PAWEŁ GOLIK, MATEUSZ JASTRZĘBIOWSKI,
ALEKSANDRA MUSZKOWSKA

# NLP course: Project 1 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<br>Warianty (2) | 3 349,00 zł<br>ARMOWA WYSYŁKA<br>dostępny | KUP TERAZ<br>IDŹ DO SKLEPU |
| 97 opinii | **Zegarek Garmin Fenix 7X Solar niebieskoszary z czarnym paskiem • RATY 0% • DOSTAWA GRATIS (paczkomat lub kurier UPS)**<br>Warianty (2) | 3 349,00 zł<br>ARMOWA WYSYŁKA<br>dostępny | KUP TERAZ<br>IDŹ DO SKLEPU |
| 2 opinie | Smartwatch GARMIN Fenix 7X Solar Czarny z czarnym paskiem 010-02541-01 ® KUP TERAZ<br>**RATY 0% I 6 M-CY NIE PŁACISZ!** | 3 349,00 zł<br>ARMOWA WYSYŁKA<br>dostępny | IDŹ DO SKLEPU |

OFFERS

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

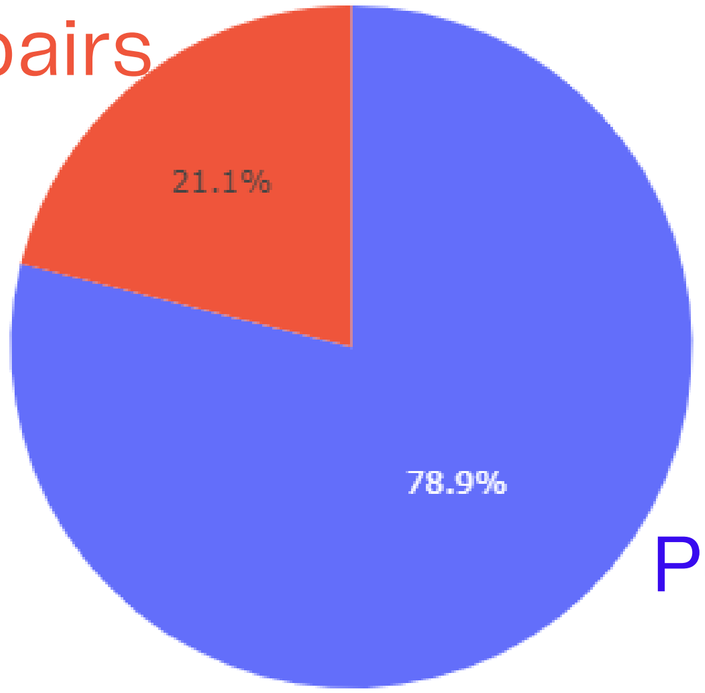
| | | | |
|---|---|---|---|
| 1 opinii | Zegarek sportowy Garmin Fenix 7X Czarny (010-02541-01)<br>**DARMOWA DOSTAWA JUŻ OD 399 ZŁ** | 3 299,00 zł<br>DARMOWA WYSYŁKA<br>dostępny | IDŹ DO SKLEPU |

# The WDC Dataset

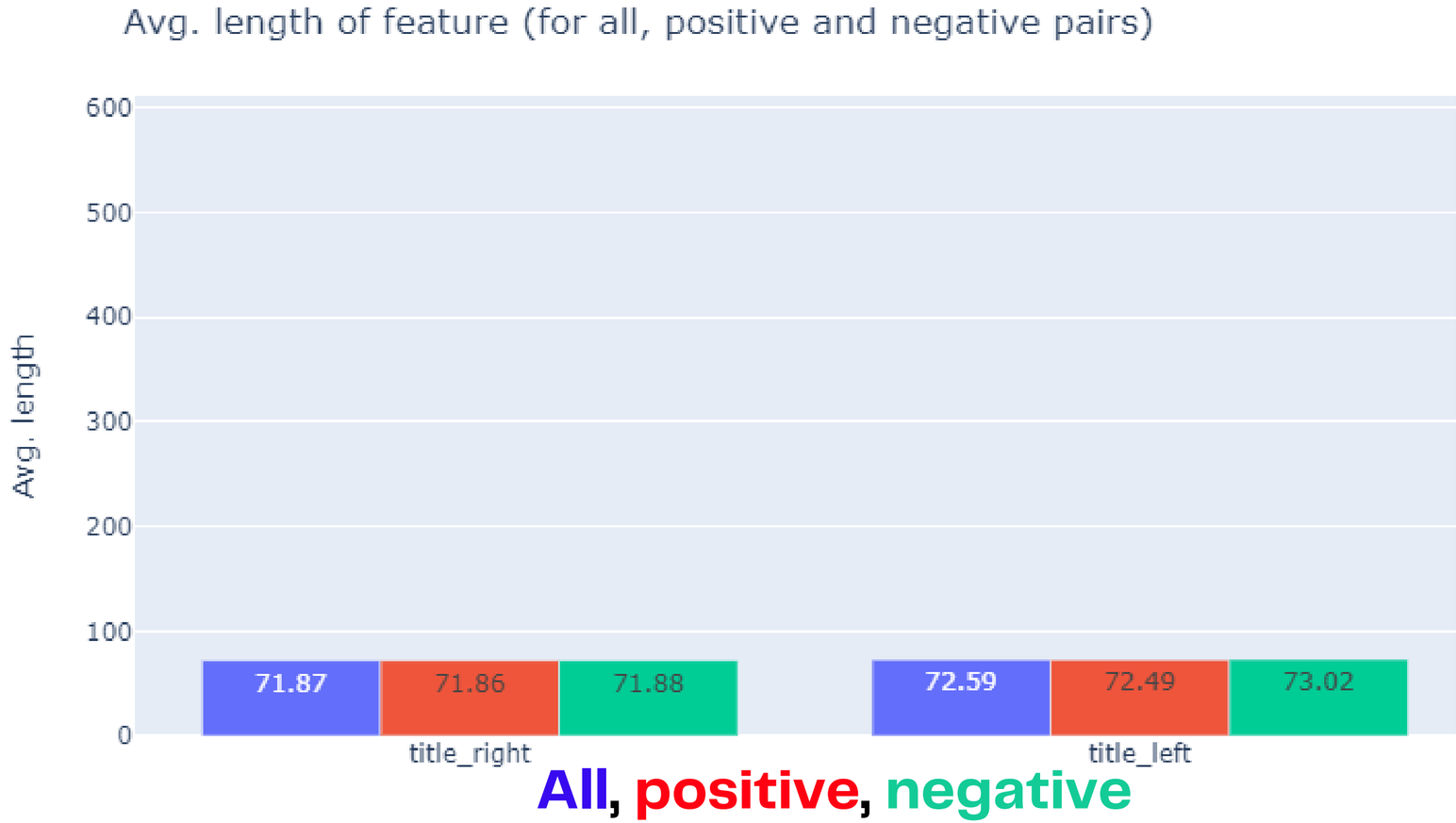
| Offer A | | Offer B | | Target (=isProdcut TheSame) |
|---|---|---|---|---|
| TitleA1 "BrandX Camera MX140" | DescA1 "..." | TitleB1 "Camera BrandX MX 140" | DescB1 "..." |  |
| TitleA2 "BrandY S1000 Digital Camera" | DescA2 "..." | TitleB2 "BrandZ Camera PRO 3F" | DescB2 "..." |  |

