

Matching

Lecture 5: NLP models in matching

Gleb Drozdov

Moscow Institute of Physics and Technology

Autumn 2023

Where are the texts?

Электроника > Телефоны и смарт-часы > Фитнес-браслеты > Xiaomi

Фитнес-браслет Xiaomi mi band 7 CN, черный

★★★★★ 89 отзывов 3 видео 22 вопроса В избранное Добавить к сравнению Поделиться

Xiaomi	
Оригинальный товар	i
Тип	Фитнес-браслет
Совместимые платформы	Android
Сенсорный экран	Да
Датчики	Пульсометр, Шагомер, Датчик уровня кислорода в крови
Мониторинг	Сна, Сердечного ритма, Уровня стресса
Защищенность	Водонепроницаемость
Время работы от аккумулятора, дни	14
Диагональ экрана, дюймы	i 1.62
Вид дисплея	AMOLED
Беспроводные интерфейсы	i Bluetooth

Перейти к описанию

Электроника > Телефоны и смарт-часы > Фитнес-браслеты > Xiaomi

Смарт-браслет Xiaomi Mi Smart Band 7 black (BHR6008GL)

★★★★★ 107 отзывов 21 вопрос В избранное Добавить к сравнению Поделиться

Xiaomi	
Оригинальный товар	i
Тип	Фитнес-браслет
Диагональ экрана, дюймы	i 1.62
Бренд	Xiaomi
Цвет	Черный
Материал браслета/ремешка	Гипоаллергенный силикон
Материал корпуса	Пластик
Страна-изготовитель	Китай
Гарантийный срок	i 12
Партномер	BHR6008GL
Цвет ремешка	Черный

Перейти к описанию

Names

Электроника > Телефоны и смарт-часы > Фитнес-браслеты > Xiaomi

Фитнес-браслет Xiaomi mi band 7 CN, черный

★★★★★ 57 отзывов 3 видео 22 вопроса в избранное + добавить к сравнению Поделиться



Xiaomi Mi Band 7

-62% Оригинал Хит

Xiaomi Оригинальный товар

Тип	Фитнес-браслет
Совместимые платформы	Android
Сенсорный экран	Да
Датчики	Пульсометр, Шагомер, Датчик уровня кислорода в крови
Мониторинг	Сна, Сердечного ритма, Уровня стресса
Защищенность	Водонепроницаемость
Время работы от аккумулятора, дни	14
Диагональ экрана, дюймы	1.62
Вид дисплея	AMOLED
Беспроводные интерфейсы	Bluetooth

Перейти к описанию

Электроника > Телефоны и смарт-часы > Фитнес-браслеты > Xiaomi

Смарт-браслет Xiaomi Mi Smart Band 7 black (BHR6008GL)

★★★★★ 107 отзывов 21 вопрос в избранное + добавить к сравнению Поделиться



-13% Оригинал Хит

Xiaomi Оригинальный товар

Тип	Фитнес-браслет
Диагональ экрана, дюймы	1.62
Бренд	Xiaomi
Цвет	Черный
Материал браслета/ремешка	Гипоаллергенный силикон
Материал корпуса	Пластик
Страна-изготовитель	Китай
Гарантийный срок	12
Партномер	BHR6008GL
Цвет ремешка	Черный

Перейти к описанию

Attributes

Электроника > Телефоны и смарт-часы > Фитнес-браслеты > Xiaomi

Фитнес-браслет Xiaomi mi band 7 CN, черный

★★★★★ 89 отзывов 3 видео 22 вопроса В избранное Добавить к сравнению Поделиться



Xiaomi	Оригинальный товар
Тип	Фитнес-браслет
Совместимые платформы	Android
Сенсорный экран	Да
Датчики	Пульсометр, Шагомер, Датчик уровня кислорода в крови
Мониторинг	Сна, Сердечного ритма, Уровня стресса
Защищенность	Водонепроницаемость
Время работы от аккумулятора, дни	14
Диагональ экрана, дюймы	1.62
Вид дисплея	AMOLED
Беспроводные интерфейсы	Bluetooth

[Перейти к описанию](#)

Электроника > Телефоны и смарт-часы > Фитнес-браслеты > Xiaomi

Смарт-браслет Xiaomi Mi Smart Band 7 black (BHR6008GL)

★★★★★ 107 отзывов 21 вопрос В избранное Добавить к сравнению Поделиться



Xiaomi	Оригинальный товар
Тип	Фитнес-браслет
Диагональ экрана, дюймы	1.62
Бренд	Xiaomi
Цвет	Черный
Материал браслета/ремешка	Гипоаллергенный силикон
Материал корпуса	Пластик
Страна-изготовитель	Китай
Гарантийный срок	12
Партномер	BHR6008GL
Цвет ремешка	Черный

[Перейти к описанию](#)

Descriptions

Электроника > Телефоны и смарт-часы > Фитнес-браслеты > Xiaomi

Фитнес-браслет Xiaomi mi band 7 CN, черный

★★★★★ 89 отзывов 3 видео 22 вопроса В избранное Добавить к сравнению Поделиться



Xiaomi Оригинальный товар

-62% Оригинал Хит

Xiaomi Mi Band 7

Перейти к описанию

Тип	Фитнес-браслет
Совместимые платформы	Android
Сенсорный экран	Да
Датчики	Пульсометр, Шагомер, Датчик уровня кислорода в крови
Мониторинг	Сна, Сердечного ритма, Уровня стресса
Защищенность	Водонепроницаемость
Время работы от аккумулятора, дни	14
Диагональ экрана, дюймы	1.62
Вид дисплея	AMOLED
Беспроводные интерфейсы	Bluetooth

Электроника > Телефоны и смарт-часы > Фитнес-браслеты > Xiaomi

Смарт-браслет Xiaomi Mi Smart Band 7 black (BHR6008GL)

★★★★★ 107 отзывов 21 вопрос В избранное Добавить к сравнению Поделиться



Xiaomi Оригинальный товар

-13% Оригинал Хит

Перейти к описанию

Тип	Фитнес-браслет
Диагональ экрана, дюймы	1.62
Бренд	Xiaomi
Цвет	Черный
Материал браслета/ремешка	Гипоаллергенный силикон
Материал корпуса	Пластик
Страна-изготовитель	Китай
Гарантийный срок	12
Партномер	BHR6008GL
Цвет ремешка	Черный

Descriptions

Are good sometimes

Фитнес-браслет Xiaomi Mi Smart Band 7, смарт часы, умные фитнес часы

★★★★★ 454 отзыва 72 вопроса Сравнить

XIAOMI MI BAND 7

219 ₽ 837 ₽ -74% **РАСПРОДАЖА**

Полноэкранные защитное стекло для смарт-часов...
★ 4.6 ⚡ 176

593 ₽ 1590 ₽ -63% **РАСПРОДАЖА**

Беспроводные наушники Pods Pro с микрофоном...
★ 4.6 ⚡ 2049

3 466 ₽ 5 990 ₽ -42% **РАСПРОДАЖА**

Наушники Redmi Buds 4 White (BHR5846GL)
Осталось 3 шт ★ 4.8 ⚡ 208

jack 3.5 ⚡ **РАСПРОДАЖА**

297 ₽ 113 ₽ -73% **РАСПРОДАЖА**

Осталось 3 шт **РАСПРОДАЖА**

3 452 ₽ 4 990 ₽ -31% **РАСПРОДАЖА**

Беспроводные наушники OPPO Enco Air 2, белые
Осталось 397 шт ★ 4.8 ⚡ 214

211 ₽ 399 ₽ -47% **РАСПРОДАЖА**

Ремешок для браслета Xiaomi Mi Band 7 Ceramic
Осталось 13907 шт ★ 4.8 ⚡ 771

Добавить в корзину

Послезавтра

Послезавтра

Послезавтра

Послезавтра

За час

Последний

XiaMall Техника и гаджеты

Добавить в корзину

Фитнес-браслет Xiaomi mi band 7 CN, черный

★★★★★ 97 отзывов 23 вопроса Сравнить

Осталось 50 шт **Пленка защитная Ceramic** 104 ₽ / шт **Осталось 22 шт**

Быстрое зарядное устройство 18W QC3.0 для... ★ 4.5 ⚡ 108 **Ремешок для фитнес - браслета Xiaomi Mi Band ...** ★ 4.7 ⚡ 186

