Autumn 2024, Part 3

Text Classification

Eliza Vialykh, @elf_lesnoy

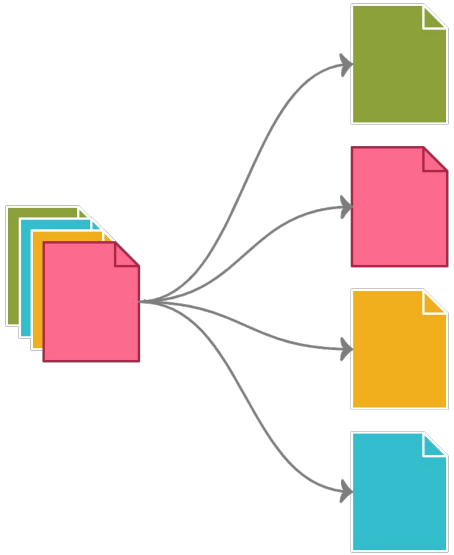
Content

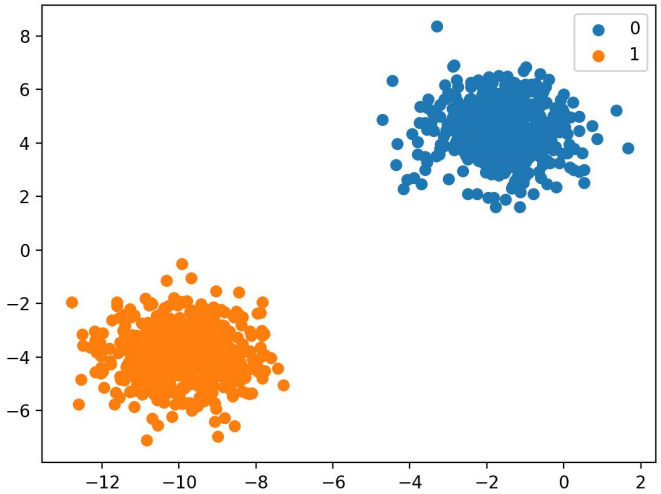
1. Classification Task
2. Text Classification

Classification Task

Classification

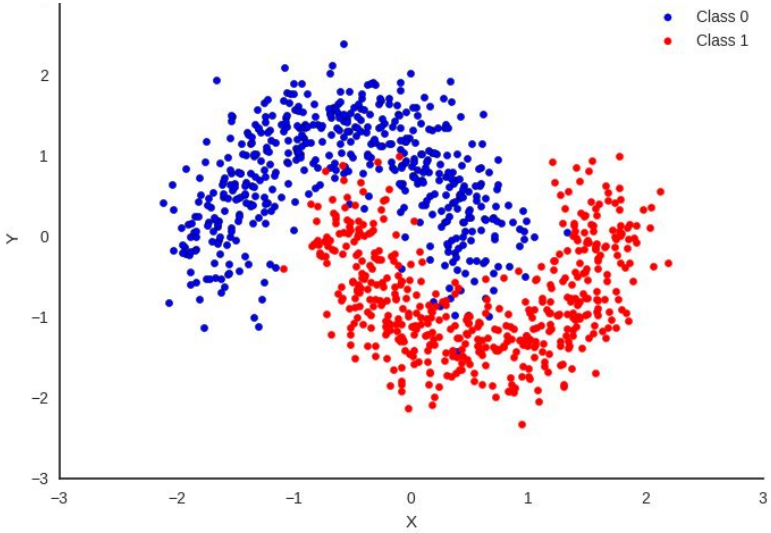
Classification lies at the heart of both human and machine intelligence. Deciding what letter, word, or image has been presented to our senses, recognizing faces or voices, sorting mail, assigning grades to homeworks

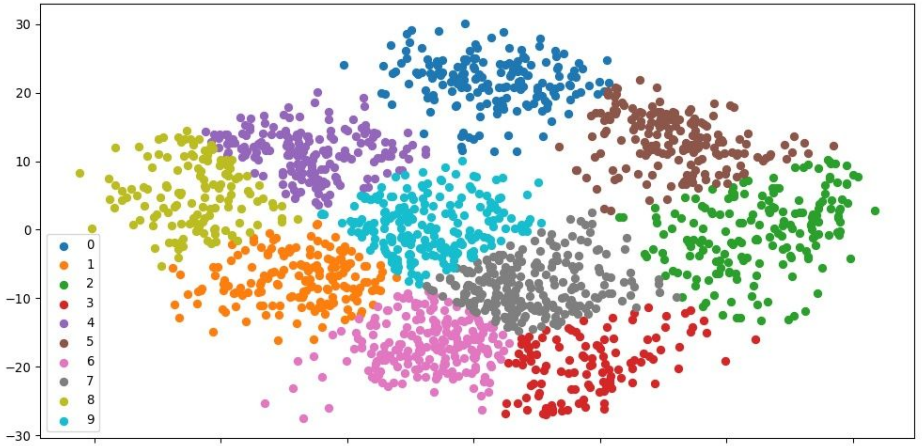




Classification

If the number of classes is two, it's called **binary classification**. If the number of classes is more than two, it's referred to as **multiclass classification**.



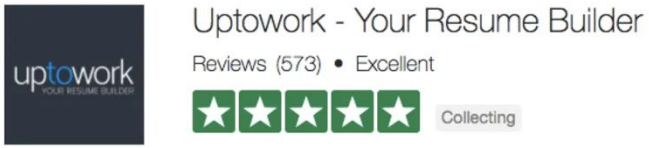


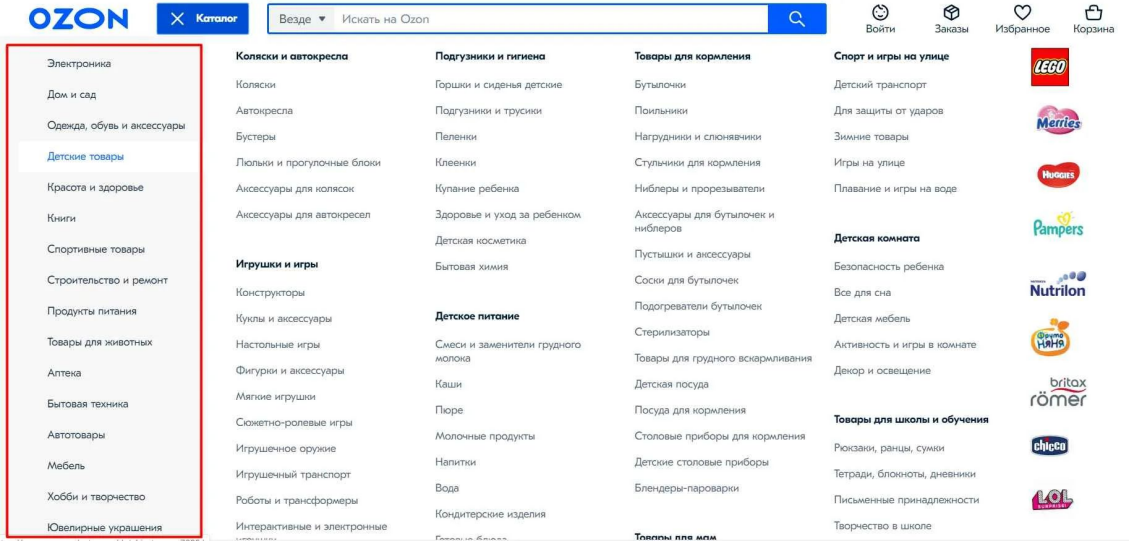
Classification

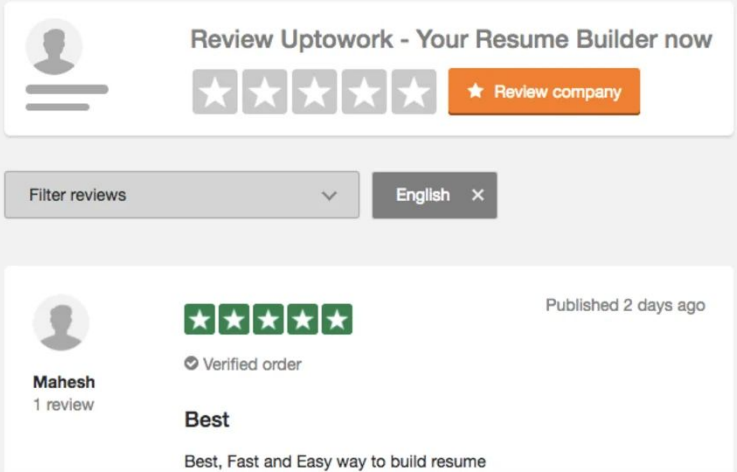
In **multilabel classification**, a document can have one or more labels/classes attached to it.

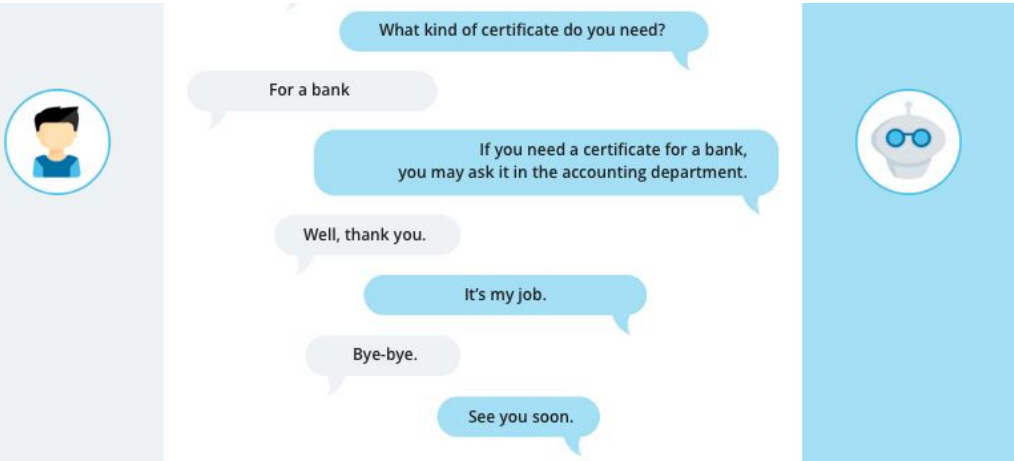
#cars
#3d
#modelling
#solidworks
#tutorial

