Low-Resource Natural Language Processing

Behnam Sabeti
Sharif University of Technology
October 2019

Who-am-i?

Behnam Sabeti

Ph.D. Candidate at Sharif University of Technology

Project Manager and NPL Expert at Miras Technologies International

Does all kind of NLP stuff specially on Persian

behnamsabeti

behnamsabeti

NLP @ Miras

- Our focus at Miras NLP team is on developing text processing services for Persian:
  - Document classification
  - Named entity recognition
  - Sentiment analysis
  - Emotion analysis
  - …
- Challenge:
  - Data!

Dataset	Size (documents)
IMDB	50 K
SST	10 K
Sentiment140	160 K
Amazon Product Data	142.8 M

Problem?

- Deep learning models are data hungry
- Persian NLP community is not large
  - We do not have enough public resources
  - Funding is also limited so we can’t afford building huge resources either

Get More Date

Get Better Data

Use Related Data

Problem Modeling

Get Better Data

Use Related Data

Problem Modeling

Solutions

- Self Supervision
  - Emotion Analysis
- Weak Supervision
  - Document Classification
- Transfer Learning
  - Named Entity Recognition
- Multi-Task Learning
  - Satire Detection
- Active Learning
  - Sentiment Analysis

Self Supervision

- Straight forward (document, label) modeling is not always your best choice.
- Model your problem in an easy-to-acquire-label setting:
  - Self-supervision
    - Labels are already in your data:
      - Language modeling
      - Word embedding
      - Emotion Analysis

Case Study: Emotion Analysis

- Emoji is a good indicator of emotion
- Instead of manually label your data use emoji
- Your dataset needs no hand-labeling effort!

$$\mathit{Emoji\ Prediction} \implies \mathit{Emotion\ Analysis}$$

I love mom's cooking	😋 49.1%	😍 8.8%	❤️ 3.1%	☺️ 3.0%	♡ 2.9%
I love how you never reply back..	😒 14.0%	😑 8.3%	😠 6.3%	😐 5.4%	💔 5.1%
I love cruising with my homies	😎 34.0%	👌 6.6%	✌️ 5.7%	😌 4.1%	💯 3.8%
I love messing with yo mind!!	😜 17.2%	😈 11.8%	😏 8.0%	😉 6.4%	🙊 5.3%
I love you and now you're just gone..	💔 39.1%	😔 11.0%	😞 7.3%	😪 5.3%	😢 4.5%
This is shit	😠 7.0%	😡 6.4%	😞 6.0%	😒 6.0%	😤 5.8%
This is the shit	🎧 10.9%	🎶 9.7%	👌 6.5%	😎 5.7%	😏 4.8%

Image: medium.com/@bjarkefelbo/what-can-we-learn-from-emojis

Modeling

DeepMoji Model

- *Predict Emoji*
- *Map Emoji to Emotion*

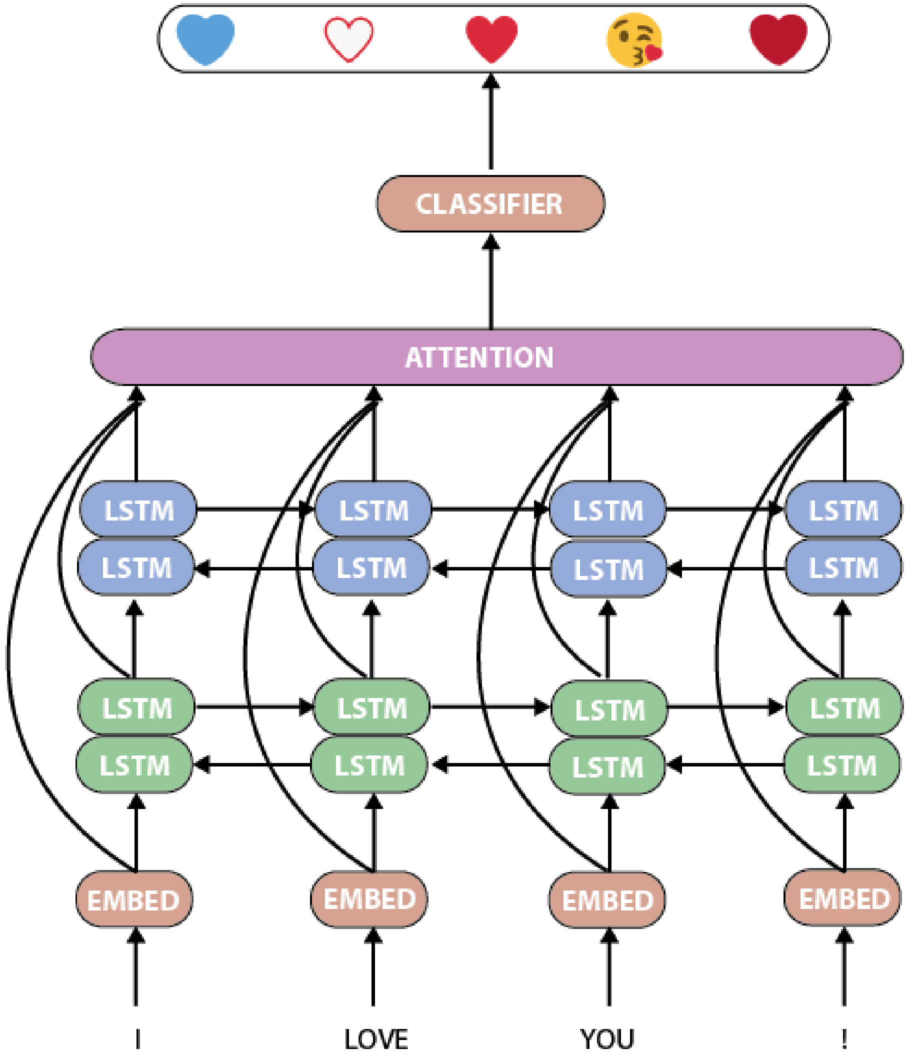
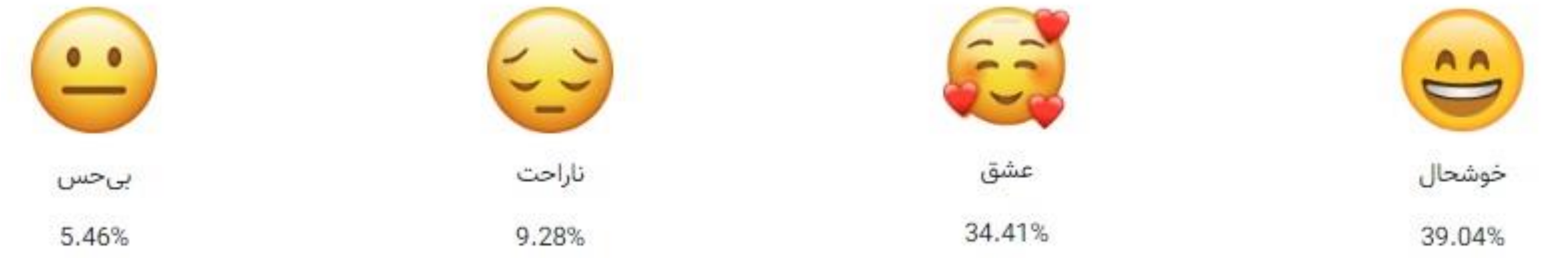


Image: medium.com/huggingface/understanding-emotions-from-keras-to-pytorch

پاییز رو خیلی دوست دارم! فصل
به این خوبی آخه! بیا زودتر...