# Number of pairs: 4147 (p) + 1108 (n) = 5255

Negative pairs




Positive pairs



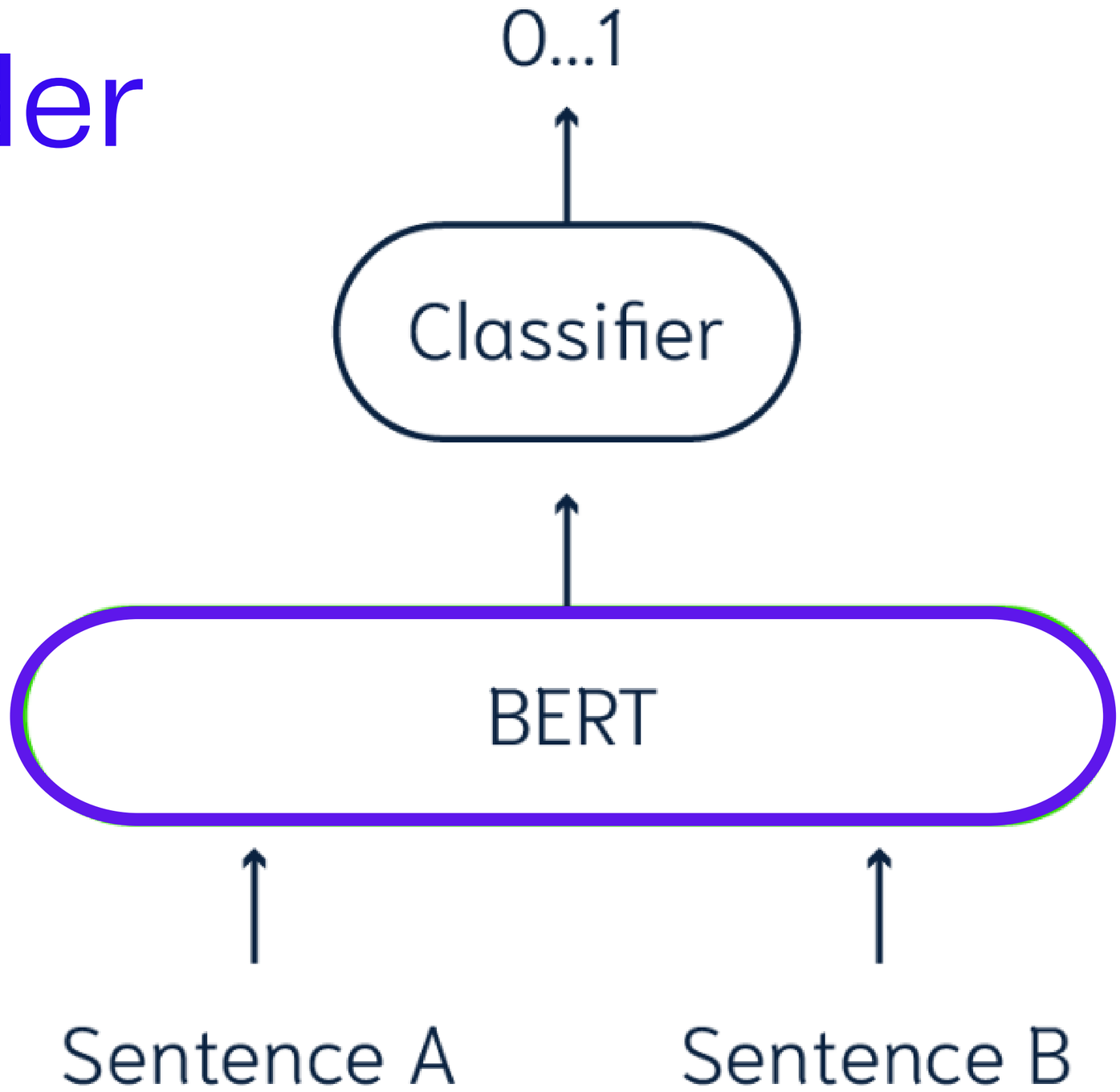

**All**, **positive**, **negative**

# Embeddings

# Cross-encoder

**[Wolf et al. 2019]**
TransferTransfo: A transfer learning approach for neural network-based conversational agents

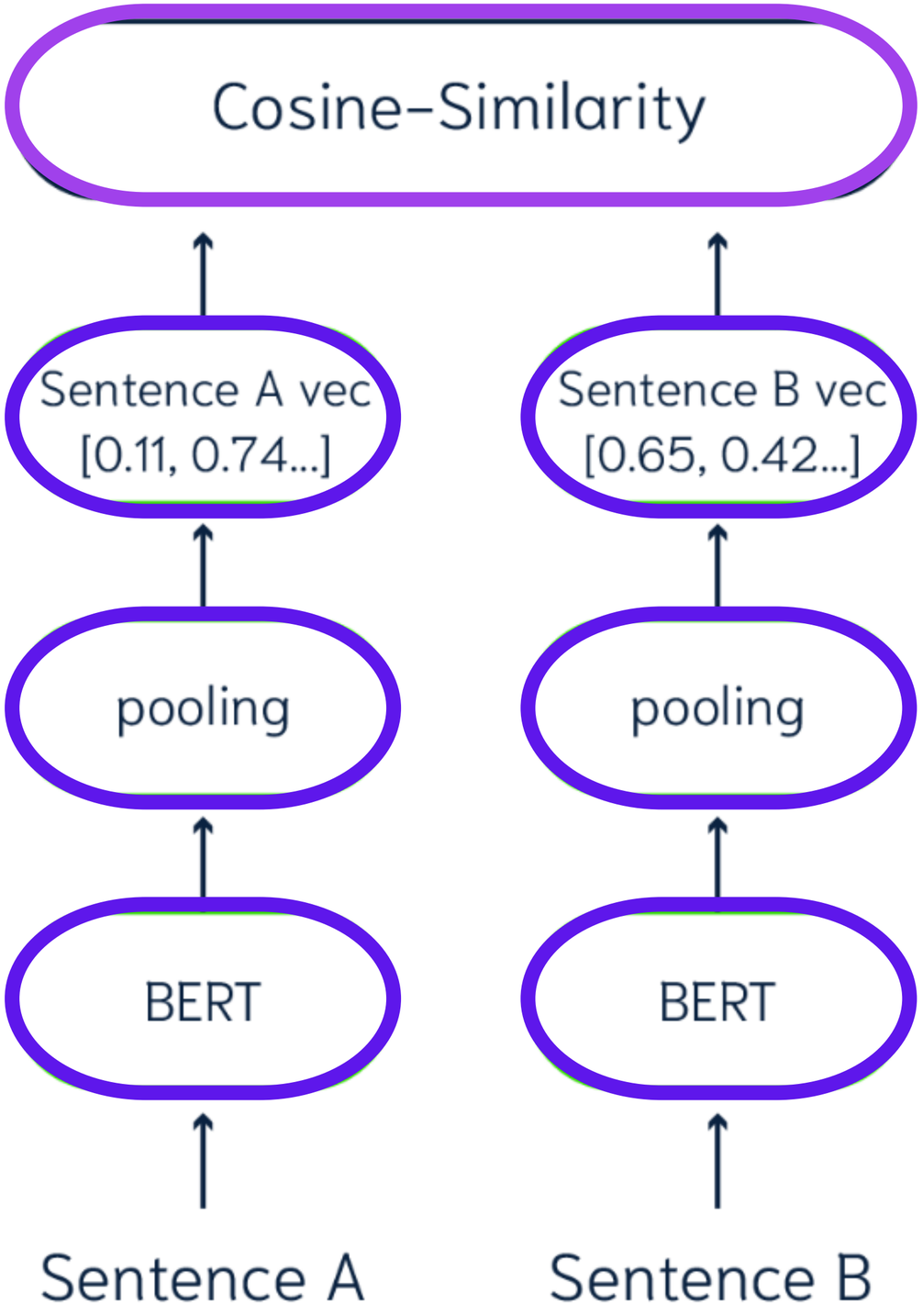
**– slower**
**– requires more data**
**– difficult to get embeddings of each sentence**
**+ more accurate**