Applications



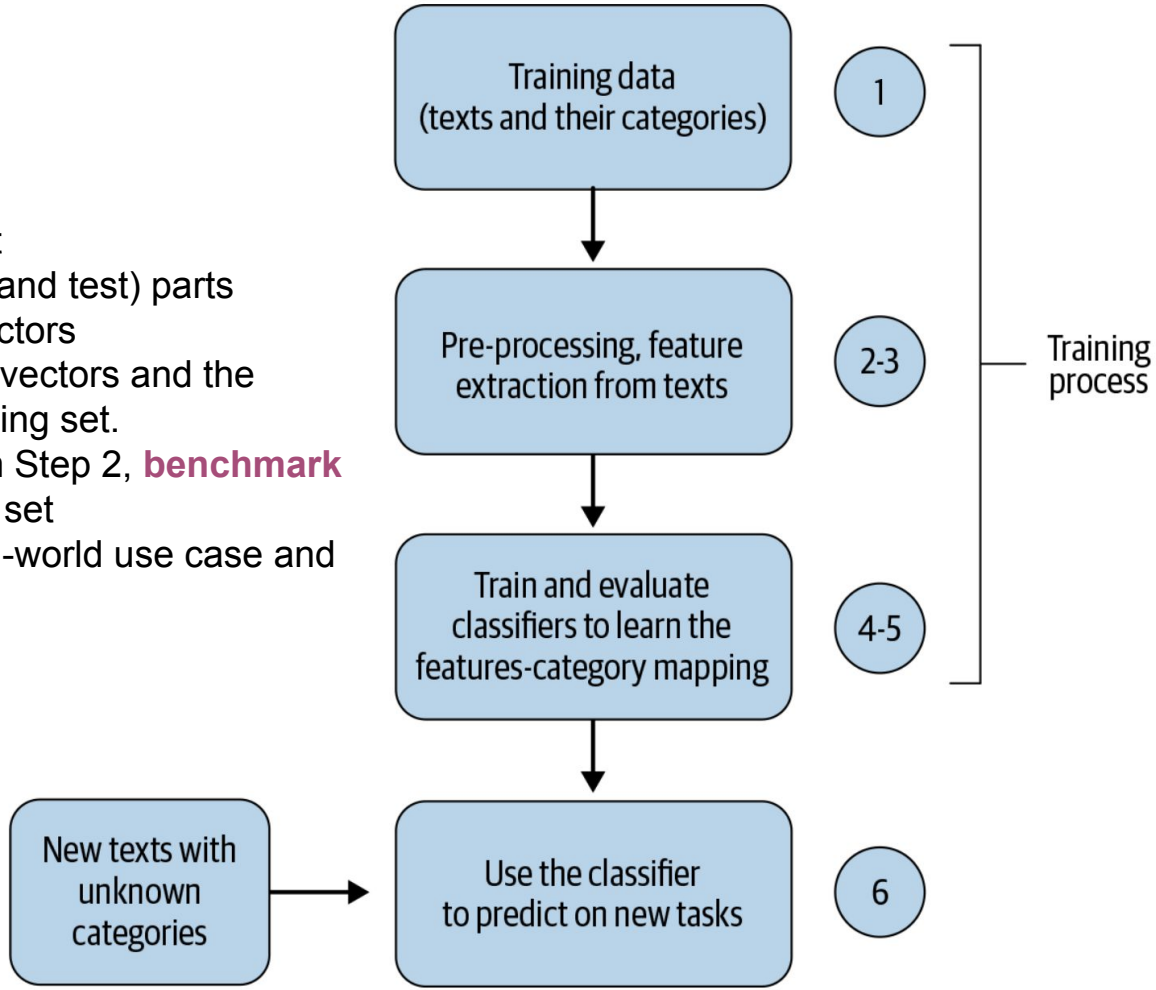






Pipelines

1. **Collect** or create a labeled dataset
2. **Split** the dataset into two (training and test) parts
3. **Transform** raw text into feature vectors
4. **Train** a classifier using the feature vectors and the corresponding labels from the training set.
5. Using the evaluation metric(s) from Step 2, **benchmark** the model performance on the test set
6. **Deploy** the model to serve the real-world use case and monitor its performance.



Datasets for Classification

Amazon Reviews Dataset
Yelp Reviews
SMS Spam Collection
IMDB Dataset
Goodreads Book Reviews
...

	label	text
0	2	Stuning even for the non-gamer: This sound tra...
1	2	The best soundtrack ever to anything.: I'm rea...
2	2	Amazing!: This soundtrack is my favorite music...
3	2	Excellent Soundtrack: I truly like this soundt...
4	2	Remember, Pull Your Jaw Off The Floor After He...
...	...	...
3599995	1	Don't do it!!: The high chair looks great when...
3599996	1	Looks nice, low functionality: I have used thi...
3599997	1	compact, but hard to clean: We have a small ho...
3599998	1	what is it saying?: not sure what this book is...
3599999	2	Makes My Blood Run Red-White-And-Blue: I agree...

Sentiment🤗😡🤢😈😻🤭

	statement	status
0	oh my gosh	Anxiety
1	trouble sleeping, confused mind, restless heart. All out of tune	Anxiety
2	All wrong, back off dear, forward doubt. Stay in a restless and restless place	Anxiety
3	I've shifted my focus to something else but I'm still worried	Anxiety
4	I'm restless and restless, it's been a month now, boy. What do you mean?	Anxiety

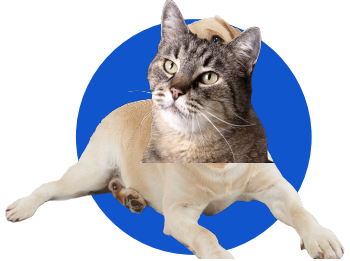
2401	Borderlands	Positive	I am coming to the borders and I will kill you all,
2401	Borderlands	Positive	im getting on borderlands and i will kill you all,
2401	Borderlands	Positive	im coming on borderlands and i will murder you all,
2401	Borderlands	Positive	im getting on borderlands 2 and i will murder you me all,

Metrics

True Positive

False Positive

True Negative



False Negative

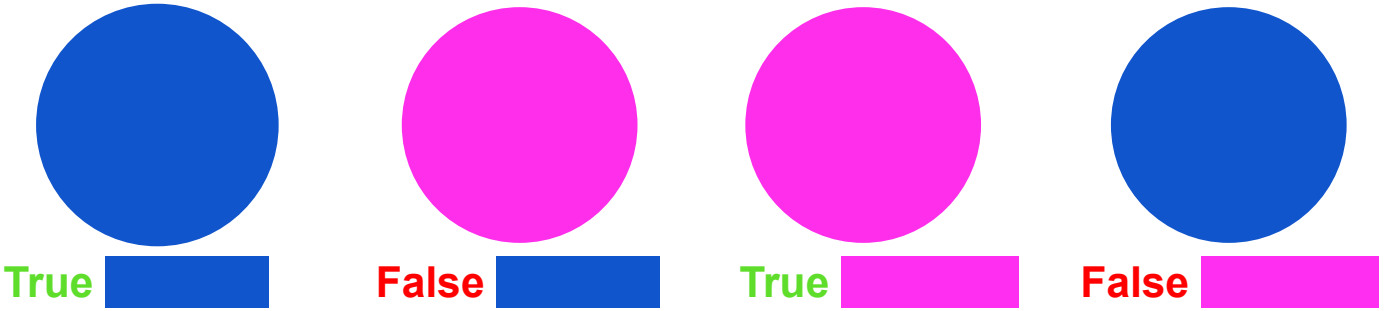
$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{\text{incorrectly classified actual negatives}}{\text{all actual negatives}} = \frac{FP}{FP + TN}$$

Metrics

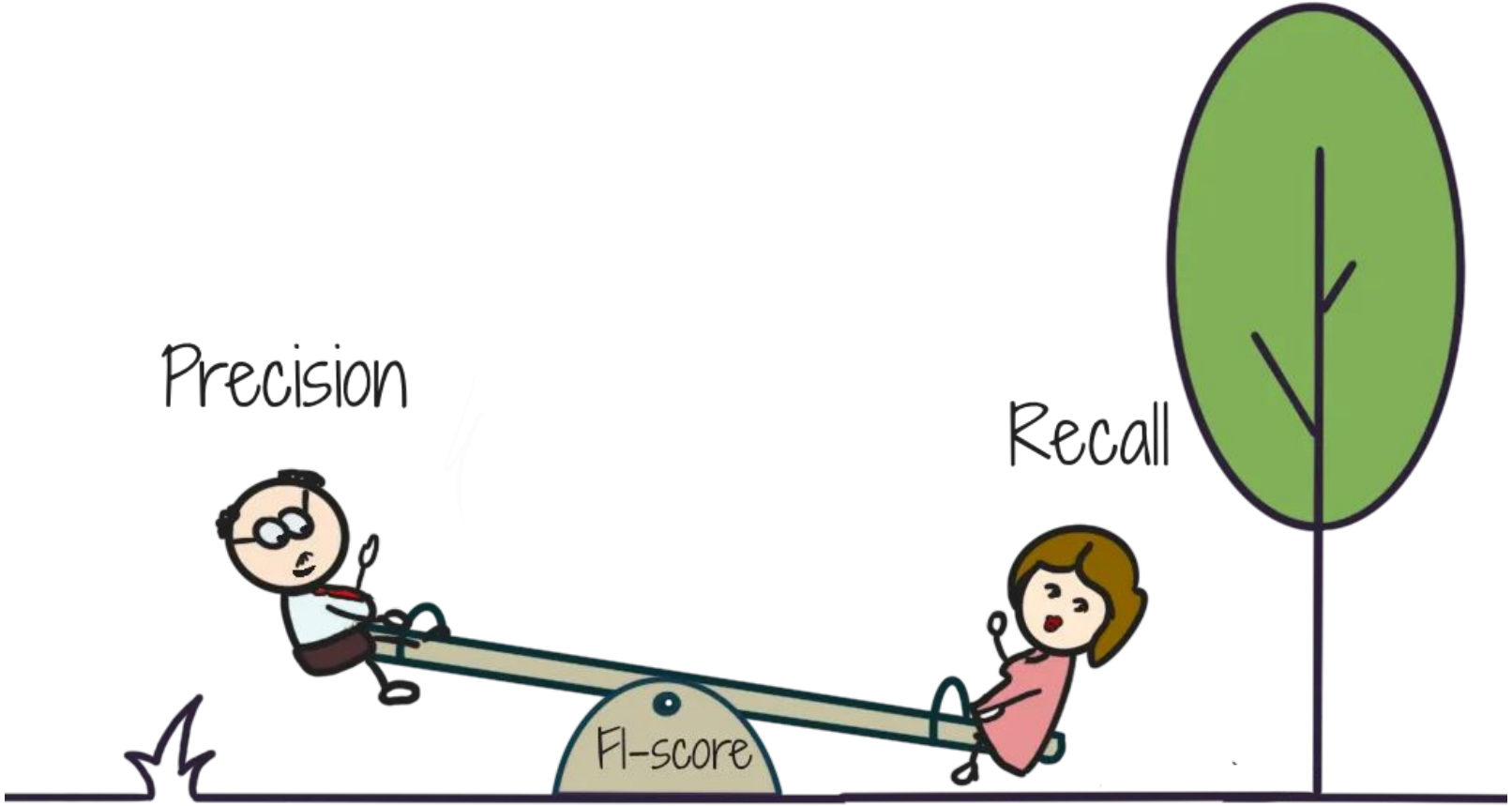


$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{\text{incorrectly classified actual negatives}}{\text{all actual negatives}} = \frac{FP}{FP + TN}$$

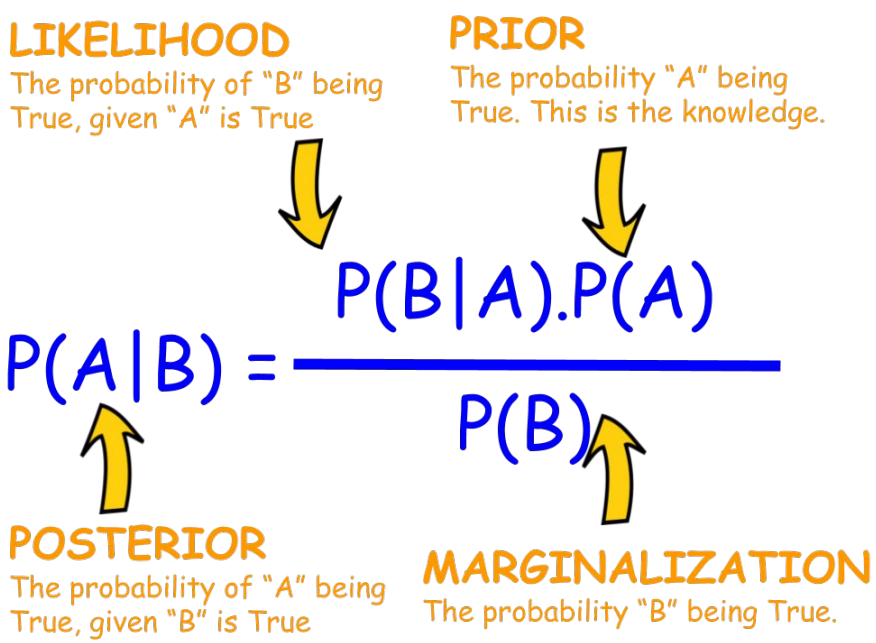


https://medium.com/@risdan.kristori/precision-vs-recall-how-i-remember-it-fd96a49db327#:~:text=Recall%20is%20the%20TP%2F%20(TP,reality%20is%20it%20is%20raining

Text Classification

Naive Bayes

Naive Bayes is a probabilistic classifier, meaning that for a document d, out of all classes c ∈ C the classifier returns the class c which has the maximum posterior probability given the document.