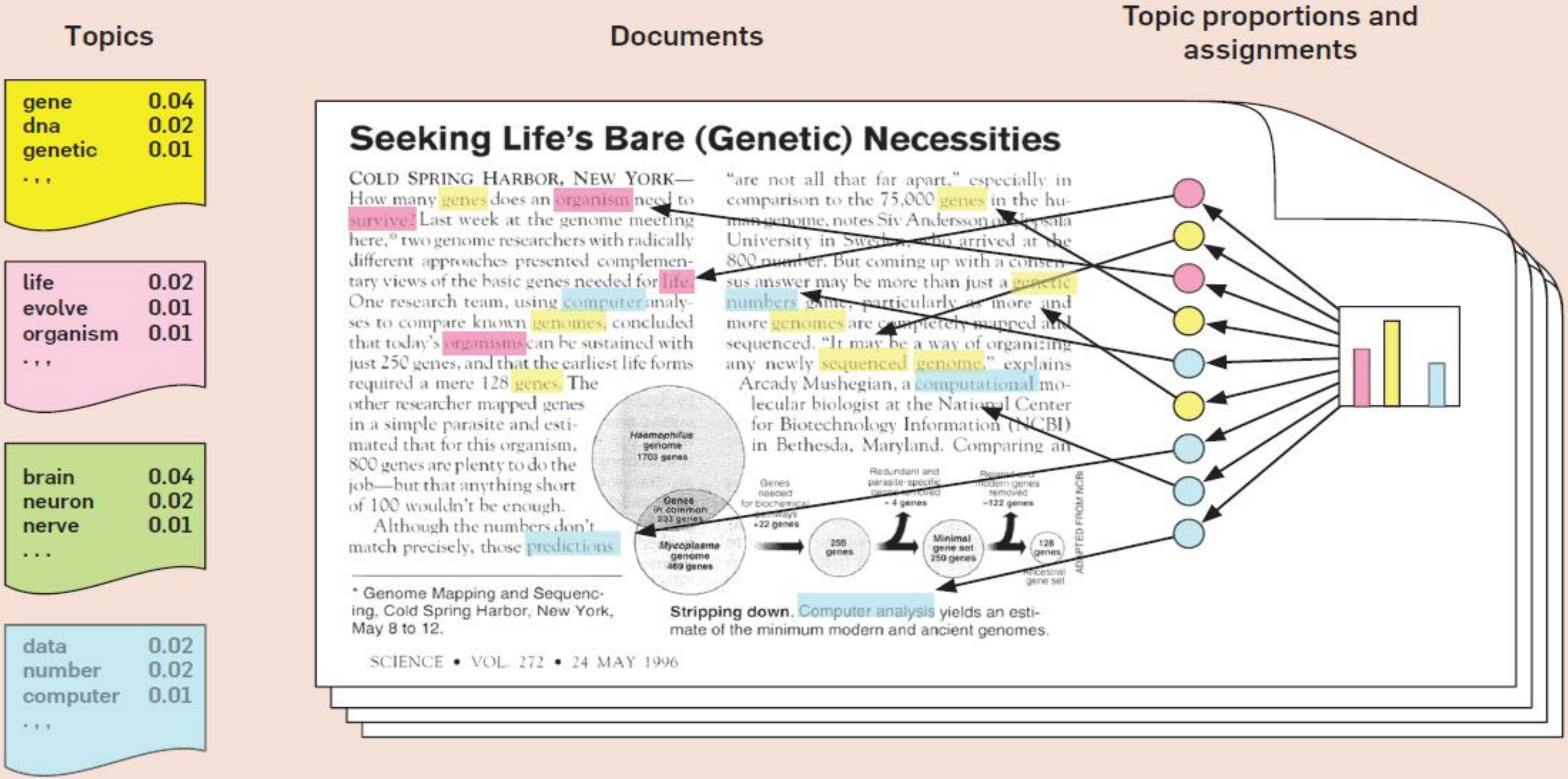
خوشحال	عشق	ناراحت	بی‌حس
39.04%	34.41%	9.28%	5.46%

Weak Supervision

- Provide noisy labels using a set of heuristics or domain knowledge
  - Use other weak classifiers
  - Constraints
  - Data transformation
- Think of a transformation on your data:
  - Reduce the effort in annotation process

Case Study: Document Classification

- Latent Dirichlet Allocation is a generative model for topic modeling:
  - computes a set of topics: each topic is a distribution on words
  - Computes the distribution of each document on topics
- Instead of manually labeling documents, annotate topics!
- With this transformation you can get a pretty good result by just labeling a handful of topics



Topics

gene	0.04
dna	0.02
genetic	0.01
. . .	

life	0.02
evolve	0.01
organism	0.01
. . .	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. . .	

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

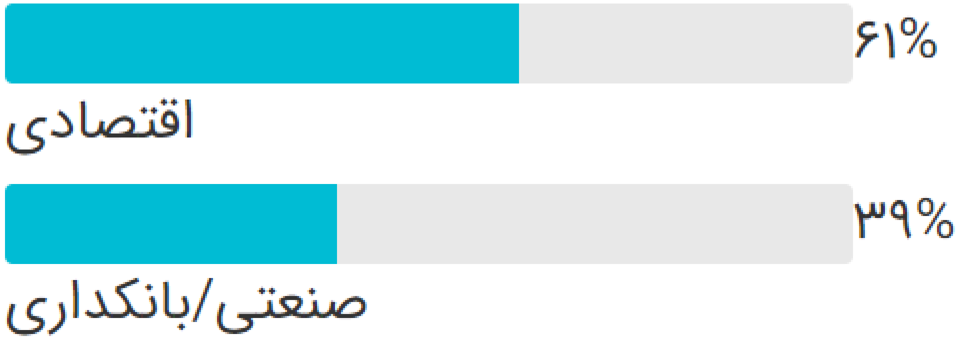
[*] Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

Image: m-cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models

در شرایط فعلی طبق گزارش واحد اطلاعاتی اکونومیست، بیشترین ریسکی که اقتصاد ایران را تهدید می‌کند، ریسک بخش بانکی و ریسک سیاسی است. عصر بانک؛ معاونت بررسی های اقتصادی اتاق بازرگانی تهران در گزارشی به واکاوی مدل ریسک کشوری پرداخته است. براساس این گزارش مدل ریسک کشوری، مدلی است که به منظور سنجش و مقایسه ریسک اعتباری کشورهای مختلف توسط واحد اطلاعاتی اکونومیست طراحی شده است. این ابزار تعاملی، امکان کمی‌سازی ریسک مبادلات مالی از جمله وام‌های بانکی، تامین مالی تجاری و سرمایه‌گذاری در اوراق بهادار را فراهم می‌کند...



Transfer Learning

- Train on a task for which you have enough data
- Fine-Tune the trained model on a new task (for which limited data is available)
- The source and target tasks need to have common characteristics:
  - Source: Language modeling, Target: Document Classification
  - Source: Emotion Detection, Target: Satire Detection
  - Source: Document Classification, Target: Sentiment Analysis
    - Document Classification: word based task
    - Sentiment Analysis: Phrase level and semantic based task

Related

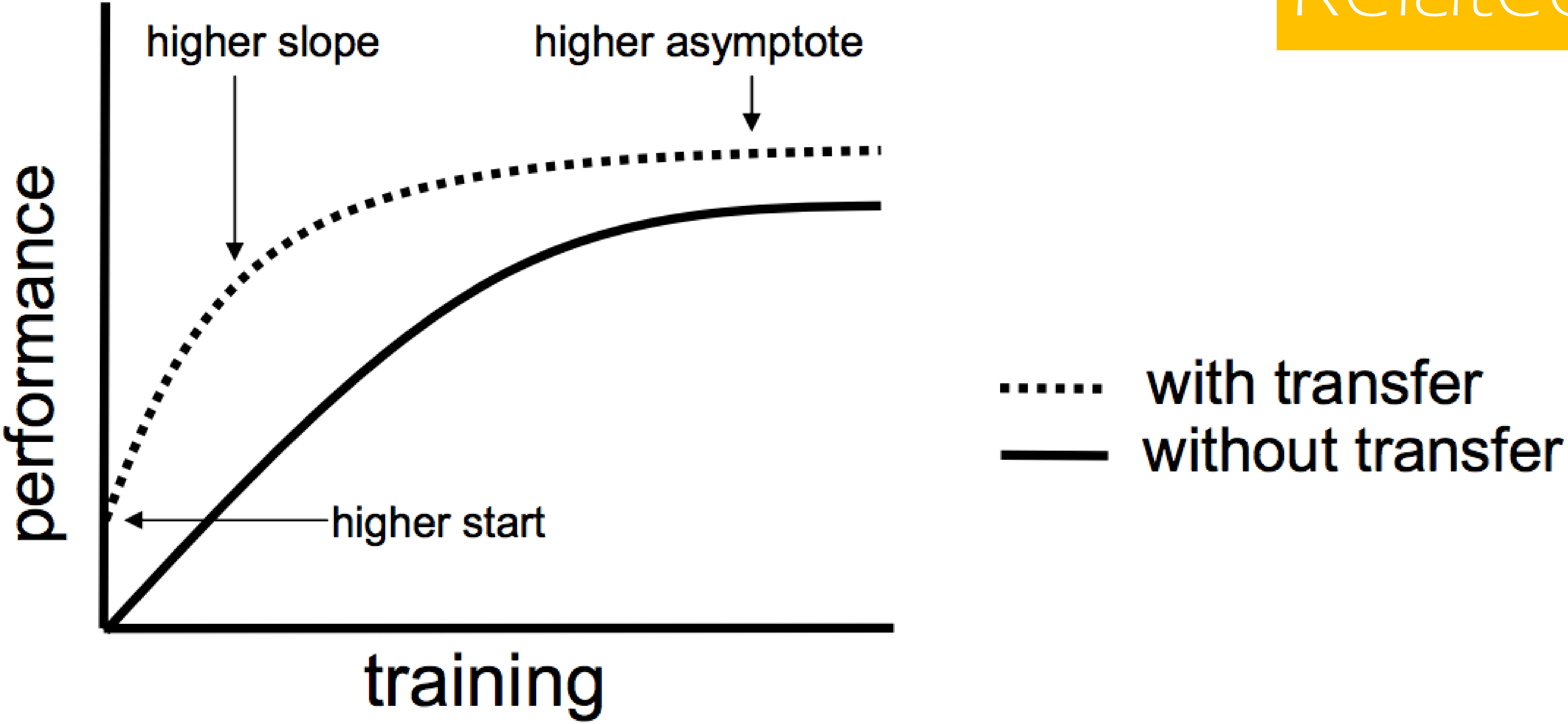


Image: machinelearningmastery.com/transfer-learning-for-deep-learning

Pre-Trained Models

- Train your own model on a source task Or use a Pre-trained model
- Pre-Trained model are a good choice because they are trained on HUGE datasets.
- Language modeling pre-trained models:
  - BERT
  - GPT
  - XLNet
  - XLM
  - CTRL
  - …

