PAWEŁ GOLIK, MATEUSZ JASTRZĘBIOWSKI,
ALEKSANDRA MUSZKOWSKA

NLP course: Project 2 Final presentation

Probing tasks for E-commerce product matching **embeddings**




GARMIN Fenix 7X Solar Czarny (100254101)

Historia cen

Smartwatche Męskie Garmin z Wi-Fi, z przekątną ekranu 1,4 cala. Opaska współpracująca z systemem Android, iOS. Oferuje takie funkcje jak: GPS, Wodoodporność. Zawiera: Pulsometr. Rodzaj paska: Silikonowy

4,8 22 opinie

3 299,00 zł

KUP TERAZ

Z wysyłką od 3308,00 zł

dostępny

WYS VAT 23%

PRODUCT

Oferty (32) | Informacje o prod | i Recenzje (28) | Zadaj pytanie | Kup lokalnie

Najlepsze oferty wybrane na

Sortuj od **najlepszych ofert** | Filtry

opinie	Fenix 7X Solar Czarny z czarnym paskiem + RATY 0%, POLSKA DYSTRYBUCJA, 3 LATA GWARANCJI, DARMOWA DOSTAWA Warianty (2)	3 349,00 zł ARMOWA WYSYŁKA dostępny	KUP TERAZ IDŹ DO SKLEPU
97 opinii	**Zegarek Garmin Fenix 7X Solar niebieskoszary z czarnym paskiem • RATY 0% • DOSTAWA GRATIS (paczkomat lub kurier UPS)** Warianty (2)	3 349,00 zł ARMOWA WYSYŁKA dostępny	KUP TERAZ IDŹ DO SKLEPU
2 opinie	Smartwatch GARMIN Fenix 7X Solar Czarny z czarnym paskiem 010-02541-01 ® KUP TERAZ **RATY 0% I 6 M-CY NIE PŁACISZ!**	3 349,00 zł ARMOWA WYSYŁKA dostępny	IDŹ DO SKLEPU

OFFERS

GARMIN Fenix 7X Solar Czarny (100254101) - Pozostałe oferty

1 opinii	Zegarek sportowy Garmin Fenix 7X Czarny (010-02541-01) **DARMOWA DOSTAWA JUŻ OD 399 ZŁ**	3 299,00 zł DARMOWA WYSYŁKA dostępny	IDŹ DO SKLEPU

The WDC Dataset

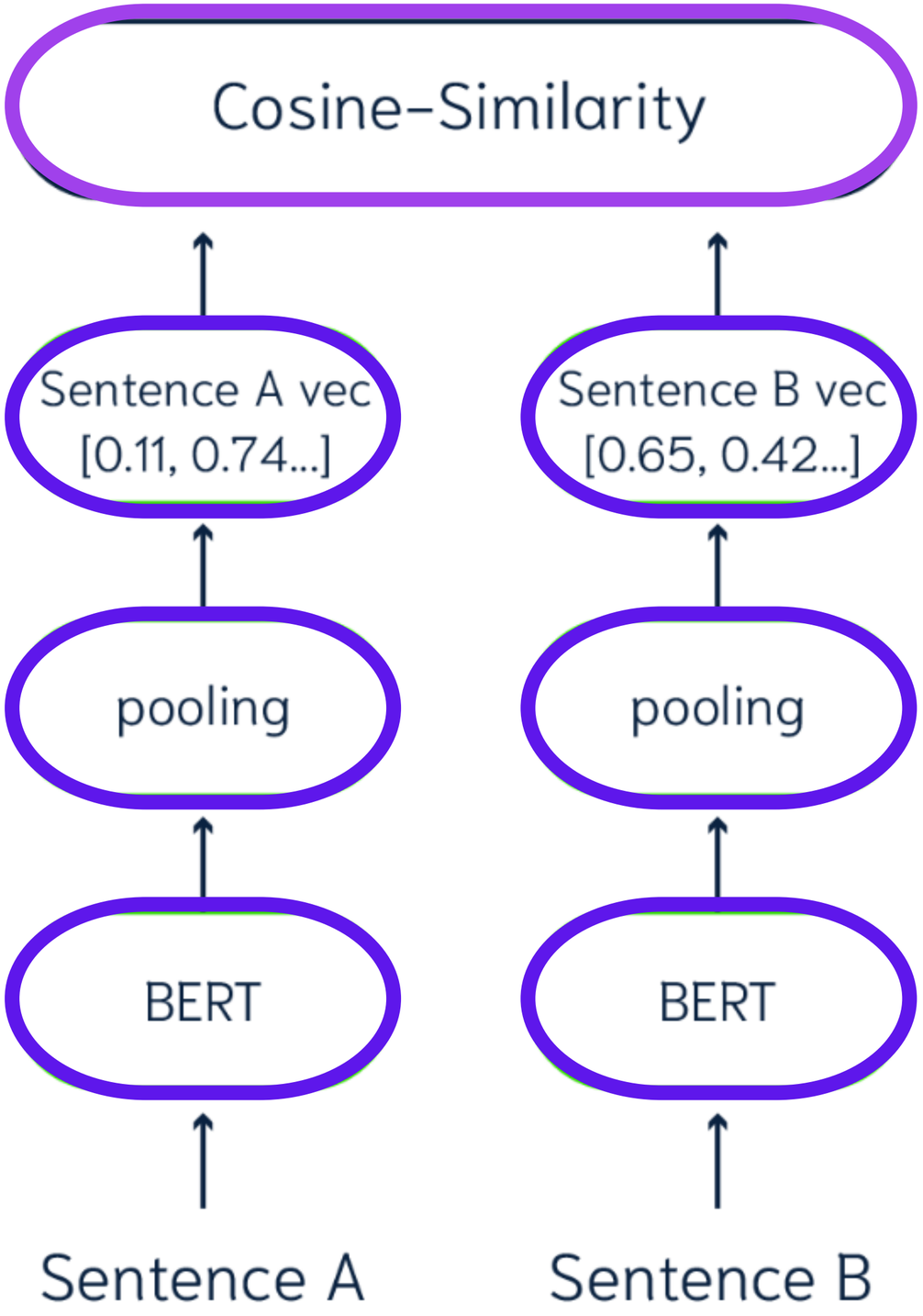
Offer A		Offer B		Target (=isProdcut TheSame)
TitleA1 "BrandX Camera MX140"	DescA1 "..."	TitleB1 "Camera BrandX MX 140"	DescB1 "..."	
TitleA2 "BrandY S1000 Digital Camera"	DescA2 "..."	TitleB2 "BrandZ Camera PRO 3F"	DescB2 "..."	

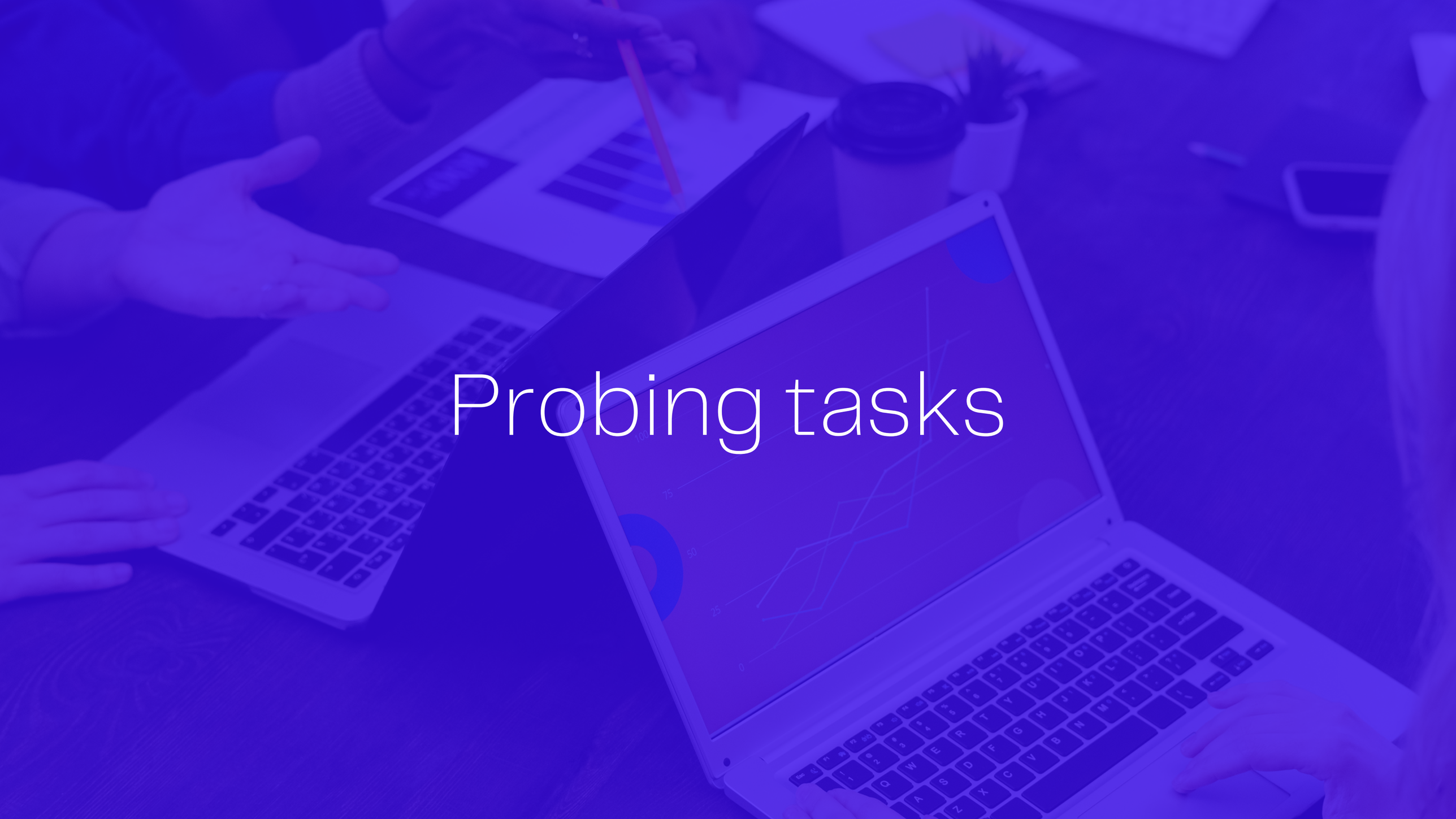
Embeddings

Bi-encoders

[Mazare et al. 2018]
Training millions of personalized dialogue agents.

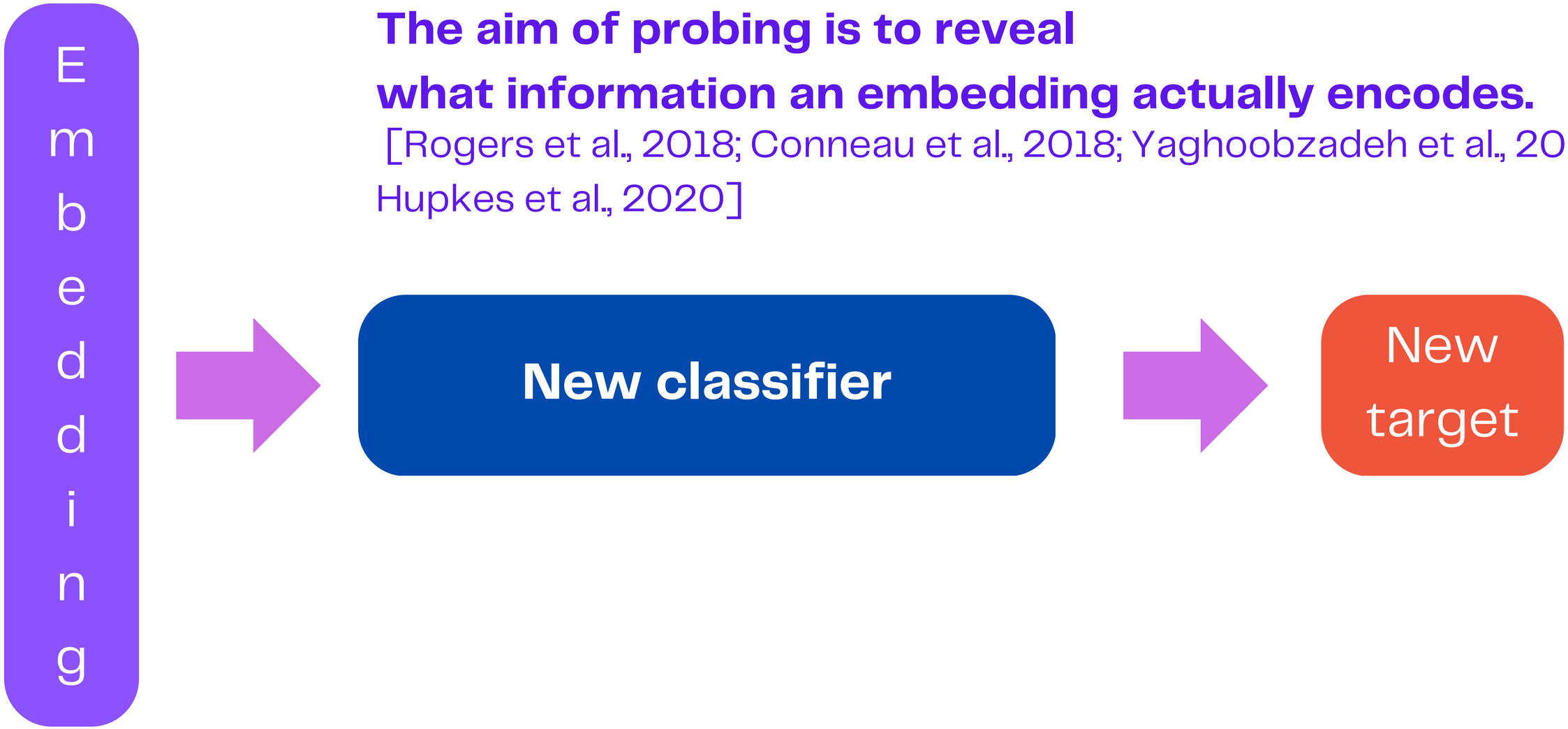
– less accurate
+ easy to get the embeddings
+ sentence embeddings – an integral part of the architecture
+ faster





Probing tasks

How probing works?



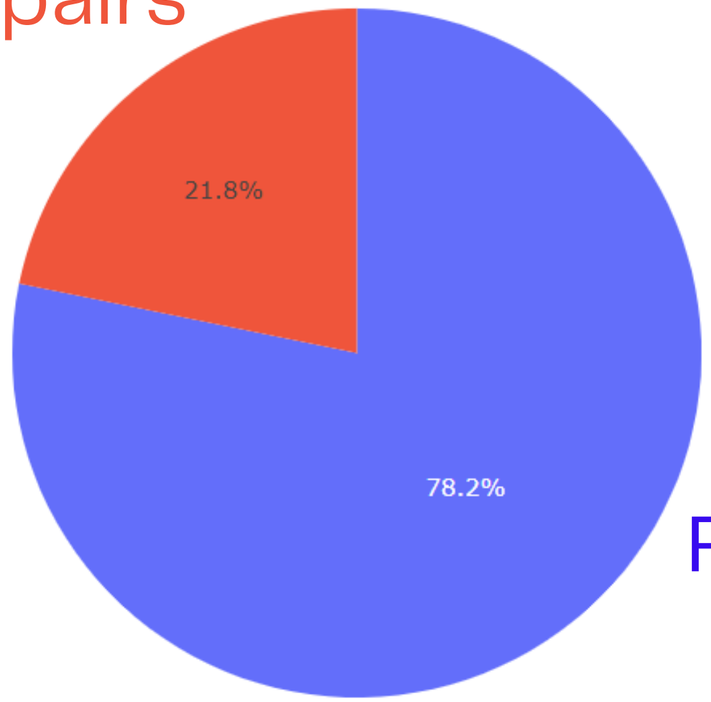
The aim of probing is to reveal
what information an embedding actually encodes.
[Rogers et al., 2018; Conneau et al., 2018; Yaghoobzadeh et al., 2019; Hupkes et al., 2020]

What is new in Project #2?