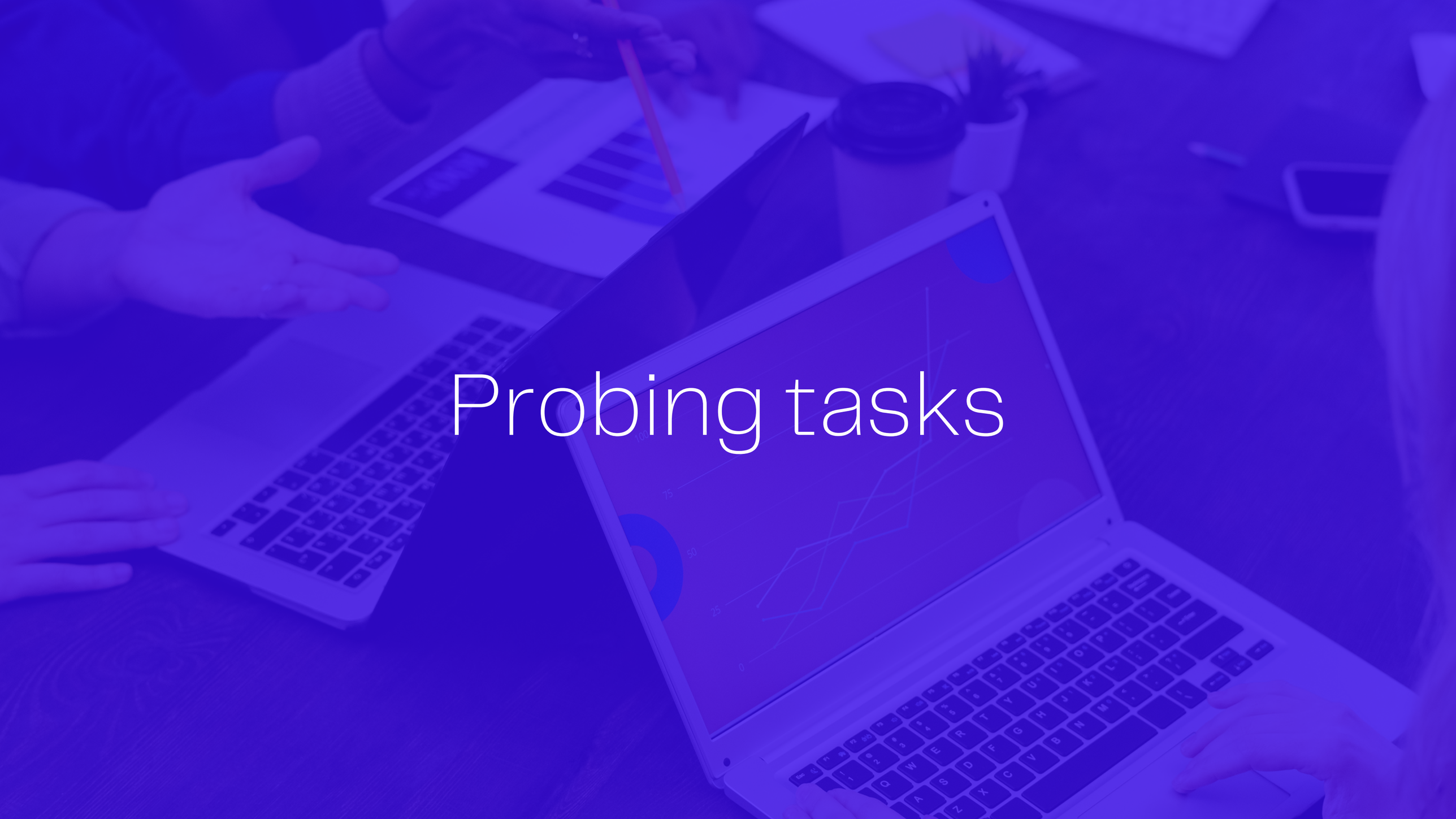


# Bi-encoders

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

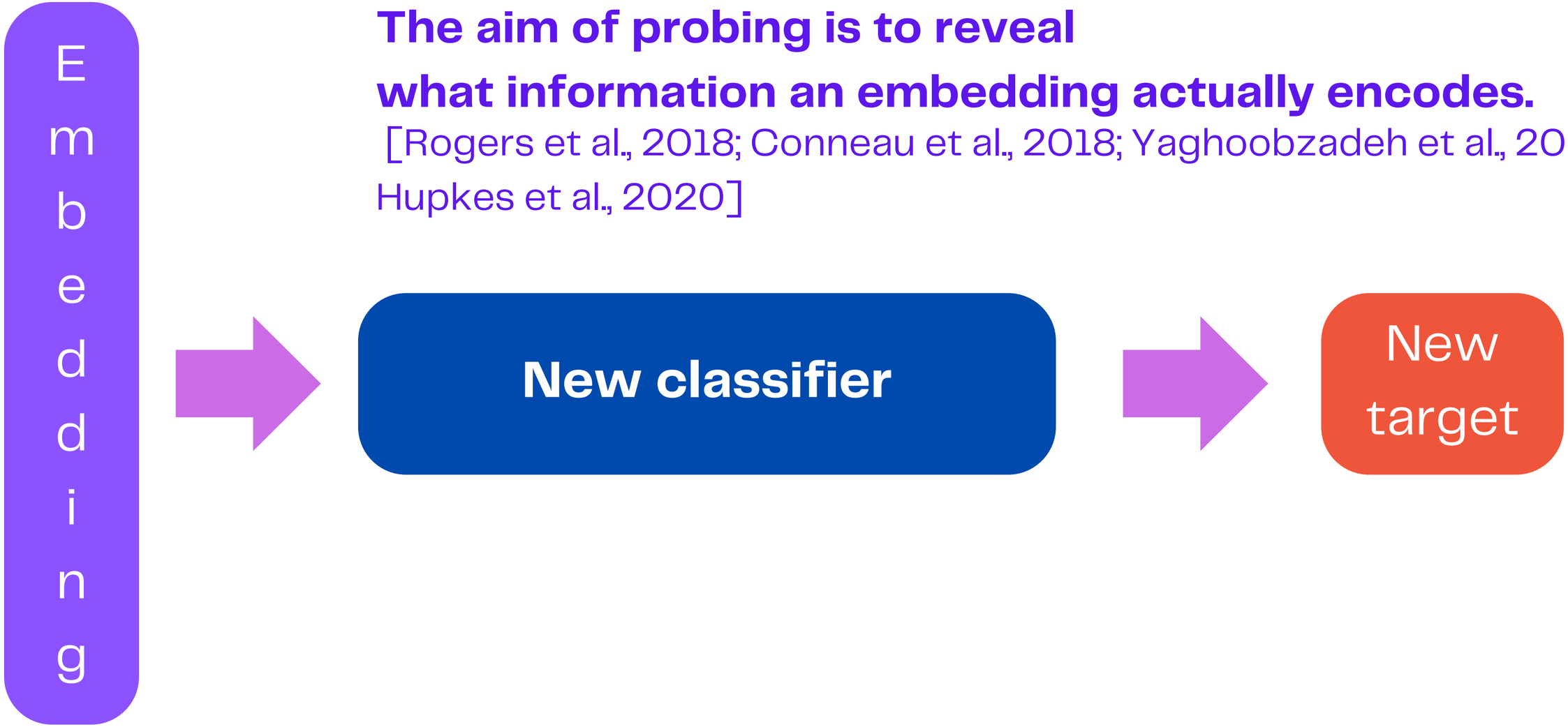
**– less accurate**
**+ easy to get the embeddings**
**+ sentence embeddings – an integral part of the architecture**
**+ faster**





# 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]

# Our inspiration

**[Lindstrom et al. 2020]** Probing Multimodal Embeddings for Linguistic Properties: the Visual-Semantic Case.

visual-semantic embeddings

1.1 A *child* holding a flowered umbrella and petting a yak.
1.2 A *checker* holding a flowered umbrella and petting a yak.

2.1 A young *man* holding an umbrella next to a herd of cattle.
2.2 A young *mime* holding an umbrella next to a herd of cattle.

3.1 a young *boy* holding an umbrella touching the horn of a cow.
3.2 a young *wad* holding an umbrella touching the horn of a cow.

4.1 A young *boy* with an umbrella who is touching the horn of a cow.