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

Randomly mask 15% of tokens

Input

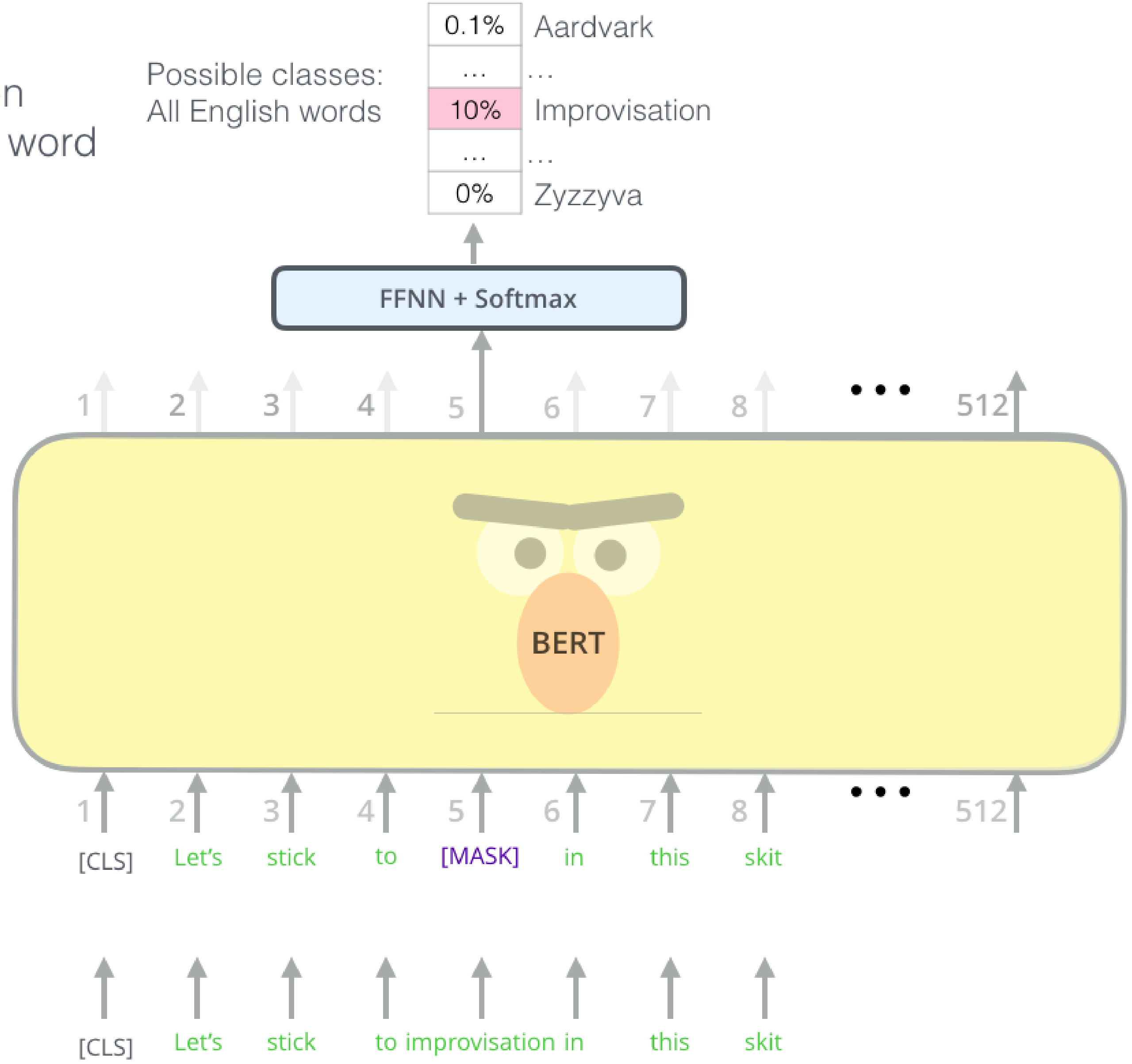
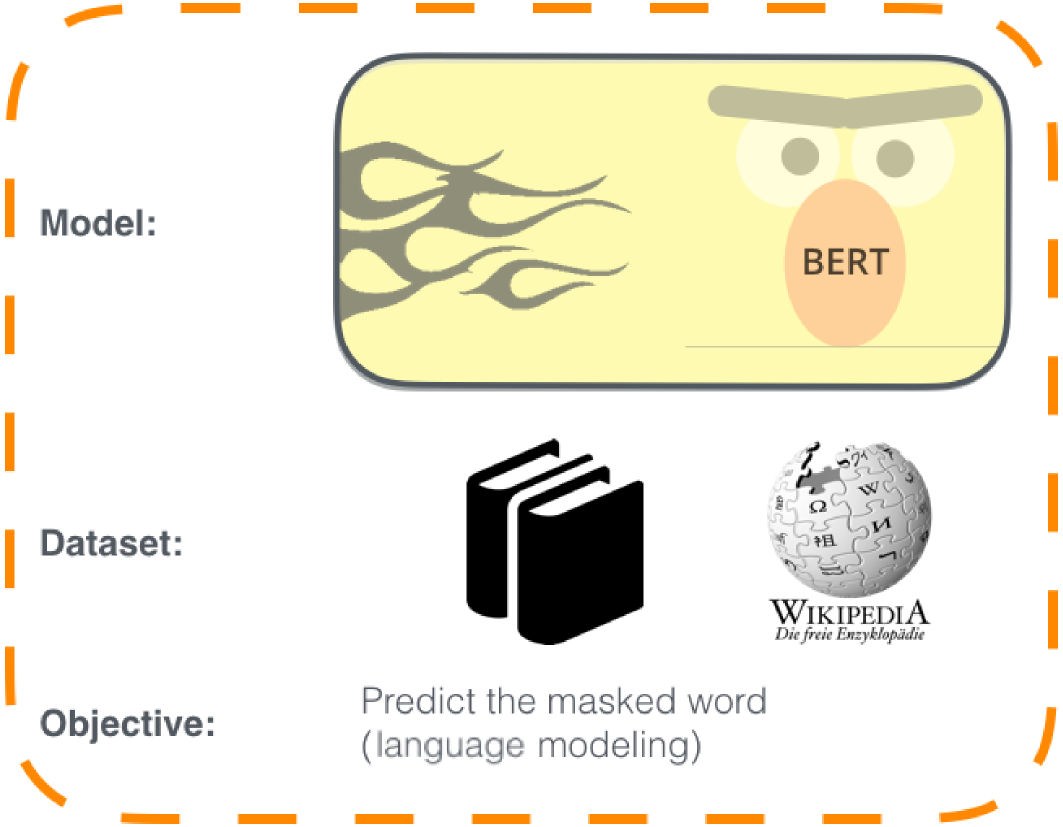