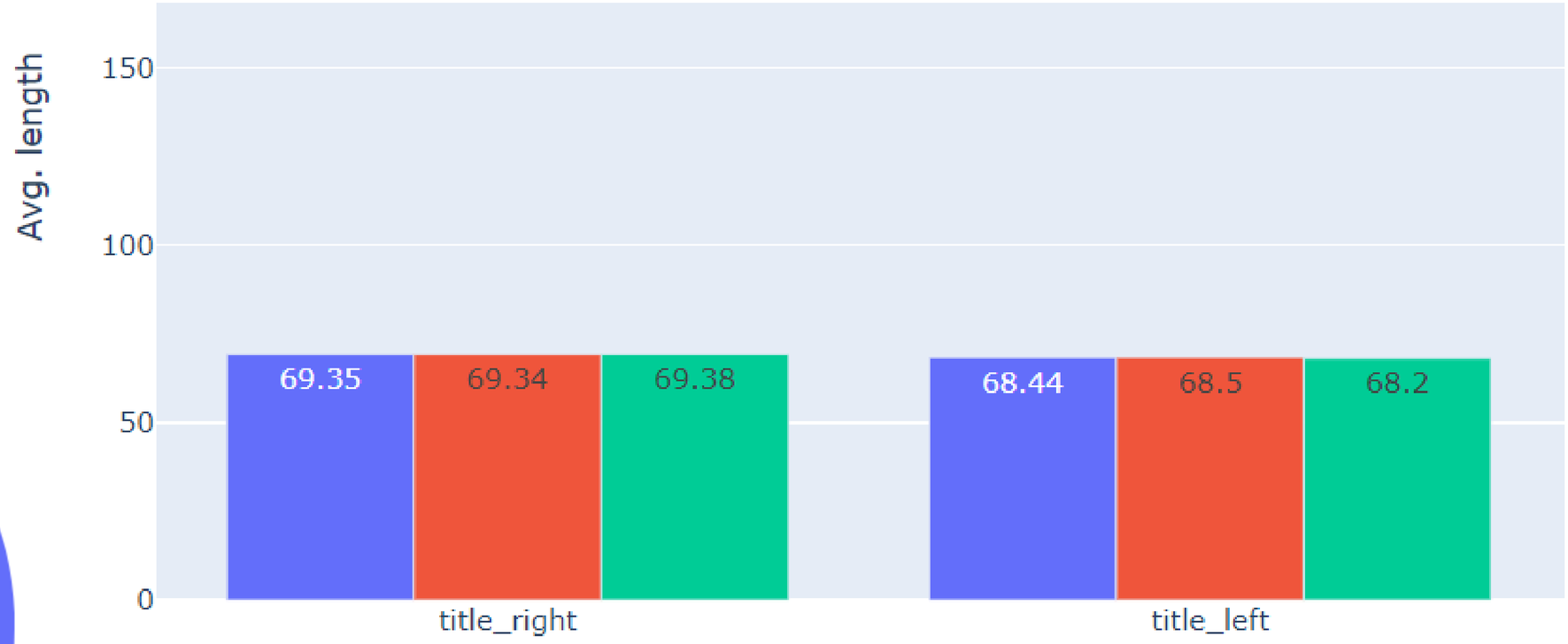
1. "Natural" dataset – a reference point
2. New probing tasks
3. Probing after and before fine-tunning
4. "Computers" category
5. Investigating SentEval

Number of pairs: 6332 (p) + 1762(n) = 8094

Negative pairs

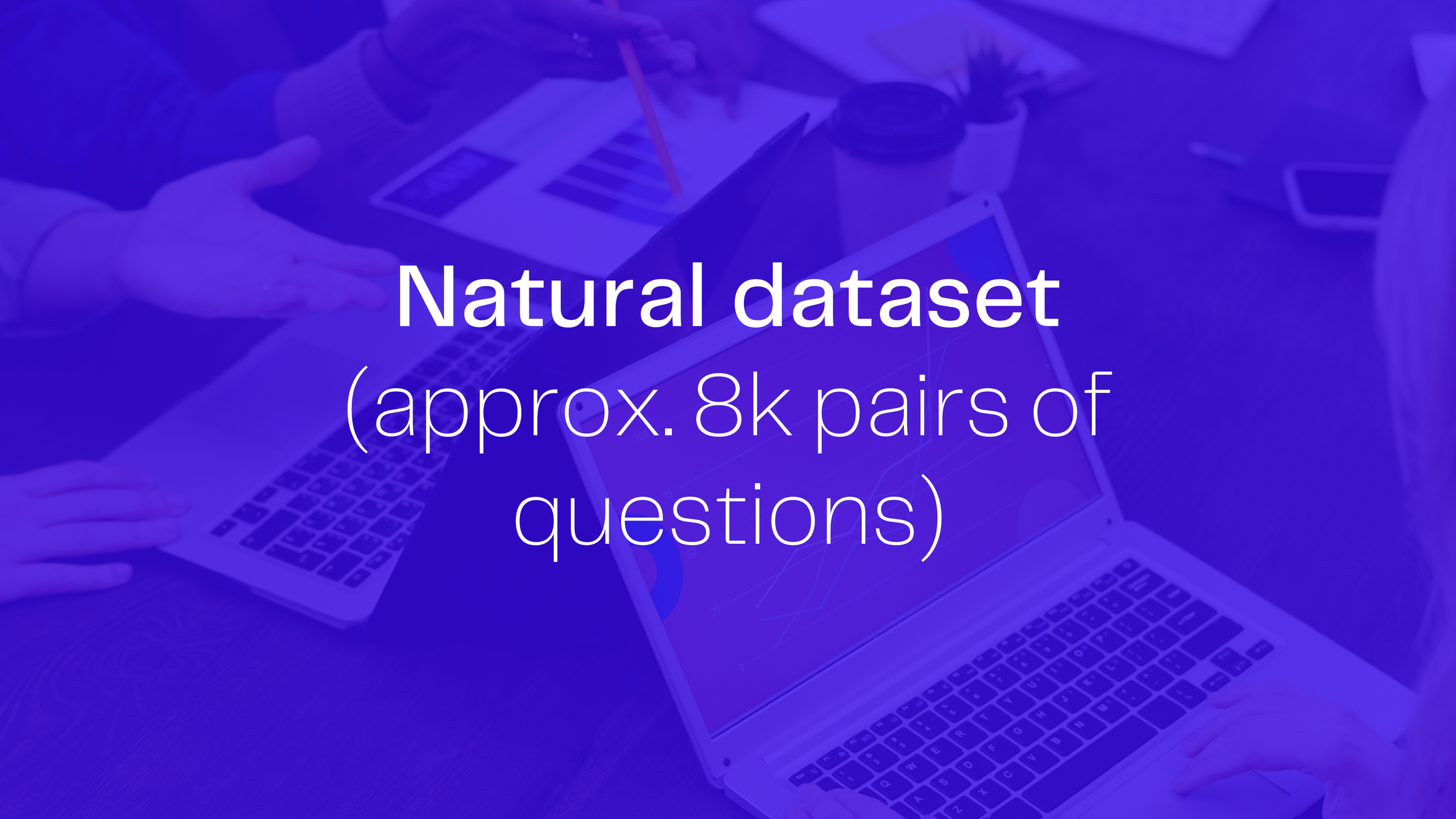


Positive pairs



All, **positive**, **negative**

Computers



Natural dataset

(approx. 8k pairs of questions)

"What are good websites for escorts?"
"How do I find a good escort?"
0

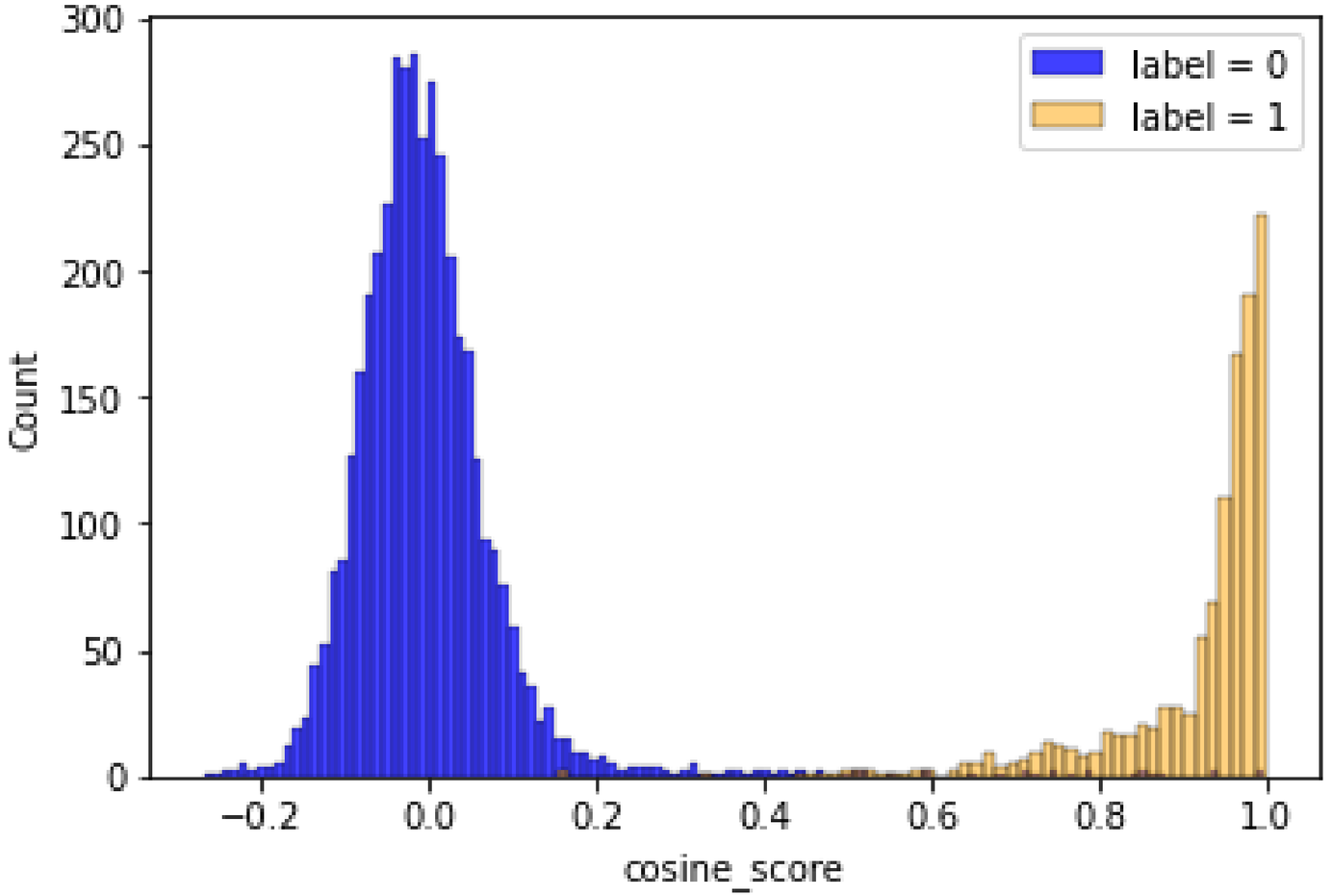
"How do I use Twitter as a business source?"
"How can I use Twitter for business?"
1

Python libraries

Hugging Face

Creating embeddings

Accuracy: 86.45%



MODEL

xlm-roberta-base

fine-tune epochs: 80

batch_size: 16

DATASET

WDC

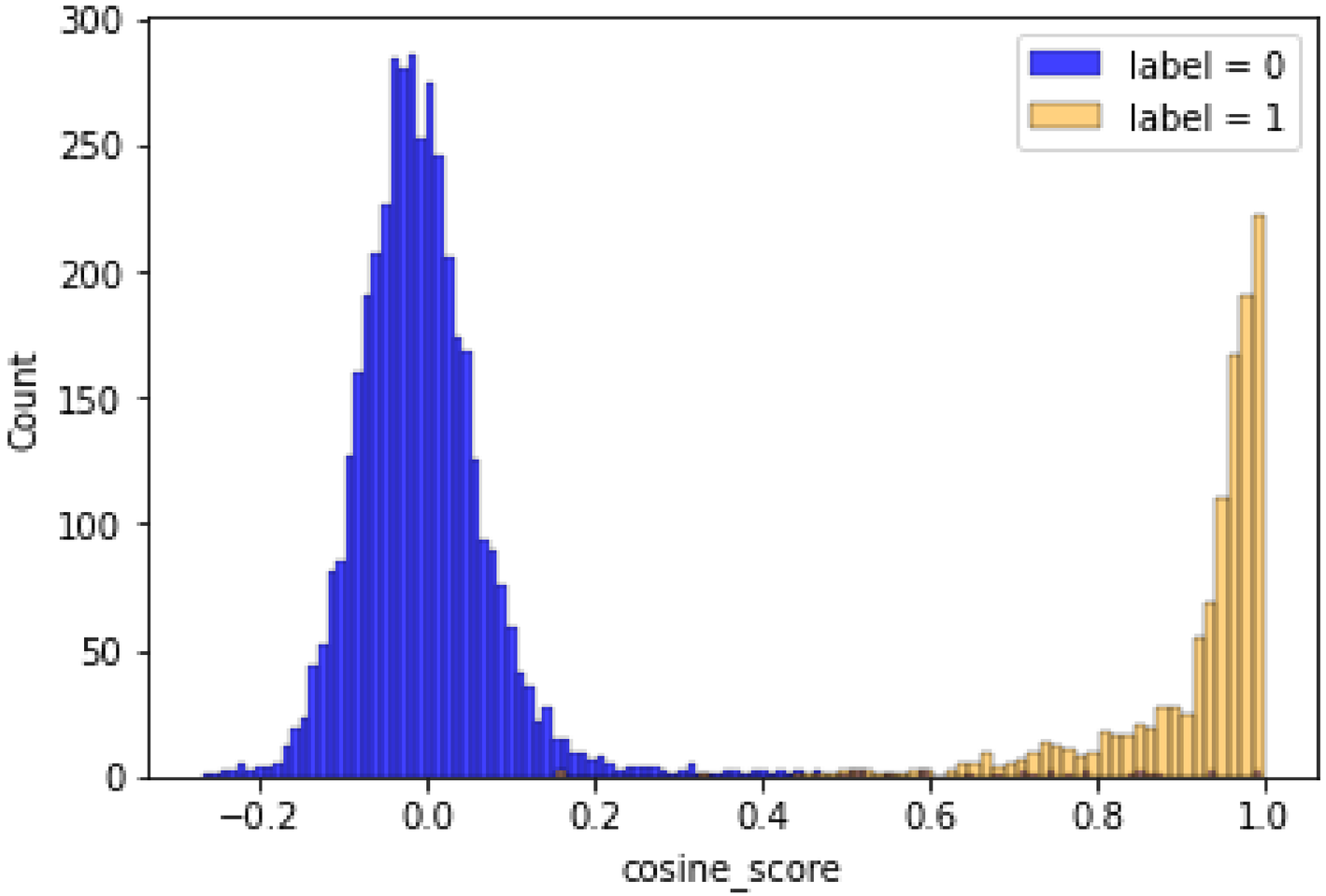
category: "Computers"

size: "medium"

features: "title" only

Creating embeddings

Accuracy: 79.34%



MODEL

xlm-roberta-base

fine-tune epochs: 20

batch_size: 16

DATASET

Natural dataset –

Quora Question Pairs

PROBING TASKS
RESULTS

Dataset

COMPUTERS

A. Pretrained Model

B. Fine-Tuned Model

Probing Tasks

COMMON WORDS

the presence of **common words** in the **title**

BRAND NAME

the presence of the **brand name in the title**

SIMILARITY MEASURES

predicting **similar(sentenceA, sentenceB)**

LENGTH OF A SENTENCE

predicting the **length** of the input sentence

1 Probing task – **common words**

COMMON WORDS ©

Influence of common words on embeddings

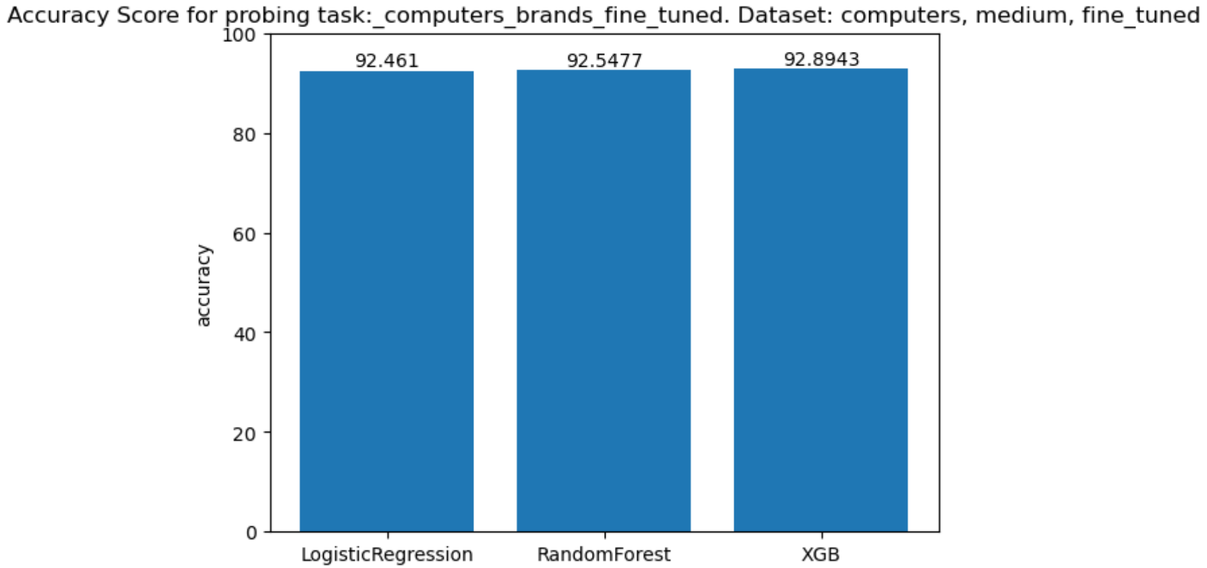
- one of common words: **['computer', 'laptop', 'processor', 'gpu', 'cpu', 'hdd', 'ssd', 'memory']**