Thomas Bayes
1702 - 1761

Naive Bayes

$$P(y_k|X) = \frac{P(y_k)P(X|y_k)}{P(X)}$$

$$P(y_k|X_1, X_2, \ldots, X_n) = \frac{P(y_k)\prod_{i=1}^{n} P(X_i|y_k)}{P(X_1, X_2, \ldots, X_n)}$$

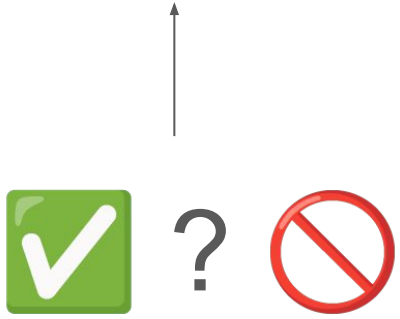
$$y_k \propto \arg\max_{y_k} P(y_k) \prod_{i=1}^{n} P(X_i|y_k)$$

Naive Bayes

$$P(C|M) \propto P(C) \prod_{i=1}^{n} P(w_i|C), \quad w_i \in M$$

word i

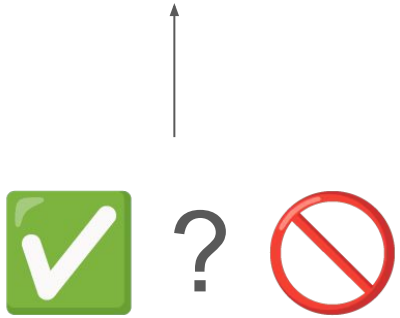
"Hi, you won a discount and you can get the prize this evening."



Word	Frequency in Not Spam	Frequency in Spam	Probability in Not Spam	Probability in Spam
Hi	1 + 1 = 2	0 + 1 = 1	2 / 28 = 0.0714	1 / 33 = 0.03
how	1 + 1 = 2	0 + 1 = 1	2 / 28 = 0.0714	1 / 33 = 0.03
are	1 + 1 = 2	0 + 1 = 1	2 / 28 = 0.0714	1 / 33 = 0.03
you	1 + 1 = 2	1 + 1 = 2	2 / 28 = 0.0714	2 / 33 = 0.06
Congratulations	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
won	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
a	0 + 1 = 1	2 + 1 = 3	1 / 28 = 0.0357	3 / 33 = 0.09
prize	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
Buy	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
the	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
product	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
now	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
	0 + 1 = 1	1 + 1 = 2	1 / 28 = 0.0357	2 / 33 = 0.06
	1 + 1 = 2	0 + 1 = 1	2 / 28 = 0.0714	1 / 33 = 0.03
	1 + 1 = 2	0 + 1 = 1	2 / 28 = 0.0714	1 / 33 = 0.03
	1 + 1 = 2	0 + 1 = 1	2 / 28 = 0.0714	1 / 33 = 0.03
	1 + 1 = 2	0 + 1 = 1	2 / 28 = 0.0714	1 / 33 = 0.03
	0 + 1 = 1	0 + 1 = 1	1 / 28 = 0.0357	1 / 33 = 0.03
	28	**33**		

Message	Class
Hi, how are you?	Not spam
Congratulations, you won a prize!	Spam
Buy the product now and get a discount!	Spam
Let's walk this evening.	Not spam

"Hi, you won a discount and you can get the prize this evening."



$$P(Spam|M) > P(Not\ \ Spam|M)$$

$\rightarrow$ ***сообщение является спамом.***

$$P(Spam|M) = 0.5 \cdot 0.03 \cdot 0.06 \cdot 0.06 \cdot 0.09 \cdot 0.06 \cdot 0.06 \cdot 0.06 \cdot 0.03 \cdot$$
$$\cdot 0.06 \cdot 0.06 \cdot 0.06 \cdot 0.03 \cdot 0.03 \approx 6.12 \cdot 10^{-18}$$

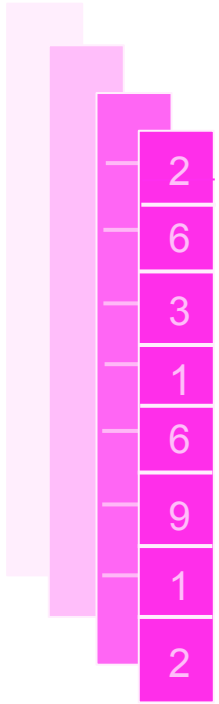
$$P(Not\ \ Spam|M) = 0.5 \cdot 0.0714 \cdot 0.0714 \cdot 0.0357 \cdot 0.0357 \cdot 0.0357 \cdot 0.0357 \cdot$$
$$\cdot 0.0714 \cdot 0.0357 \cdot 0.0357 \cdot 0.0357 \cdot 0.0357 \cdot 0.0714 \cdot 0.0714 \approx 2.45 \cdot 10^{-18}$$

Start

Истоки КЛ восходят к исследованиям известного американского лингвиста Н. Хомского по формализации структуры естественного языка [6], к первым экспериментам по машинному переводу, выполненным программистами и математиками, а также к разработанным в области искусственного интеллекта первым программам понимания естественного языка (например, [28]).

Поскольку в КЛ объектом обработки выступают тексты естественного языка, ее развитие невозможно без базовых знаний в области общей лингвистики (языкознания) [32]. Лингвистика изучает общие законы естественного языка — его структуру и функционирование, и включает такие области:

- **фонология** — изучает звуки речи и правила их соединения при формировании речи;
- **морфология** — занимается внутренней структурой и внешней формой слов речи, включая части речи и их категории;
- **синтаксис** — изучает структуру предложений, правила сочетаемости и порядка следования слов в предложении, а также общие его свойства как единицы языка.
- **семантика** и **прагматика** — тесно связанные области: семантика занимается смыслом слов, предложений и других единиц речи, а прагматика — особенностями выражения этого смысла в связи с конкретными целями общения;



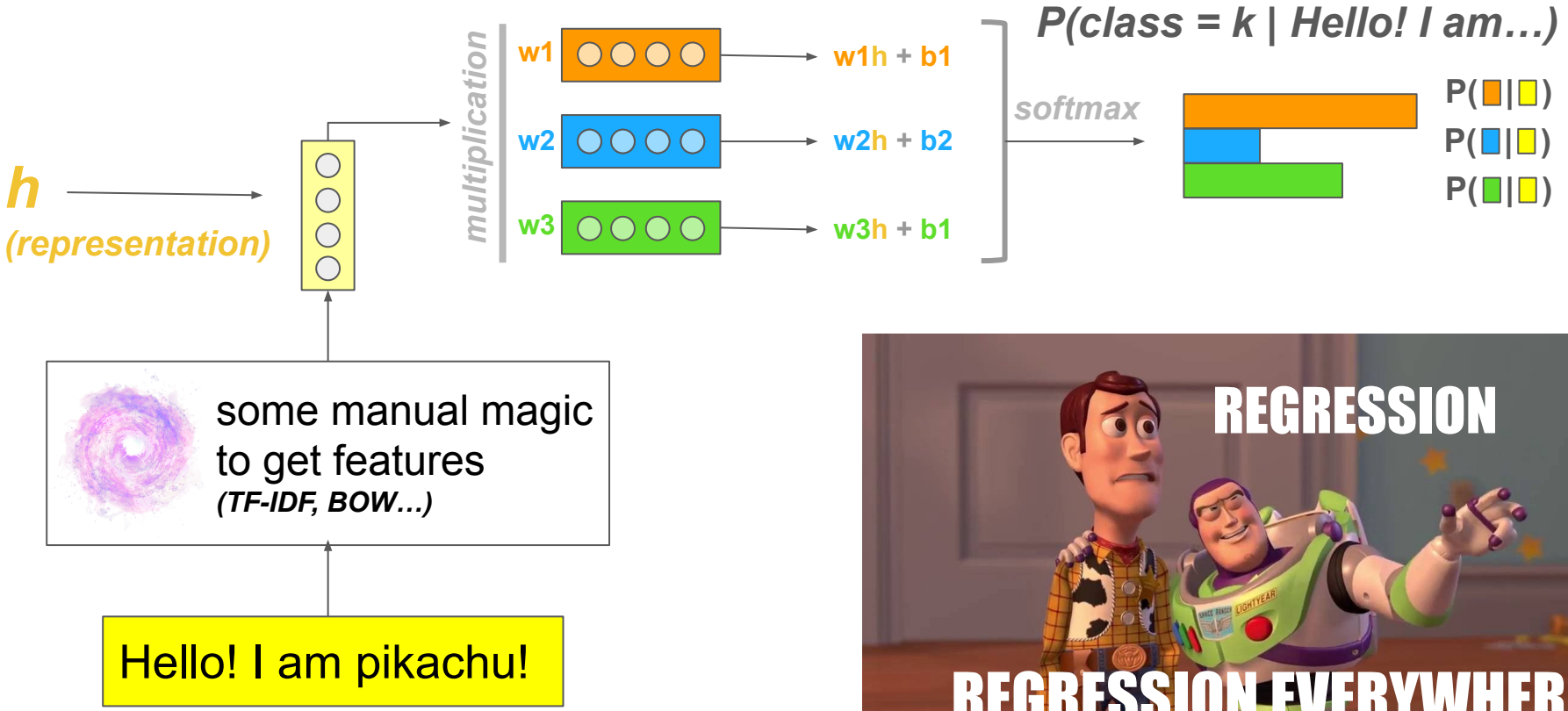
Word2Vec,
GloVe

Algorithm

Text

Text Embeddings

Logistic Regression



Logistic Regression

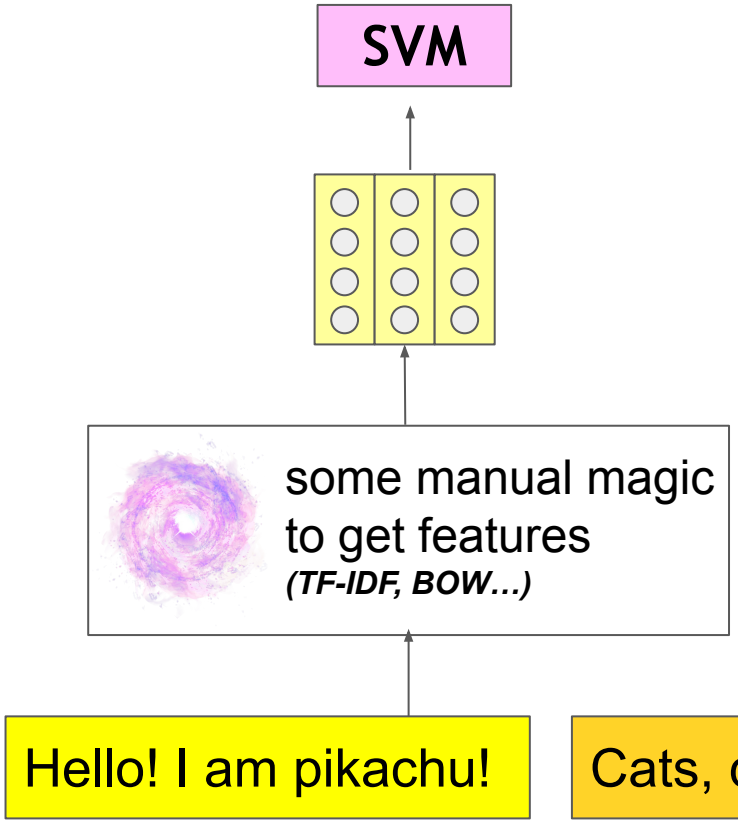
feature representation of the input text