★★ 4.6 ⚡ 730 **Пленка защитная Ceramic** 104 ₽ / шт **Осталось 22 шт**

★★ 4.7 ⚡ 2259 **Задняя пленка для Xiaomi Mi Smart Band 7...** ★ 4.7 ⚡ 2259

20 октября

Послезавтра

20 октября

Послезавтра

Послезавтра

Послезавтра

Послезавтра

Послезавтра

Добавить в корзину

Описание

• Основные характеристики

Размеры: 46,5 мм x 20,7 мм x 12,25 мм

Дисплей: Сенсорный дисплей AMOLED диагональю 1,62 дюйма

Датчики: Высокоточный 6-осевой датчик и 6-осевой датчик частоты сердечных сокращений ФПГ: 3-осевой акселерометр и 3-осевой гироскоп с низким энергопотреблением, датчик ФПГ для измерения частоты сердечных сокращений

Аккумулятор: Стандартное время использования: ≥14 дн; Ёмкость аккумулятора: 180 мАч

Фитнес: 10+ режимов тренировок; 5 режимов автоматического определения: бег на открытом воздухе, ходьба, беговая дорожка, гребля, эллиптический тренажер

Здоровье: Мониторинг частоты сердечных сокращений

Отслеживание уровня кислорода в крови (SpO₂)

Технические характеристики: Водонепроницаемость 5 атм ;Bluetooth 5.2 BLE

Операционная система: Android 6.0 или выше, iOS 10.0 или выше

Описание

Фитнес-браслет Xiaomi Mi Band 7 оснащен большим AMOLED-дисплеем с диагональю 1,62 дюйма, который занимает еще больше места на корпусе, четко показывая не только время, но и сообщения, вызовы, уведомления и спортивные данные.

Функция постоянного наблюдения за качеством сна и дыхания в Mi Band 7 наблюдает за изменением сатурации, а также анализирует качество дыхания во время сна.

Встроенный в Mi Band 7 биодатчик PPG круглосуточно наблюдает за сердцебиением и его изменениями. А в случае обнаружения аномалии он автоматически выдает предупреждение с вибрацией.

Xiaomi Mi Band 7 стойко выдерживает давление воды до 5 атмосфер, так что даже любители погружаться могут отправляться с ним на исследования подводных красот. Фитнес-браслет Mi Band 7 может автоматически распознавать до 120 видов активности, позволяя зафиксировать каждое достижение без лишних сложностей

Одного заряда батареи Mi Band 6 хватит на 15 дней стандартной работы, а при активном использовании заряда хватит на 9 дней.

Комплектация

Фитнес-браслет Xiaomi Mi Band 7 - 1 шт., руководство пользователя - 1 шт., магнитная зарядка - 1 шт.

Description

Are noisy mostly

 **Фитнес-браслет Xiaomi Mi Band 7, черный**
★★★★★ 255 отзывов [?](#) 35 вопросов [≡+](#) Сравнить

стекло для смарт-часов... [★★ 4.6 ⏱ 176](#) [Послезавтра](#) 

Защитная гидрогелевая пленка для смарт часов... [★★ 4.7 ⏱ 2788](#) [Послезавтра](#)

Доска с гвоздями для ног [★★ 4.9 ⏱ 44](#) [Послезавтра](#)

Садху для начинающих,... [★★ 4.9 ⏱ 44](#) [Послезавтра](#)

стекло для смарт-часов... [★★ 4.6 ⏱ 513](#) [Послезавтра](#) 

фитнес брас... [★★ 4.8 ⏱ 5797](#) [Послезавтра](#)

Описание

Модель M2129B1

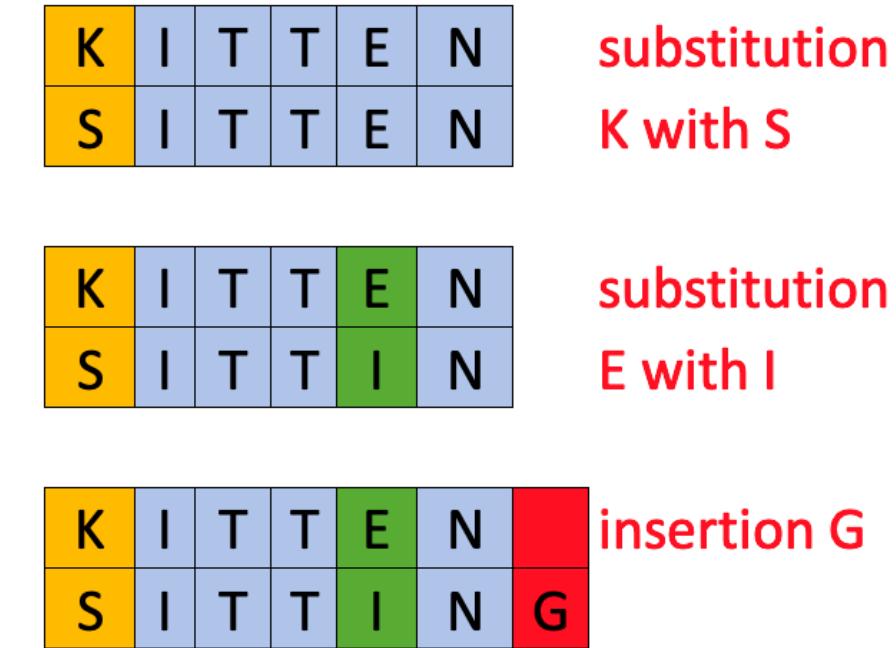
Комплектация

Фитнес-браслет - 1 шт, Силиконовый ремешок - 1 шт, Зарядное устройство - 1 шт, Инструкция - 1 шт

Simple algorithms on strings

Levenstein distance

- Minimum number of operations required to change one word into the other
- Operations: insert/delete/substitute.
- $\text{lev_dist}(\text{«kitten»}, \text{«sitting»}) = 3$
- $\text{lev_dist_normalized}(s, t) = \frac{\text{lev_dist}(s, t)}{\max(|s|, |t|)}$;



Ratcliff-Obershelp algorithm

1. «I love to eat apple». , «I do not like to eat pineapple.»
2. Find the Longest Common Subsequences (LCS).
3. We get all the matching blocks (contain the length of all the matches) as we can see in the output `Match(a=0, b=0, size=2)`, `Match(a=2, b=9, size=1)`, `Match(a=5, b=12, size=9)`, `Match(a=14, b=25, size=6)`.
4. We sum up the match sizes, $\text{matches} = 2 + 1 + 9 + 6$, which equals 18.
5. The total length of both the sequences will come out to be 51 as it is shown in the match output `Match(a=20, b=31, size=0)`.
6. Ratio = $2 * 18 / 51$

Others

- [thefuzz](#)
- [difflib](#)

Simple algorithms on strings

Pros:

- Straightforward and easily implemented
- Fast

Cons:

- Too naive, especially for long strings.

String A = «Фитнес-браслет Xiaomi mi band 7 CN, черный»

String B = «Смарт-браслет Xiaomi Mi Smart Band 7 black (BHR6008GL)»