Image:
jalammar.github.io/illustrated-bert

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

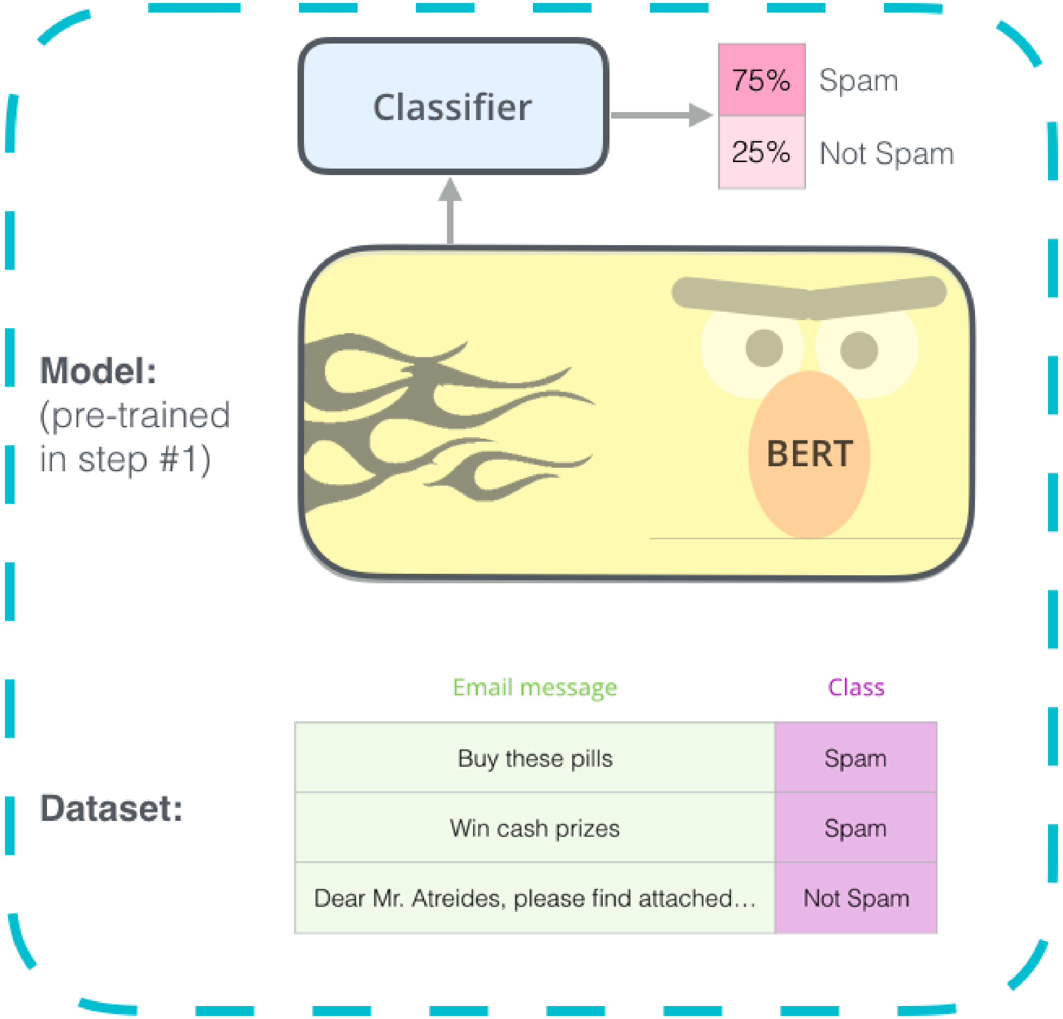
Semi-supervised Learning Step



Objective: Predict the masked word (language modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

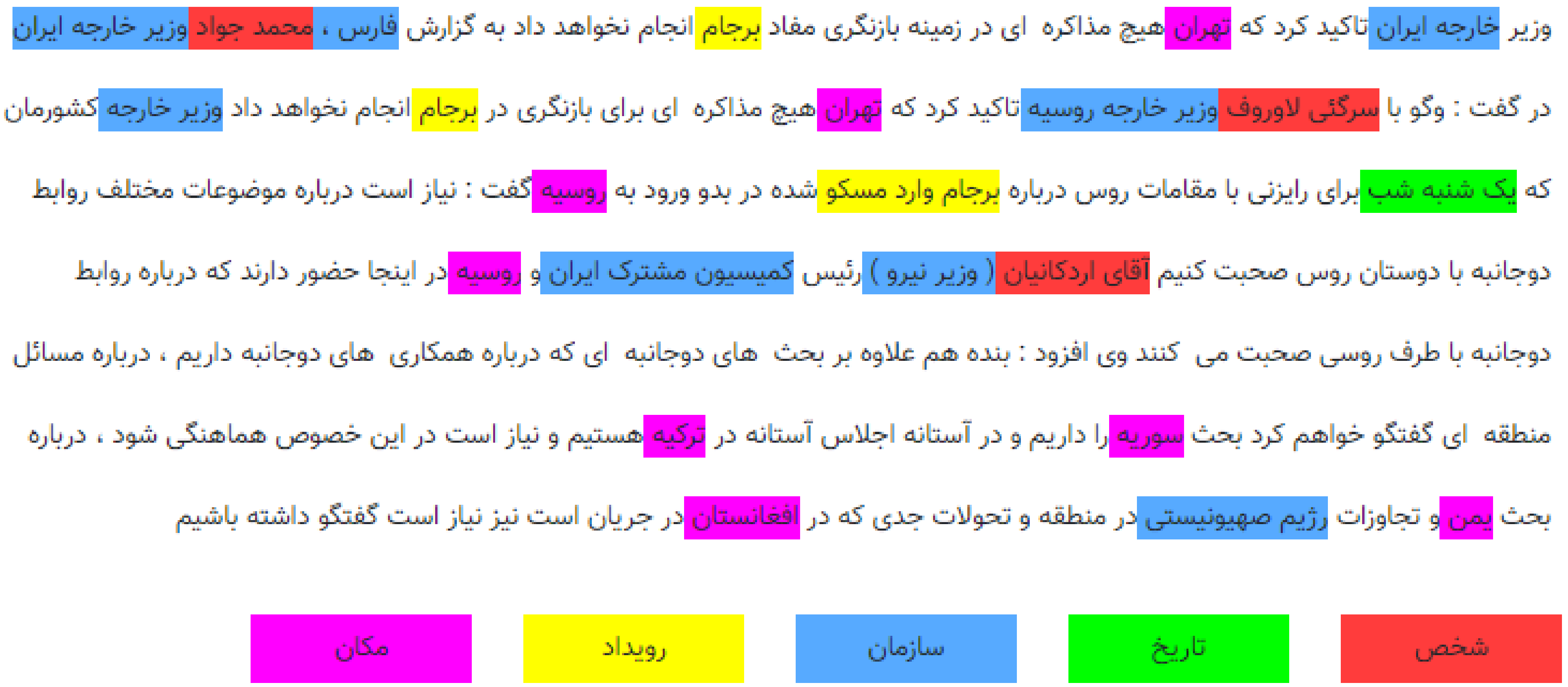


Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Image: medium.com/huggingface/introducing-fastbert-a-simple-deep-learning-library-for-bert-models

Case Study: Named Entity Recognition

- Target Task: Named Entity Recognition
  - Extract locations, persons, organizations, events and times from text
- Source: Multilingual BERT model
- Data: 50K hand labeled sentences with NER tags



وزیر خارجه ایران تاکید کرد که تهران هیچ مذاکره ای در زمینه بازنگری مفاد برجام انجام نخواهد داد به گزارش فارس ، محمد جواد وزیر خارجه ایران
در گفت : وگو با سرگئی لاوروف وزیر خارجه روسیه تاکید کرد که تهران هیچ مذاکره ای برای بازنگری در برجام انجام نخواهد داد وزیر خارجه کشورمان
که یک شنبه شب برای رایزنی با مقامات روس درباره برجام وارد مسکو شده در بدو ورود به روسیه گفت : نیاز است درباره موضوعات مختلف روابط
دوجانبه با دوستان روس صحبت کنیم آقای اردکانیان (وزیر نیرو) رئیس کمیسیون مشترک ایران و روسیه در اینجا حضور دارند که درباره روابط
دوجانبه با طرف روسی صحبت می کنند وی افزود : بنده هم علاوه بر بحث های دوجانبه ای که درباره همکاری های دوجانبه داریم ، درباره مسائل
منطقه ای گفتگو خواهم کرد بحث سوریه را داریم و در آستانه اجلاس آستانه در ترکیه هستیم و نیاز است در این خصوص هماهنگی شود ، درباره
بحث یمن و تجاوزات رژیم صهیونیستی در منطقه و تحولات جدی که در افغانستان در جریان است نیز نیاز است گفتگو داشته باشیم

🟥 شخص　🟩 تاریخ　🟦 سازمان　🟨 رویداد　🟪 مکان

Multi-Task Learning

- Train multiple tasks together
  - More data
  - Synergic effects in training
    - Tasks: tweet reconstruction + emoji prediction + satire detection
      - General features
      - Emotion features
      - Satire features
- Entails multi objective loss functions