4.2 A young *bear* with an umbrella who is touching the horn of a cow.

5.1 A *boy* holding an umbrella while standing next to livestock.
5.2 A *fry* holding an umbrella while standing next to livestock.

Figure 2: In task *SemanticCongruence*, the objective is to recognise semantically implausible captions.

# Probing tasks in [Lindstrom et al. 2020]:

- **ObjectCategories –** which of the 80 MS-COCO **object categories** are **present** in a given image,
- **NumObjects –** to estimate the **number of objects** in an image,
- **SemanticCongruence –** whether a **caption** has been **modified**

# Aim of the project

Create and test new probing tasks:

## COMMON WORDS

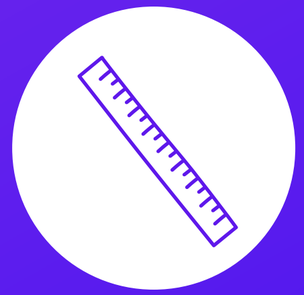
the presence of **common words** ('camera', 'len', 'digital') in the **title**

## BRAND NAME

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

## LEVENSHTEIN SIMILARITY

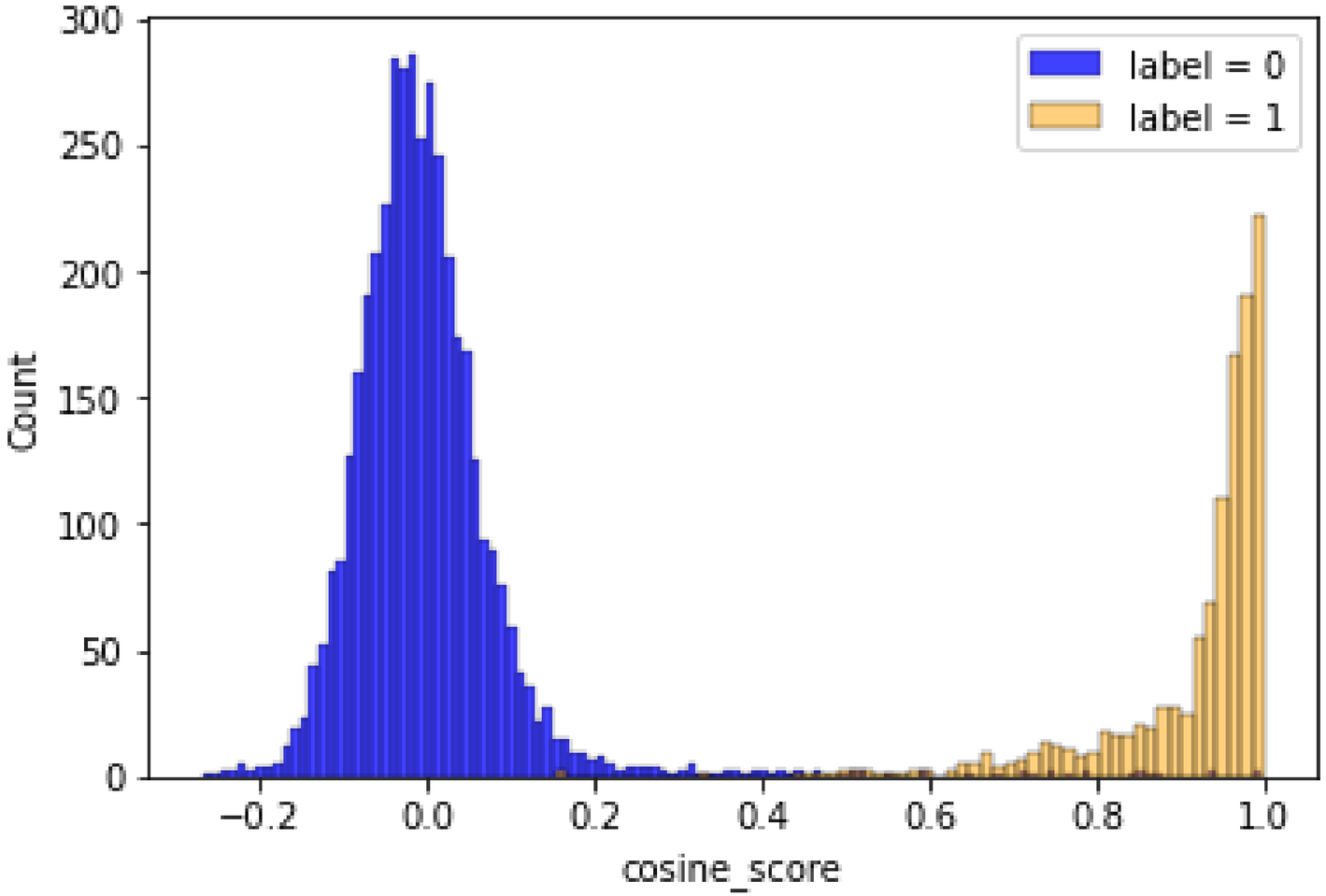
predicting **lev_dist(sentenceA, sentenceB)**