lev_dist_norm = 0.55

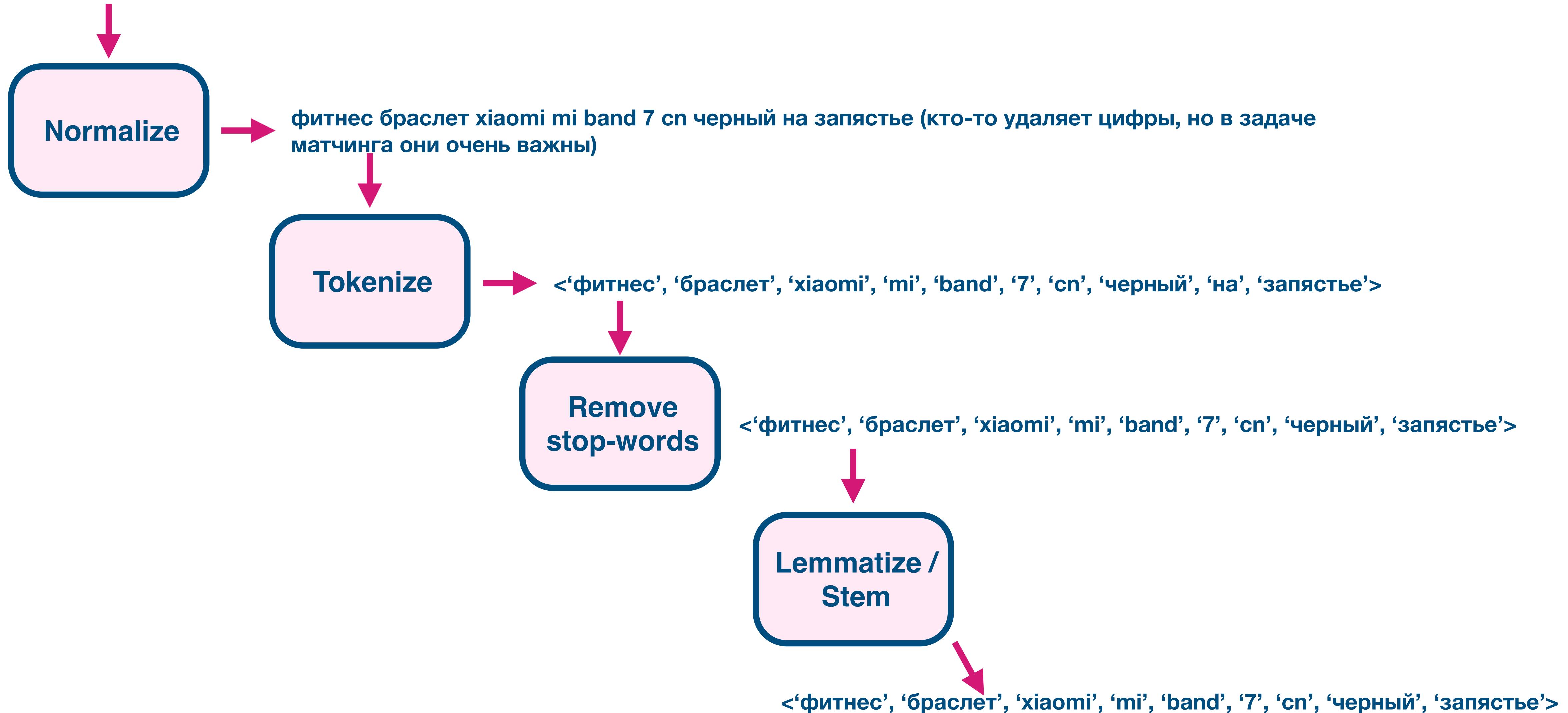
Ratcliff-Obershelp = 0.45

Algorithms on tokens

1. Jaccard e. t. c.
2. BOW
3. TF-IDF
4. BM-25

Text Preprocessing

Фитнес-браслет Xiaomi mi band 7 CN, черный, на запястье



Text Preprocessing

Stemming vs Lemmatization

Stemming

- Process of reducing a word to its base/stem form.
 - May not consider context and might not return an actual word.
 - Typically faster as it employs algorithmic methods.
 - Example: "running" → "run", "flies" → "fli".
-

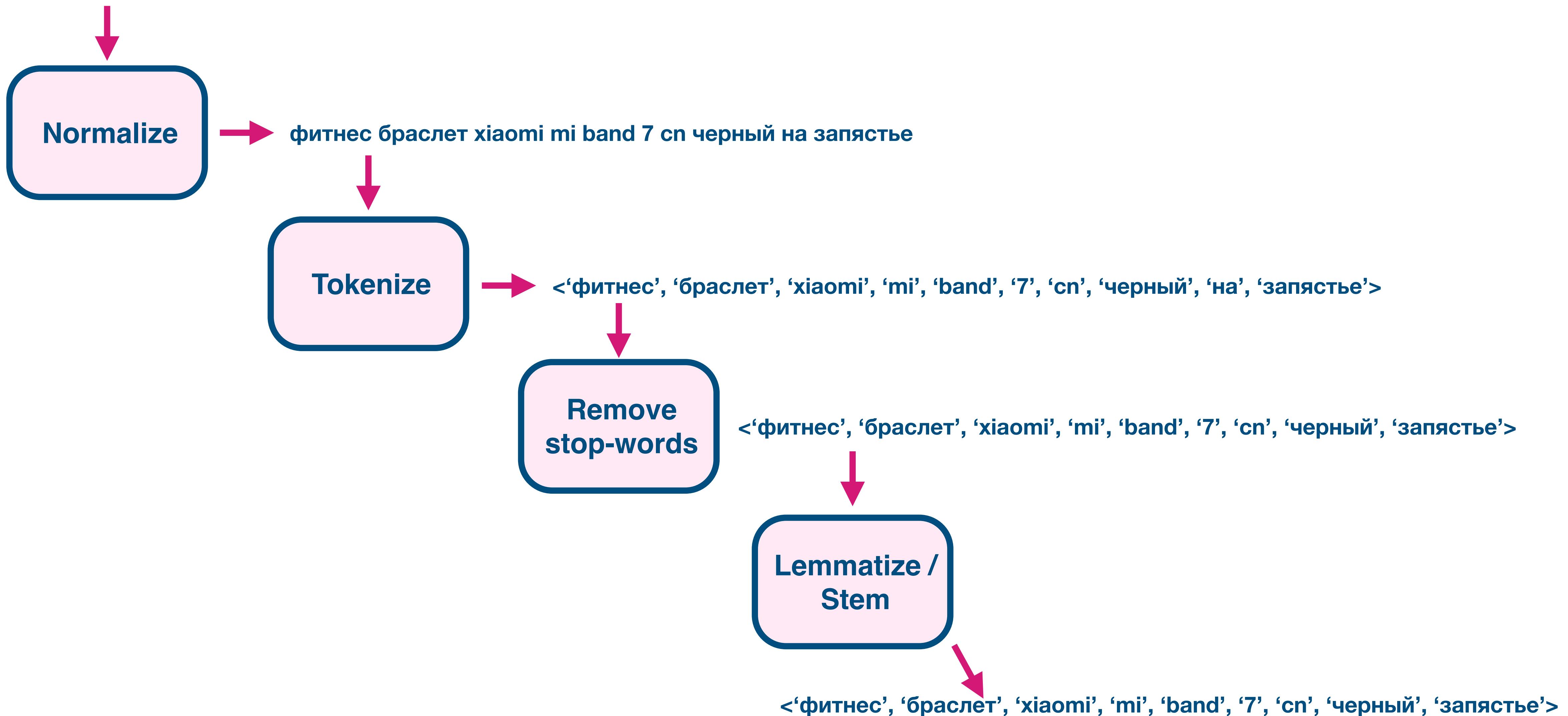
- Купальник -> куп
- Купивший -> куп

Lemmatization

- Process of converting a word to its base or dictionary form
 - Considers context and uses morphological analysis.
 - Generally slower as it might use dictionaries and databases.
 - Example: "running" → "run", "flies" → «fly».
-

- Купальник -> Купальник
- Купивший -> купить

Features on tokens



Algorithms on tokens

Jaccard index

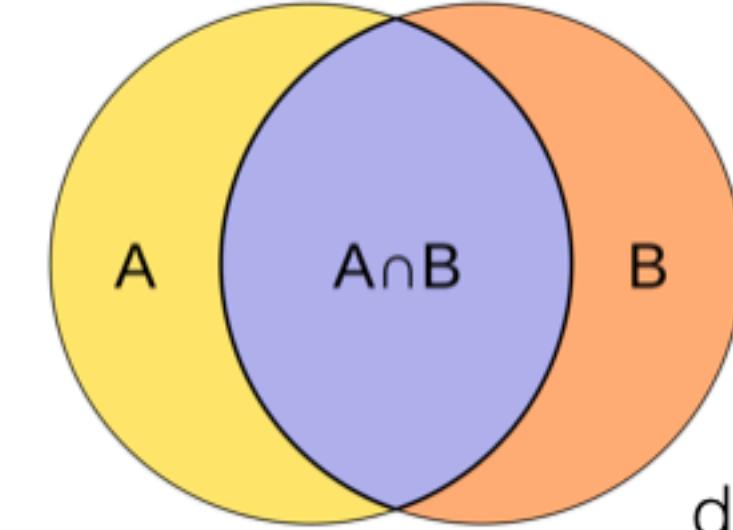
- Jaccard index

- quantify the dissimilarity between two sets

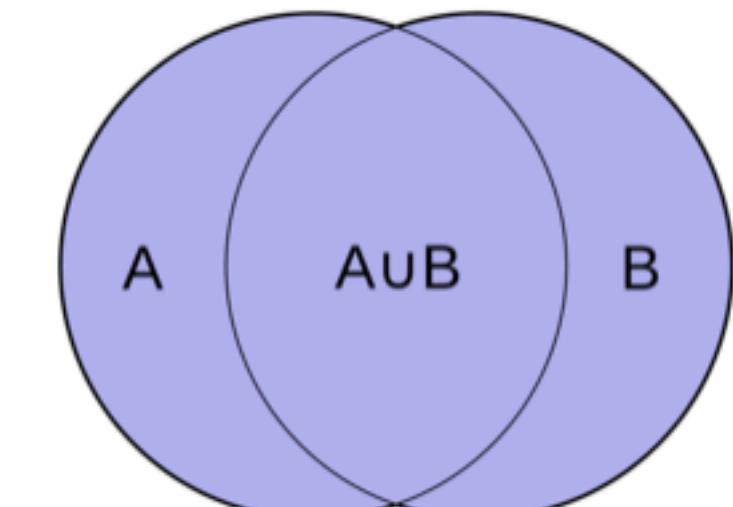
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