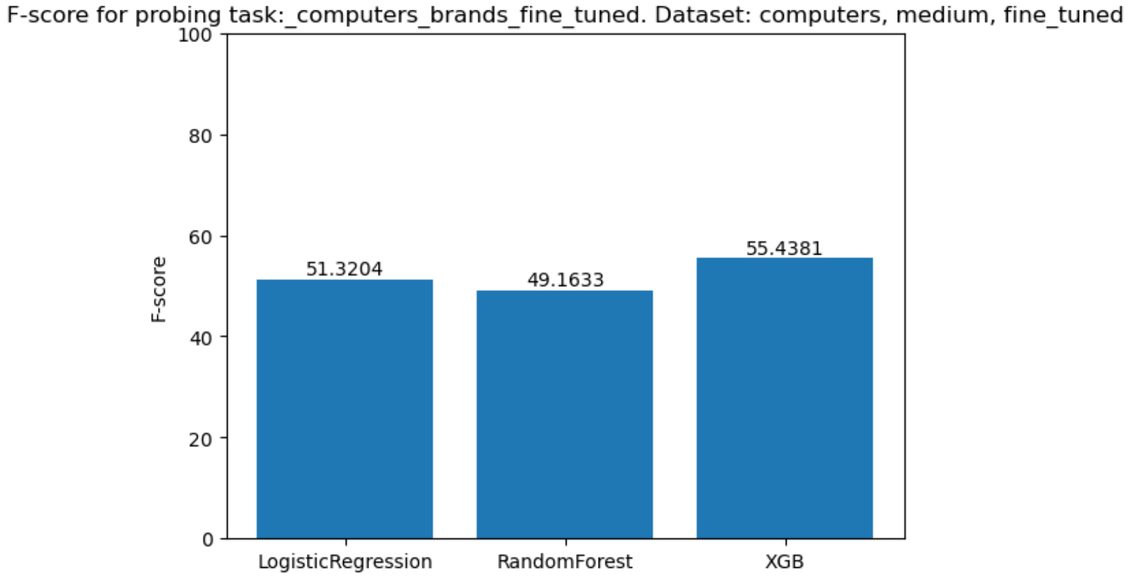
Classes

without common words	2701
with comon words	1145

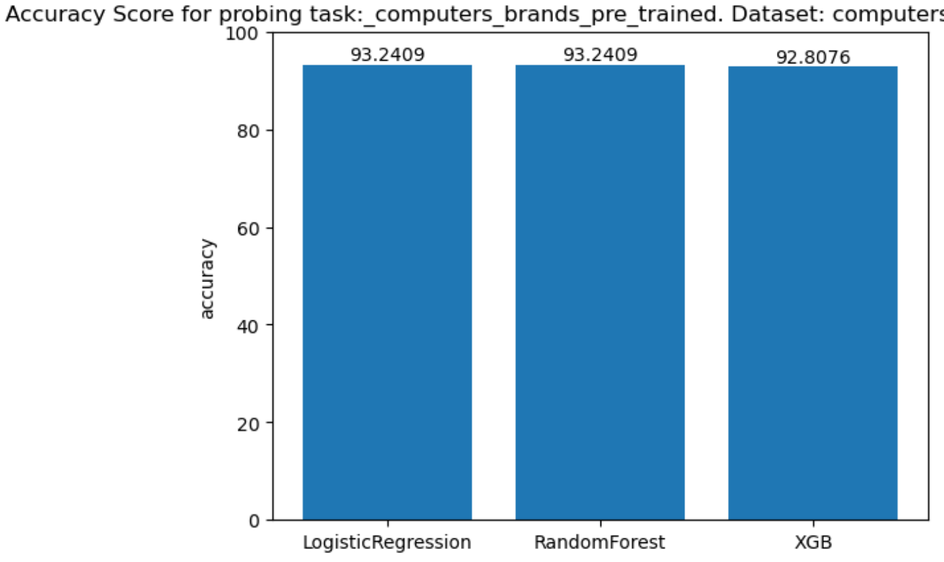
Results COMMON WORDS ©

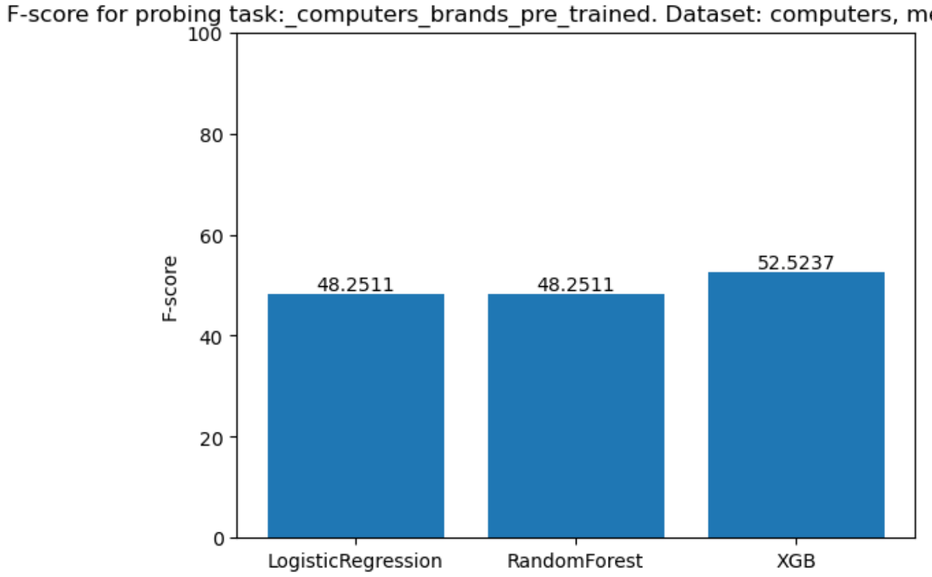
FINE TUNED





PRE TREINED





2 Probing task – **brand name**

BRAND NAME

New dataset

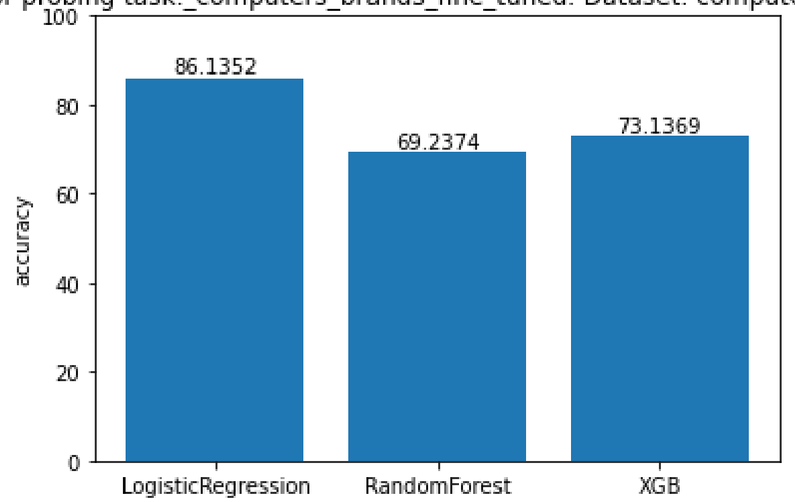
- without a brand name (HP, Samsung, ...) in title
- balanced classes

without brand name	1926
with brand name	1920

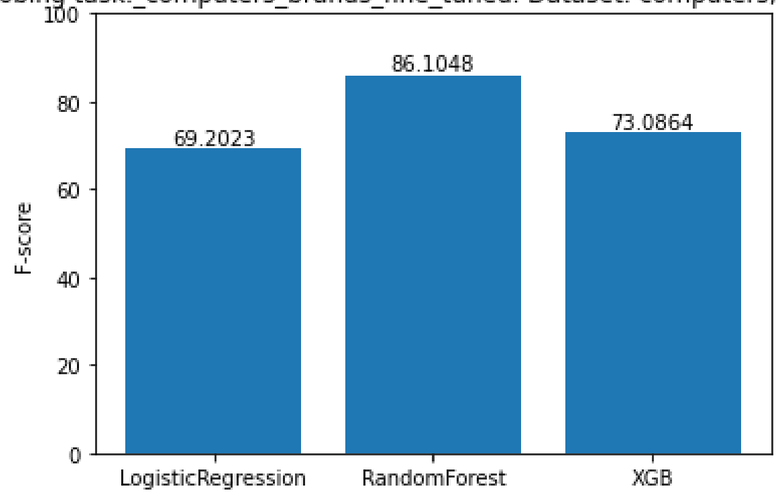
RESULTS **BRAND NAME**

FINE TUNED

curacy Score for probing task:_computers_brands_fine_tuned. Dataset: computers, medium, fine_tuned

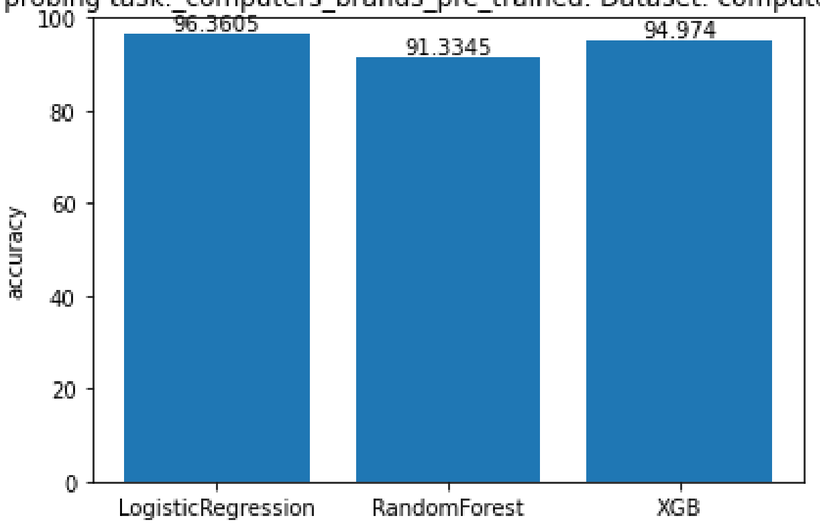


F-score for probing task:_computers_brands_fine_tuned. Dataset: computers, medium, fine_tuned

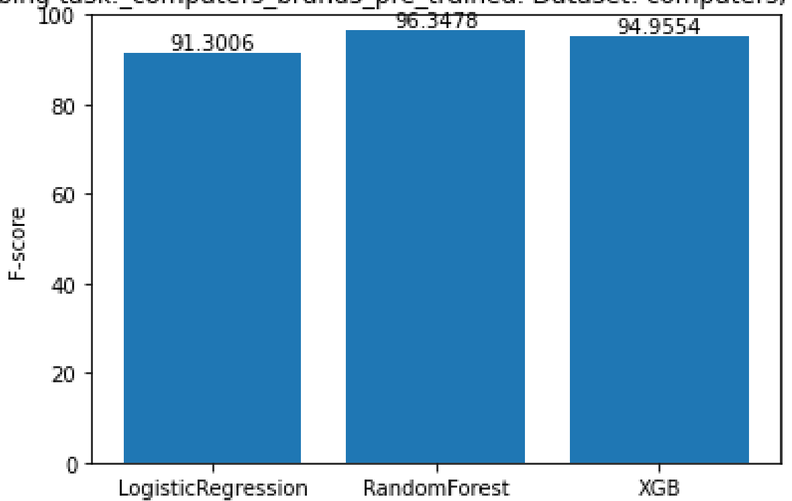


PRE TREINED

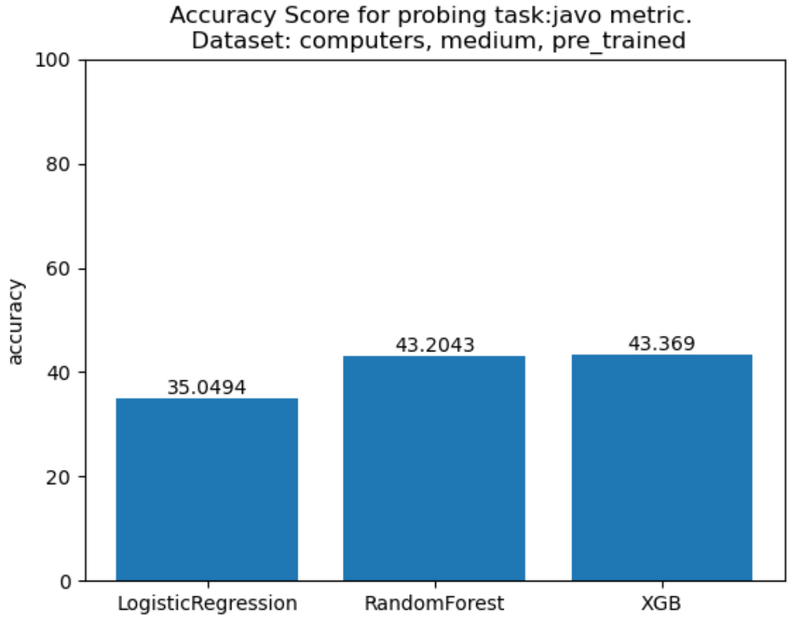
Accuracy Score for probing task:_computers_brands_pre_trained. Dataset: computers, medium, pre

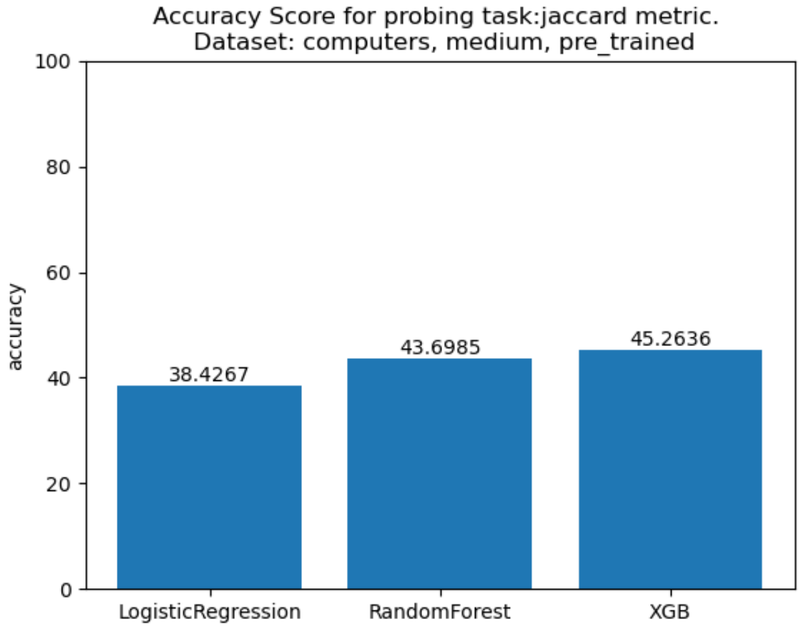


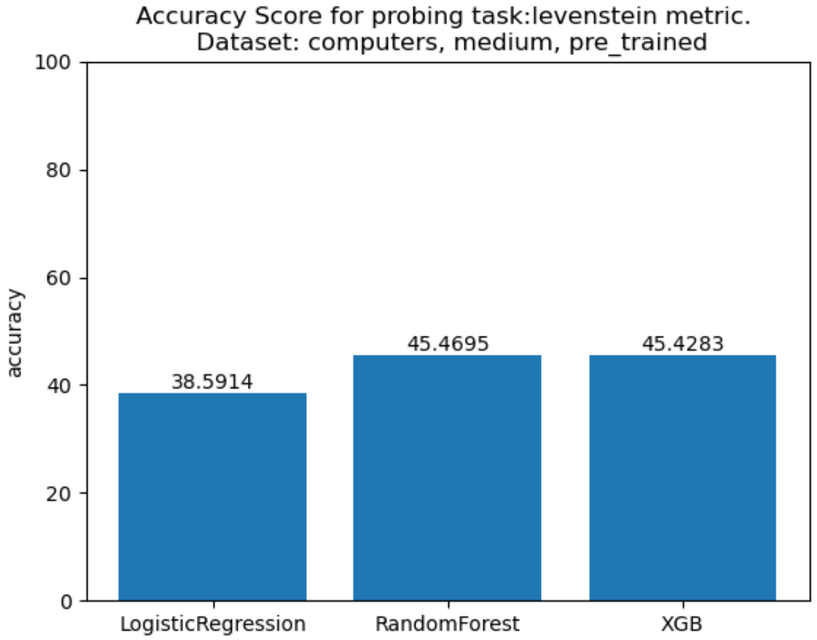
F-score for probing task:_computers_brands_pre_trained. Dataset: computers, medium, pre_train