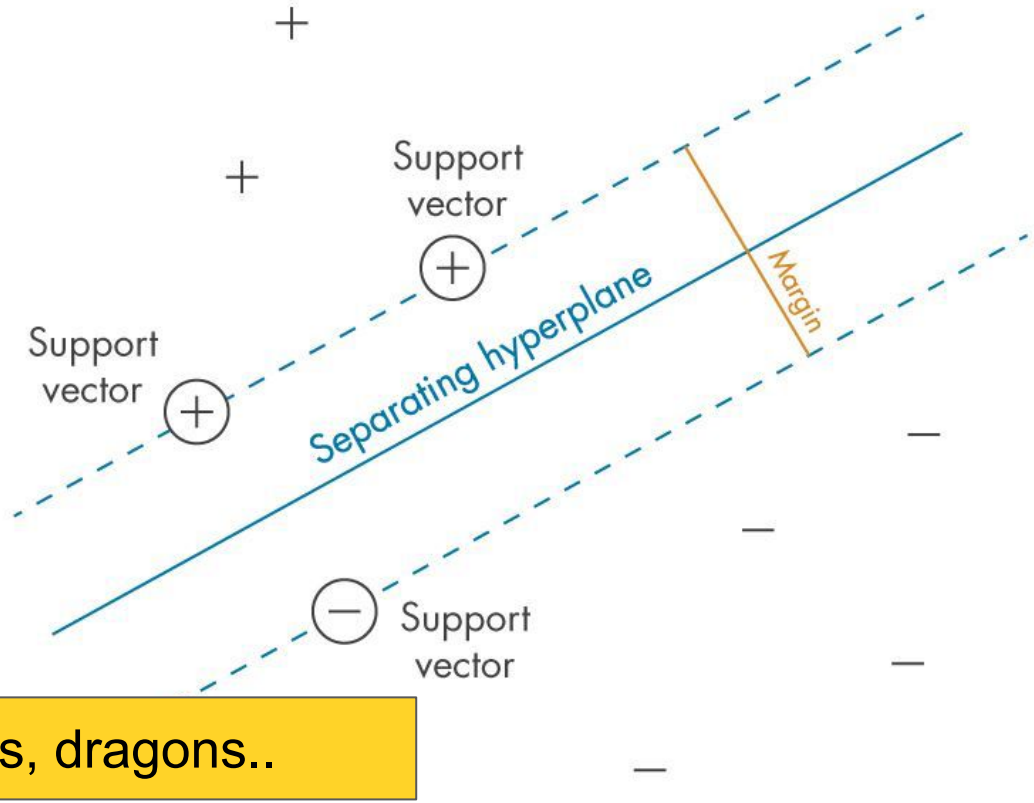
$$w^{(k)}h = w_1^{(k)} \cdot f_1 + \cdots + w_n^{(k)} \cdot f_n, \qquad k = 1, \ldots, K.$$