J(A,B) =

The intersect of A & B



The union of A & B



- Overlap coefficient

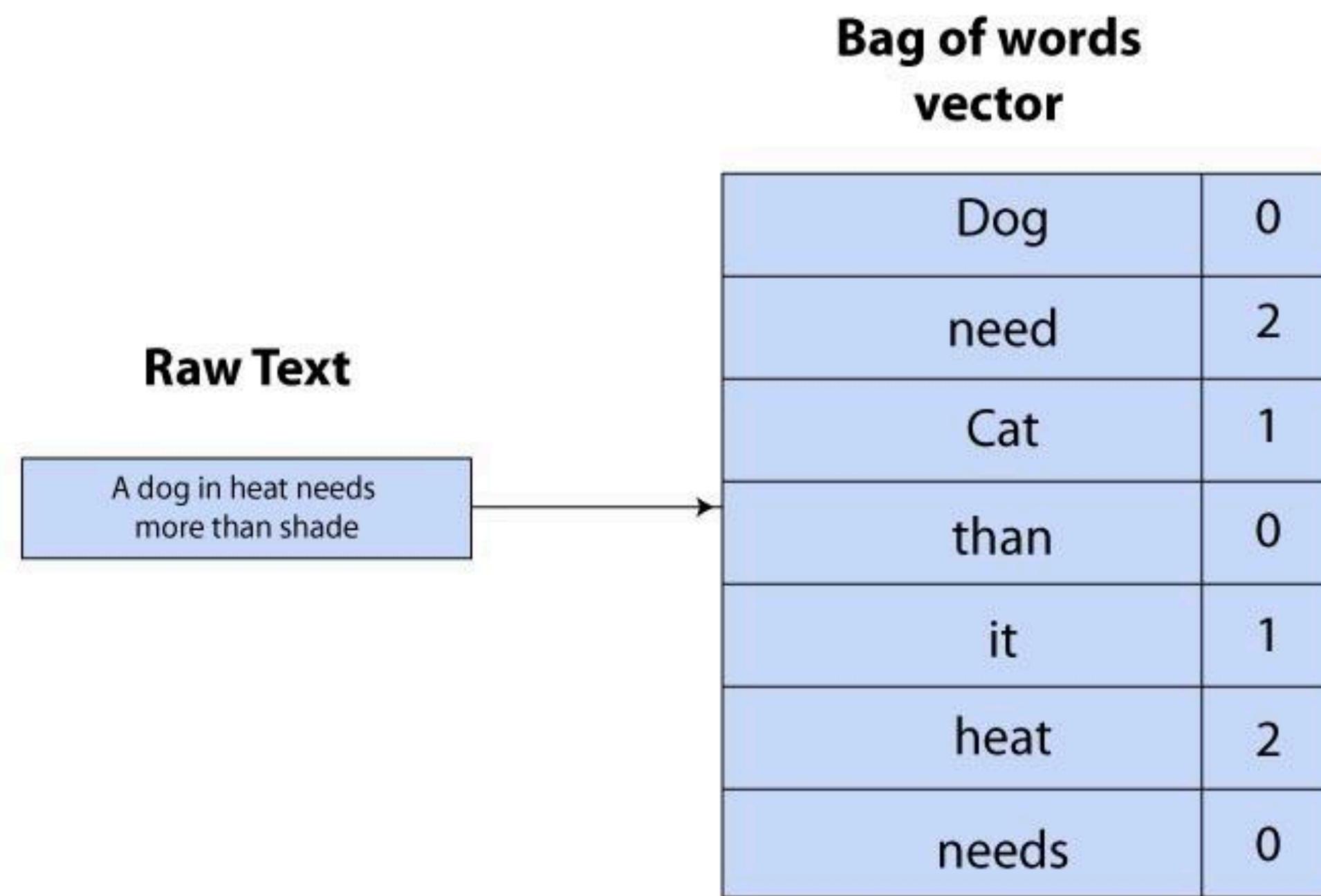
- it measures the proportion of A that is shared with B

$$J_{\text{left}}(A, B) = \frac{|A \cap B|}{|A|}$$

Algorithms on tokens

Bag of words (BOW)

- Represents text as a collection of individual words and their frequencies.
- Ignores word order and grammar;
- Features
 - calculate the cosine similarity between embeddings
 - feed embeddings into gradient boosting



Algorithms on tokens

TF-IDF

- Define the importance of a keyword or phrase within a document
- Differentiates important words from frequent but non-informative words
- Features
 - Cosine distance on names and descriptions

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

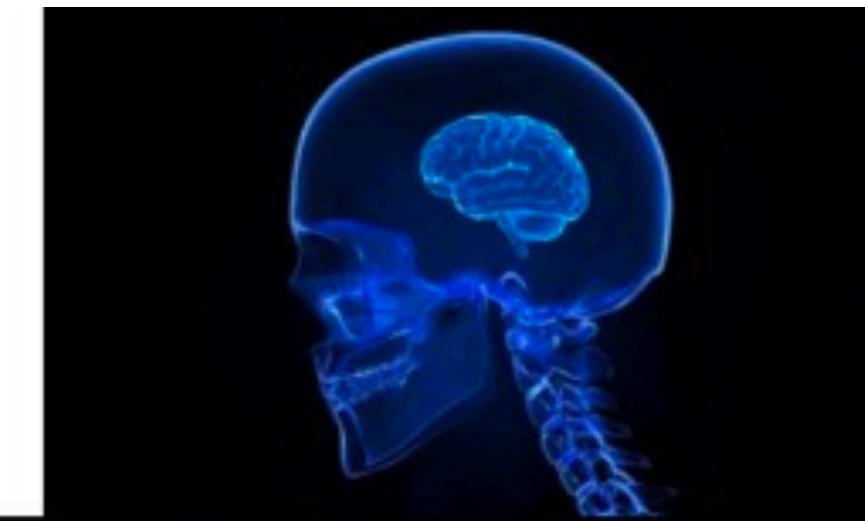
df_x = number of documents containing x

N = total number of documents

text_features is all you need

- 1. BOW**
- 2. Naive Bayes**
- 3. BM-25**

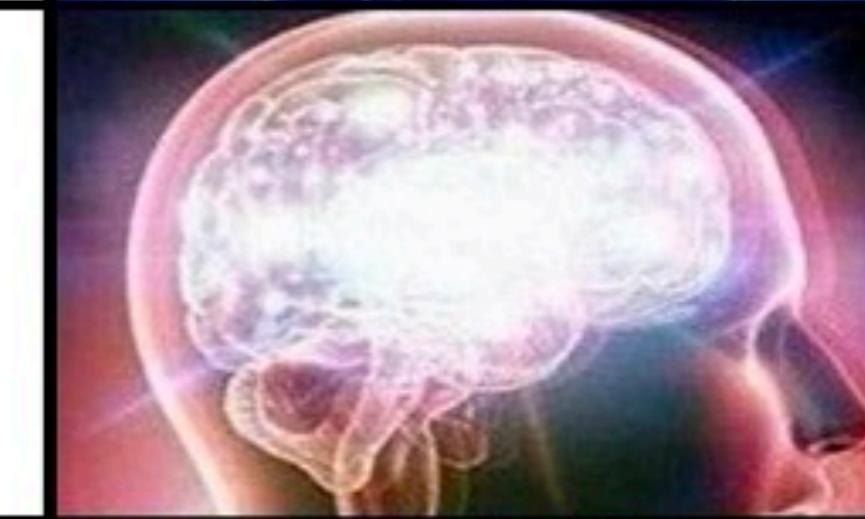
**СОБИРАТЬ
ЧИСТЫЙ ДАТАСЕТ**



ОБУЧАТЬ ВЕРТ



ИСПОЛЬЗОВАТЬ ЧАТГРТ



CATBOOST:TEXT_FEATURES



Algorithms on tokens

Pros:

- Do not require high computational power
- Train quickly
- Show high quality for most tasks

Cons:

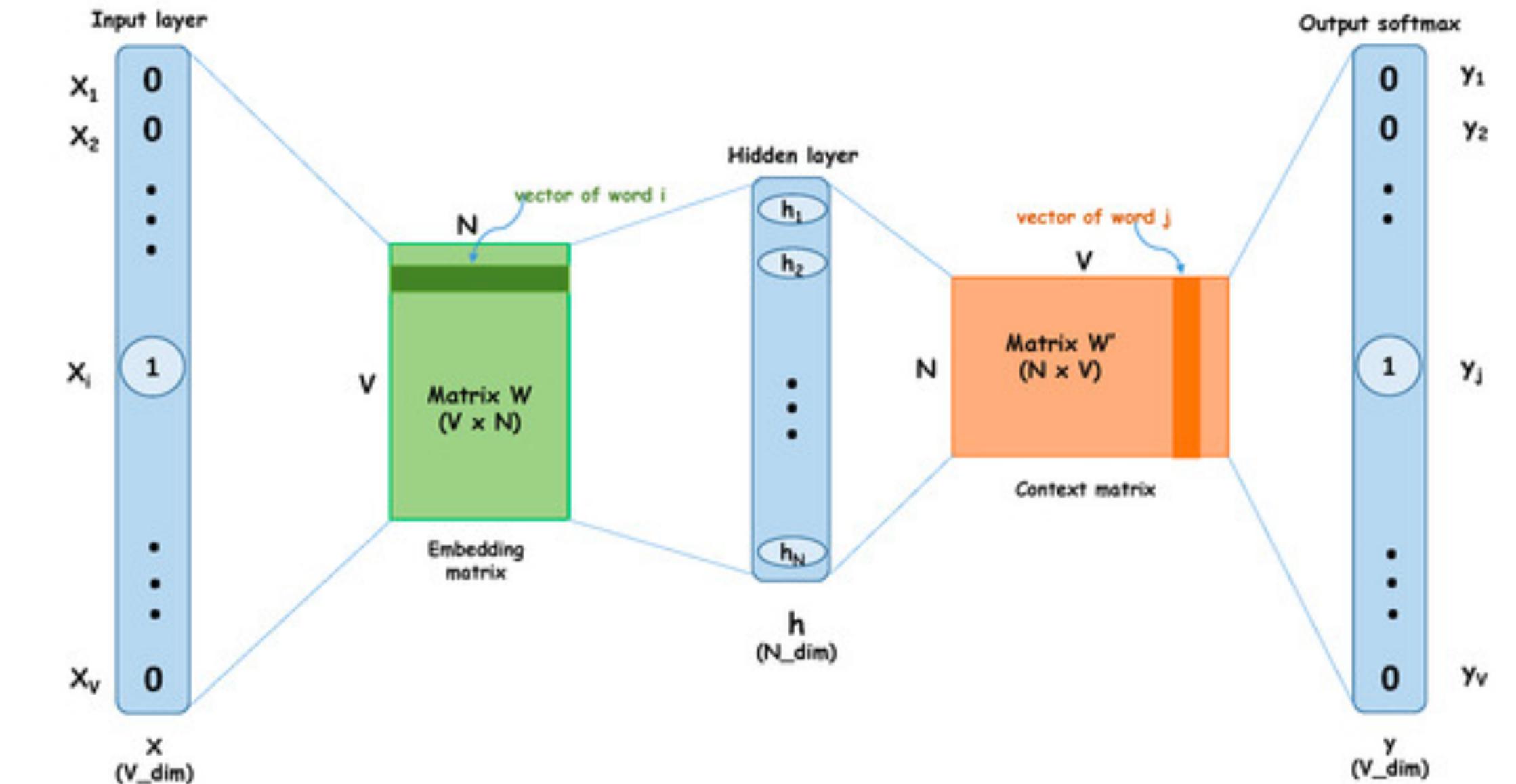
- Poor at capturing context
- Do not detect synonymous words
- Struggle with out of vocab words

Neural Nets

- 1. Word2Vec**
- 2. Fasttext**
- 3. RNN**
- 4. Transformers**

Word2Vec

1. Capture semantic meaning based on words' context in the text
2. Two Architectures
 1. CBOW: Predicts target words from context.
 2. Skip-gram: Predicts context words from a target word.



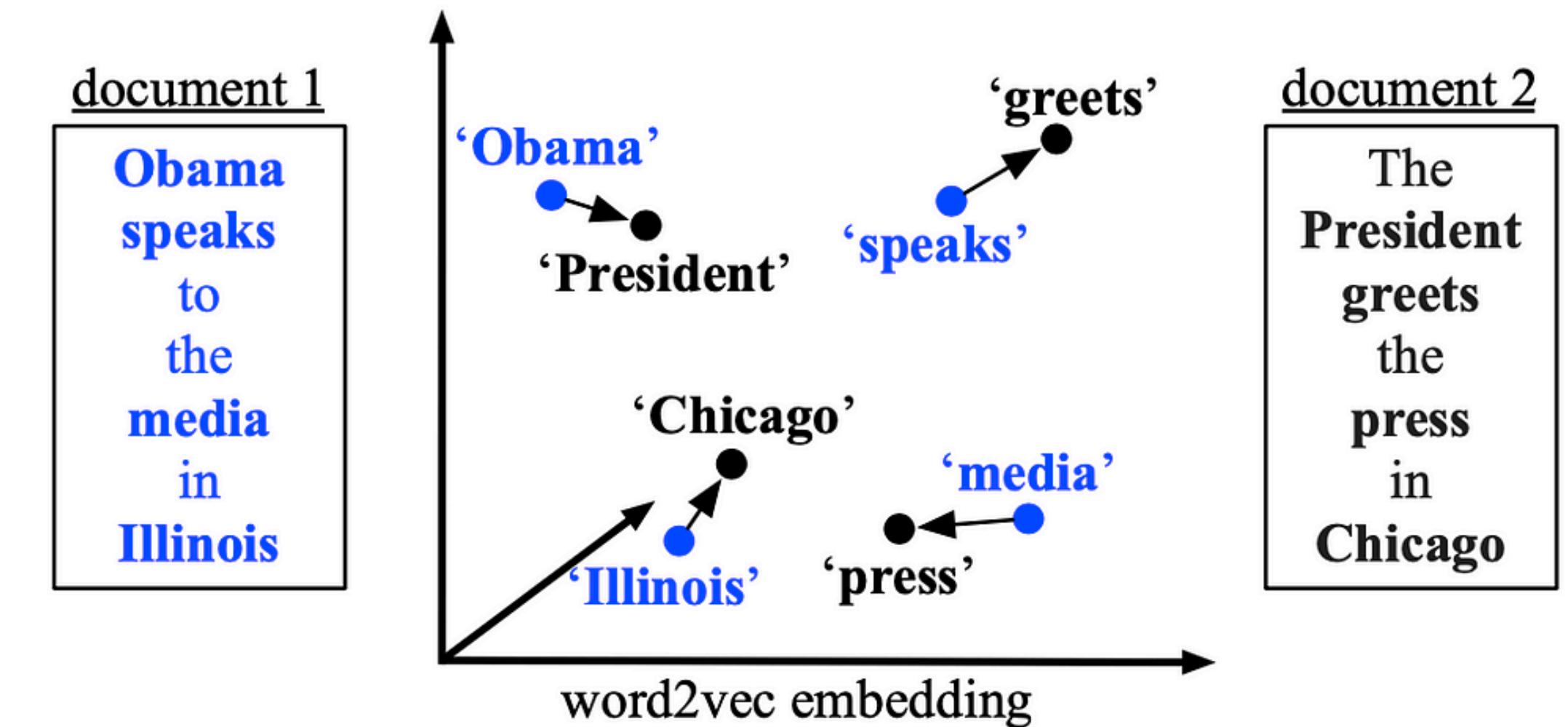
FastText

1. **Considers character n-grams to create word embeddings.**
2. **Improve word representations by capturing morphological information and better handling of rare words.**
3. **Can work with OOV words**