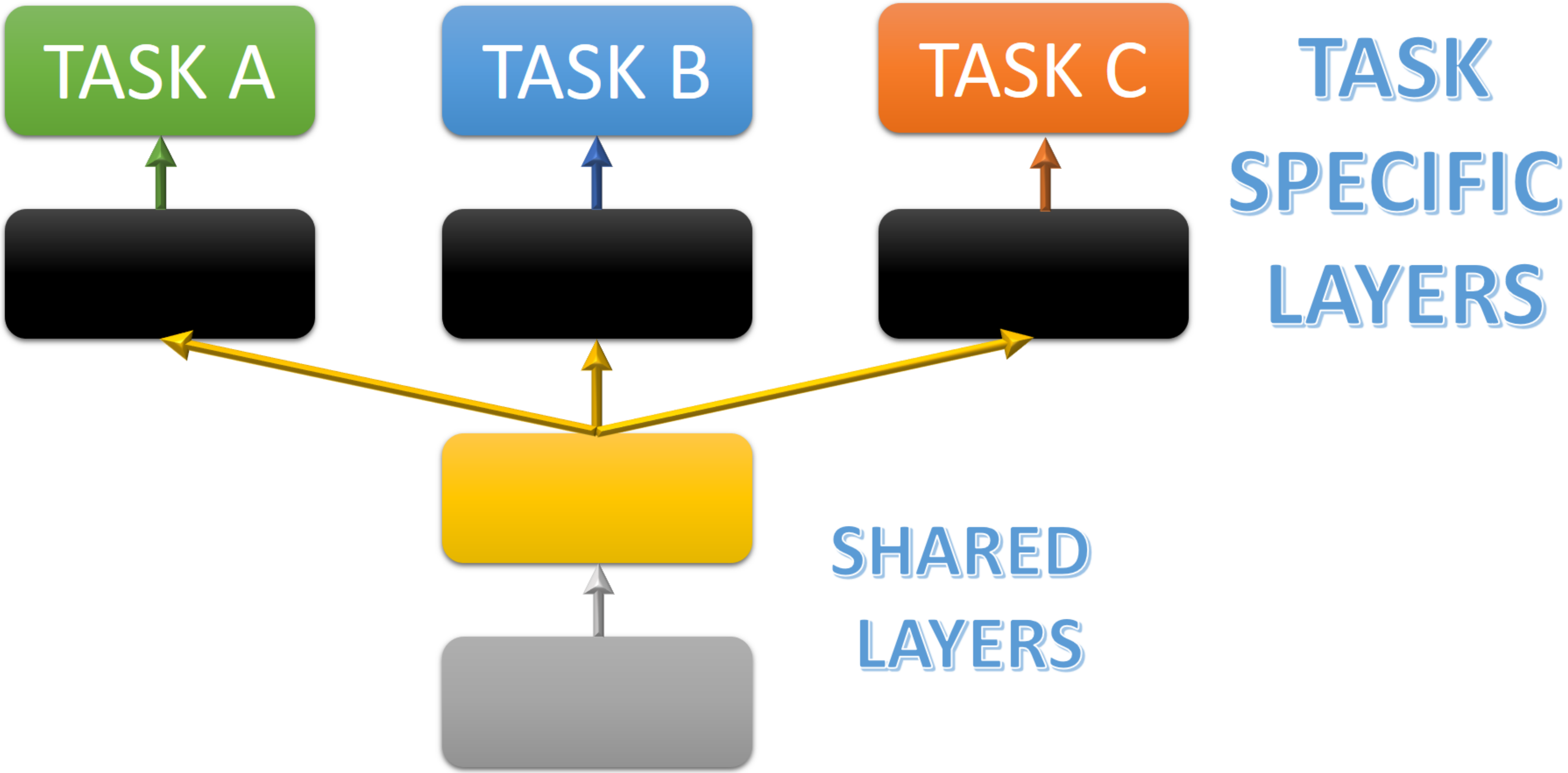
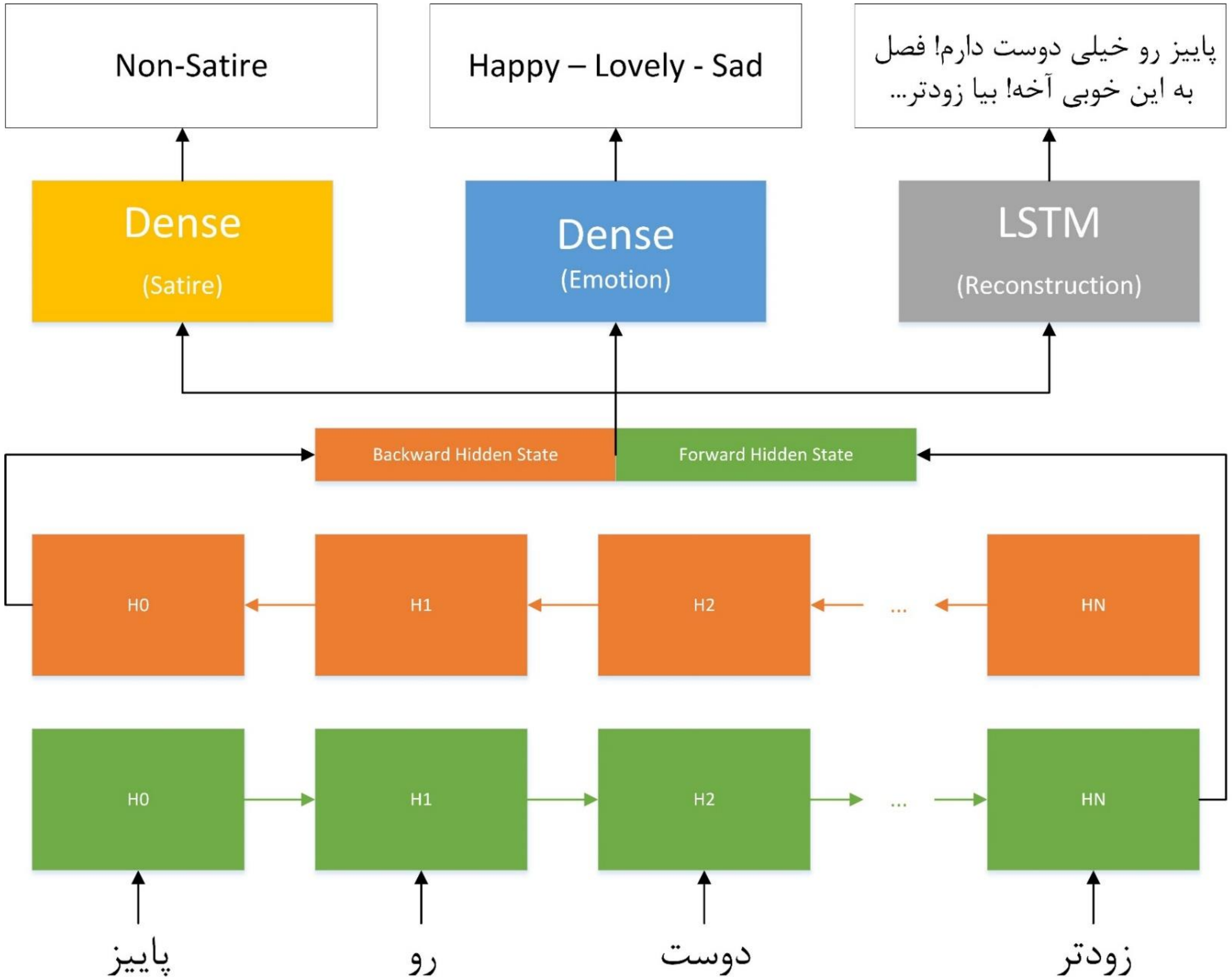


Image: medium.com/manash-en-blog/multi-task-learning-in-keras-implementation-of-multi-task-classification-loss

Case Study: Satire Detection

- Satire dataset: 2K tweets
- Emotion dataset: 300K tweets
- Reconstruction Tweets: as much as you have! (200M)

Related



Non-Satire

Happy – Lovely - Sad

پاییز رو خیلی دوست دارم! فصل
به این خوبی آخه! بیا زودتر...

Dense

(Satire)

Dense

(Emotion)

LSTM

(Reconstruction)

Backward Hidden State

Forward Hidden State

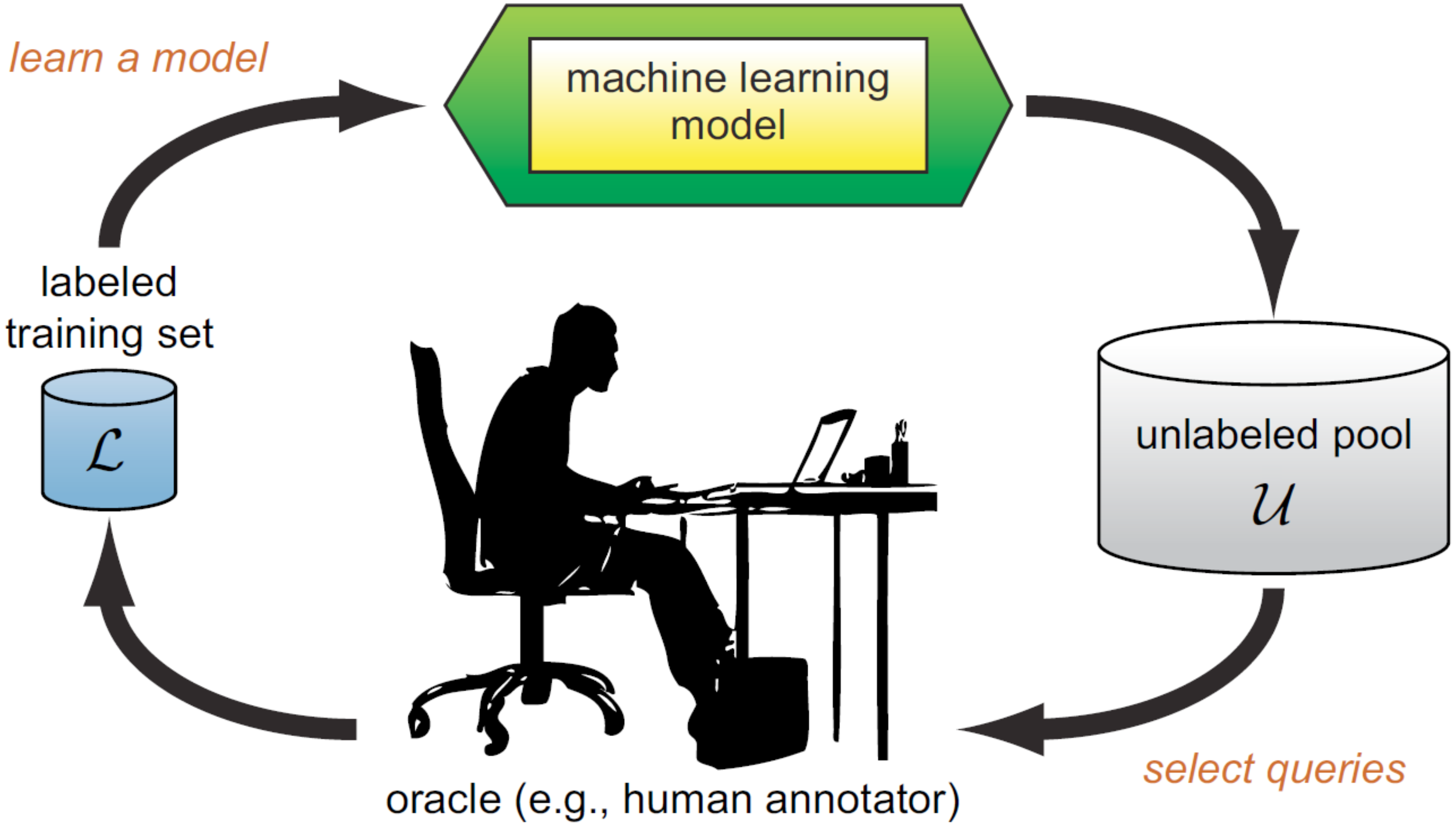
Satire Model	Performance (F1)
Single task	55 %
Multi task	68 %