$$P(class = k|h) = \frac{\exp(w^{(k)}h)}{\sum_{i=1}^{K} \exp(w^{(i)}h)}$$

$$w^* = \arg\max_w \sum_{i=1}^{N} \log P(y = y^i|x^i)$$

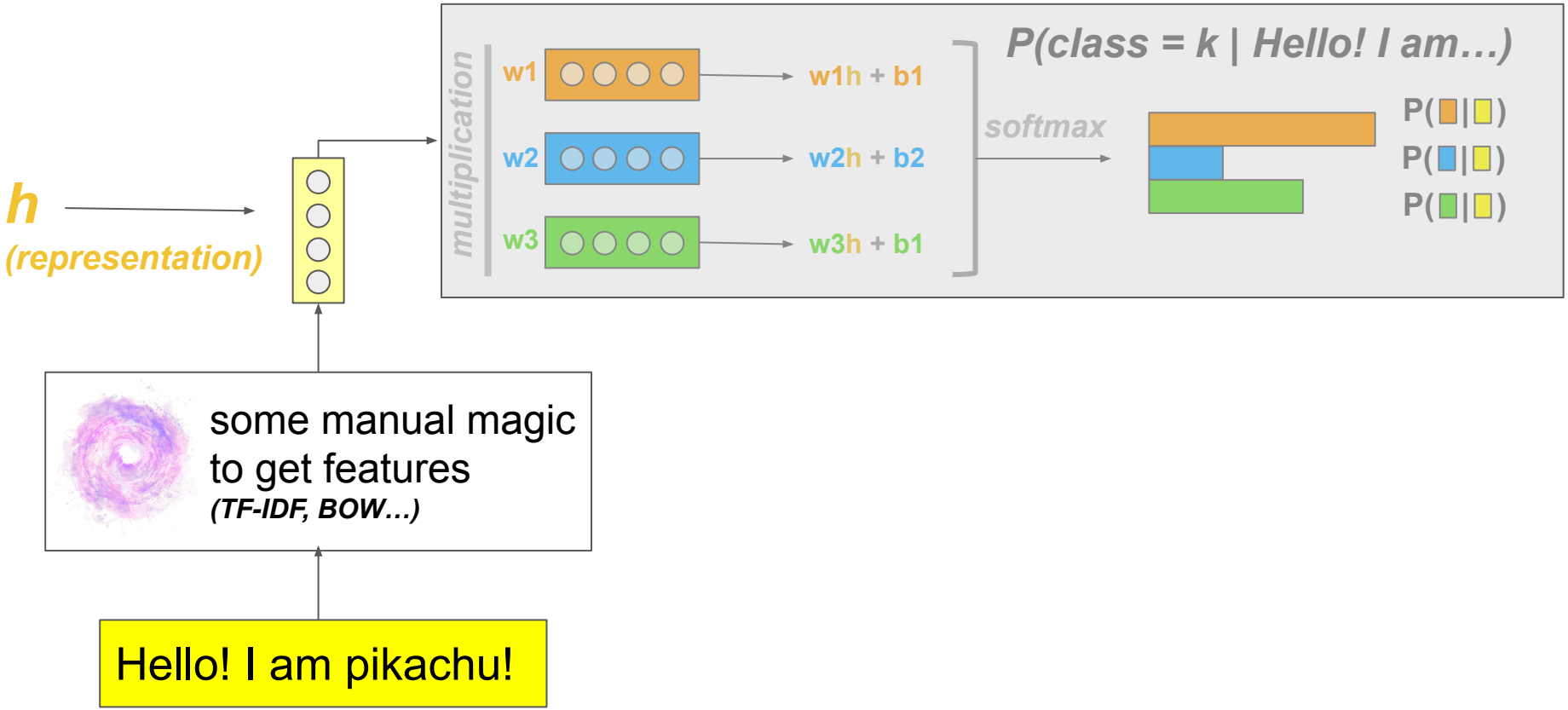
Support Vector Machine





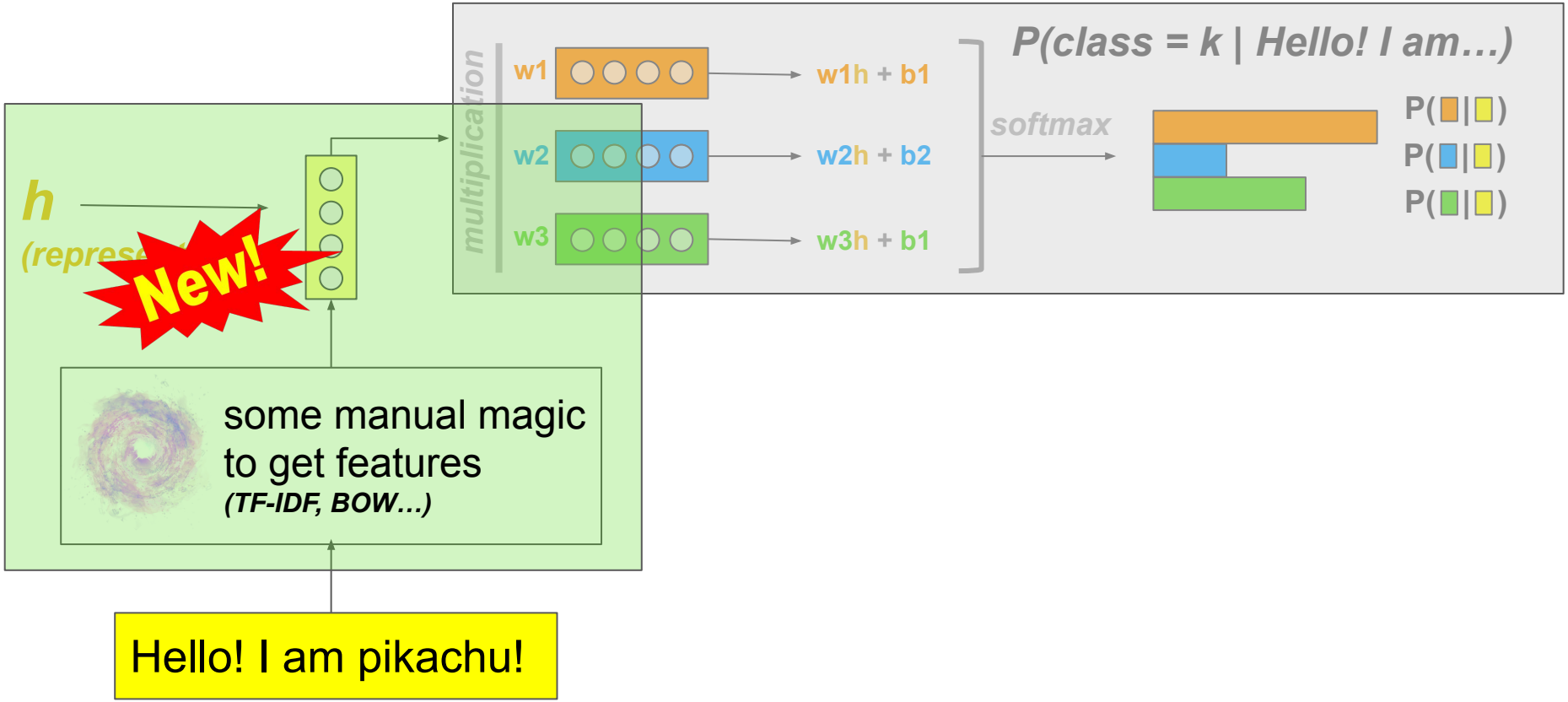
Neural network

The classification module remains the same

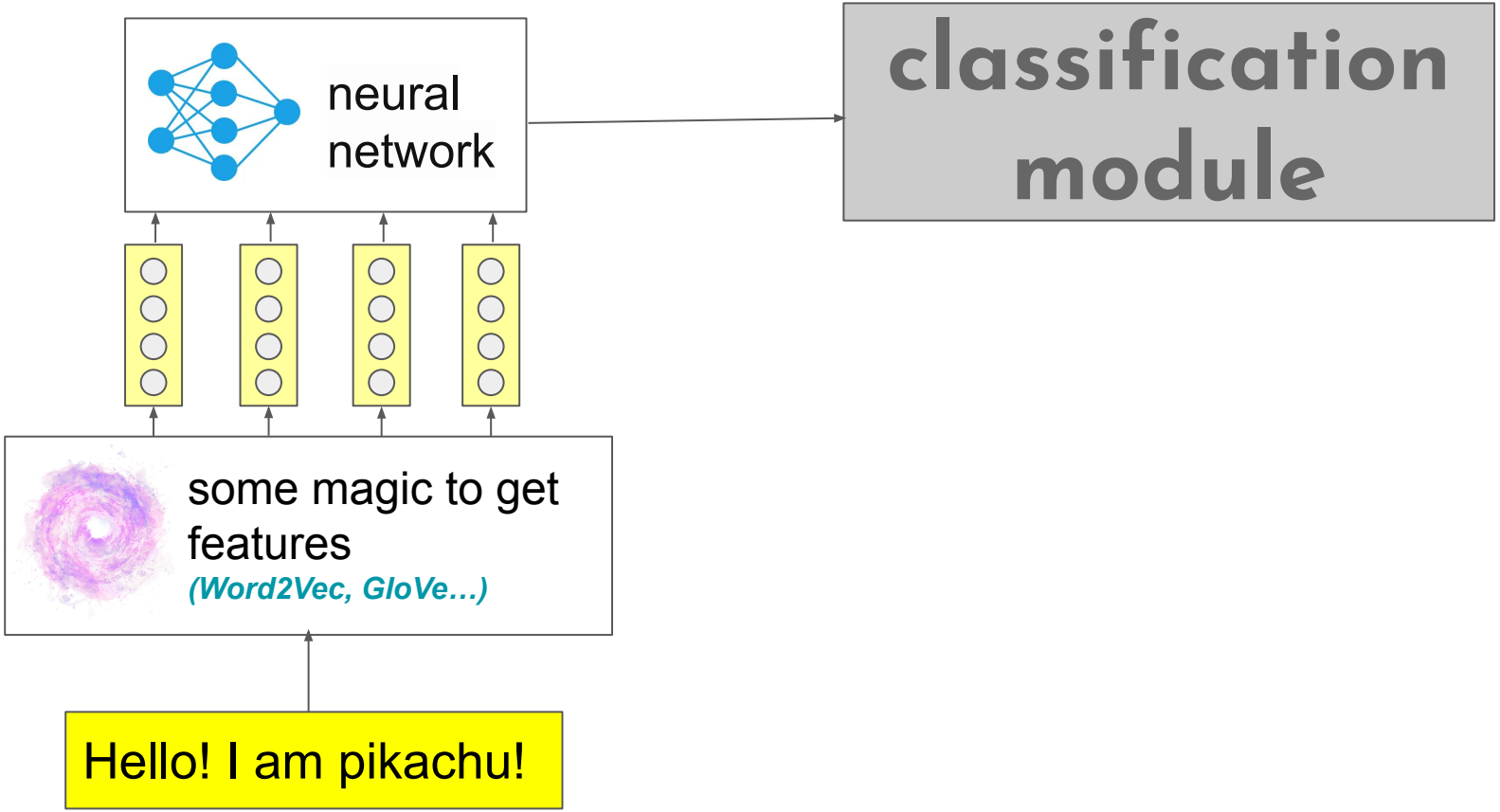


Neural network

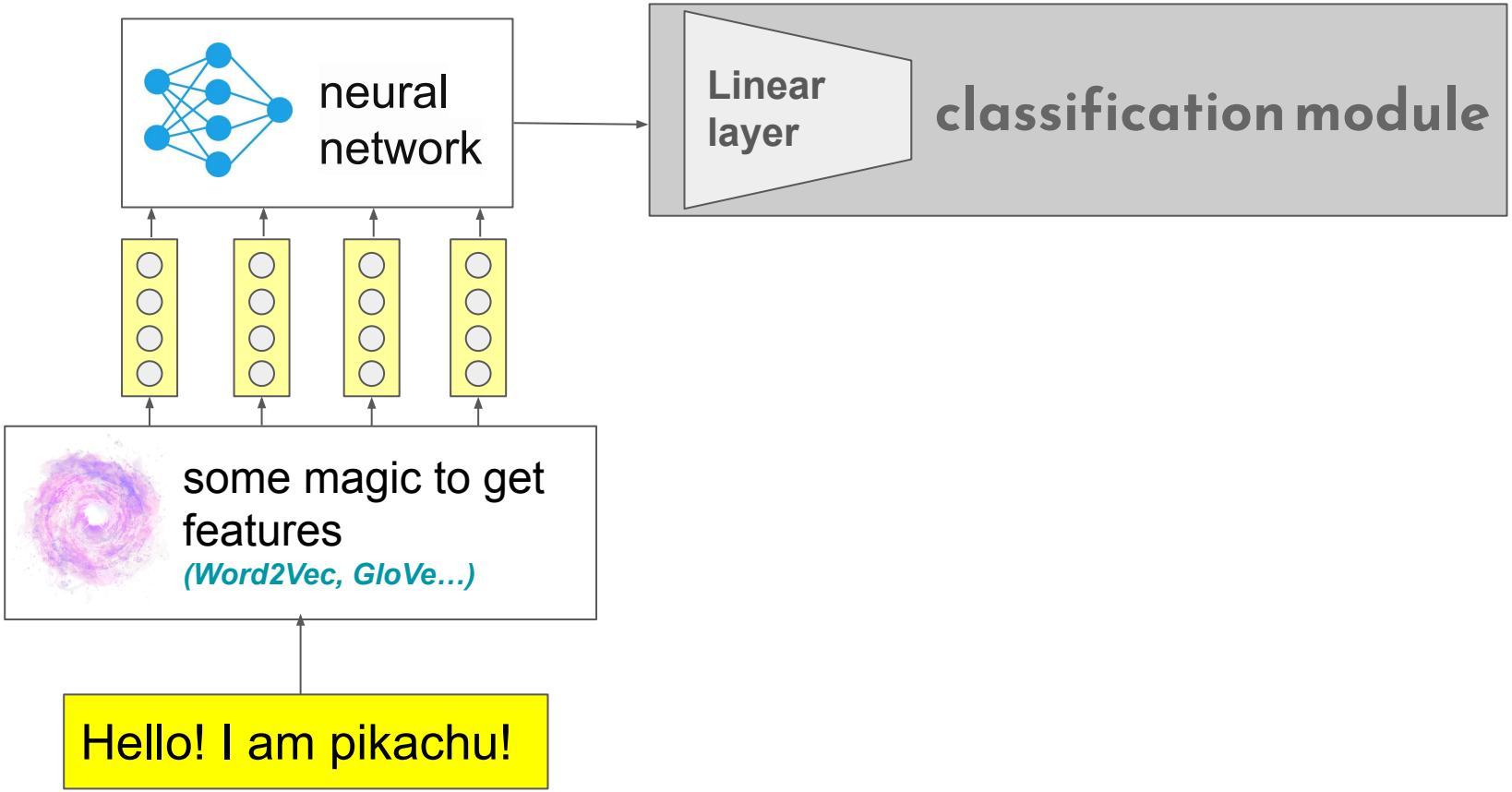
The classification module remains the same



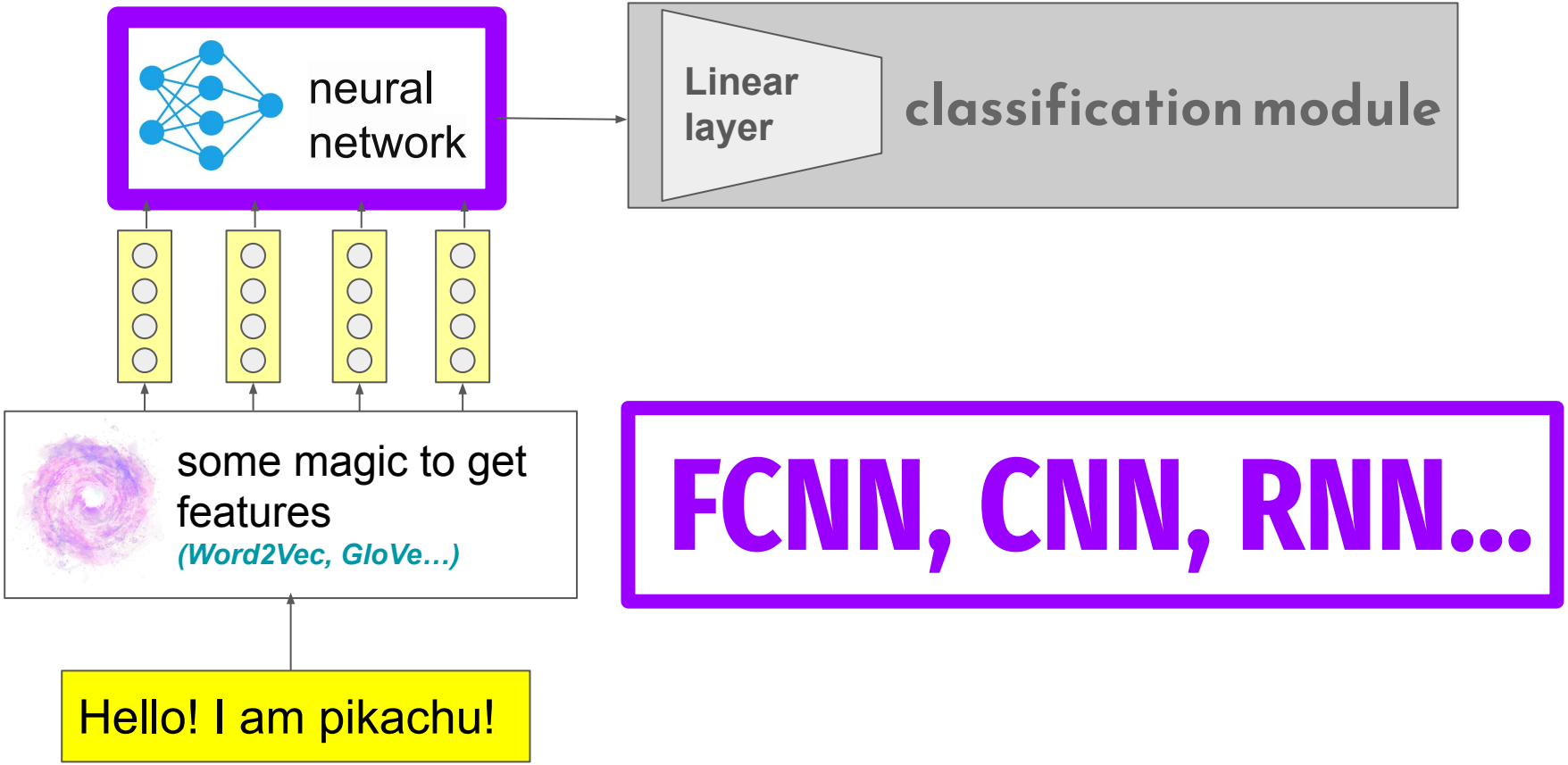
Neural network



Neural network



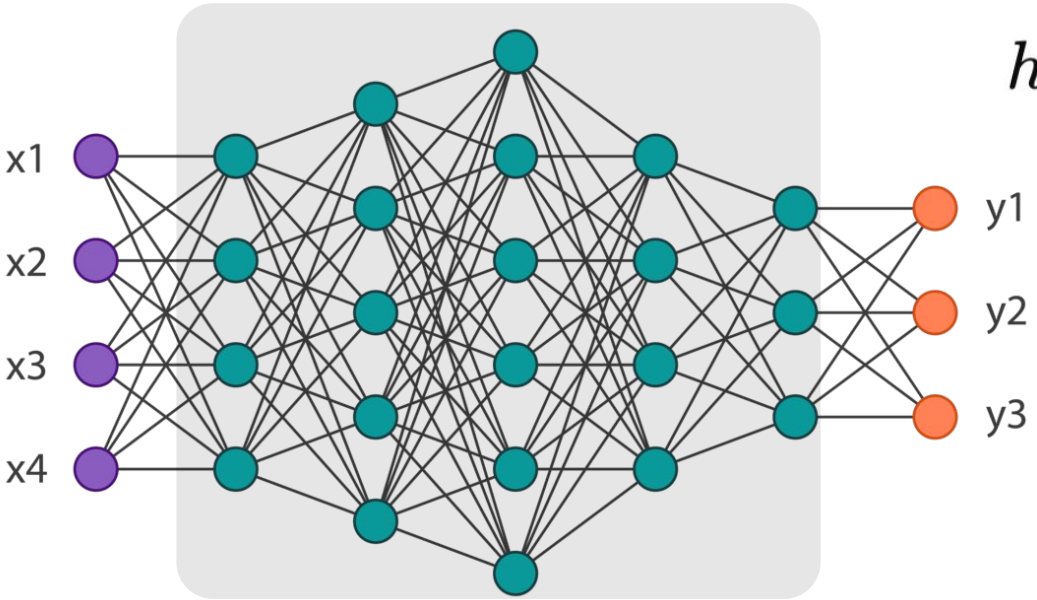
Neural network



Fully connected neural network (FCN)