## LENGTH OF A SENTENCE

predicting the **length** of the input sentence

# Creating embeddings




## MODEL

bert-base-multilingual-uncased
fine-tune epochs: 200
batch_size: 16

## DATASET

WDC
category: "Cameras"
size: "medium"
features: "title" only

# Python libraries

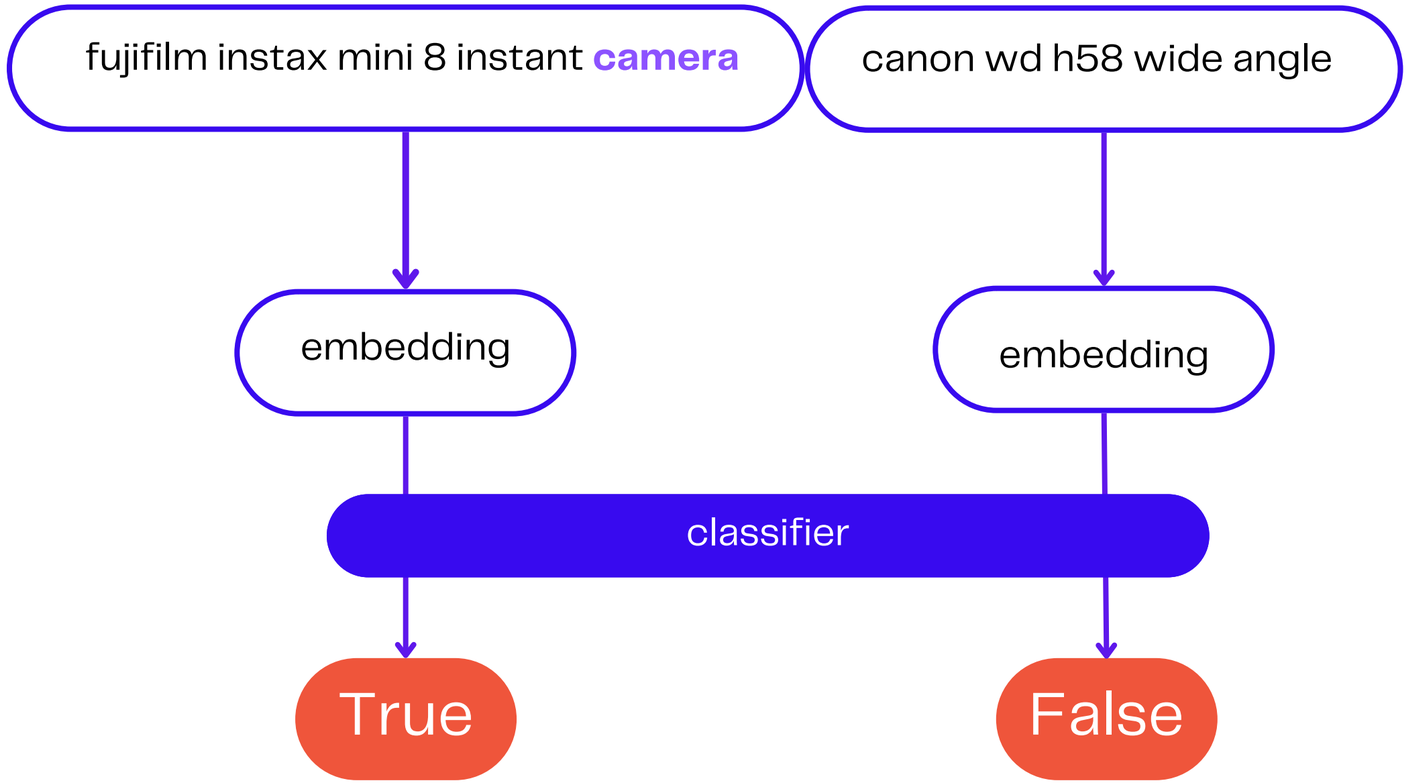

Hugging Face

# 1 Probing task – **common words**

# COMMON WORDS ©

## Influence of common words on embeddings

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