Better

Active Learning

- How to select samples for annotation?
  - Random
    - Annotate as much as you can
  - Smart
    - Annotate "Better" samples

Better


learn a model
machine learning model
labeled training set
$\mathcal{L}$
unlabeled pool
$\mathcal{U}$
oracle (e.g., human annotator)
select queries

Better

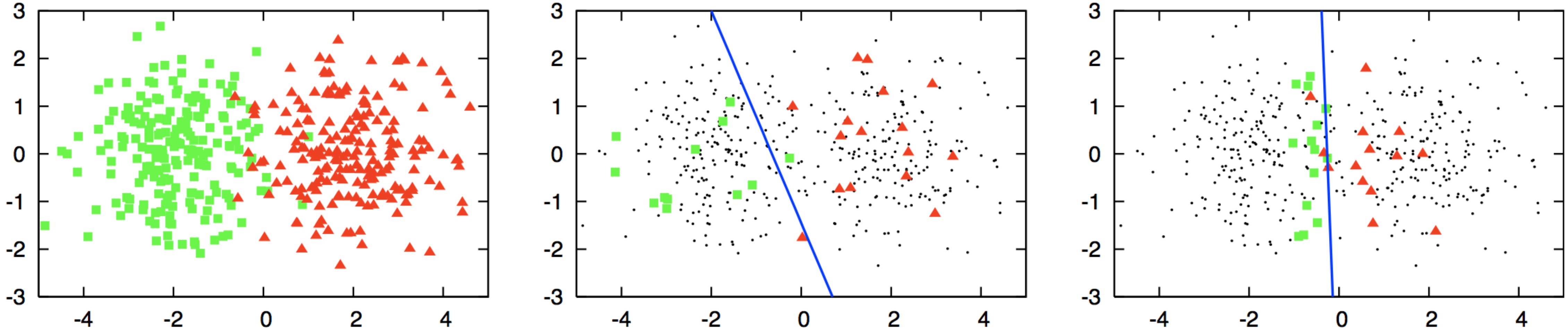


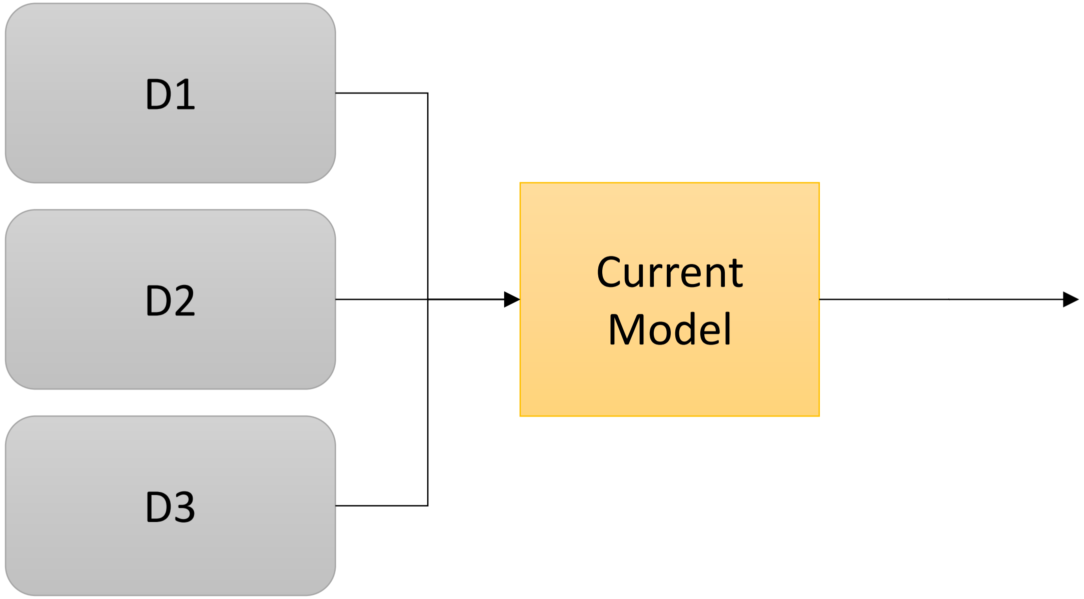
Image: www.datacamp.com/community/tutorials/active-learning

Better

Active Learning

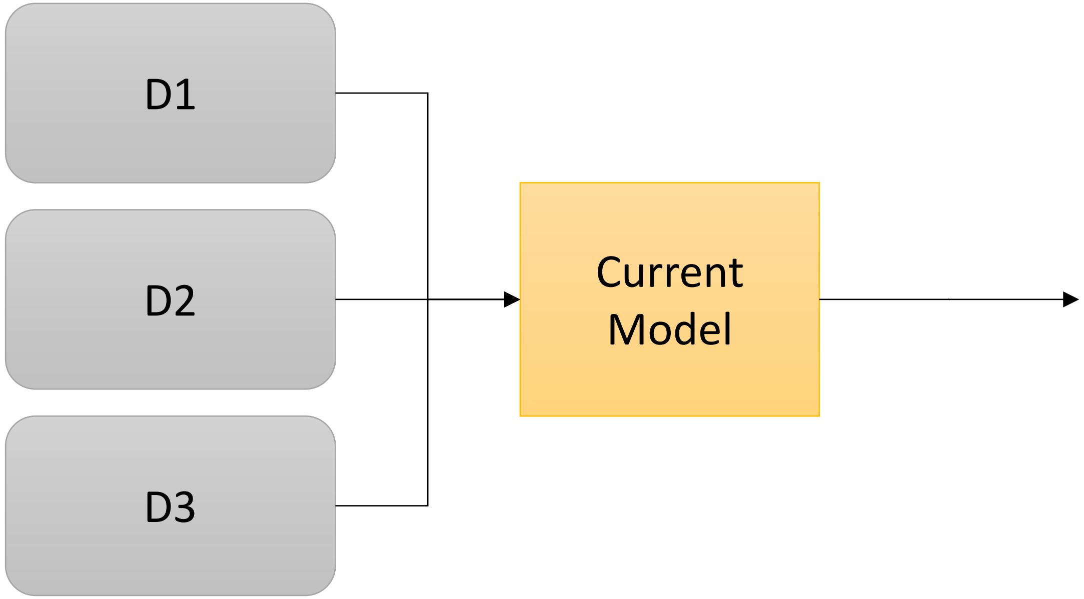
- How to select samples for annotation
  - Random
  - Smart
    - Select samples that current model is uncertain about (LC)
    - Select samples with low margin between category labels (Margin)
    - Select samples with the highest entropy (Entropy)
- Get better performance with fewer samples