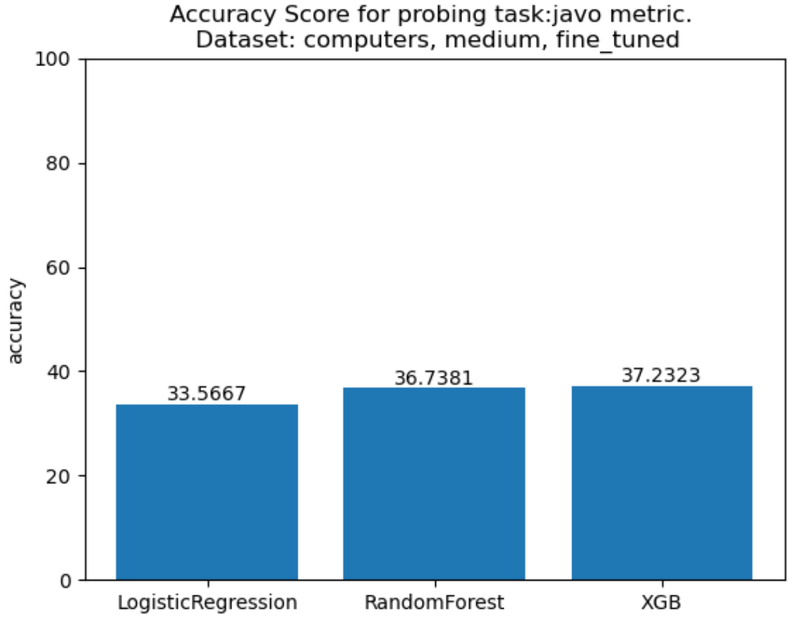


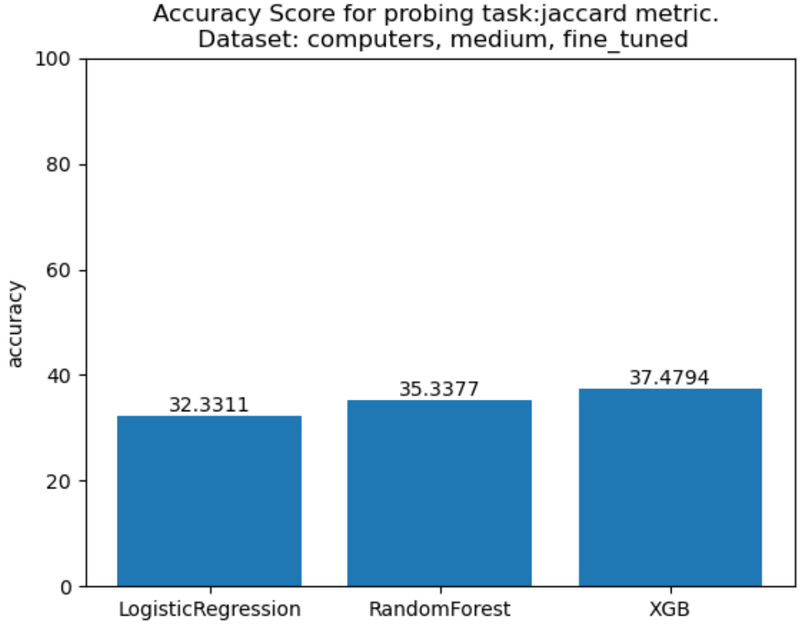
3 Probing task –
Text similarity measures

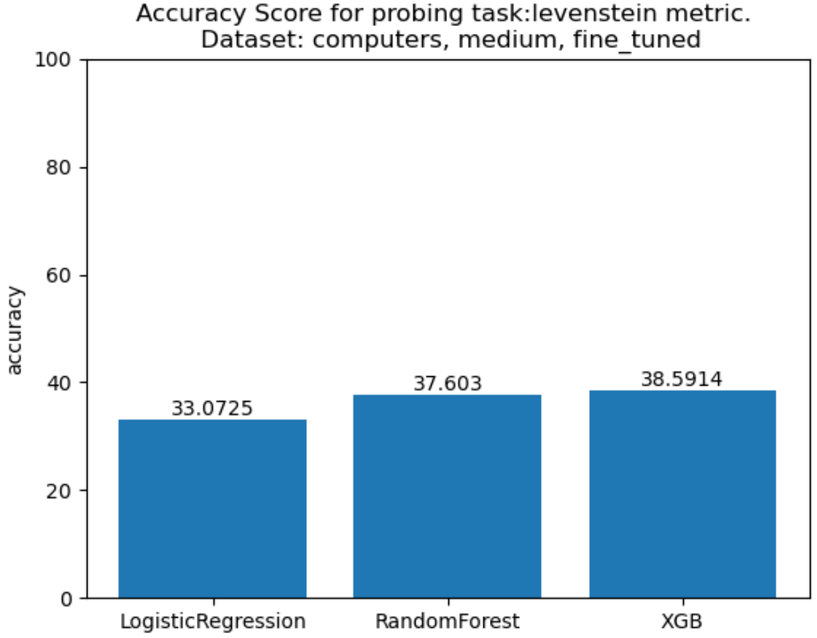

Accuracy Score for probing task:javo metric.
Dataset: computers, medium, pre_trained
accuracy
0
20
40
60
80
100
35.0494
43.2043
43.369
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:jaccard metric.
Dataset: computers, medium, pre_trained
accuracy
0
20
40
60
80
100
38.4267
43.6985
45.2636
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:levenstein metric.
Dataset: computers, medium, pre_trained
accuracy
0
20
40
60
80
100
38.5914
45.4695
45.4283
LogisticRegression
RandomForest
XGB

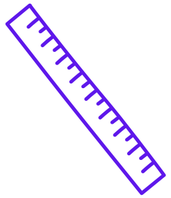

Accuracy Score for probing task:javo metric.
Dataset: computers, medium, fine_tuned
accuracy
0
20
40
60
80
100
33.5667
36.7381
37.2323
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:jaccard metric.
Dataset: computers, medium, fine_tuned
accuracy
0
20
40
60
80
100
32.3311
35.3377
37.4794
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:levenstein metric.
Dataset: computers, medium, fine_tuned
accuracy
0
20
40
60
80
100
33.0725
37.603
38.5914
LogisticRegression
RandomForest
XGB

4 Probing task –
Length of the sentence

Length of the sentence



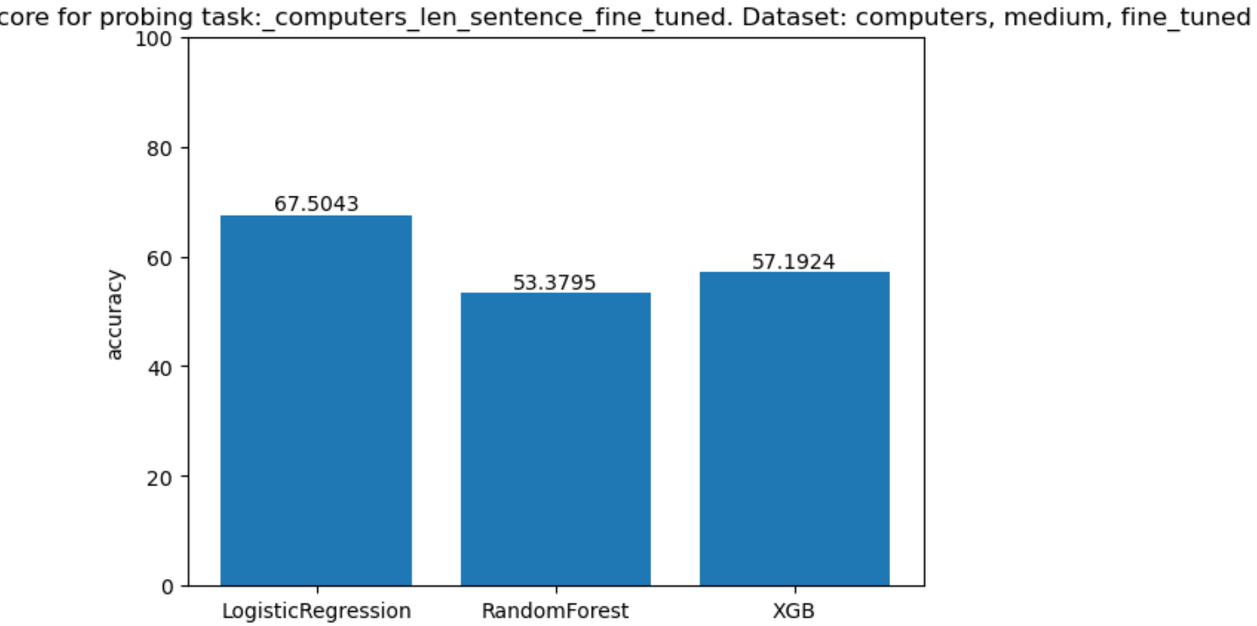
Influence of length of the sentence on embeddings.

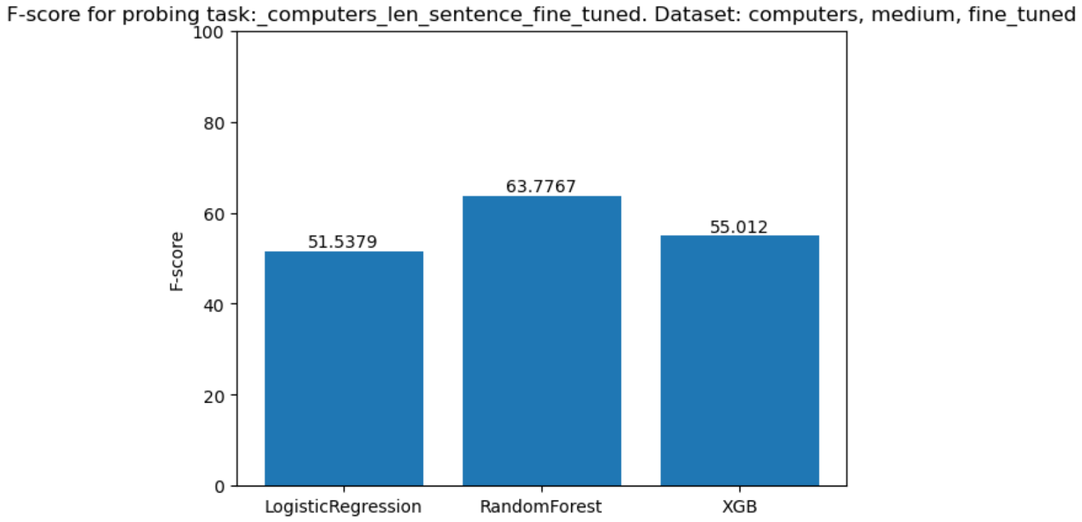
Classes

sentence length	nr of offers
[0,10)	1725
[10,15)	1122
[15,20)	763
[20, 100]	236

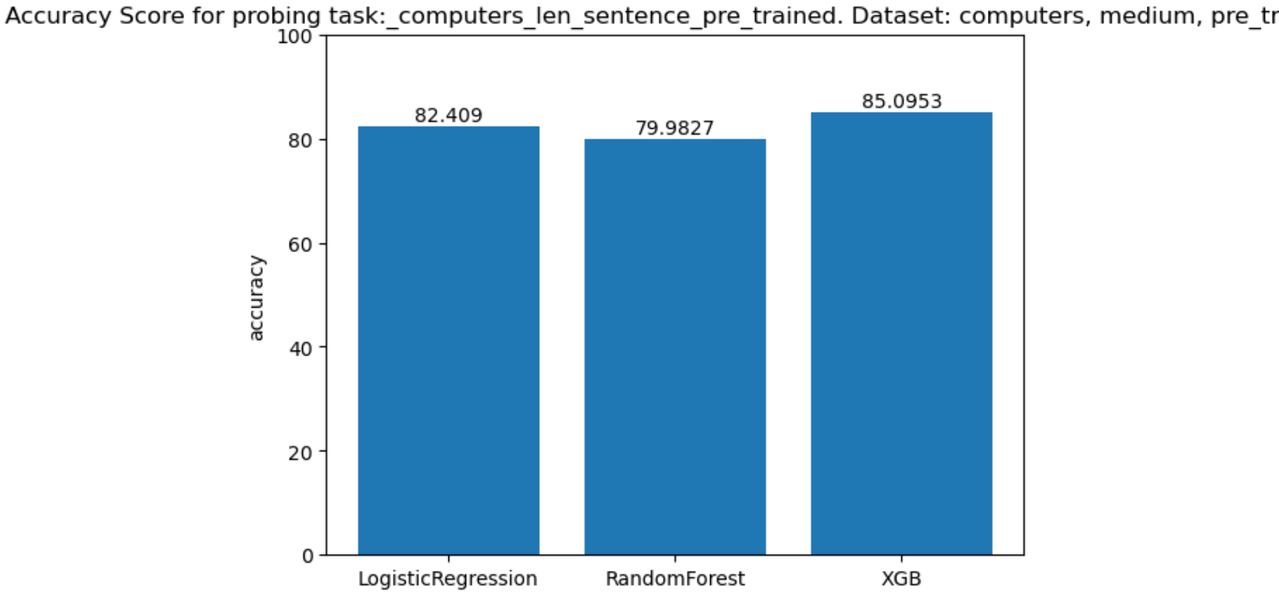
Results **Length of the sentence**

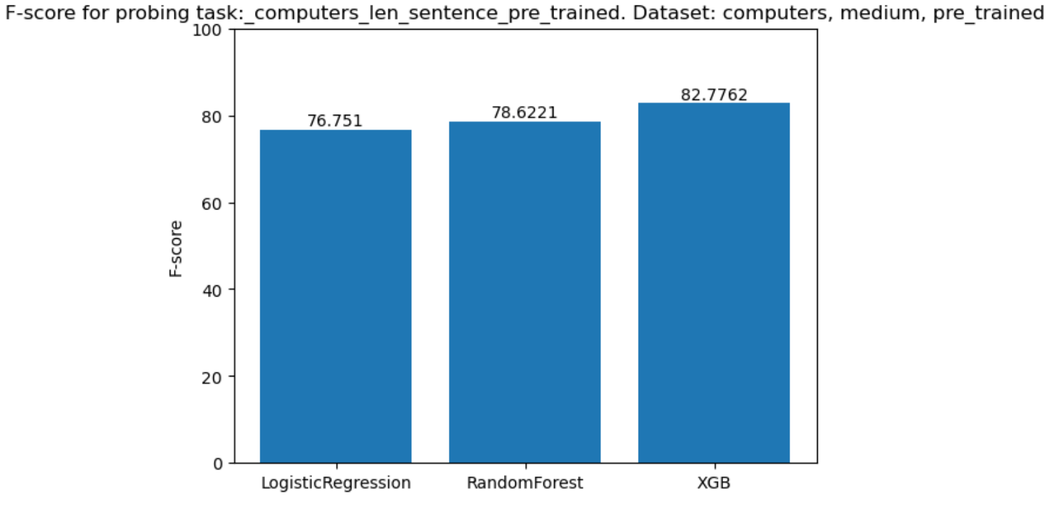
FINE TUNED





PRE TREINED





Dataset

Natural

A. Pretrained Model

B. Fine-Tuned Model

Probing Tasks

WH WORDS

the presence of **wh-words** in the title

NAME ENTITY RECOGNITION

the presence of the **name entity in the title**

SIMILARITY MEASURES

predicting **similar(sentenceA, sentenceB)**

LENGTH OF A SENTENCE

predicting the **length** of the input sentence

1 Probing task – **wh-words**

WH-WORDS ©

Influence of wh-words on embeddings

- wh-words: **what, where, who, why**

Well balanced classes

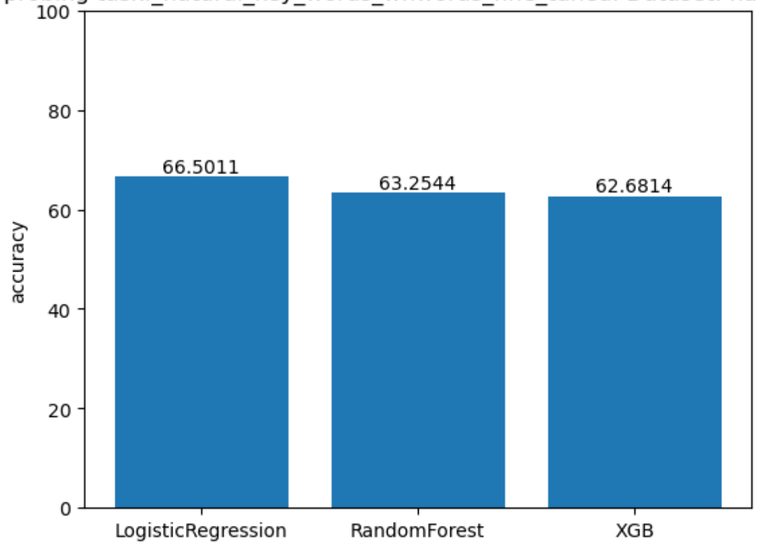
without wh-words	1522
with wh-words	1260