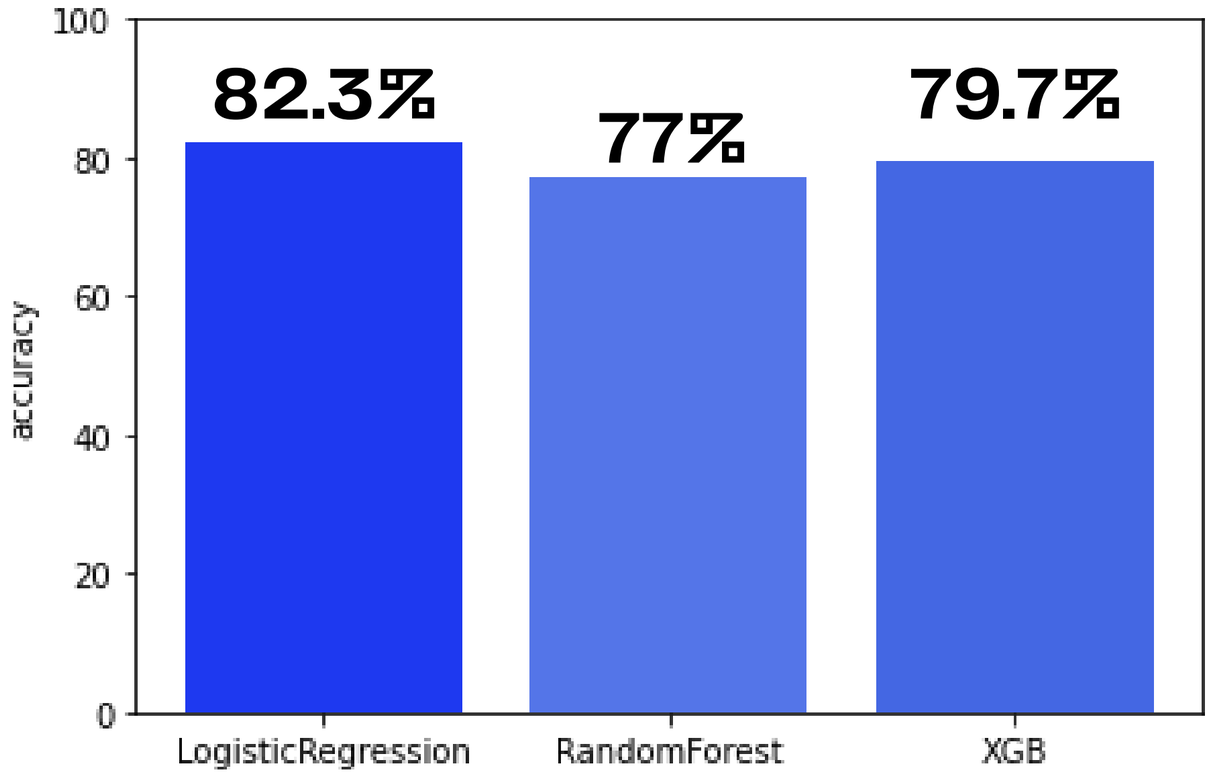


# COMMON WORDS ©

## Results

### Well balanced classes

| | |
|---|---|
| without common words | 1522 |
| with comon words | 1260 |

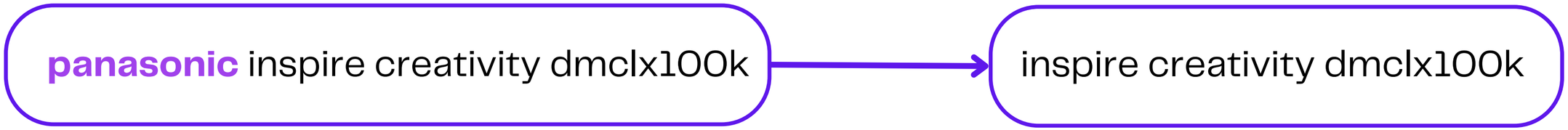


# 2 Probing task – **brand name**

# BRAND NAME

## New dataset

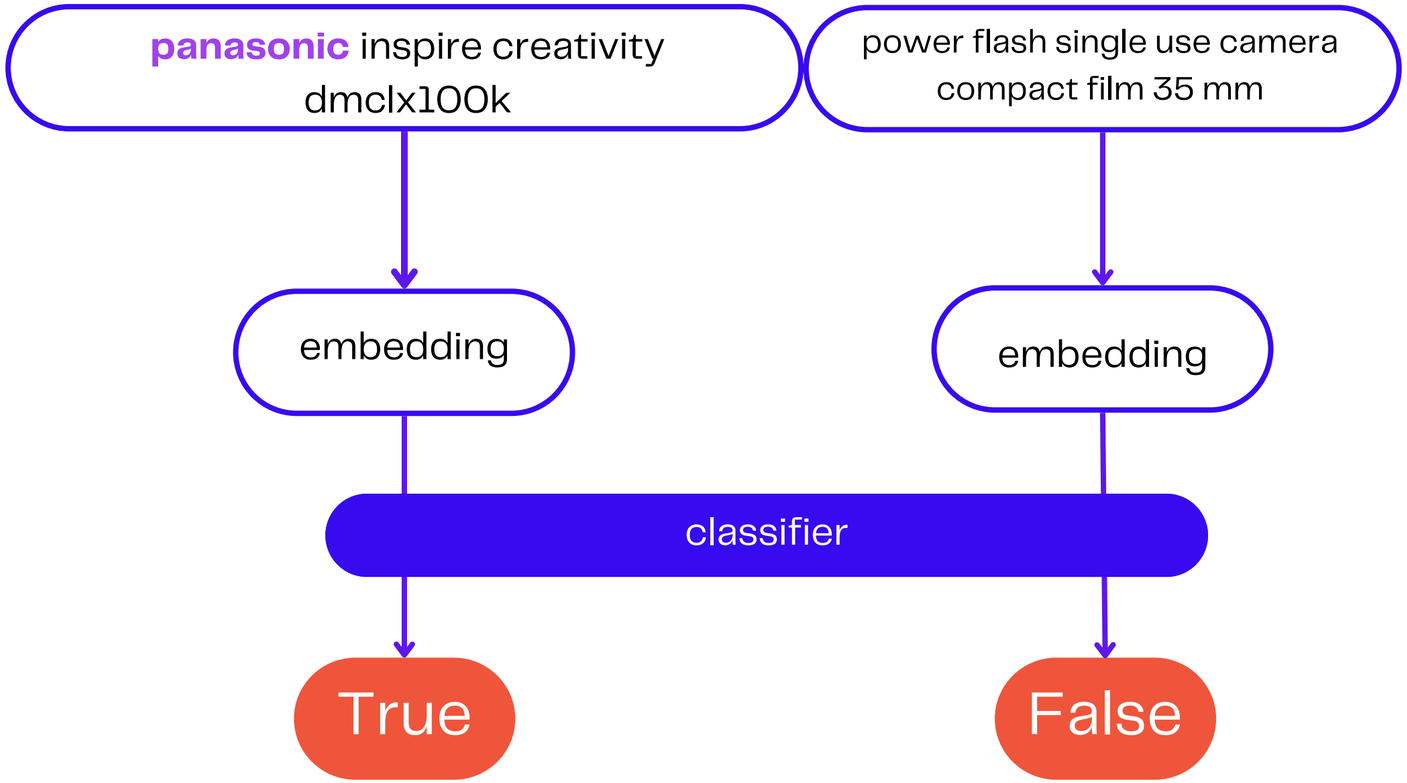
- without a brand name (Canon, Samsung, ...)
- balanced classes



