
Removed 2: 0.6572

Only random 197: -0.0159
Only 0: 0.2033
Only 1: 0.3368
Only 2: 0.0384

… and more layers

Layer 5

All features: 0.5655
Best score (only 427):0.6030

PCA (226)
All features: 0.7418
Removed random 64: 0.7367
Removed 0: 0.4544
Removed 1: 0.1567
Removed 2: 0.5778

Only random 64: -0.0146
Only 0: 0.1768
Only 1: 0.2959
Only 2: 0.0823

Layer 6

All features: 0.5959
Best score (only 427): 0.5994

PCA (114)
All features: 0.7417
Removed random 128: 0.7420
Removed 0: 0.4389
Removed 1: 0.3774
Removed 2: 0.4224

Only random 128: -0.0161
Only 0: 0.1797
Only 1: 0.1806
Only 2: 0.1664

Layer 7

All features: 0.5735
Best score (only 427): 0.4952

PCA (214)
All features: 0.7315
Removed random 69: 0.7317
Removed 0: 0.5694
Removed 1: 0.48169
Removed 2: 0.47295
Removed 3: 0.46717

Only random 69: -0.0140
Only 0: 0.1014
Only 1: 0.1244
Only 2: 0.1371
Only 3: 0.1514

Layer 8

All features: 0.5263
Best score (only 427): 0.3972

PCA (195)
All features: 0.7379
Removed random 112: 0.7387
Removed 0: 0.5977
Removed 1: 0.6650
Removed 2: 0.6923
Removed 4: 0.6364
Removed 6: 0.4544

Only random 112: -0.0168
Only 0: 0.0868
Only 1: 0.0229
Only 2: 0.0093
Only 4: 0.0562
Only 6: 0.1775

… and more layers

Layer 5

All features: 0.5655
Best score (only 427):0.6030

PCA (226)
All features: 0.7418

Layer 6

All features: 0.5959
Best score (only 427): 0.5994

PCA (114)
All features: 0.7417

Layer 7

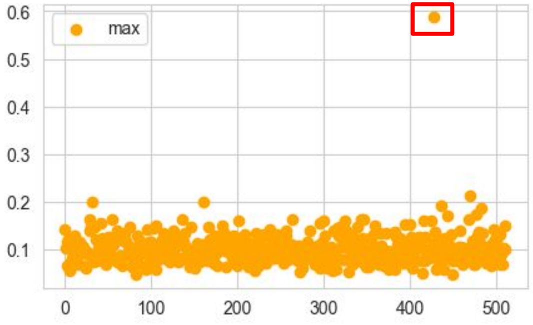
All features: 0.5735
Best score (only 427): 0.4952

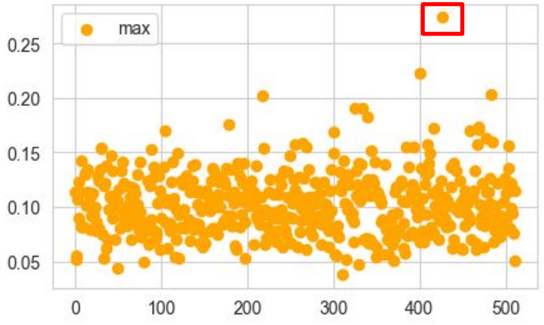
PCA (214)
All features: 0.7315

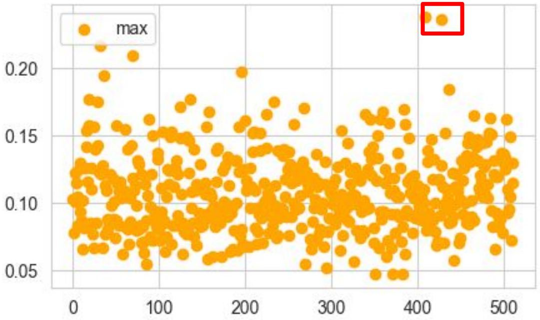
Layer 8

All features: 0.5263
Best score (only 427): 0.3972

PCA (195)
All features: 0.7379







Only 4: 0.0562
Only 6: 0.1775

Thanks for your attention!

References

- **Syntactic features extraction**: Profiling-UD: a Tool for Linguistic Profiling of Texts” Brunato D., Cimino A., Dell’Orletta F., Montemagni S., Venturi G. (2020).
- **Original BERT**: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. Devlin J., Chang M., Lee K., Toutanova K. (2019)
- **Smaller BERTs**: “Well-Read Students Learn Better: On the Importance of Pre-training Compact Models”. Turc I., Chang M., Lee K., Toutanova K. (2019)
- **Probings on BERT**: “Linguistic Profiling of a Neural Language Model”. Miaschi A., Brunato D., Dell’Orletta F, Venturi G. (2020)
- **LASSO for most important BERT’s features**: “How Do BERT Embeddings Organize Linguistic Knowledge?”. Puccetti G., Miaschi A., Dell’Orletta F. (2020)
- **Finding outliers within BERT’s features**: “Outliers dimensions that disrupt transformers are driven by frequency”. Puccetti G., Rogers A., Drozd A., Dell'Orletta F. (2022)