Results

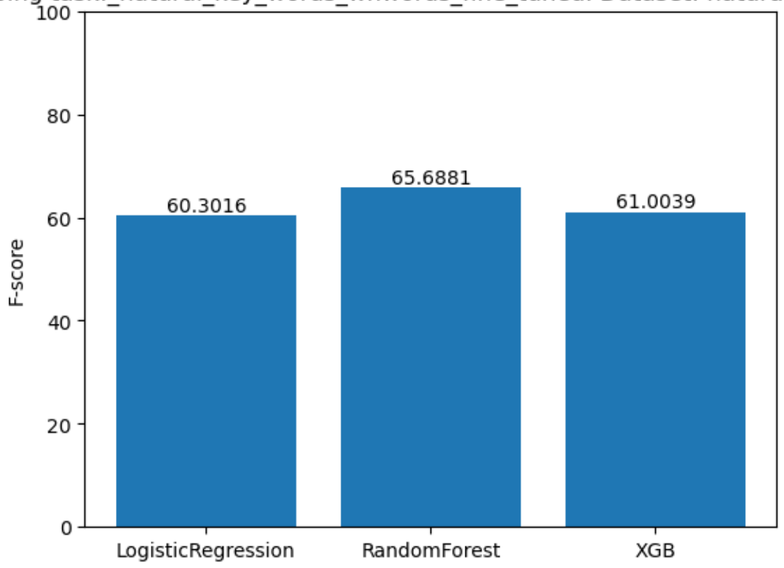
WH-WORDS

FINE TUNED

Accuracy Score for probing task:_natural_key_words_whwords_fine_tuned. Dataset: natural, medium, fine_tuned

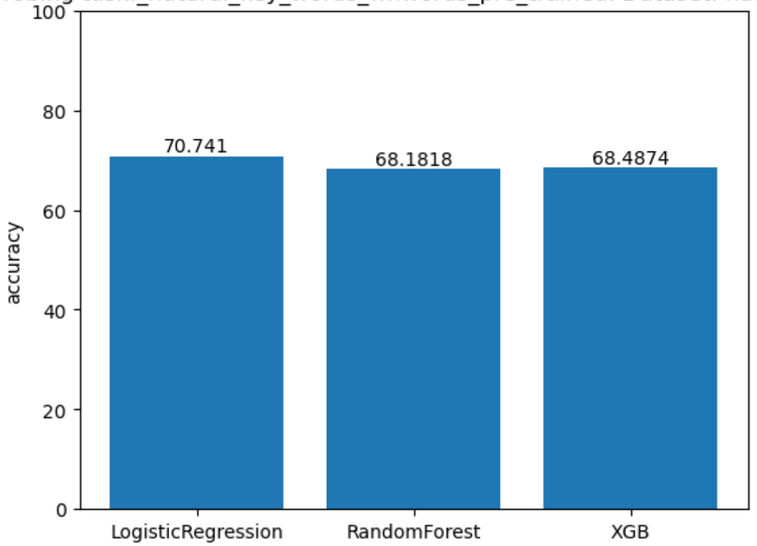


F-score for probing task:_natural_key_words_whwords_fine_tuned. Dataset: natural, medium, fine_tuned

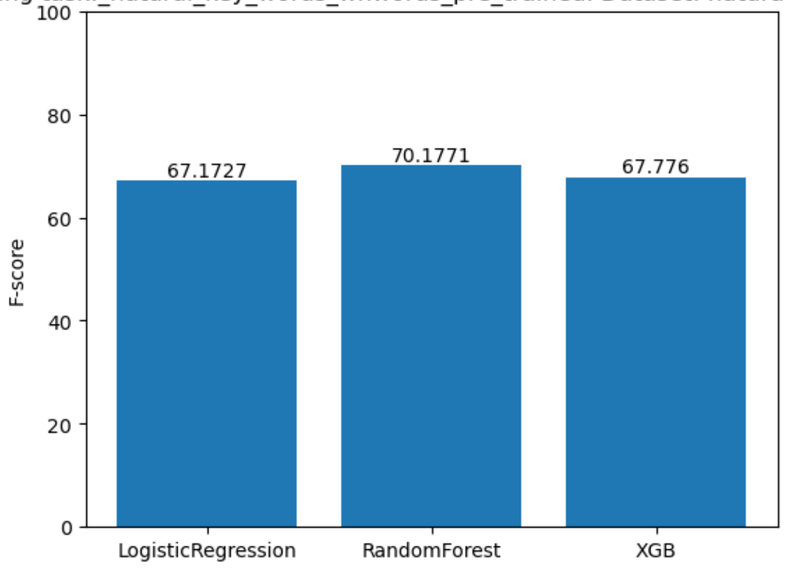


PRE TREINED

Accuracy Score for probing task:_natural_key_words_whwords_pre_trained. Dataset: natural, medium, pre_trained



F-score for probing task:_natural_key_words_whwords_pre_trained. Dataset: natural, medium, pre_trained



2 Probing task – **named entity**

NAMED ENTITY

Dataset

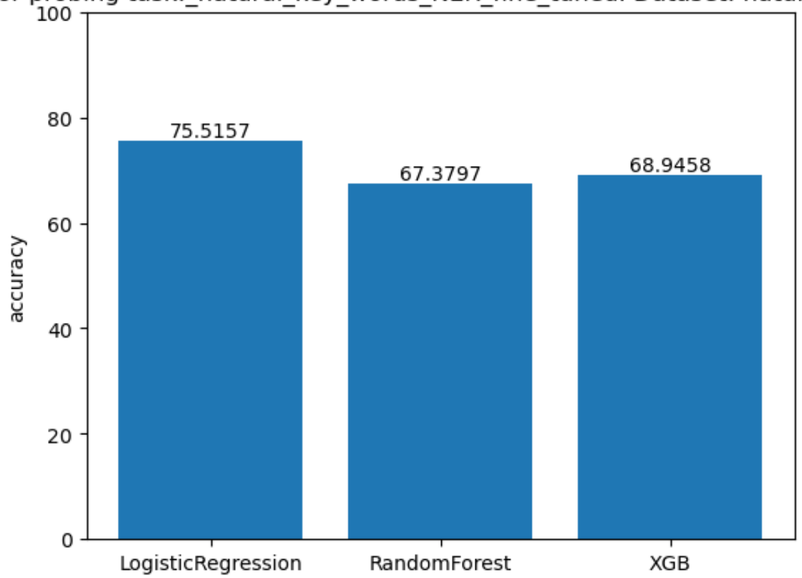
- without a named entity (Goggle, Harry Potter, ...)
- balanced classes
- bert-base-NER – Named Entity Recognition – Library

without amed entity	4905
with amed entity	3821

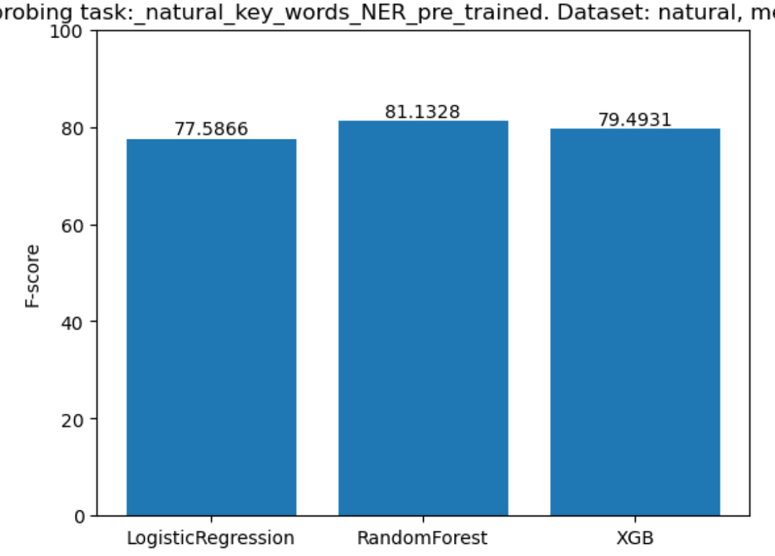
RESULTS **NAMED ENTITY**

FINE TUNED

Accuracy Score for probing task:_natural_key_words_NER_fine_tuned. Dataset: natural, medium, fine_tuned

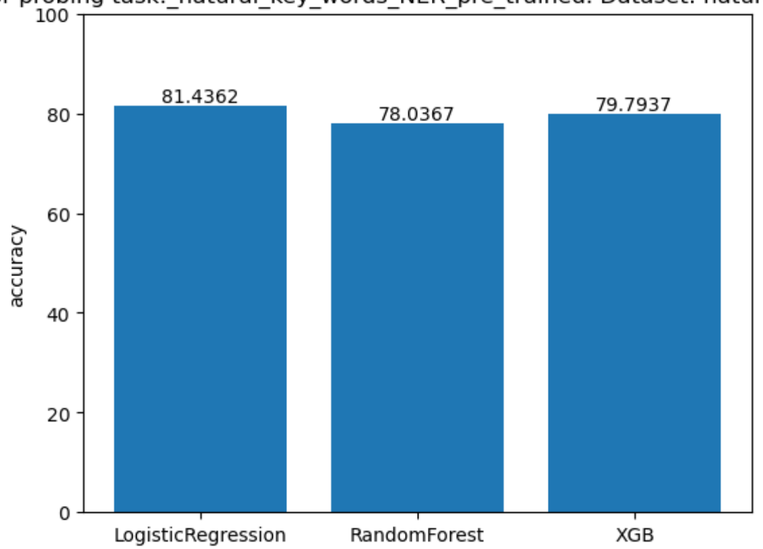


F-score for probing task:_natural_key_words_NER_pre_trained. Dataset: natural, medium, pre_trained

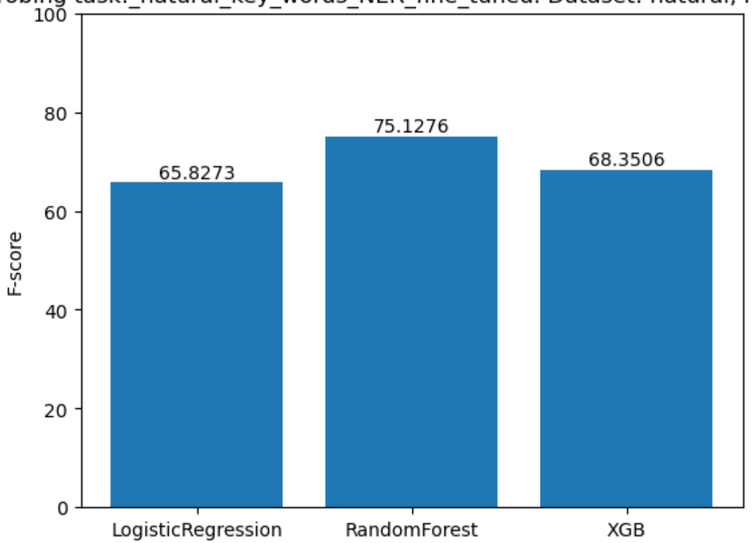


PRE TREINED

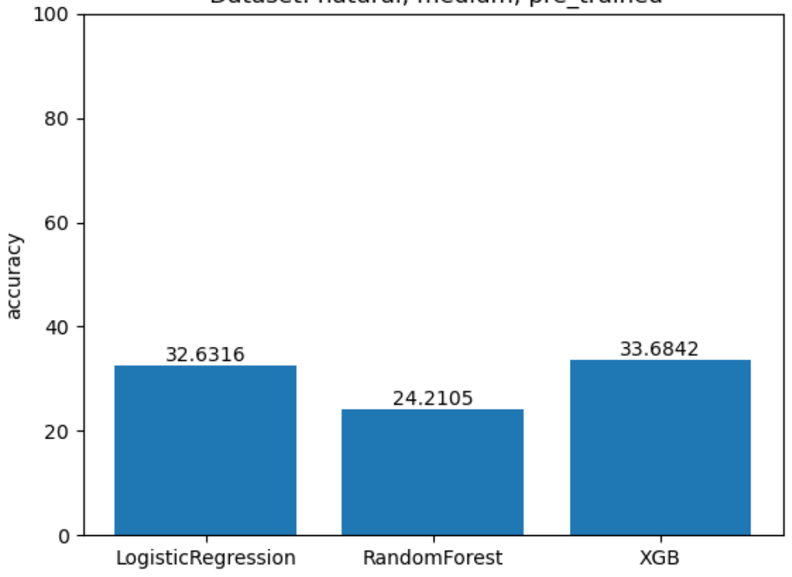
Accuracy Score for probing task:_natural_key_words_NER_pre_trained. Dataset: natural, medium, pre_trained

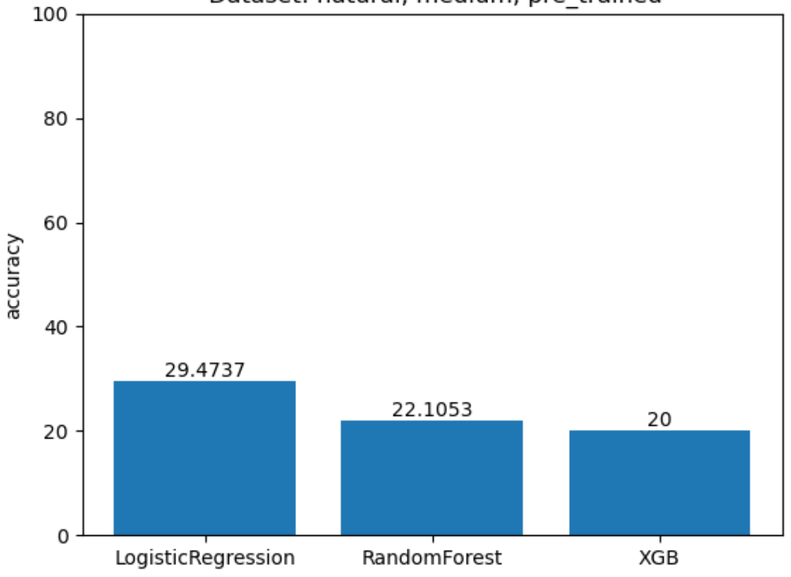


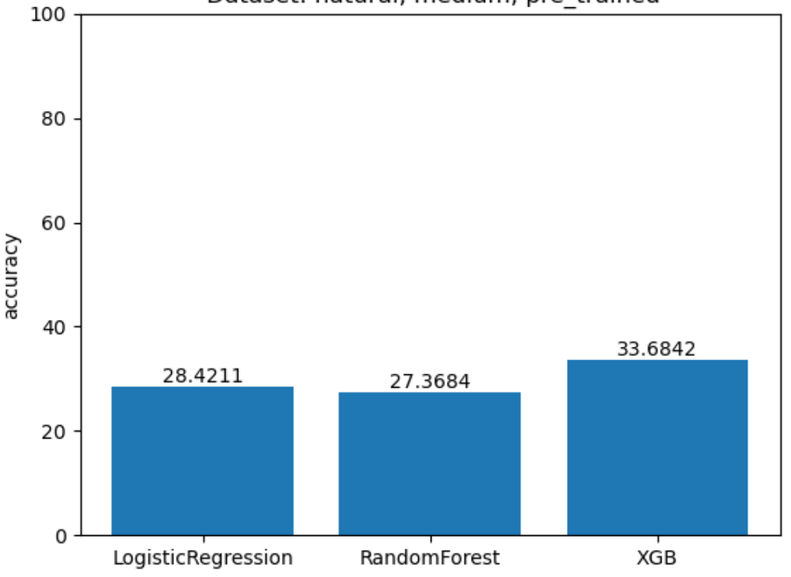
F-score for probing task:_natural_key_words_NER_fine_tuned. Dataset: natural, medium, fine_tuned