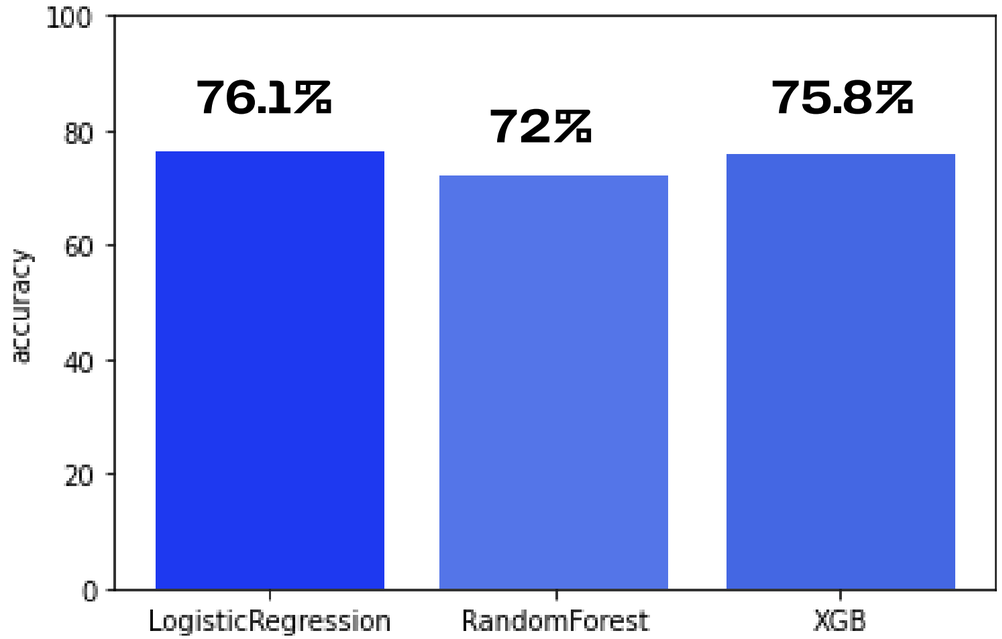

| | |
|---|---|
| without brand name | 1521 |
| with brand name | 1261 |

# BRAND NAME

## MODEL




## RESULTS



# 3 Probing task – **the Levenshtein distance**

# Levenshtein distance

String metric for measuring the difference between two sequences

# Idea

1. Calculate Levenshtein distance between offers & discretization.

**lev_dist("BrandX Camera MX140", "Camera BrandX MX 140") -> 13**

⋮

**lev_dist("BrandY S1000 Digital Camera", "BrandZ Camera PRO 3F") -> 19**

**Labels:**

| Similar | Quite Sililar | . . . | Not Similar |
|---|---|---|---|

**Discretization**

# Levenshtein distance

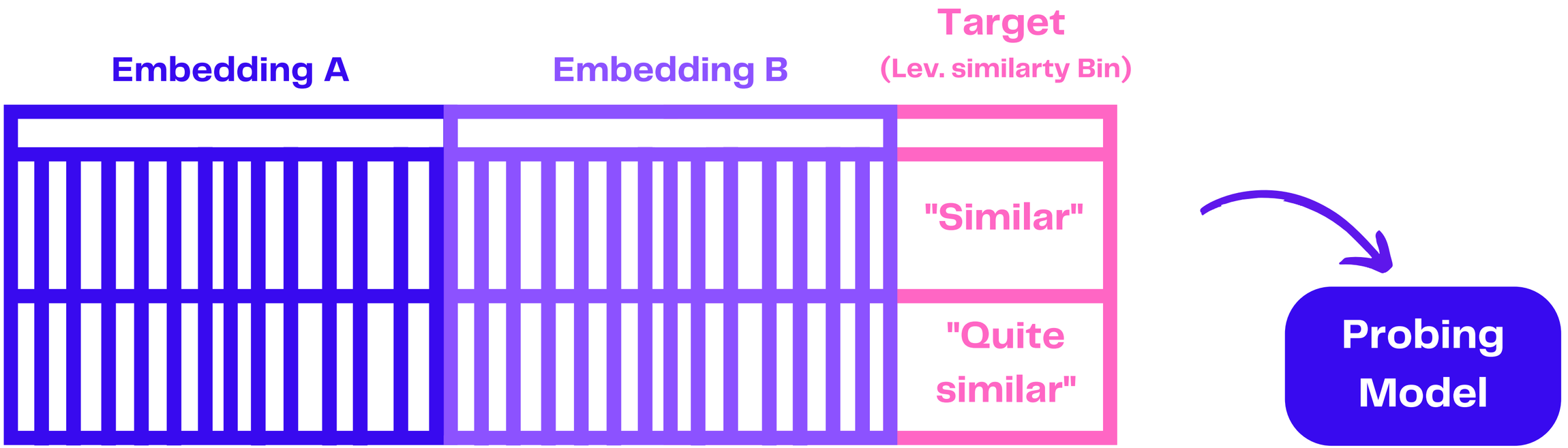
2. Classification model – Probing

Input: **Embeddings**

Output: **Lev. similarity bin**

**lev_dist("BrandX Camera MX140", "Camera BrandX MX 140") -> "Similar"**

**lev_dist("BrandY S1000 Digital Camera", "BrandZ Camera PRO 3F") -> "Quite Similar"**



# Levenshtein distance

**5 