Features on word vectors

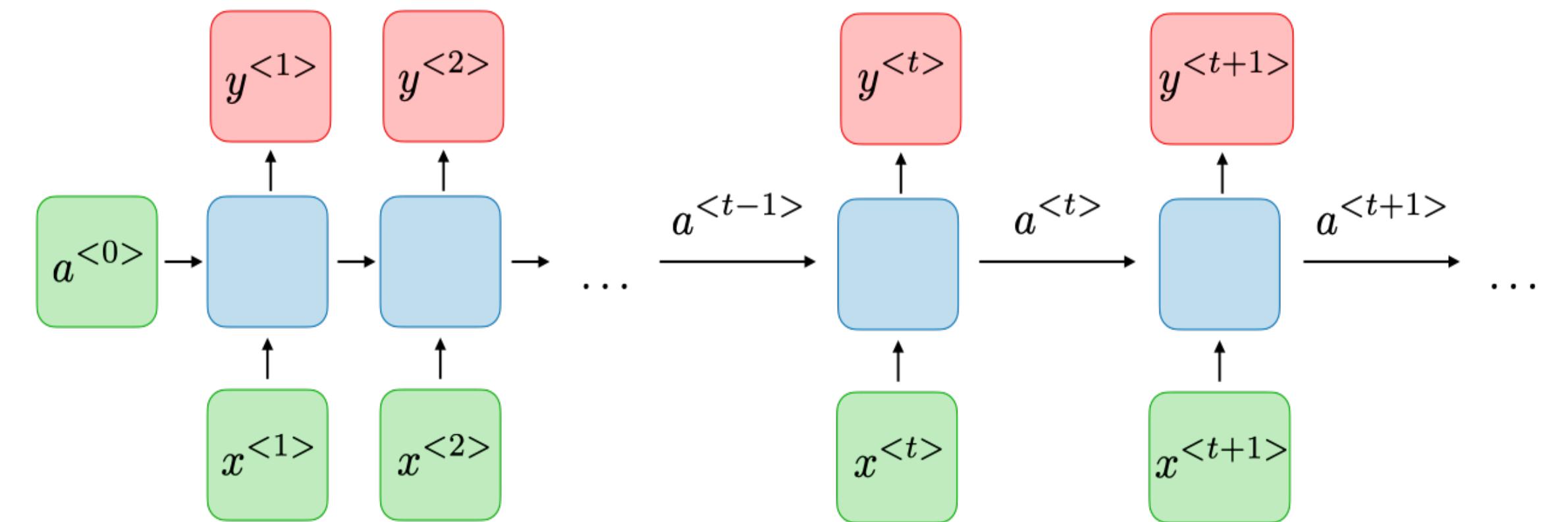
1. Euclidean/Cosine distances between mean of word vectors. $u_T := \frac{1}{n} \sum_{i=1}^n u_{w_i}$
2. Weighted distance on means vector_{doc} = $\sum_{w \in \text{doc}} \text{IDF}(w) \times \text{vector}_w$
3. Word movers distance (WMD)
 - measures how far you'd have to "move" the words from one document to match the words in another, capturing their semantic difference.