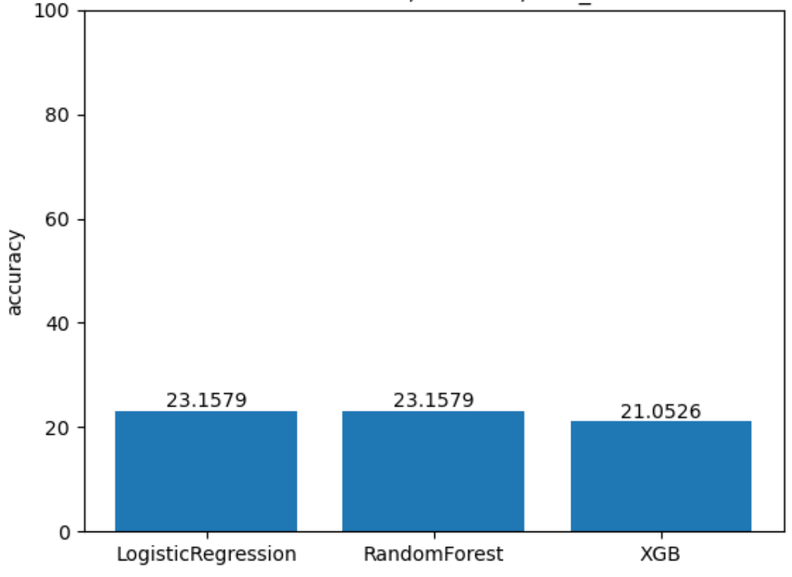


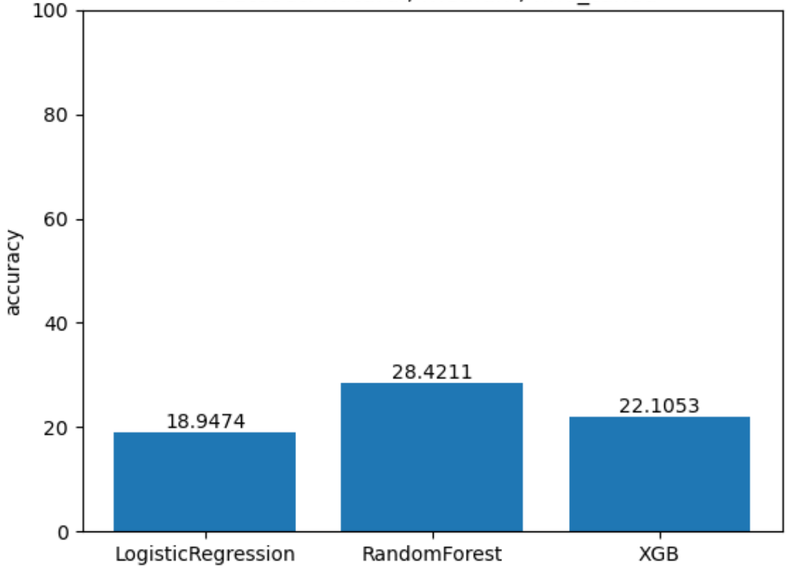
3 Probing task –
Text similarity measures

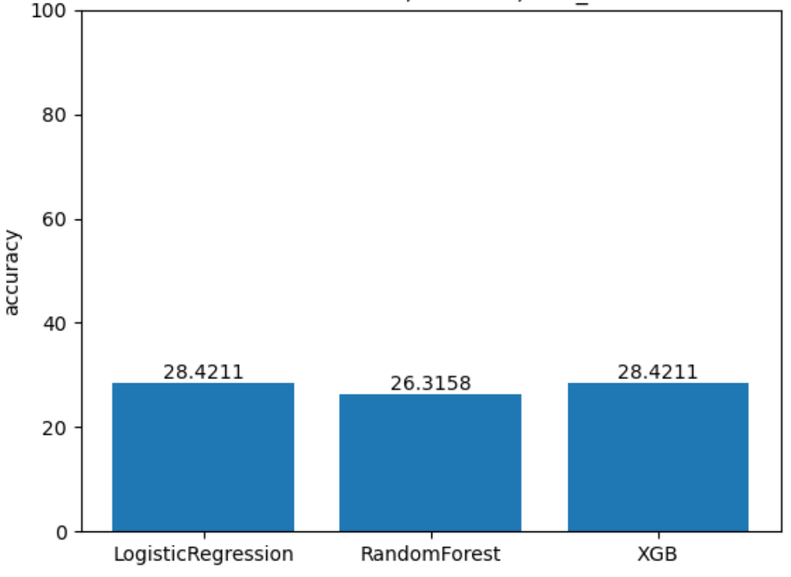

Accuracy Score for probing task:javo metric.
Dataset: natural, medium, pre_trained
accuracy
100
80
60
40
20
0
32.6316
24.2105
33.6842
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:jaccard metric.
Dataset: natural, medium, pre_trained
accuracy
100
80
60
40
20
0
29.4737
22.1053
20
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:levenstein metric.
Dataset: natural, medium, pre_trained
accuracy
100
80
60
40
20
0
28.4211
27.3684
33.6842
LogisticRegression
RandomForest
XGB

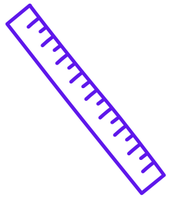

Accuracy Score for probing task:javo metric.
Dataset: natural, medium, fine_tuned
accuracy
100
80
60
40
20
0
23.1579
23.1579
21.0526
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:jaccard metric.
Dataset: natural, medium, fine_tuned
accuracy
100
80
60
40
20
0
18.9474
28.4211
22.1053
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:levenstein metric.
Dataset: natural, medium, fine_tuned
accuracy
100
80
60
40
20
0
28.4211
26.3158
28.4211
LogisticRegression
RandomForest
XGB

4 Probing task –
Length of the sentence

Length of the sentence



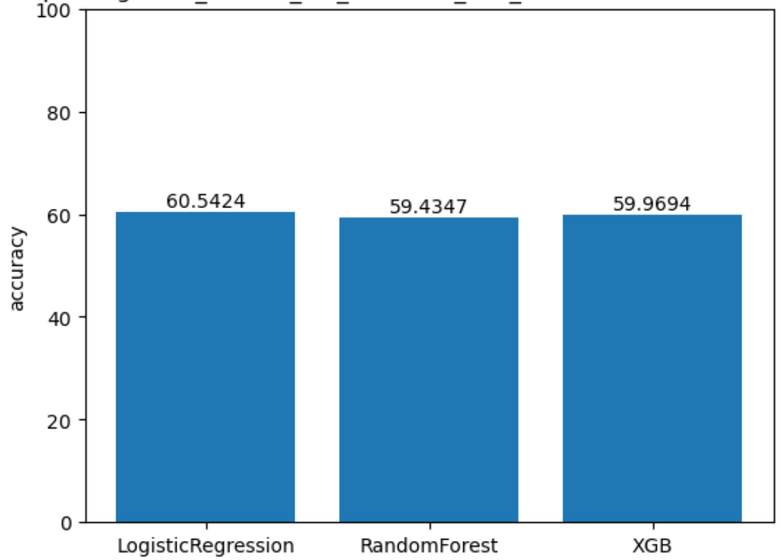
Influence of length of the sentence on embeddings.

sentence length	nr of offers
[0,10)	5104
[10,15)	2317
[15,20)	429
[20, 100]	503

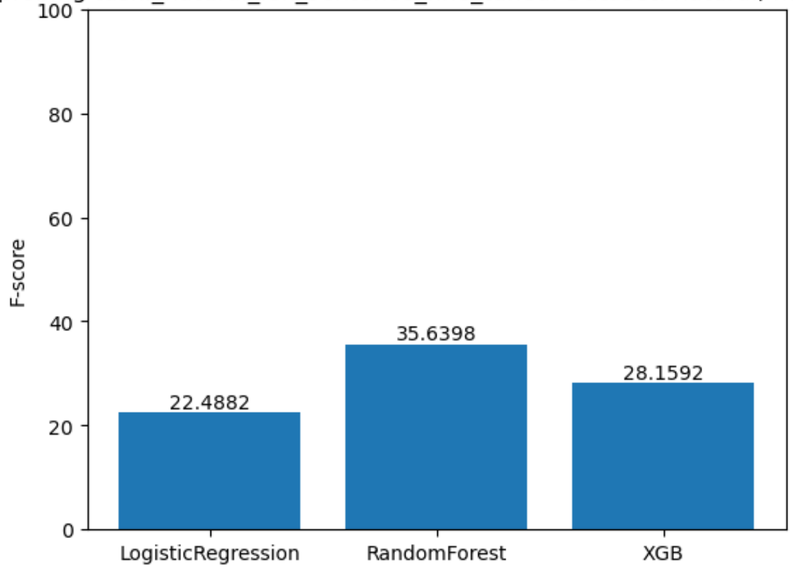
Results **Length of the sentence**

FINE TUNED

Accuracy Score for probing task:_natural_len_sentence_fine_tuned. Dataset: natural, medium, fine_tuned

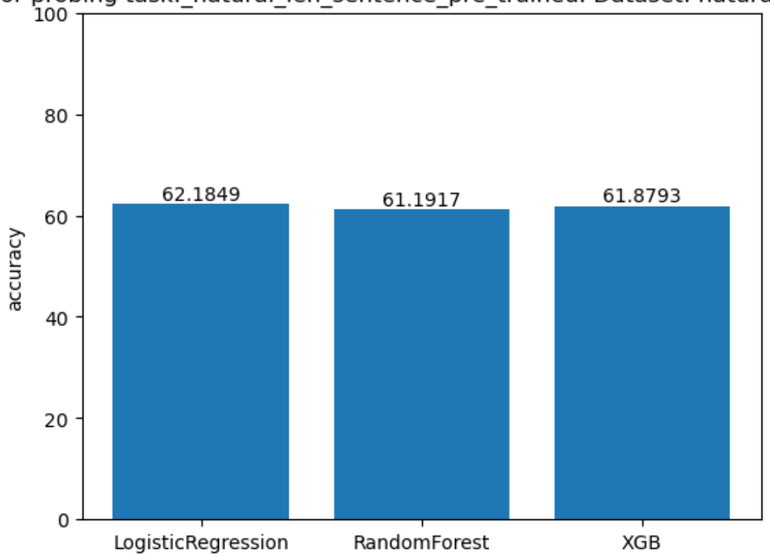


F-score for probing task:_natural_len_sentence_fine_tuned. Dataset: natural, medium, fine_tuned

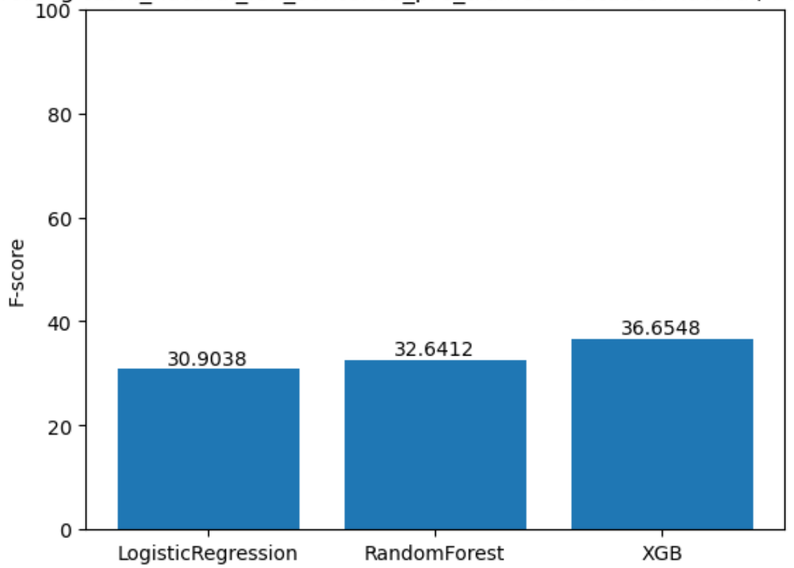


PRE TREINED

Accuracy Score for probing task:_natural_len_sentence_pre_trained. Dataset: natural, medium, pre_trained



F-score for probing task:_natural_len_sentence_pre_trained. Dataset: natural, medium, pre_trained



Dataset

CAMERAS

A. Pretrained Model

B. Fine-Tuned Model

Probing Tasks

COMMON WORDS

the presence of **common words** ('camera', 'len', 'digital') in the **title**

BRAND NAME

the presence of the **brand name in the title**

SIMILARITY MEASURES

predicting similar**(sentenceA, sentenceB)**

LENGTH OF A SENTENCE

predicting the **length** of the input sentence

1 Probing task – **common words**

COMMON WORDS ©

Influence of common words on embeddings

- one of common words: **camera, len, digital**

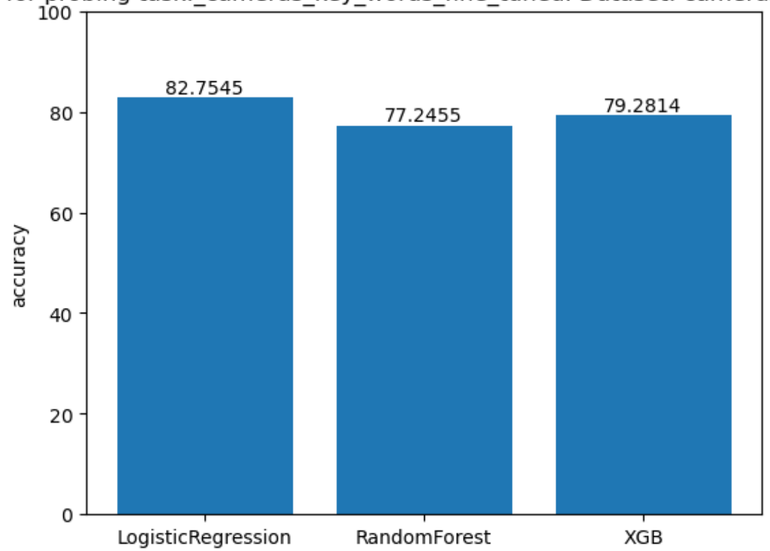
Well balanced classes

without common words	1522
with comon words	1260

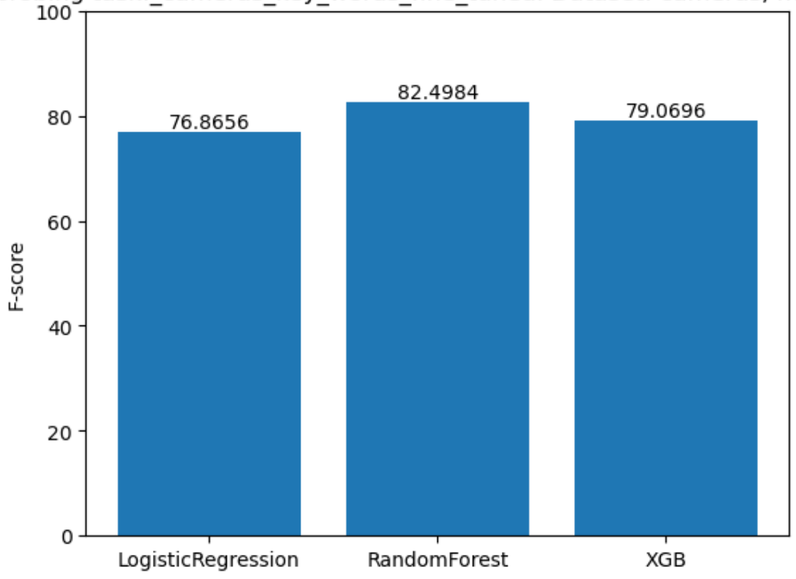
Results COMMON WORDS ©

FINE TUNED

Accuracy Score for probing task:_cameras_key_words_fine_tuned. Dataset: cameras, medium, fine_tuned