Better



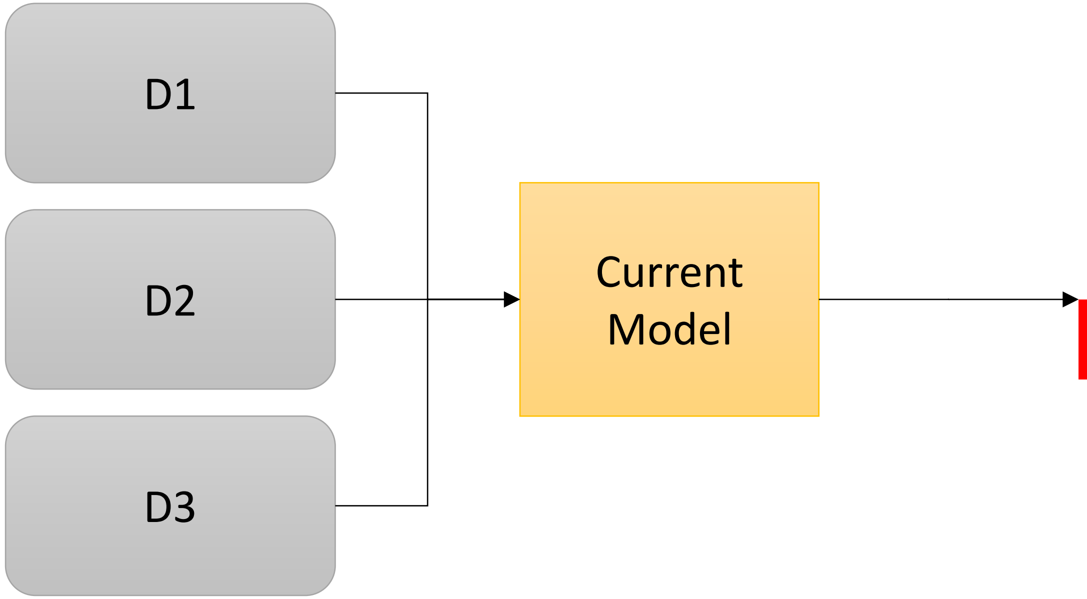
Positive	Neutral	Negative
0.5	0.45	0.05
0.4	0.3	0.3
0.95	0.05	0

Better



Positive	Neutral	Negative
0.5	0.45	0.05
0.4	0.3	0.3
0.95	0.05	0

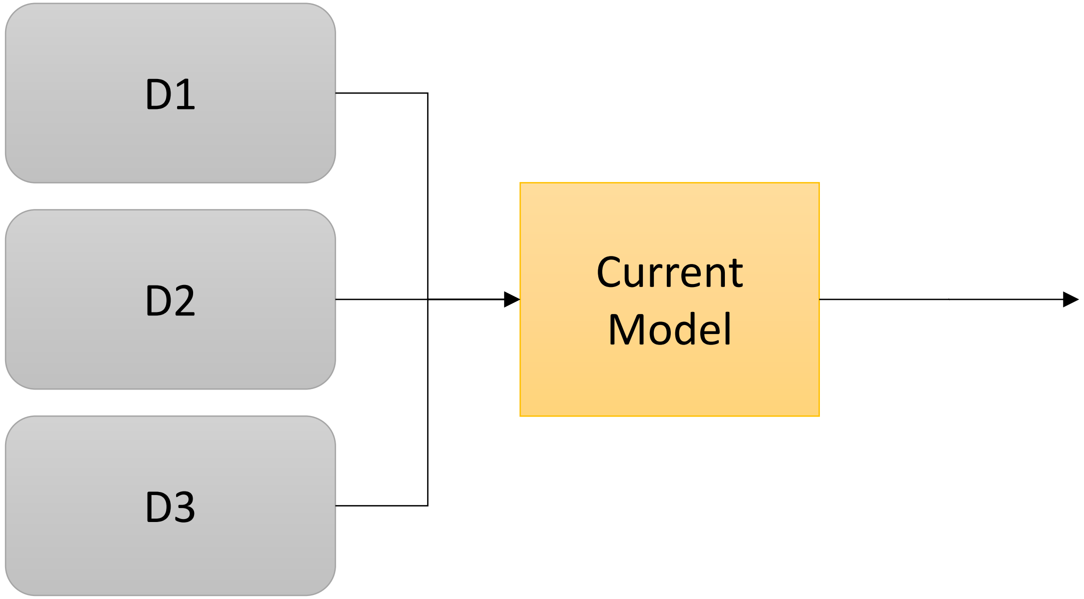
Better



Positive	Neutral	Negative
0.5	0.45	0.05
0.4	0.3	0.3
0.95	0.05	0

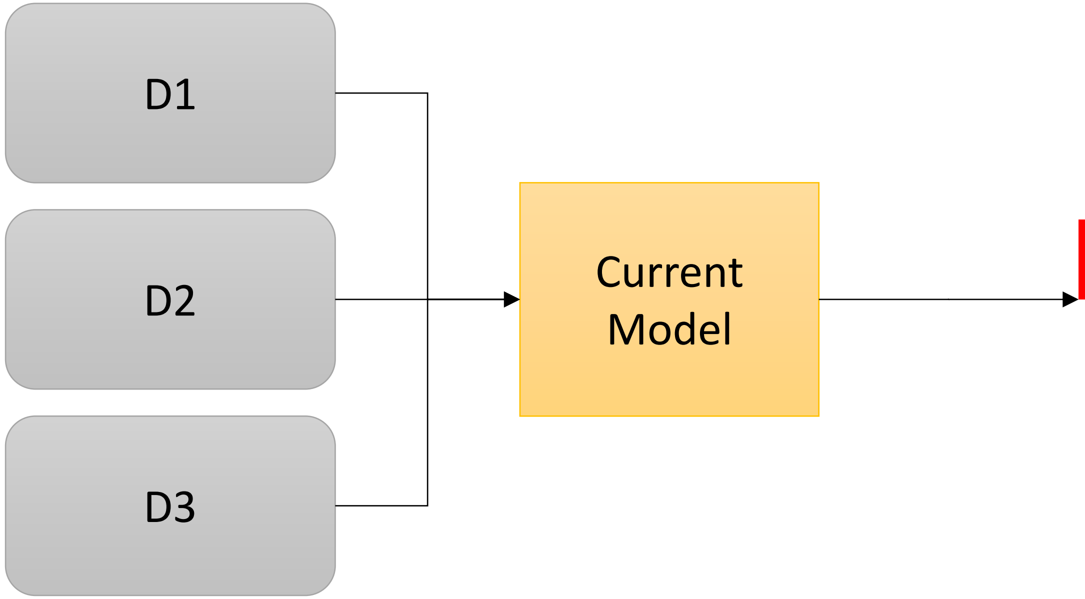
Least Confident

Better



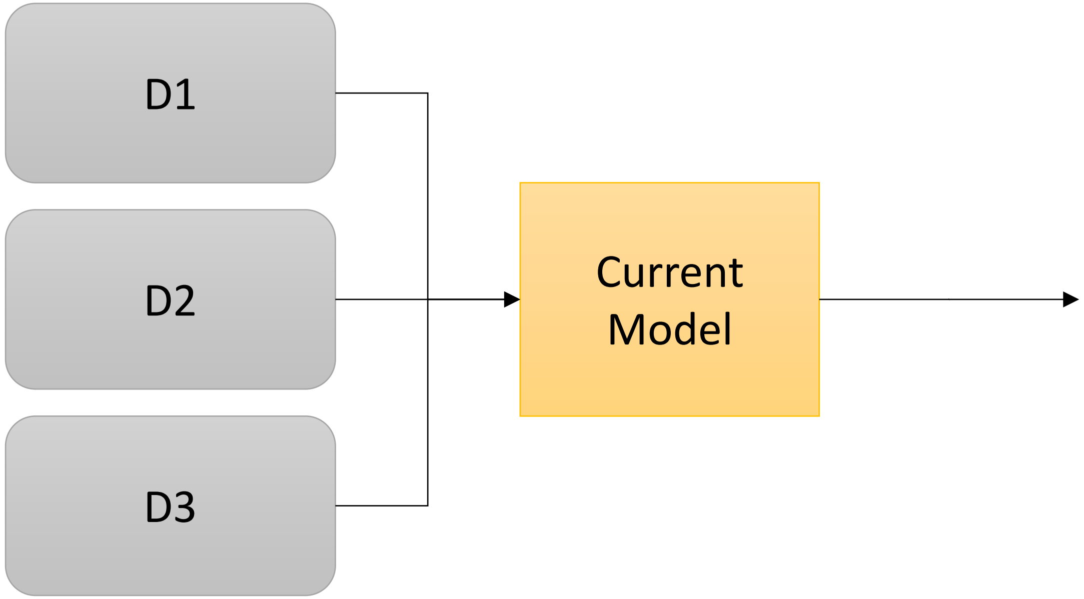
Positive	Neutral	Negative
0.5	**0.45**	0.05
0.4	**0.3**	0.3
0.95	**0.05**	0

Better



Positive	Neutral	Negative	
0.5	**0.45**	0.05	Margin
0.4	**0.3**	0.3	
0.95	**0.05**	0	