classes (Levenshtein distance – discretized):**

**[Similar, Quite similar, Neutral,  Hardly similar, Not Similar]**

## Results

**LogisticRegression:**

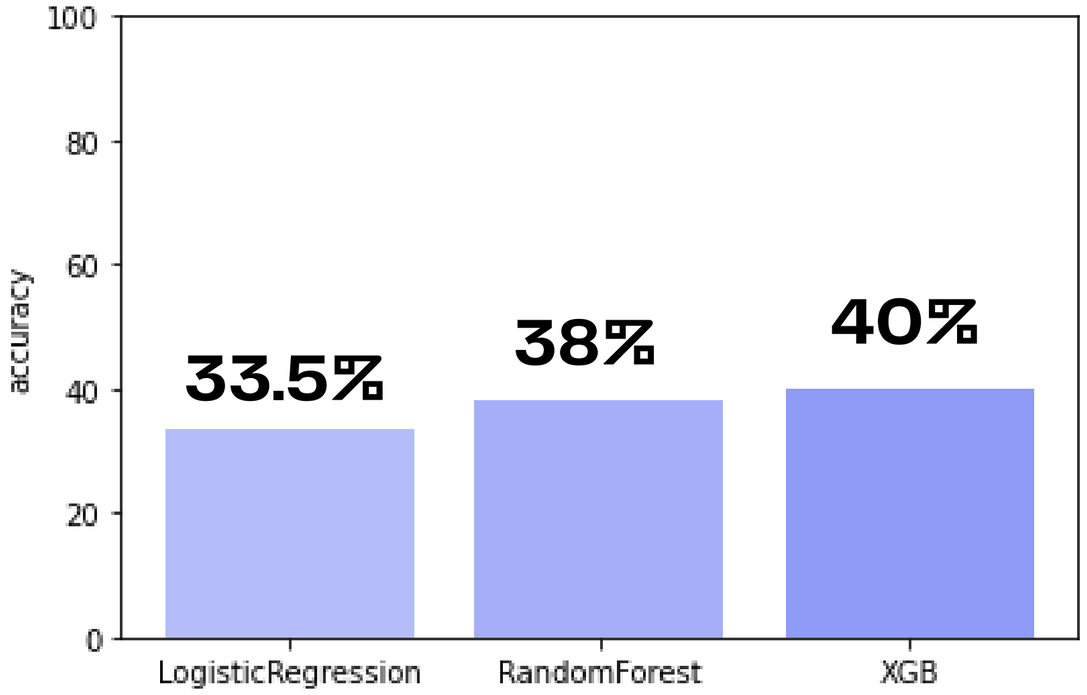
**Accuracy: 33.5%,**
**f_score: 0.10**

**RandomForest:**

**Accuracy: 38%,**
**f_score: 0.29**

**XGBoost:**

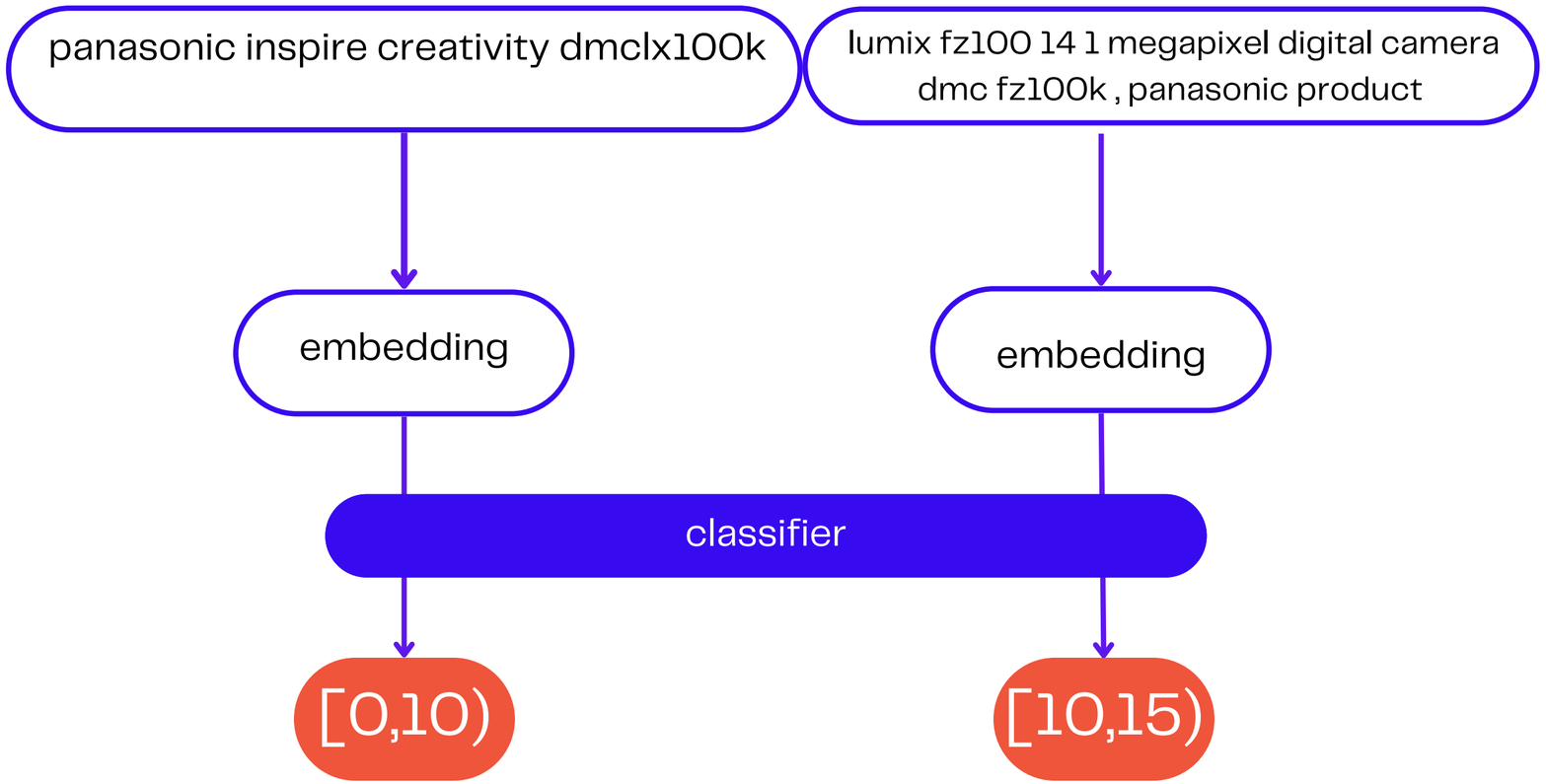
**Accuracy: 40%,**
**f_score: 0.34**



# 4 Probing task –
# **Length of the sentence**

# Length of the sentence

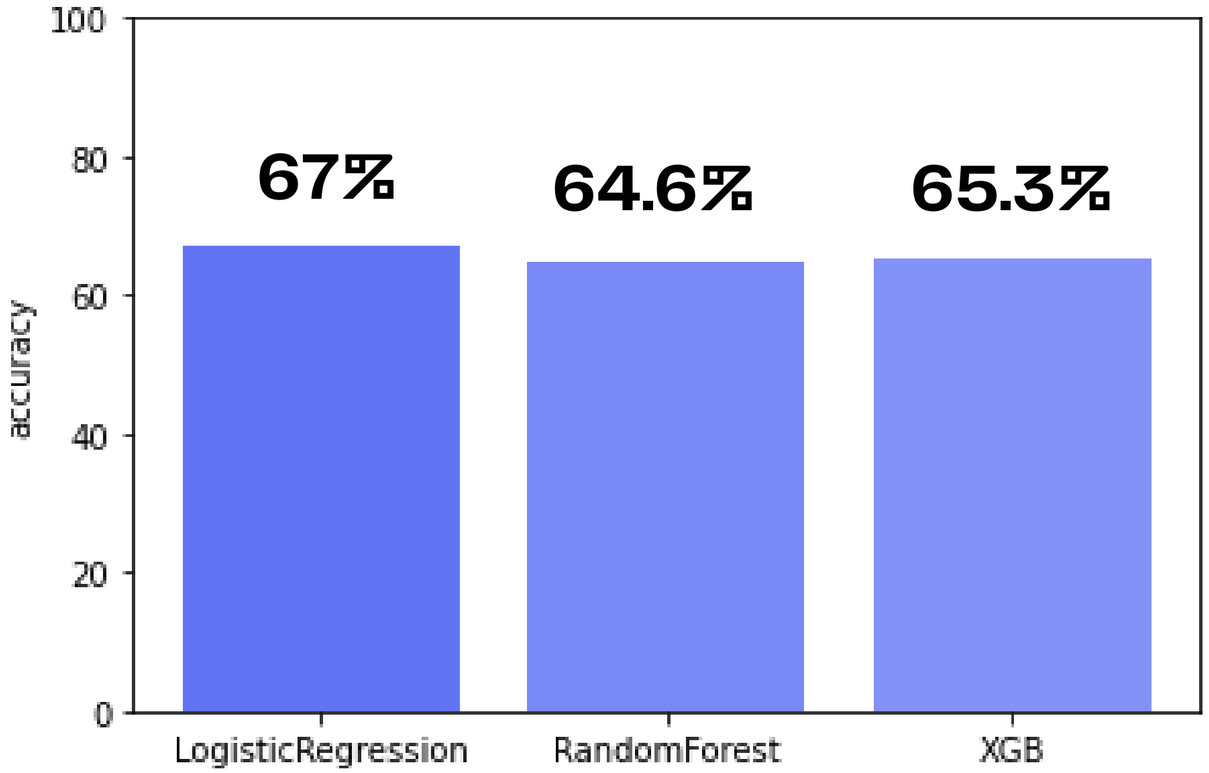
Influence of length of the sentence on embeddings.




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

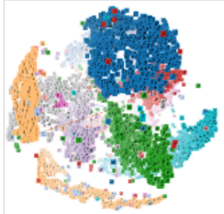
# Length of the sentence

## Results




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

# EMBEDDINGS VISUALIZATION



**Embedding projector - visualization of high-dimensional data**

Visualize high dimensional data.

tensorflow.org /

# References:


[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