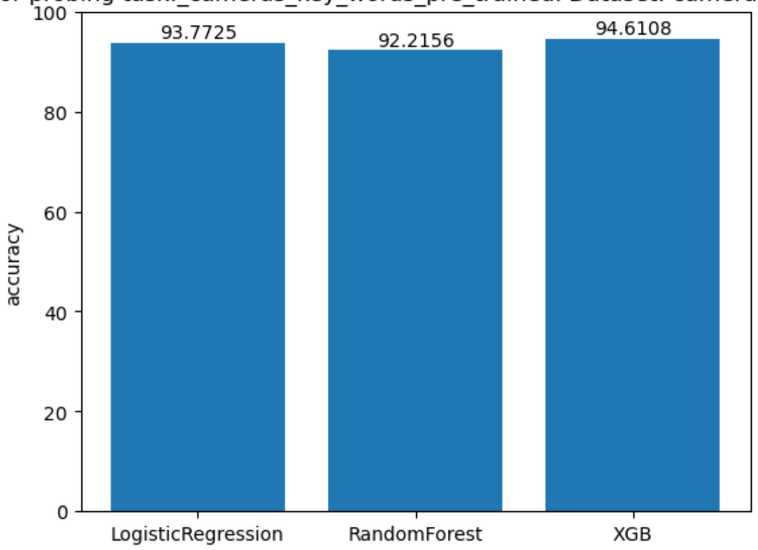


F-score for probing task:_cameras_key_words_fine_tuned. Dataset: cameras, medium, fine_tuned

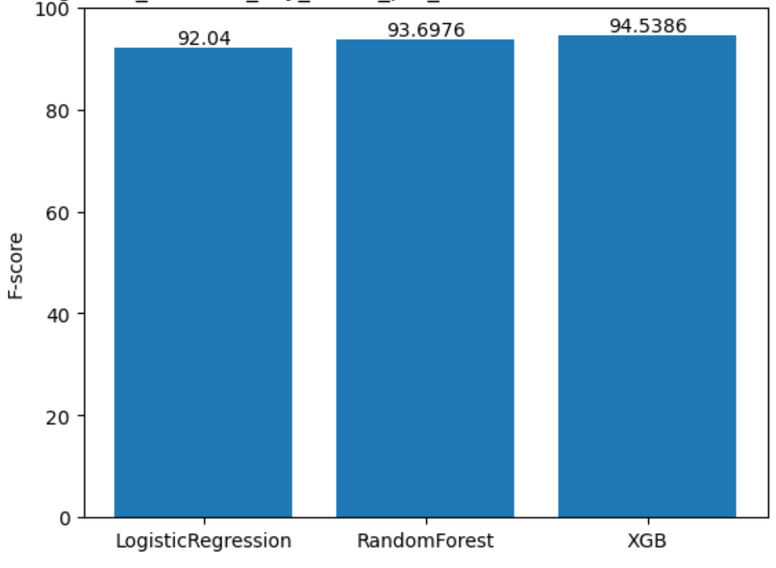


PRE TREINED

Accuracy Score for probing task:_cameras_key_words_pre_trained. Dataset: cameras, medium, pre_trained



F-score for probing task:_cameras_key_words_pre_trained. Dataset: cameras, medium, pre_trained

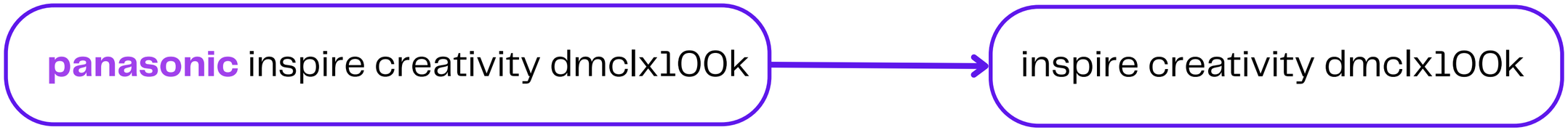


2 Probing task – **brand name**

BRAND NAME

New dataset

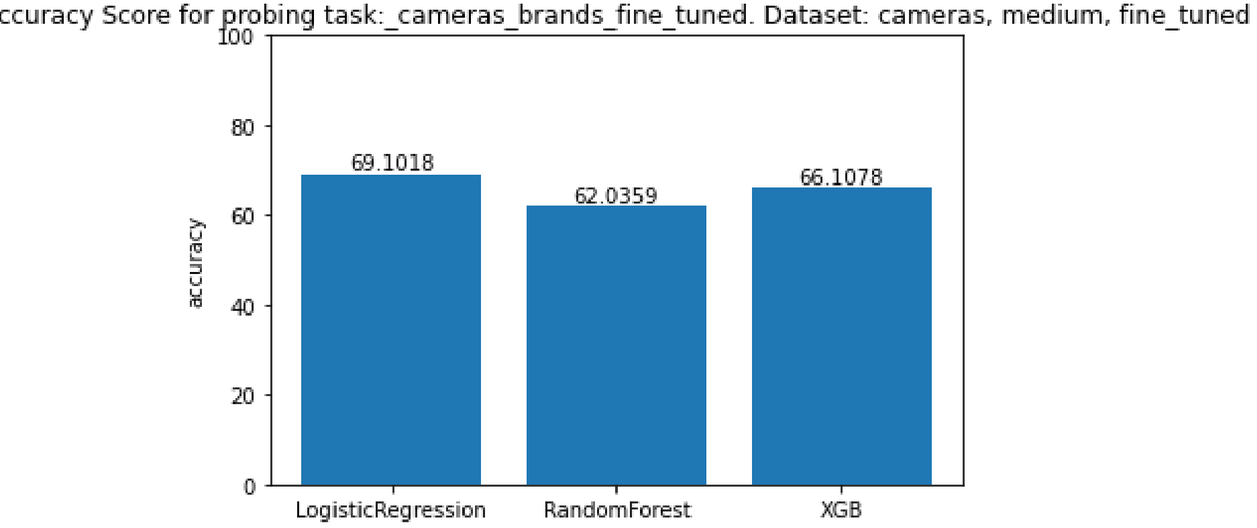
- without a brand name (Canon, Samsung, ...)
- balanced classes