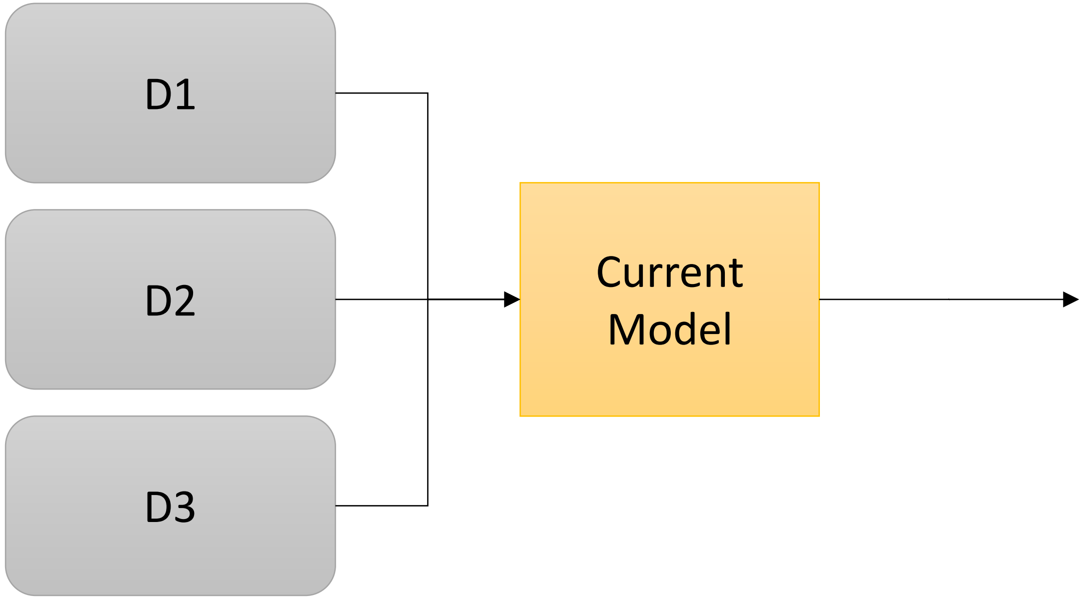
Better



$$Entropy(P) = -\sum_{i} p_i log p_i$$

Positive	Neutral	Negative
0.5	0.45	0.05
0.4	0.3	0.3
0.95	0.05	0

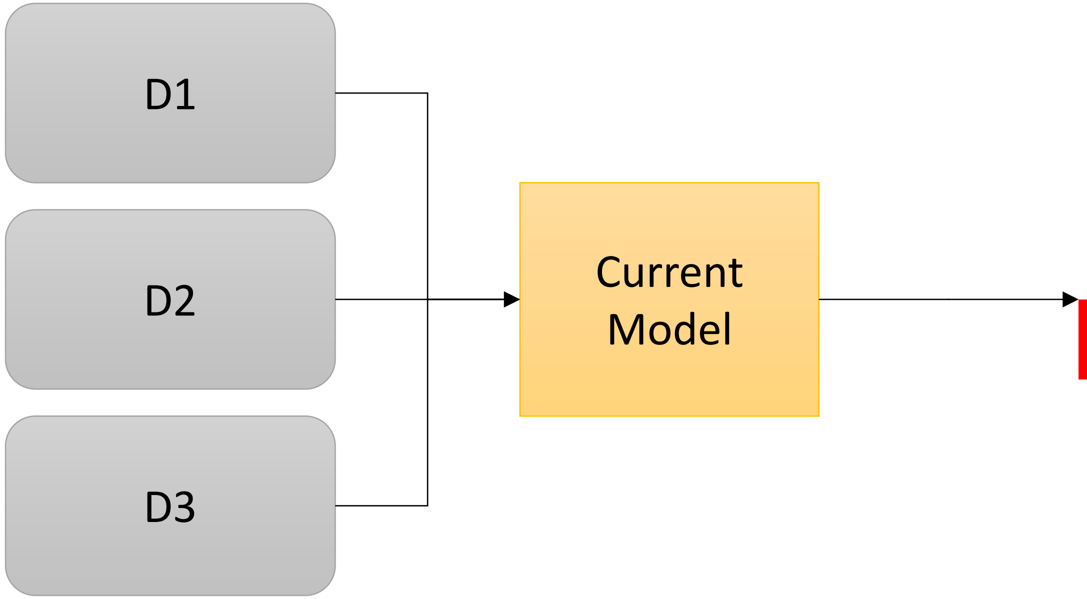
Better



$$Entropy(P) = -\sum_{i} p_i log p_i$$

Positive	Neutral	Negative	
0.5	0.45	0.05	1.23
0.4	0.3	0.3	**1.57**
0.95	0.05	0	0.29

Better



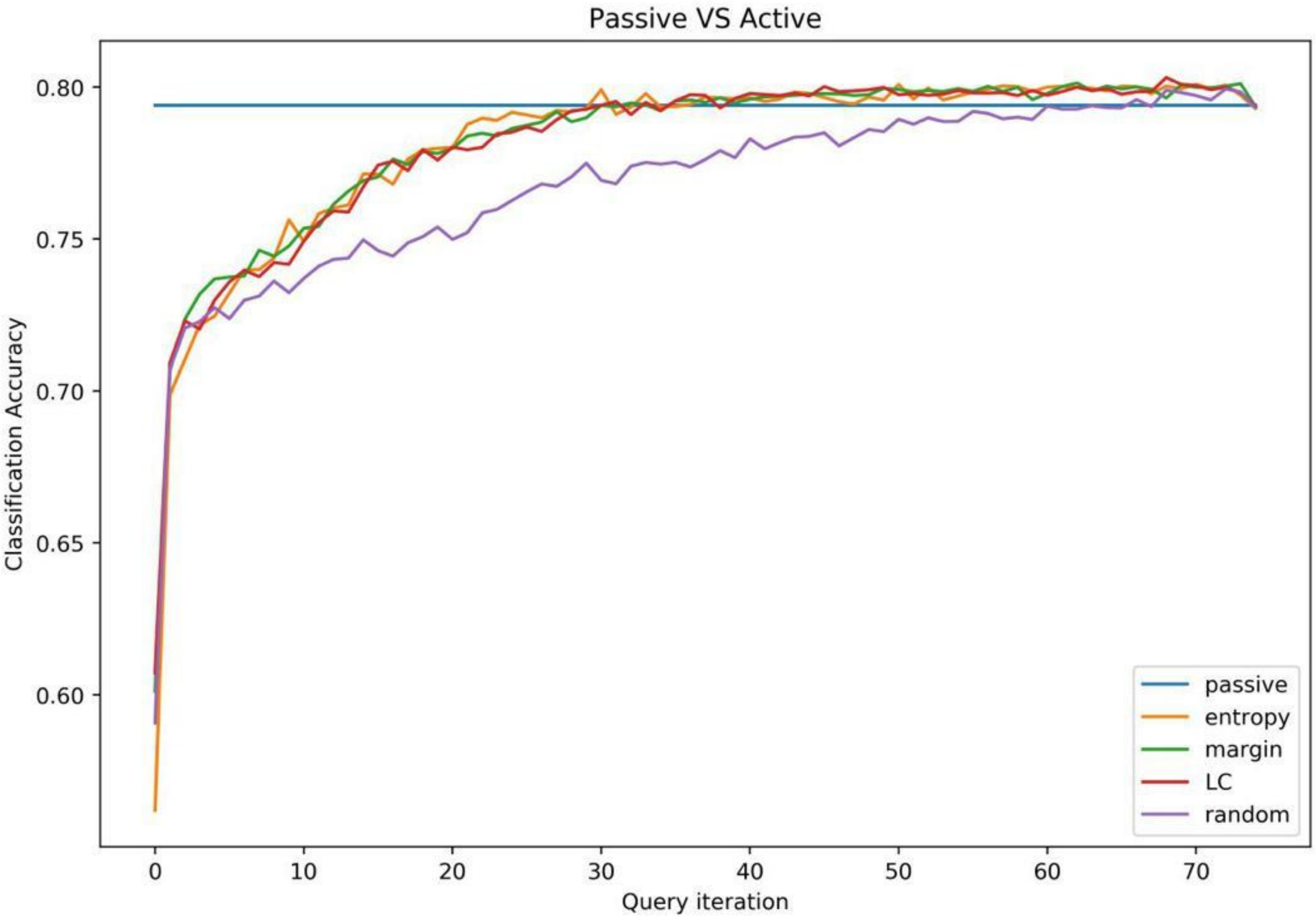
$$Entropy(P) = -\sum_{i} p_i log p_i$$

Positive	Neutral	Negative	
0.5	0.45	0.05	
0.4	0.3	0.3	Entropy
0.95	0.05	0	

Case Study: Sentiment Analysis

- Model: LSTM
- Embedding: embedding layer
- Data: 100K Hand-Labeled Digikala comments (positive, neutral, negative)
- Test Scenarios:
    - Train on all data
    - Active Learning
        - Entropy
        - Margin
        - Least Confident
        - Random

Better


Passive VS Active
Classification Accuracy
0.80
0.75
0.70
0.65
0.60
Query iteration
0
10
20
30
40
50
60
70
passive
entropy
margin
LC
random

Summary

- Problem Modeling
  - Self Supervision
    - Model your problem in a way that labels are easy to get (usually available alongside your data)
  - Weak Supervision
    - Transform data into a new space for less annotation effort
- Use Related Data
  - Transfer Learning
    - Transfer model knowledge between tasks
  - Multi-Task Learning
    - Use related tasks for more data and synergic effects
- Get Better Data
  - Active Learning
    - Smart selection of samples for annotation

ZAAL

Natural Language Processing Services for Persian

www.getzaal.com

Thank You