Sentence vectors

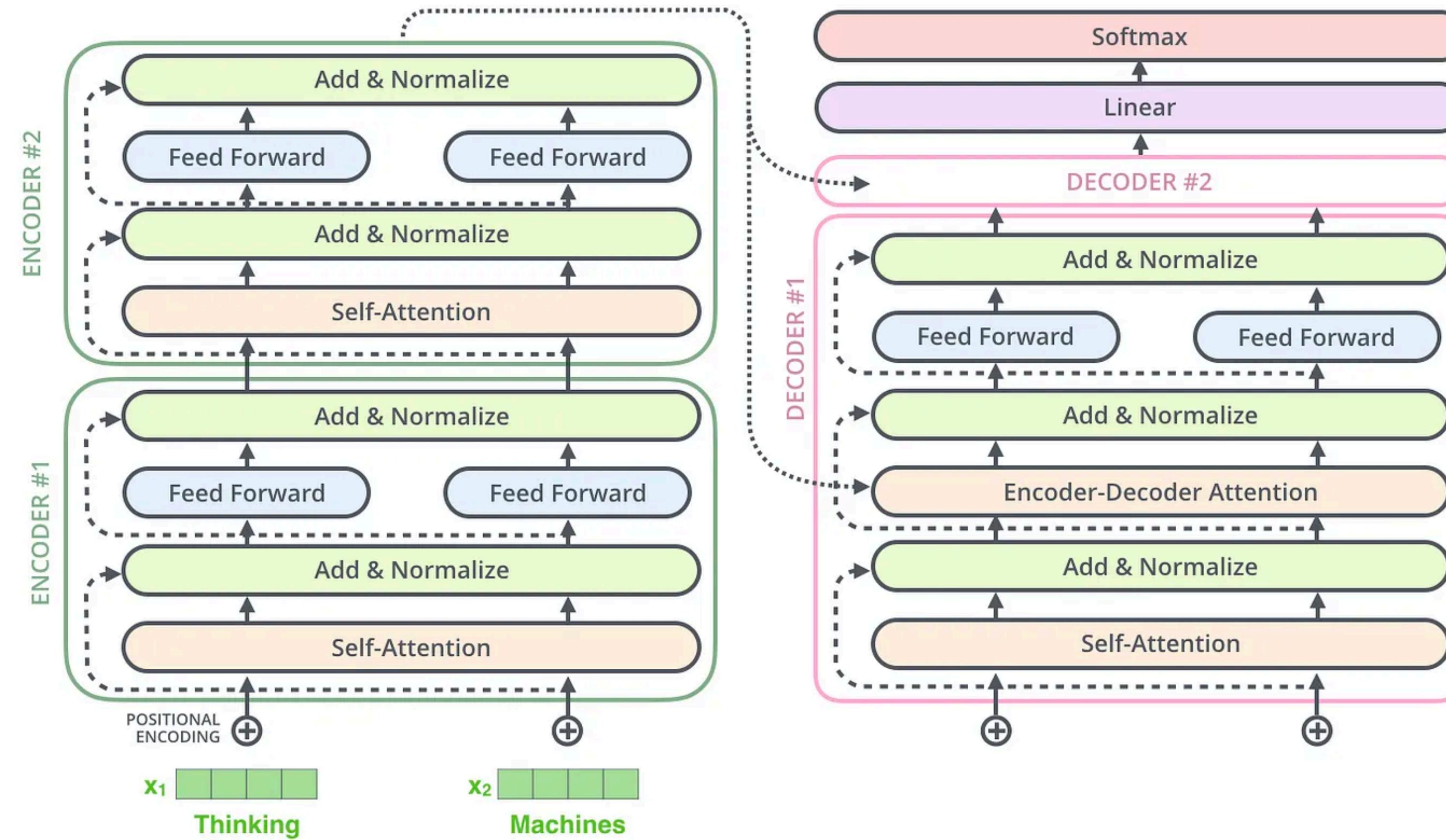
RNN

1. Vanishing & Exploding Gradients
2. Difficult to parallelize
3. Short-Term Memory



Sentence vectors

Transformers



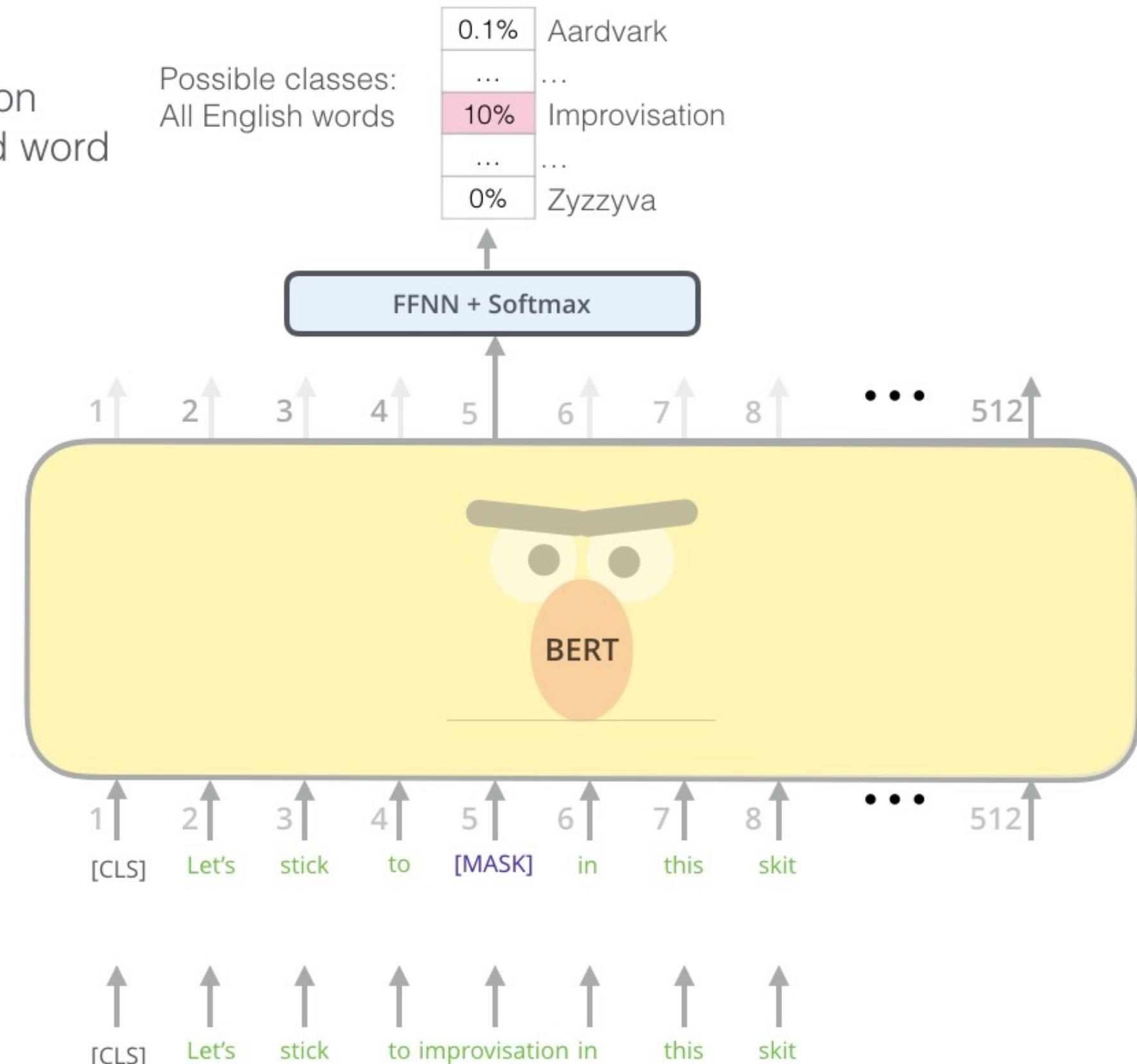
Sentence vectors

BERT

Use the output of the masked word's position to predict the masked word

Randomly mask 15% of tokens

Input



BERT

Train for matching

1. Loss

1. TripletLoss

2. Contrastive

$$L(x_1, x_2, y) = (1 - y) \times \frac{1}{2} D^2 + y \times \frac{1}{2} \times \max(0, m - D)^2$$

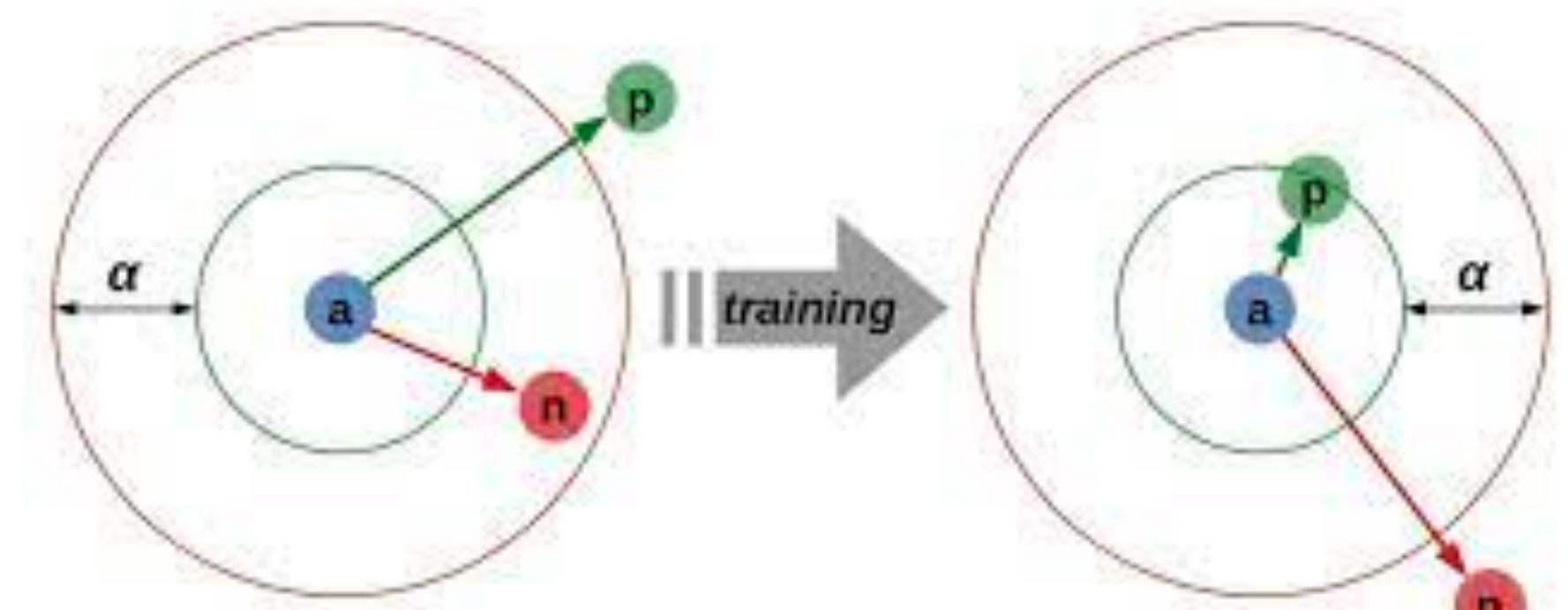
3. Cross-Entropy

$$\mathcal{L} = - (y \log(p) + (1 - y) \log(1 - p))$$

2. Dataset

1. Toloka is not always enough

2. Collect hard negatives



$$\mathcal{L}(A, P, N) = \max (0, \|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha)$$