$$h_i = f(W_i \cdot x + b_i)$$

$$ReLU(x) = max(0, x)$$

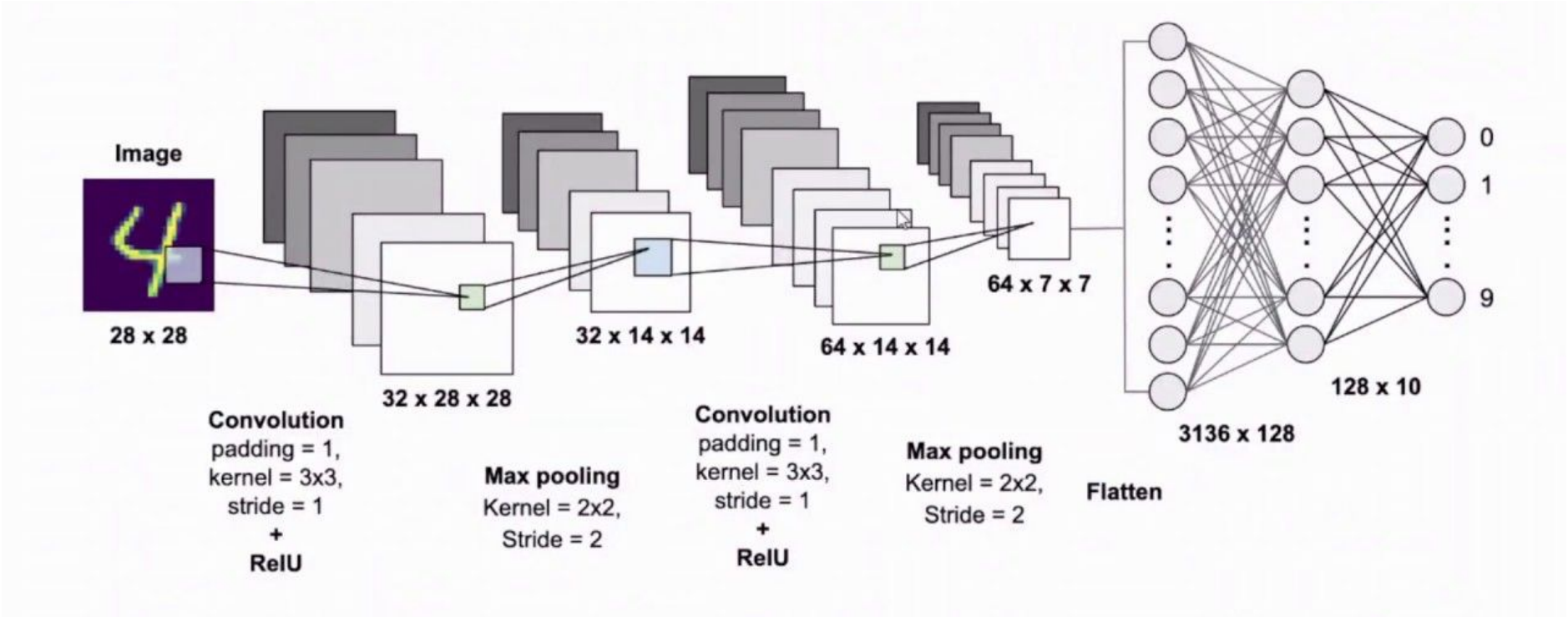


$$h_j^{(i)} = \max(0, W_j \cdot x_i + b_j)$$

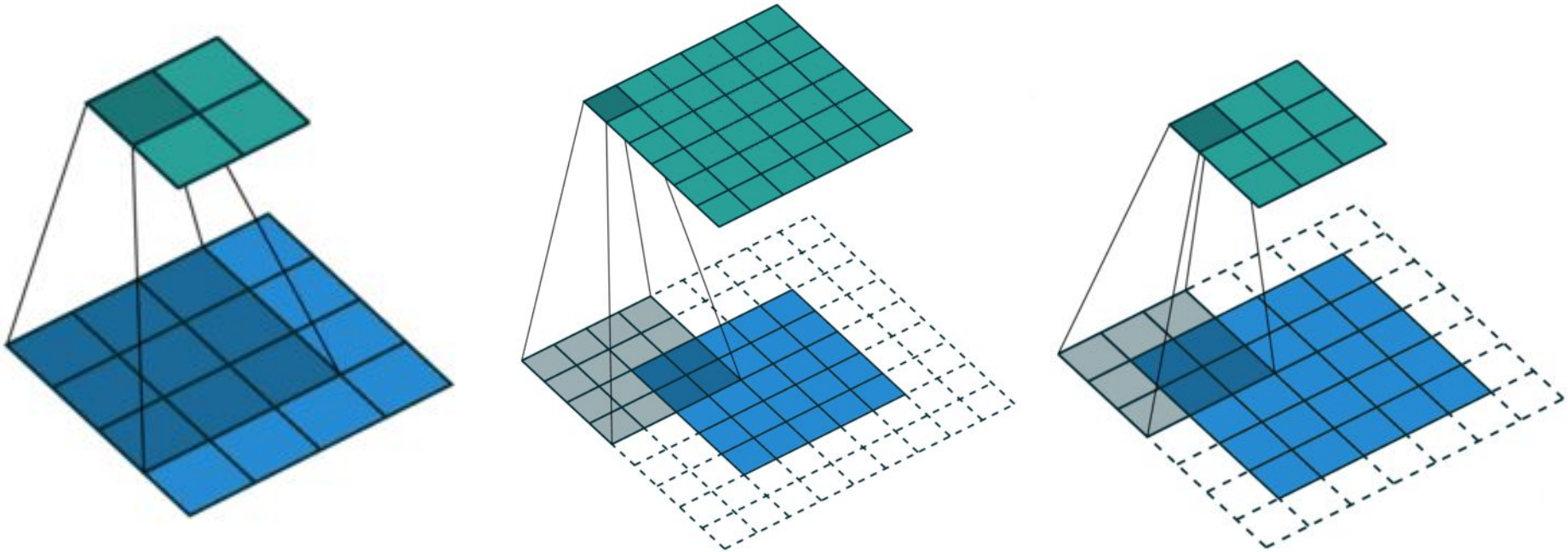
$$P(C_k|x_i) = \frac{\exp(o_k)}{\sum_{j=1}^{3} \exp(o_j)}$$

$$\hat{C} = \arg\max_k P(C_k|x_i)$$

Convolutional Neural Network



Convolutional Neural Network

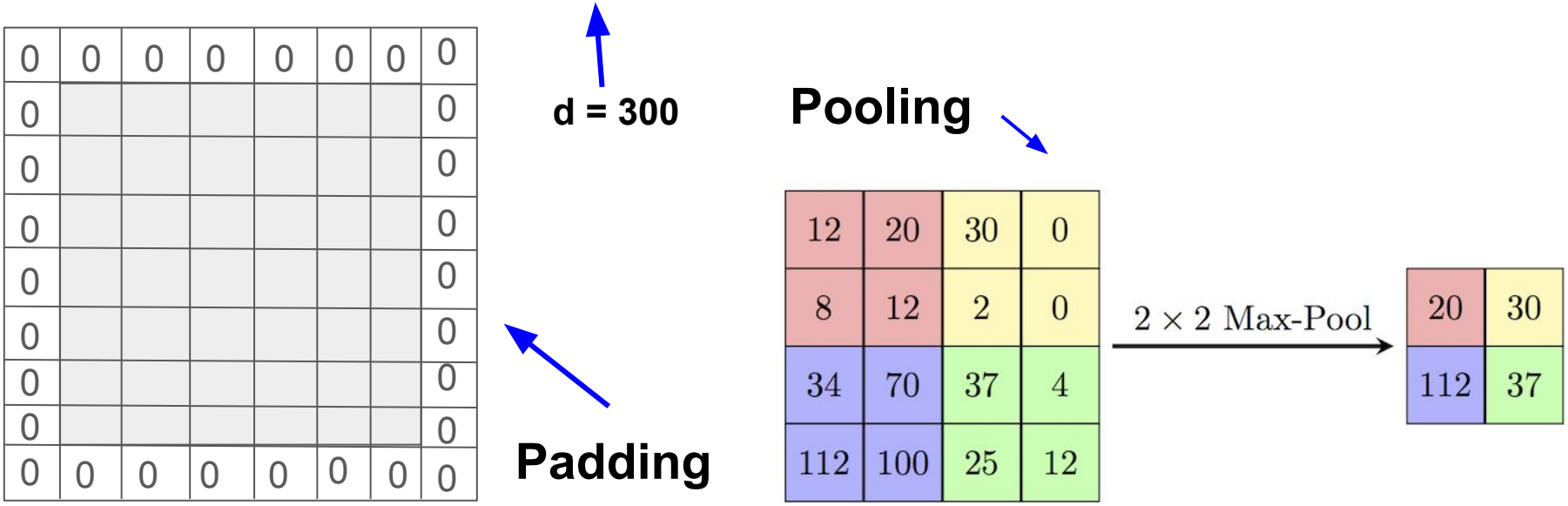


https://github.com/vdumoulin/conv_arithmetic https://arxiv.org/pdf/1603.07285

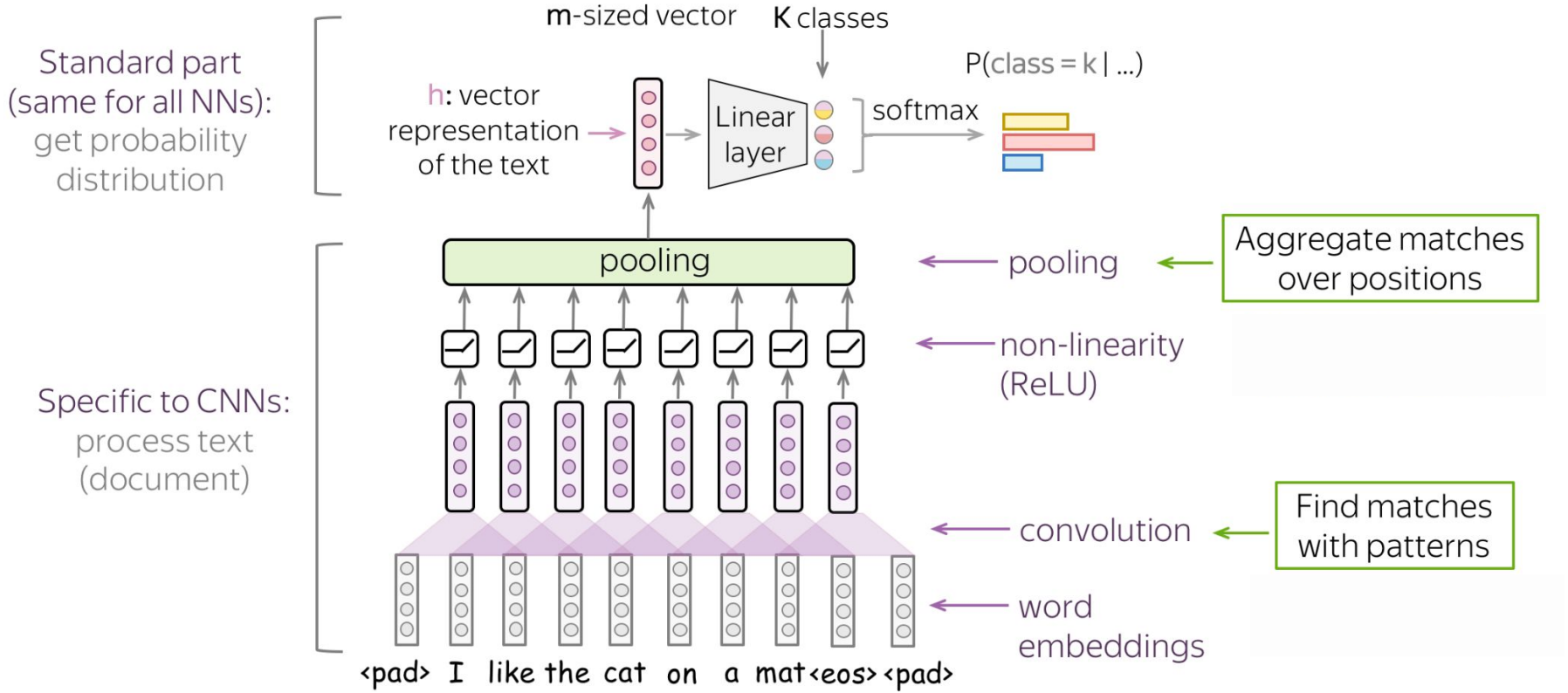
Convolutional Neural Network

Stride: Size of the step filter moves every instance of time.