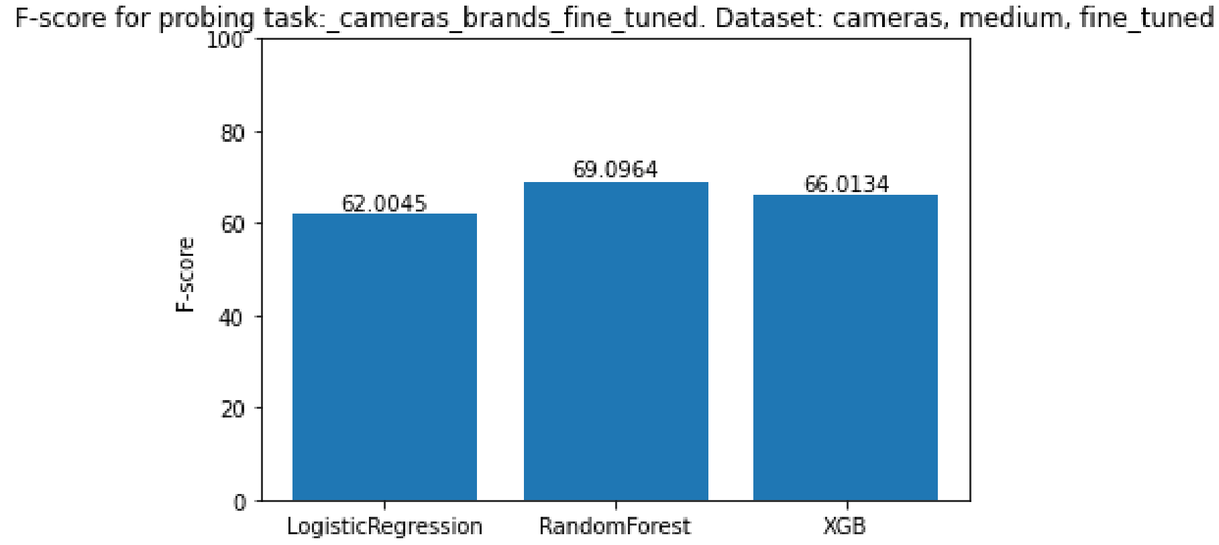


without brand name	1521
with brand name	1261

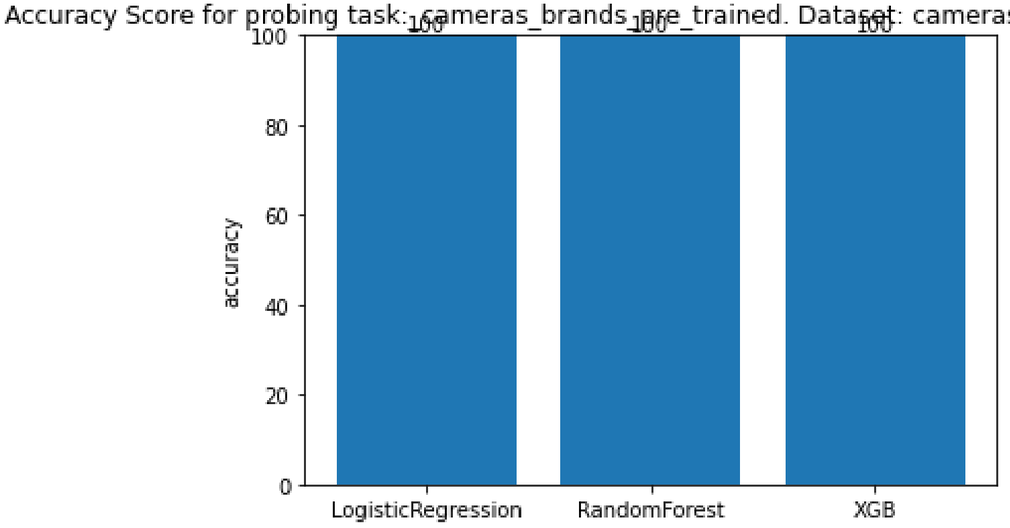
RESULTS **BRAND NAME**

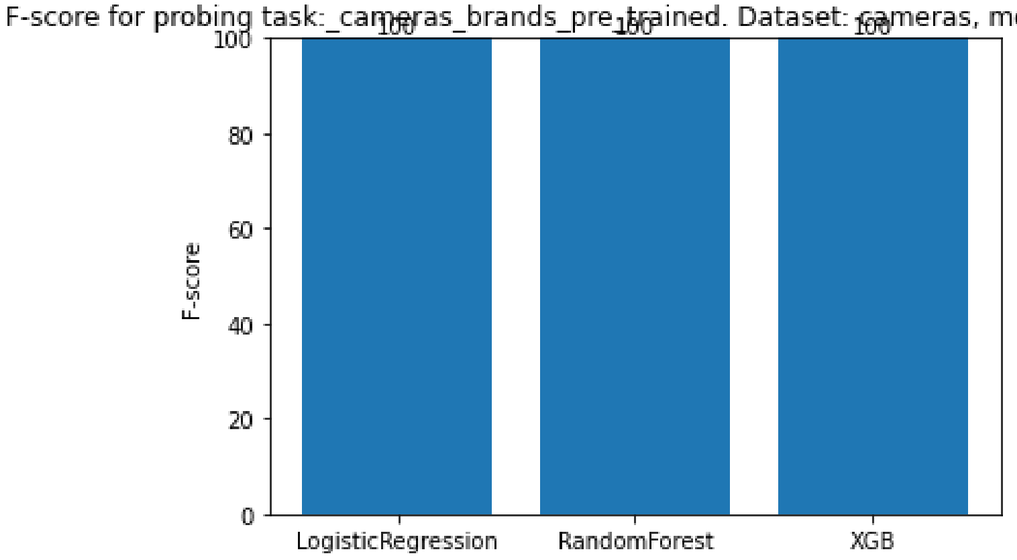
FINE TUNED



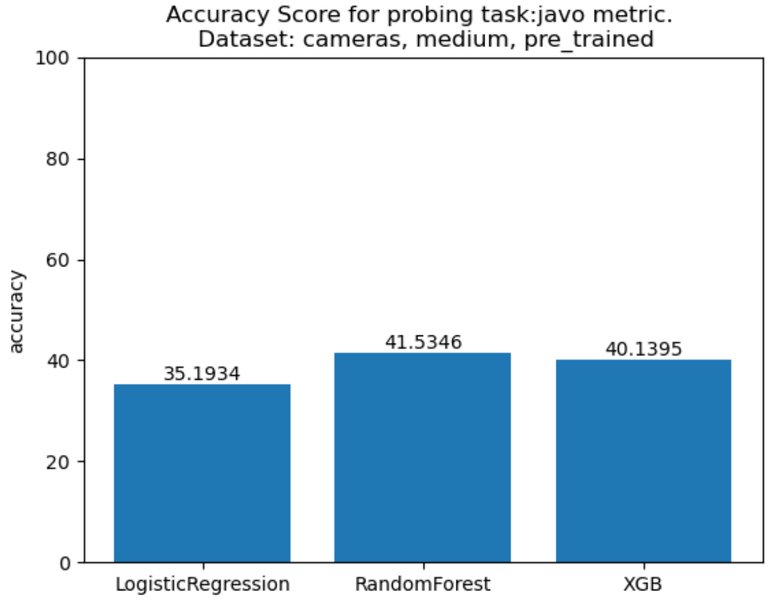


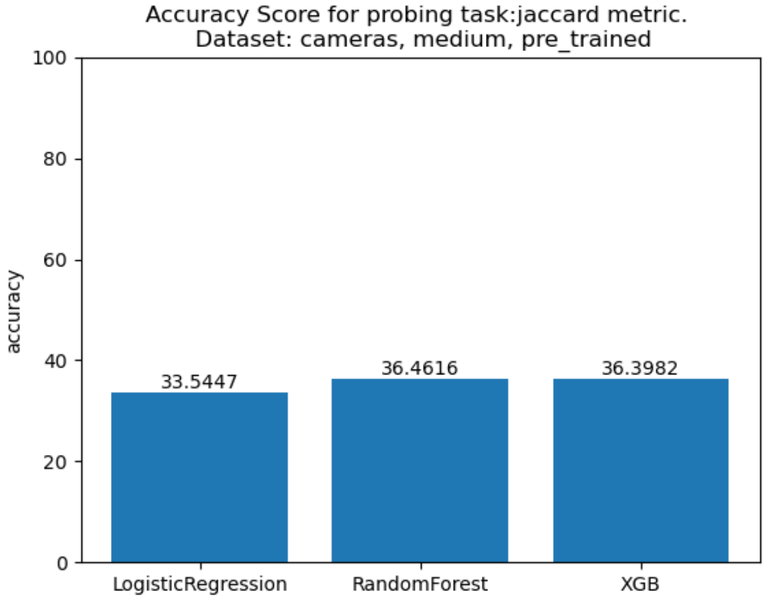
PRE TREINED

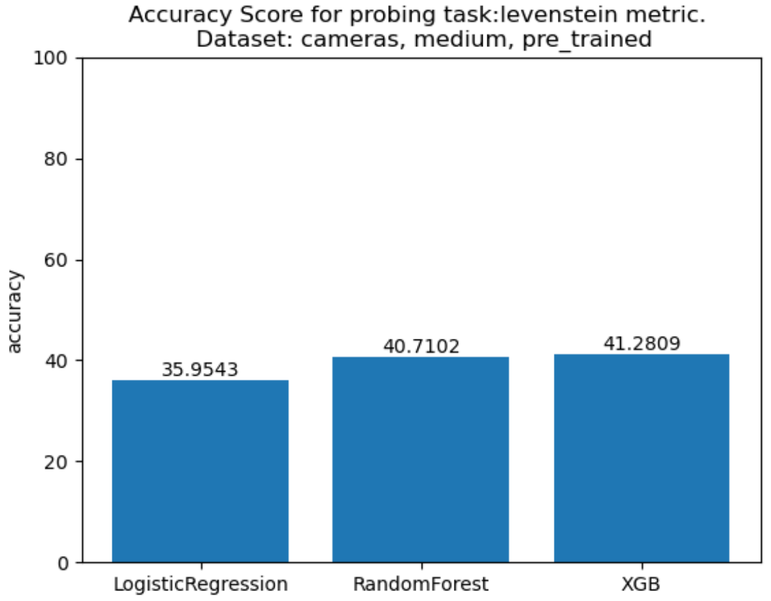


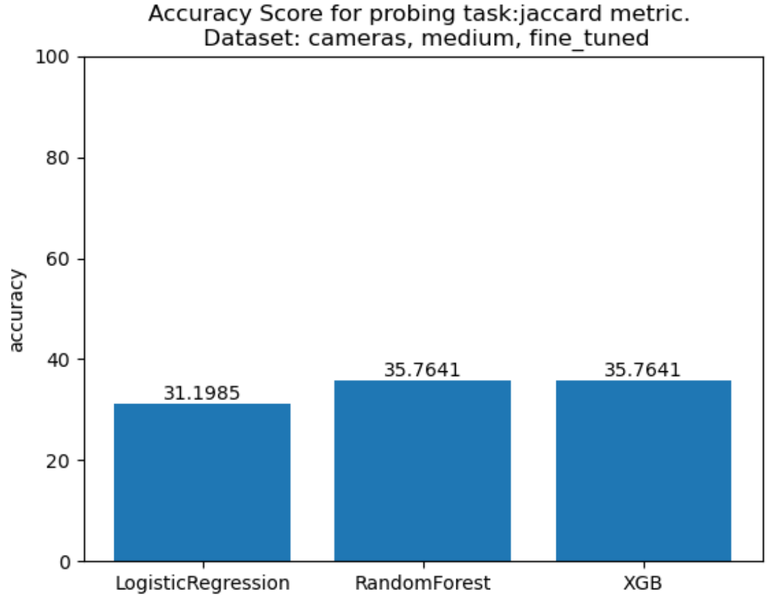


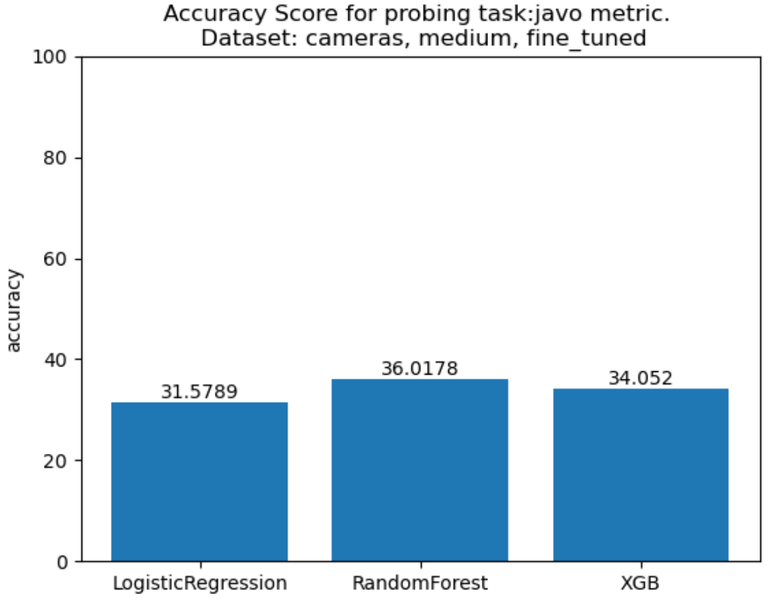
3 Probing task –
Text similarity measures

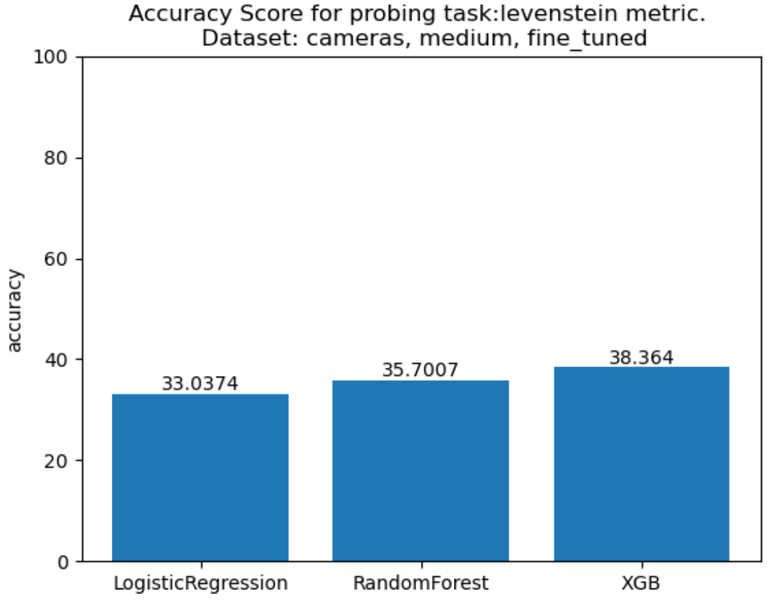

Accuracy Score for probing task:javo metric.
Dataset: cameras, medium, pre_trained
accuracy
100
80
60
40
20
0
35.1934
41.5346
40.1395
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:jaccard metric.
Dataset: cameras, medium, pre_trained
accuracy
100
80
60
40
20
0
33.5447
36.4616
36.3982
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:levenstein metric.
Dataset: cameras, medium, pre_trained
accuracy
100
80
60
40
20
0
35.9543
40.7102
41.2809
LogisticRegression
RandomForest
XGB

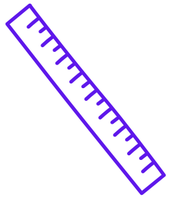

Accuracy Score for probing task:jaccard metric.
Dataset: cameras, medium, fine_tuned
accuracy
100
80
60
40
20
0
31.1985
35.7641
35.7641
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:javo metric.
Dataset: cameras, medium, fine_tuned
accuracy
100
80
60
40
20
0
31.5789
36.0178
34.052
LogisticRegression
RandomForest
XGB


Accuracy Score for probing task:levenstein metric.
Dataset: cameras, medium, fine_tuned
accuracy
100
80
60
40
20
0
33.0374
35.7007
38.364
LogisticRegression
RandomForest
XGB

4 Probing task –
Length of the sentence

Length of the sentence

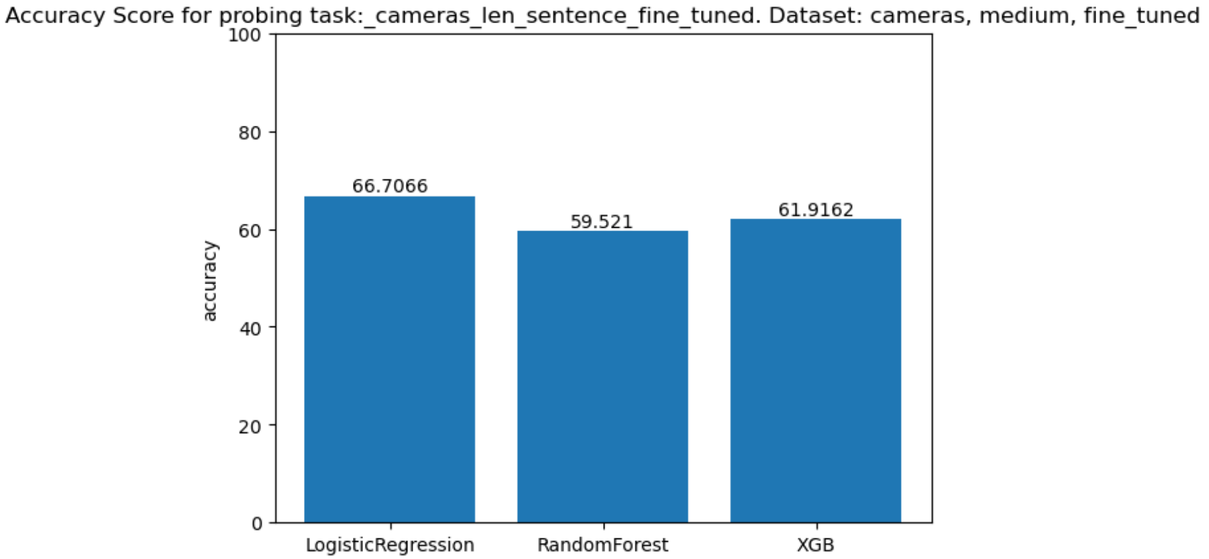


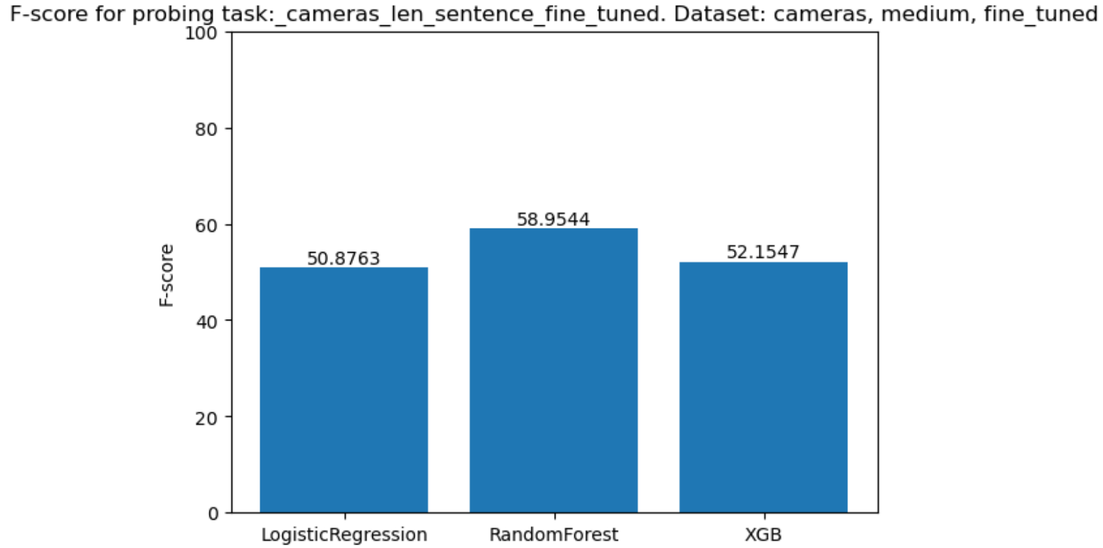
Influence of length of the sentence on embeddings.

sentence length	nr of offers
[0,10)	1088
[10,15)	1055
[15,20)	429
[20, 100]	201

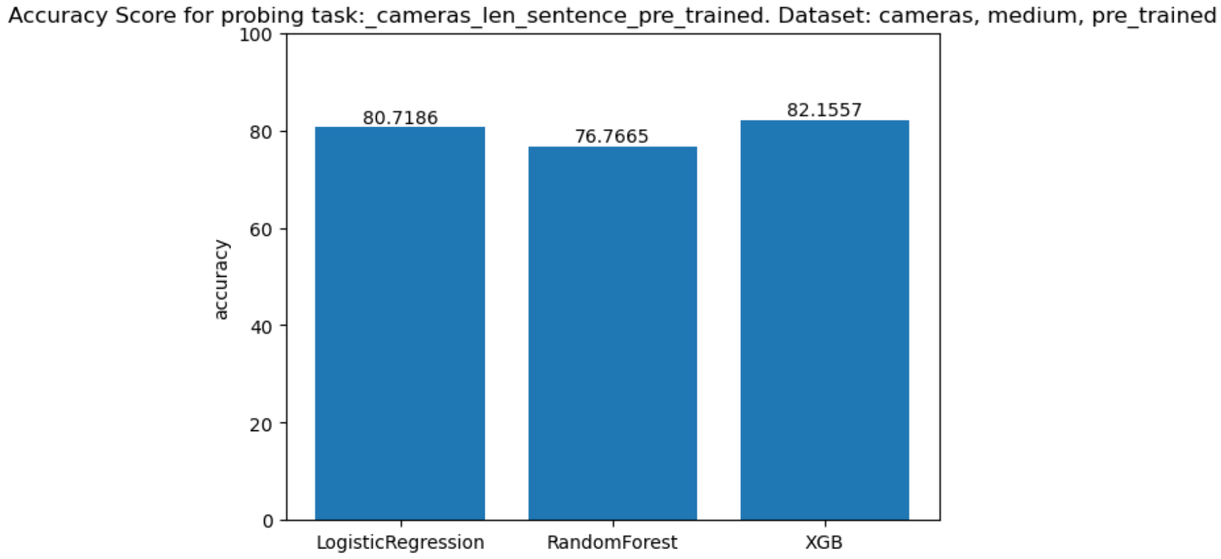
Results **Length of the sentence**

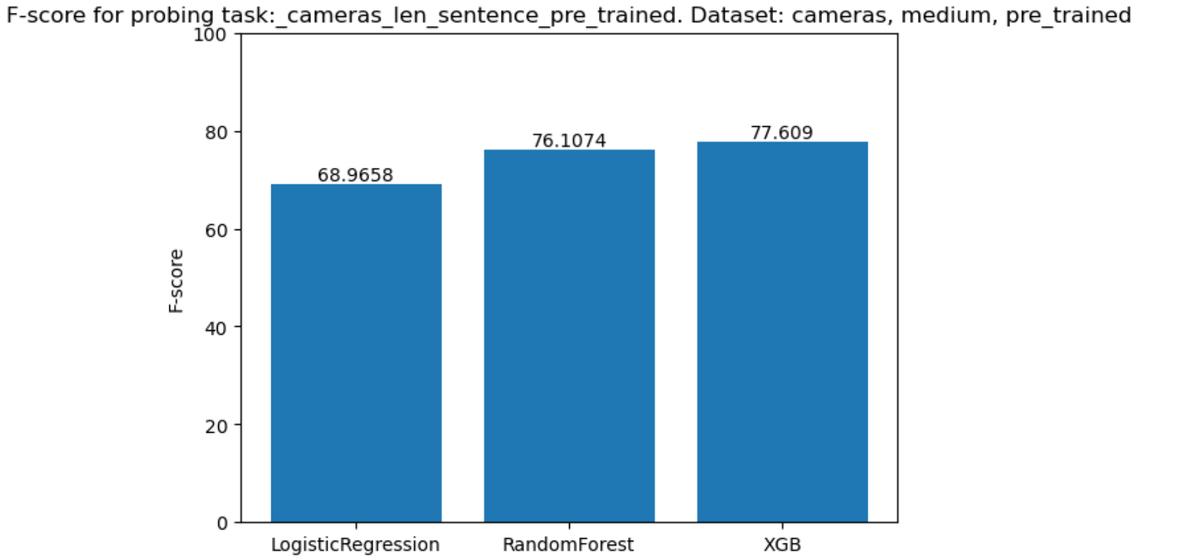
FINE TUNED





PRE TREINED





SentEval

Task	Type	#train	#test	needs_train	set_classifier
SentLen	Length prediction	100k	10k	1	1
WC	Word Content analysis	100k	10k	1	1
TreeDepth	Tree depth prediction	100k	10k	1	1
TopConst	Top Constituents prediction	100k	10k	1	1
BShift	Word order analysis	100k	10k	1	1
Tense	Verb tense prediction	100k	10k	1	1
SubjNum	Subject number prediction	100k	10k	1	1
ObjNum	Object number prediction	100k	10k	1	1
SOMO	Semantic odd man out	100k	10k	1	1
CoordInv	Coordination Inversion	100k	10k	1	1

[1] A. Conneau, D. Kiela, SentEval: An Evaluation Toolkit for Universal Sentence Representations

References:

[1] A. Conneau, D. Kiela, SentEval: An Evaluation Toolkit for Universal Sentence Representations[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.

[2] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network-based conversational agents.

[3] Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents.

[4] Nils Reimers and Iryna Gurevych. SentenceBERT: Sentence Embeddings using Siamese BERTNetworks.

5] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring.

[6] Nandan Thakur, Nils Reimers, Johannes Daxenberge, and Iryna Gurevych. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks.

[7] Giambattista Amati. BM25, Springer US, Boston, MA.

[8] Lindstr¨om, Adam & Bj¨orklund,
Johanna & Bensch, Suna & Drewes, Frank. 2021.
Probing Multimodal Embeddings for Linguistic
Properties: the Visual-Semantic Case.

The End
THANK YOU