BERT

Hard negatives

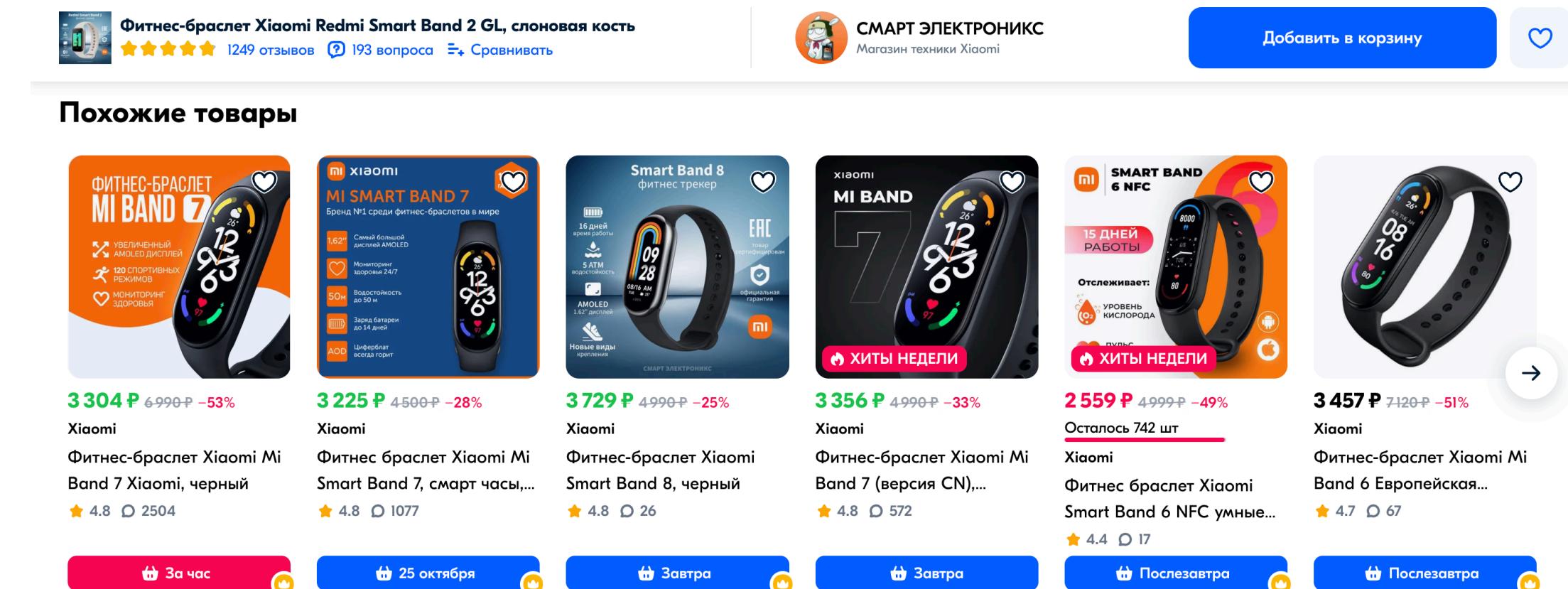
1. Toloka

2. Different cat4, but same cat3.

1. Cat3 = ‘Телефон’

2. Cat4 = ‘смартфон apple’,
‘смартфон samsung’

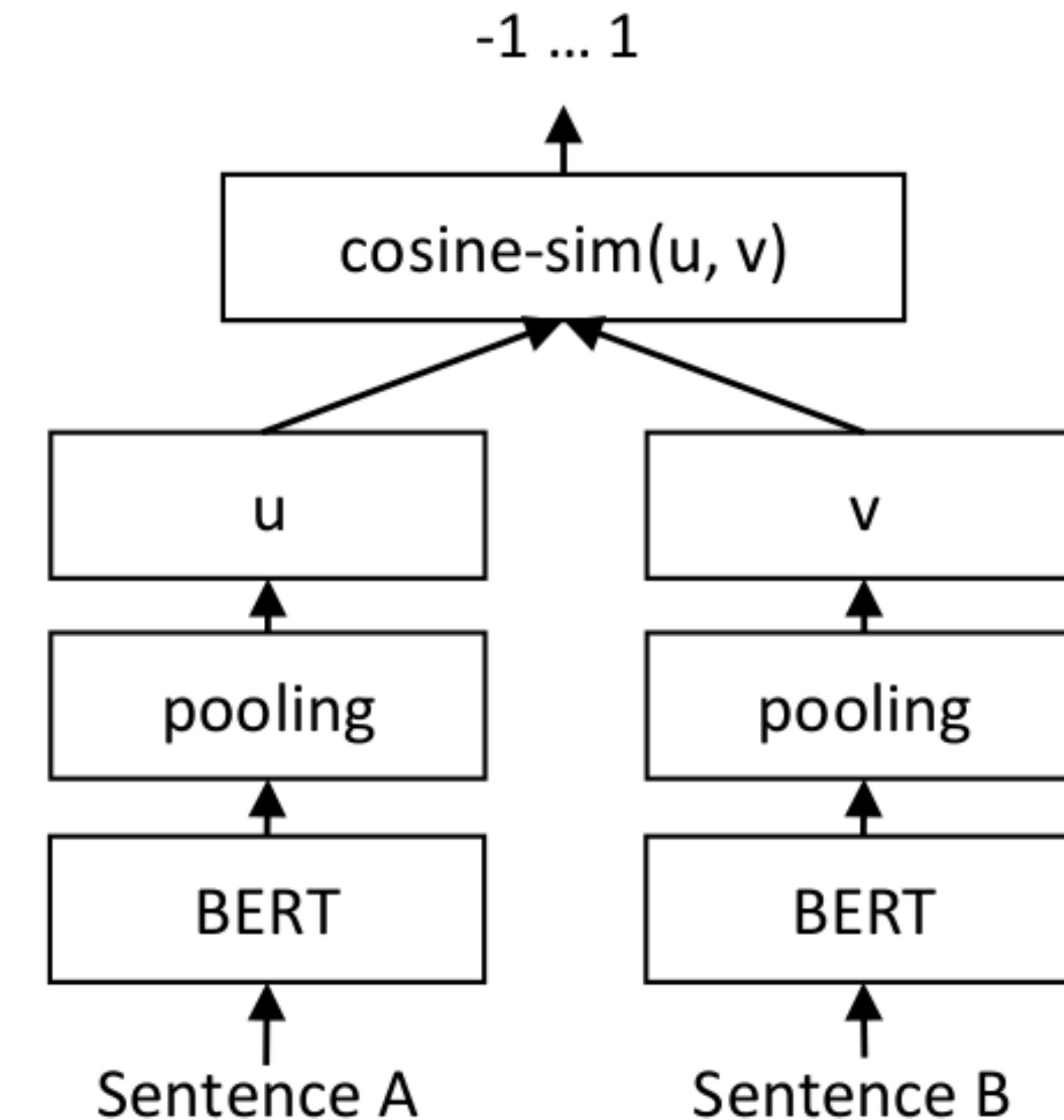
3. Recommendations



BERT

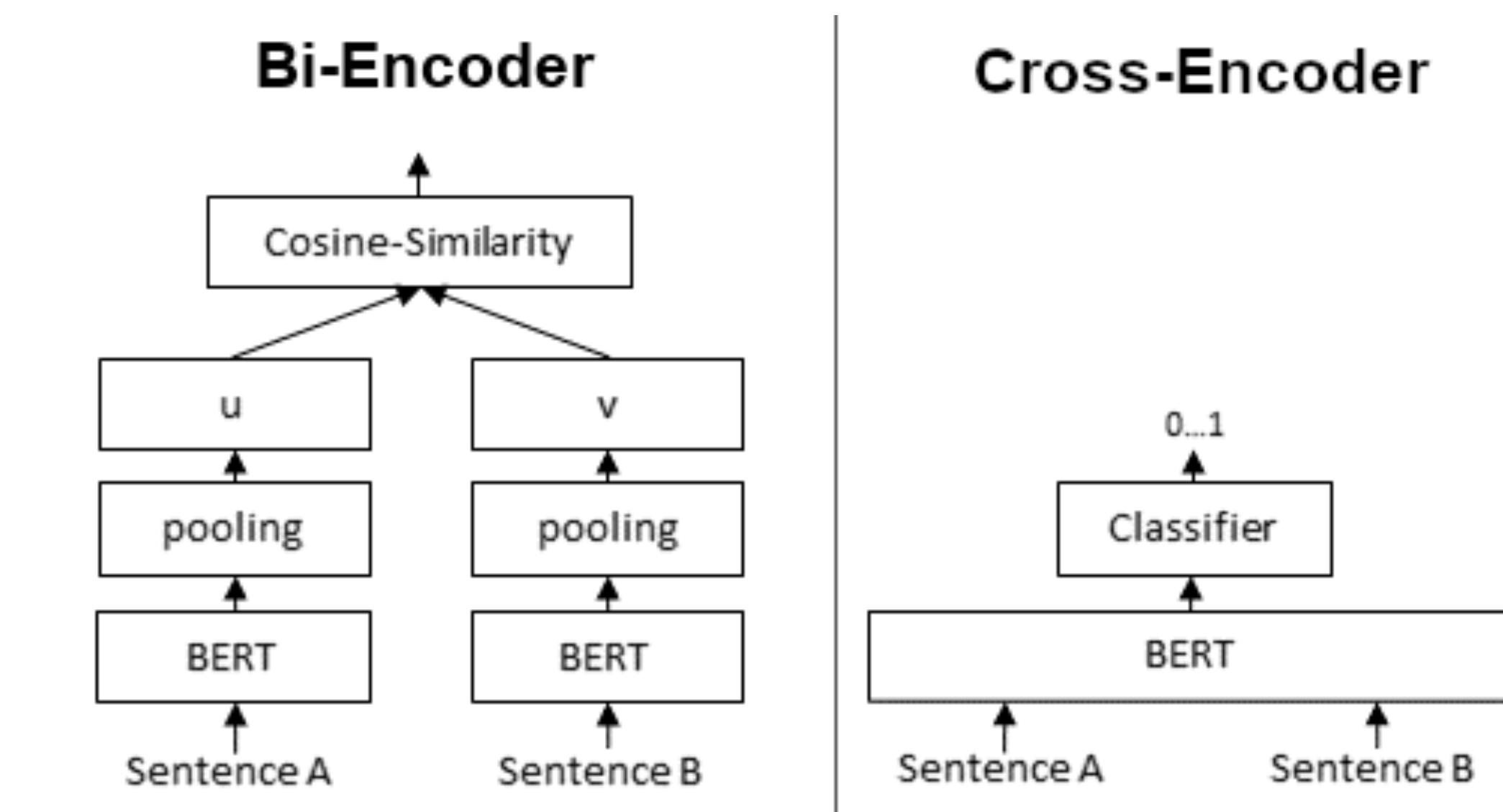
Train and some tips

- 1. Mean pooling and then Triplet Loss**
- 2. Train your own tokenizer**
- 3. Train MLM model**



Bi-Encoder vs Cross-encoder

1. Bi-encoder faster at inference
2. Cross-encoder is more accurate



Bi-Encoder vs Cross-encoder

How to speed up inference?

1. Distillation, e.g. onnxruntime
2. Linear time attention, e.g
FlashAttention
3. Cache in service, e.g. Redis

NN not always better than classic

Secrets, tricks and folk wisdom

- First of all think about data and then about SOTA models
- Classic approaches sometimes are better then NN
- If possible, train the model from scratch along with MLM and the tokenizer