Filter count: Number of filters we want to use

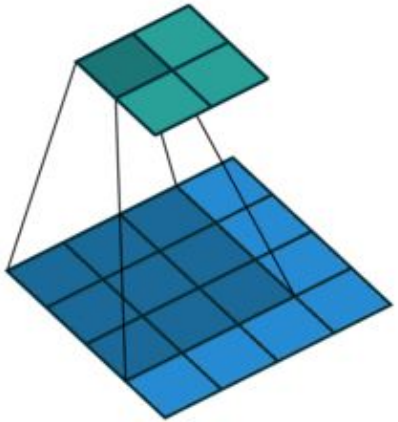


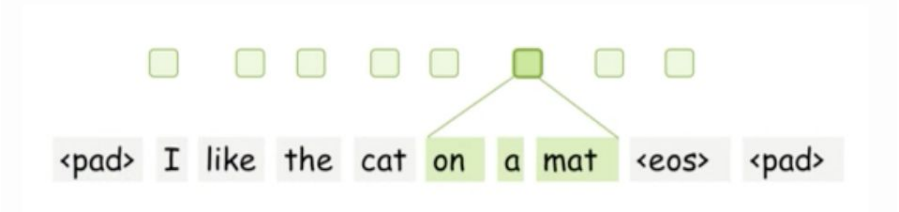
How it works?



https://lena-voita.github.io/nlp_course/text_classification.htm

Convolutional Neural Network





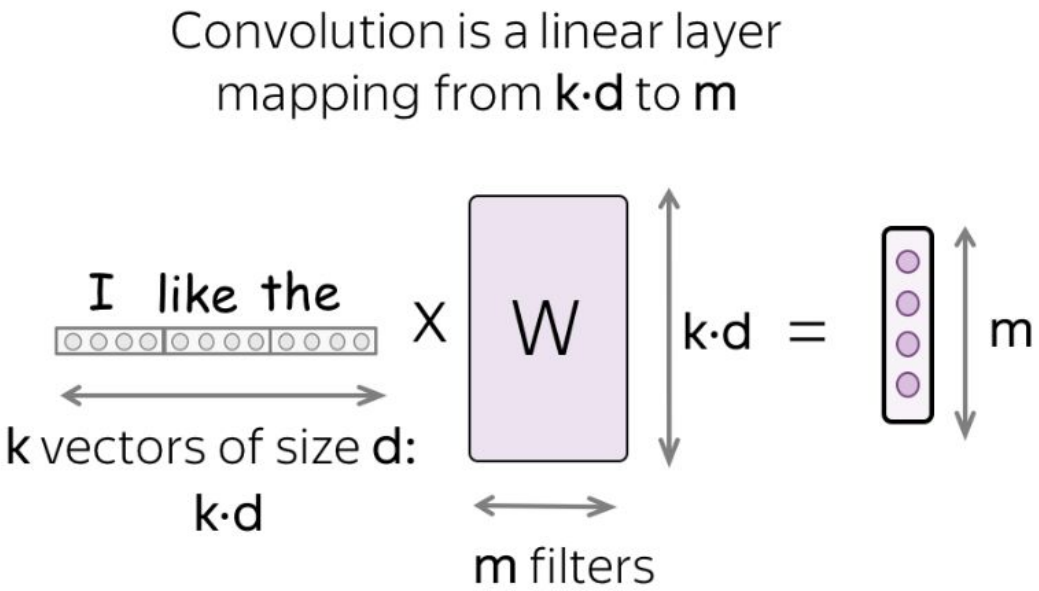
two dimensions

one dimension

https://github.com/vdumoulin/conv_arithmetic

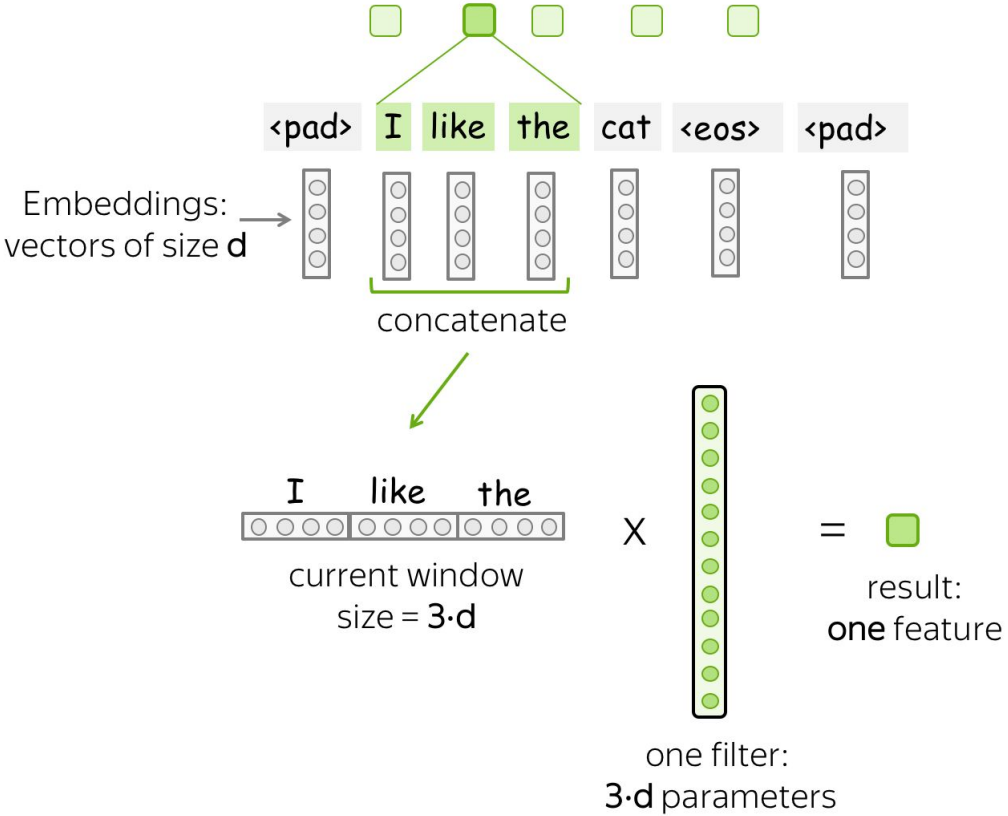
https://lena-voita.github.io/nlp_course/text_classification.htm

Convolutional Neural Network



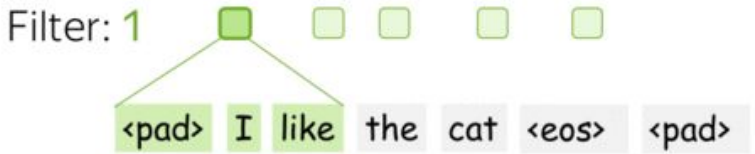
https://lena-voita.github.io/nlp_course/text_classification.htm

Convolutional Neural Network



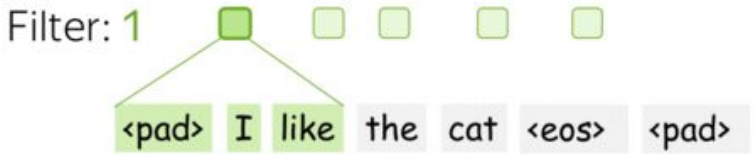
https://lena-voita.github.io/nlp_course/text_classification.htm

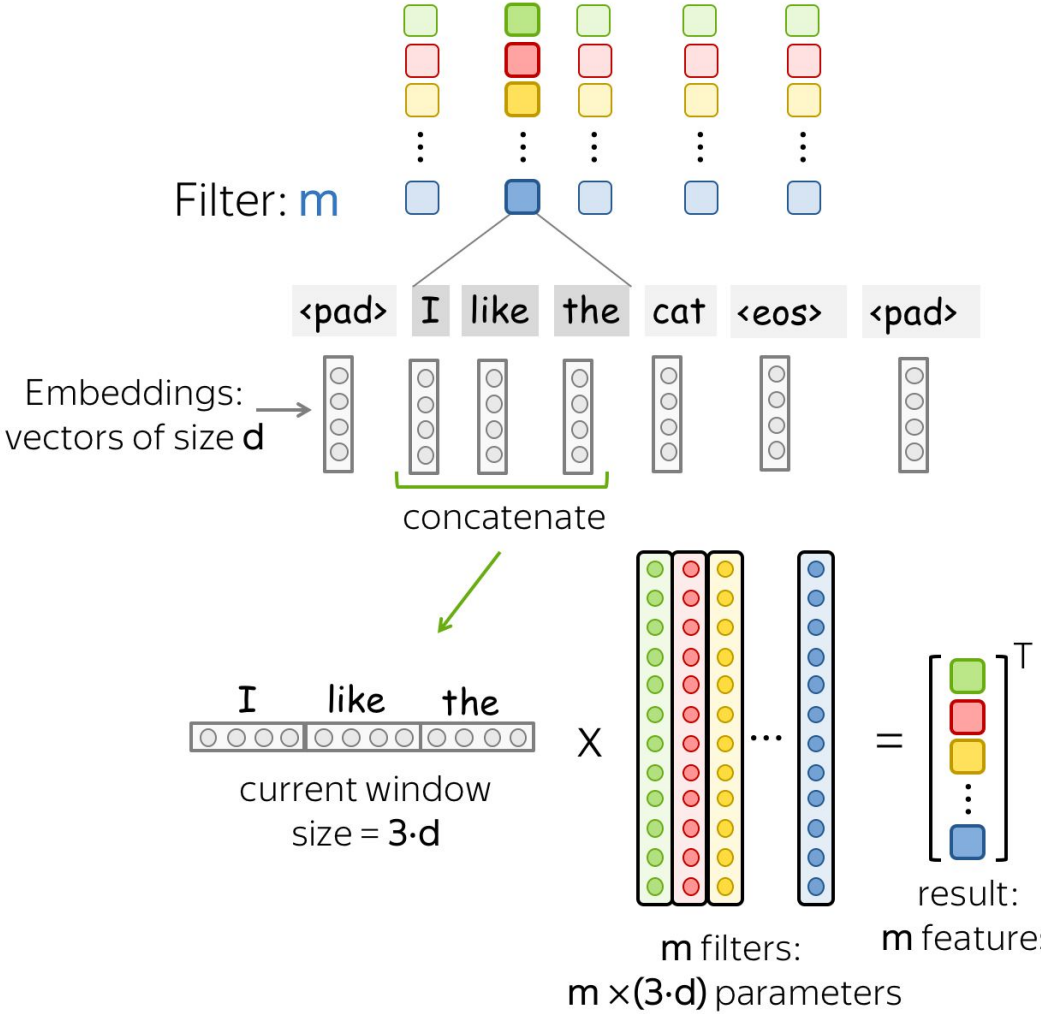
m filters: m feature extractors



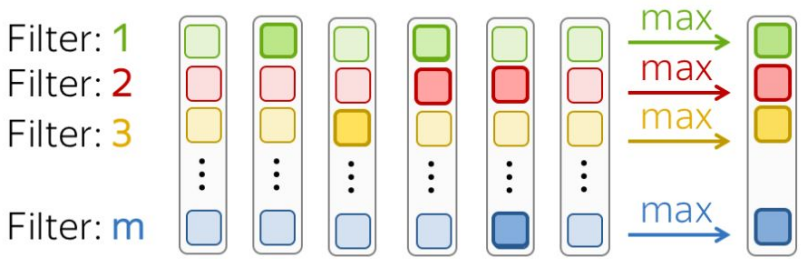
https://lena-voita.github.io/nlp_course/text_classification.htm

m filters: m feature extractors

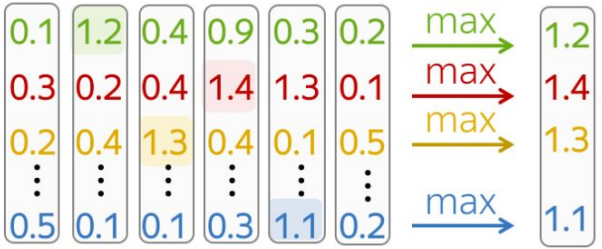




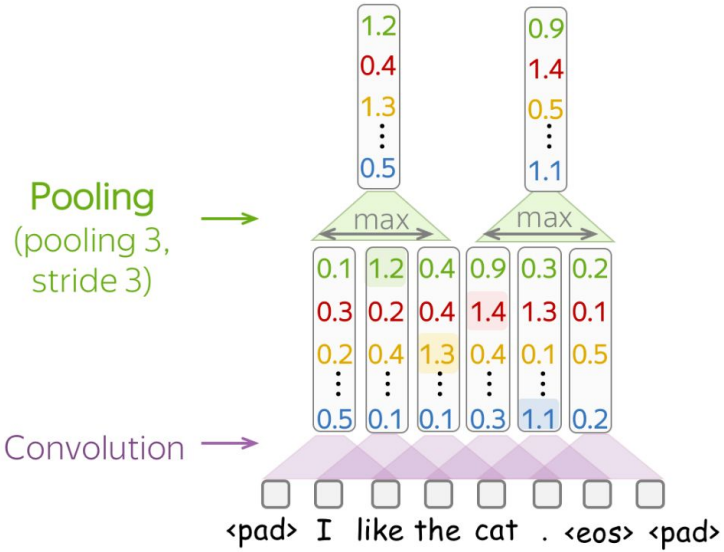
https://lena-voita.github.io/nlp_course/text_classification.htm



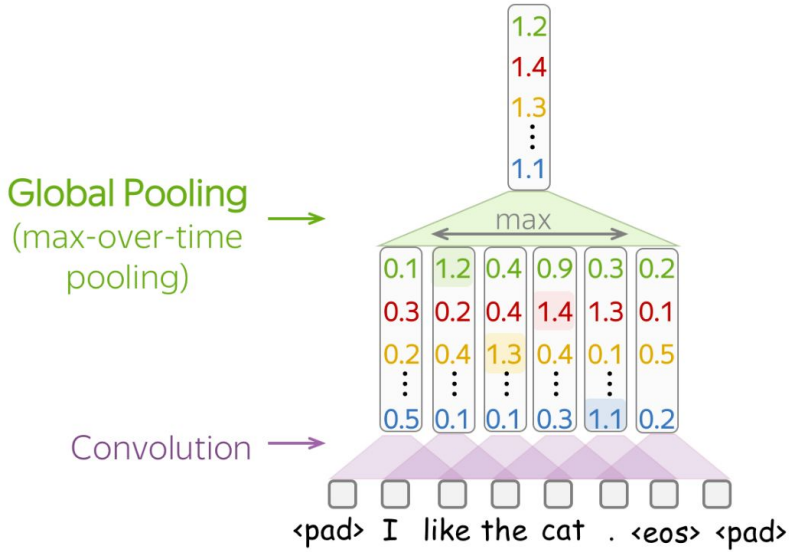
Max pooling: maximum for each dimension (feature)



Pooling



Global Pooling



https://lena-voita.github.io/nlp_course/text_classification.htm

this can be discussed endlessly, but there is a base, so let's move on...

Augmentation

Data augmentation are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data.



Synonym Replacement

Randomly replace words in the text with their synonyms.

Original: "The quick brown fox jumps over the lazy dog."

Augmented: "The fast brown fox leaps over the lazy dog."

Random Insertion

Insert random words from the sentence at random positions:

Original: "The quick brown fox jumps over the lazy dog."

Augmented: "The quick brown fox jumps over the dog lazy dog."

Random Deletion

Randomly remove words from the sentence with a certain probability.

Original: "The quick brown fox jumps over the lazy dog."

Augmented: "The quick fox over lazy dog."

Random Swap

Randomly swap the positions of two words in the sentence.

Original: "The quick brown fox jumps over the lazy dog."

Augmented: "The brown quick fox jumps over the lazy dog."

Back-Translation

Translate the text into another language and then back to English, creating a paraphrased version.

Original: "The quick brown fox jumps over the lazy dog."

Translated to French: "Le rapide renard brun saute par-dessus le chien paresseux."

Back to English: "The fast brown fox jumps over the lazy dog."

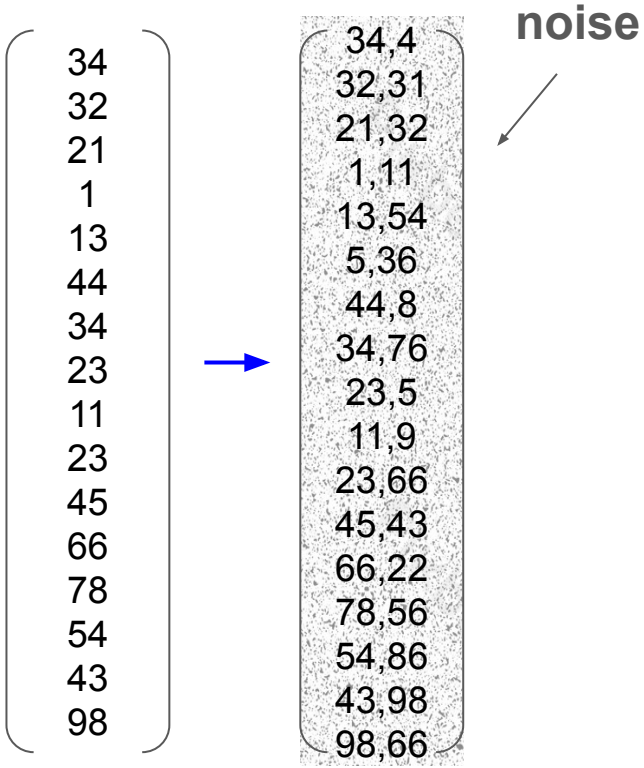
...more with text

- Text Shuffling
- Noise Injection
- Sentence Paraphrasing
- Sentence Splitting and Merging

without augmentations

augmentations

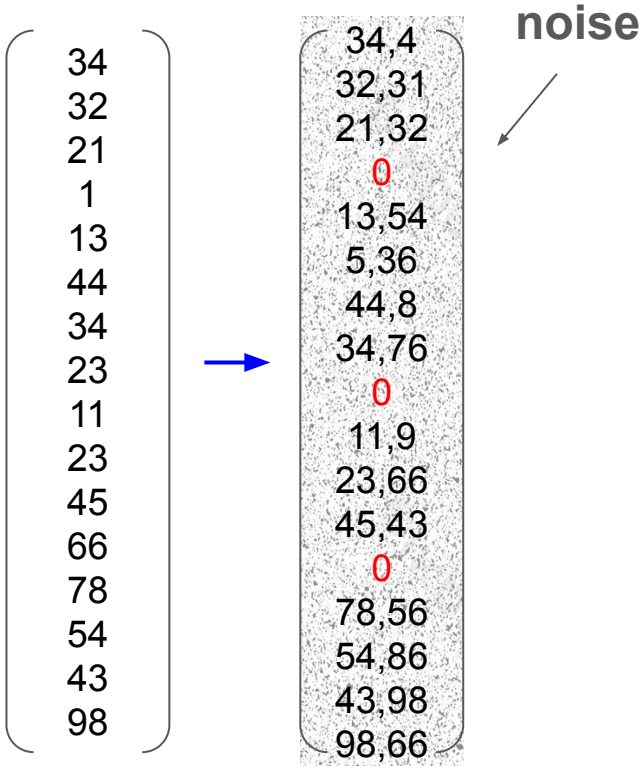
Embedding-level augmentation



- Noise Injection
- Interpolation
- Affine Transformations
- Embedding Dropout
- Contextual Augmentation

.....

Embedding-level augmentation



- Noise Injection
- Interpolation
- Affine Transformations
- Embedding Dropout
- Contextual Augmentation

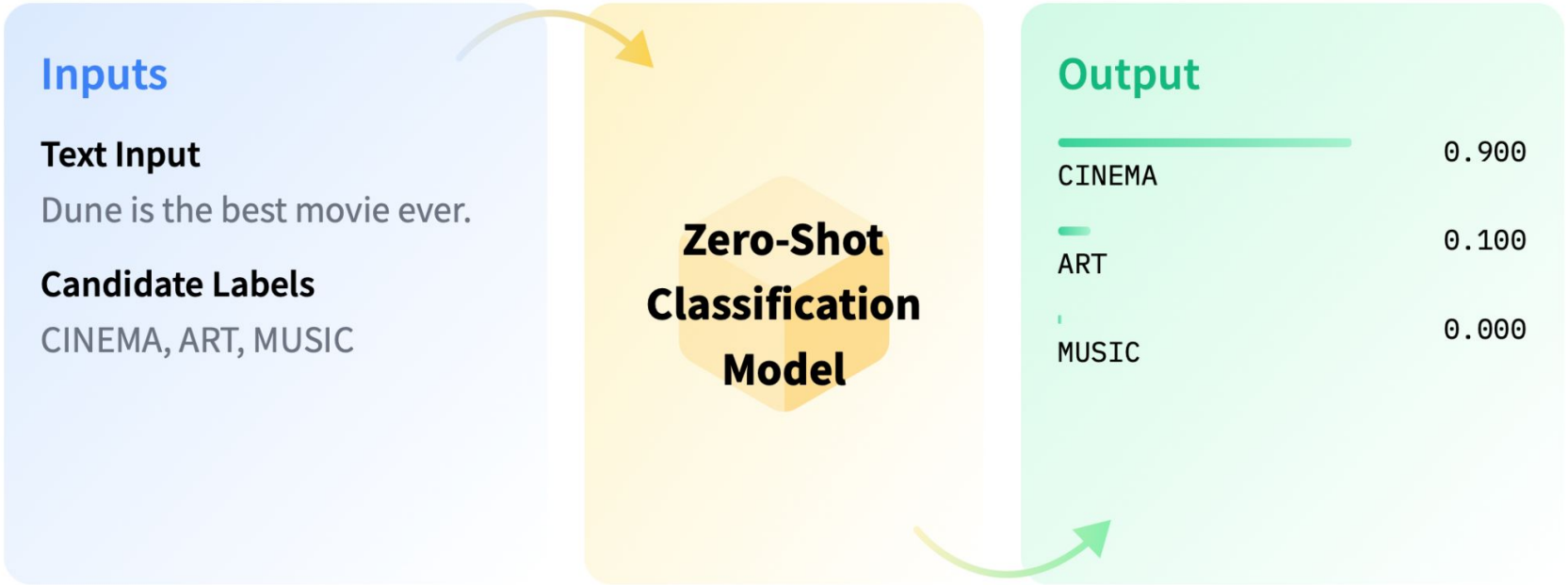
.....

Tools

- nlpaug
- Albumentations
- EDA (Easy Data Augmentation)
- TextAugment

Zero-shot classification

Zero Shot Classification is the task of predicting a class that wasn't seen by the model during training.



Zero-shot classification

Classify the following input text into one of the following three categories: [positive, negative, neutral]

Input Text: I haven't slept well for 2 days, it's like I'm restless. why huh :([].

Sentiment: Anxiety

GPT-3

T5