Deep Text Mining of Instagram Data Without Strong Supervision

WI 2018 Santiago | International Conference on Web intelligence

Kim Hammar, Shatha Jaradat, Nima Dokoohaki, and Mihhail Matskin

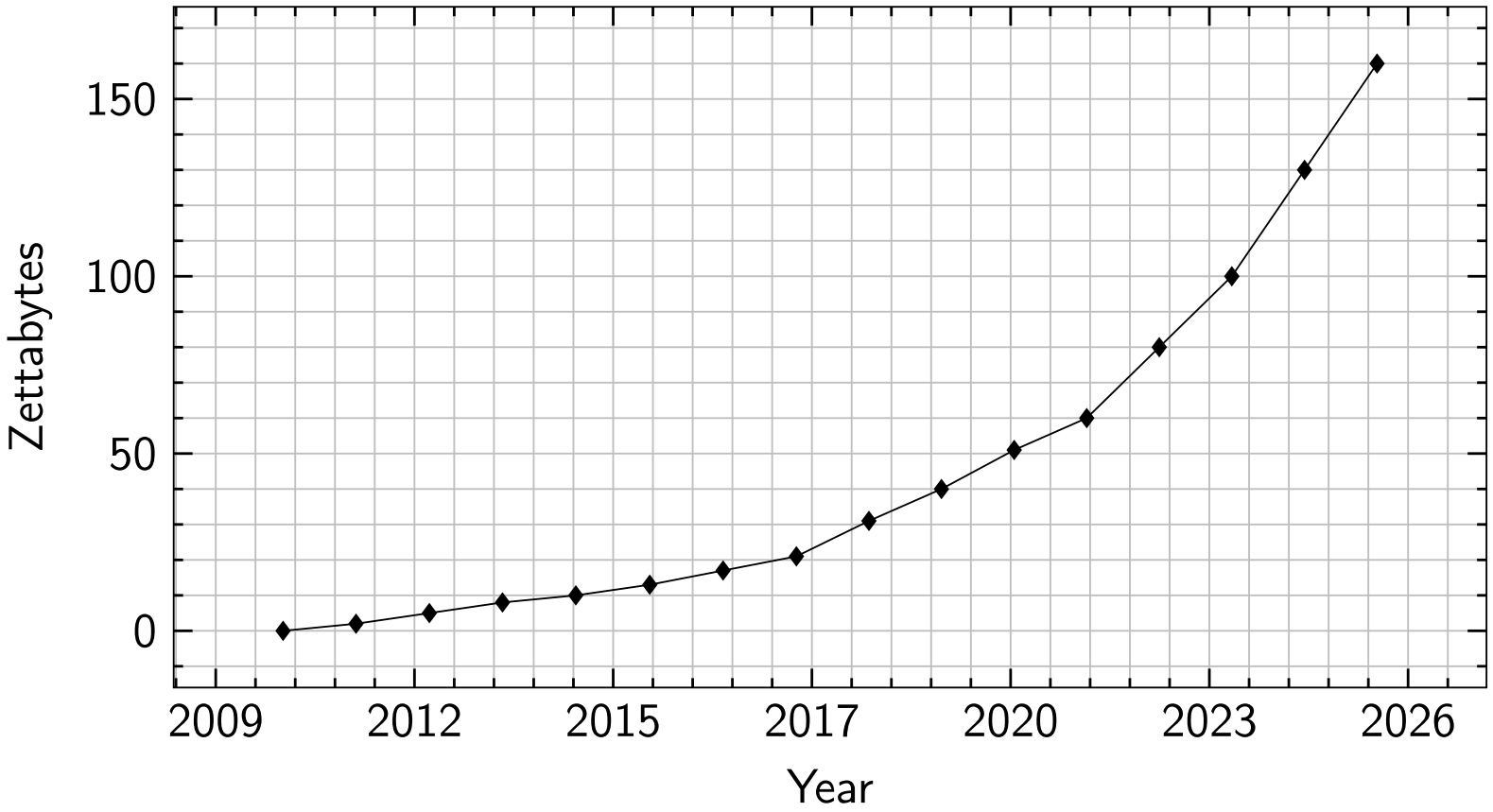
KTH Royal Institute of Technology

kimham@kth.se

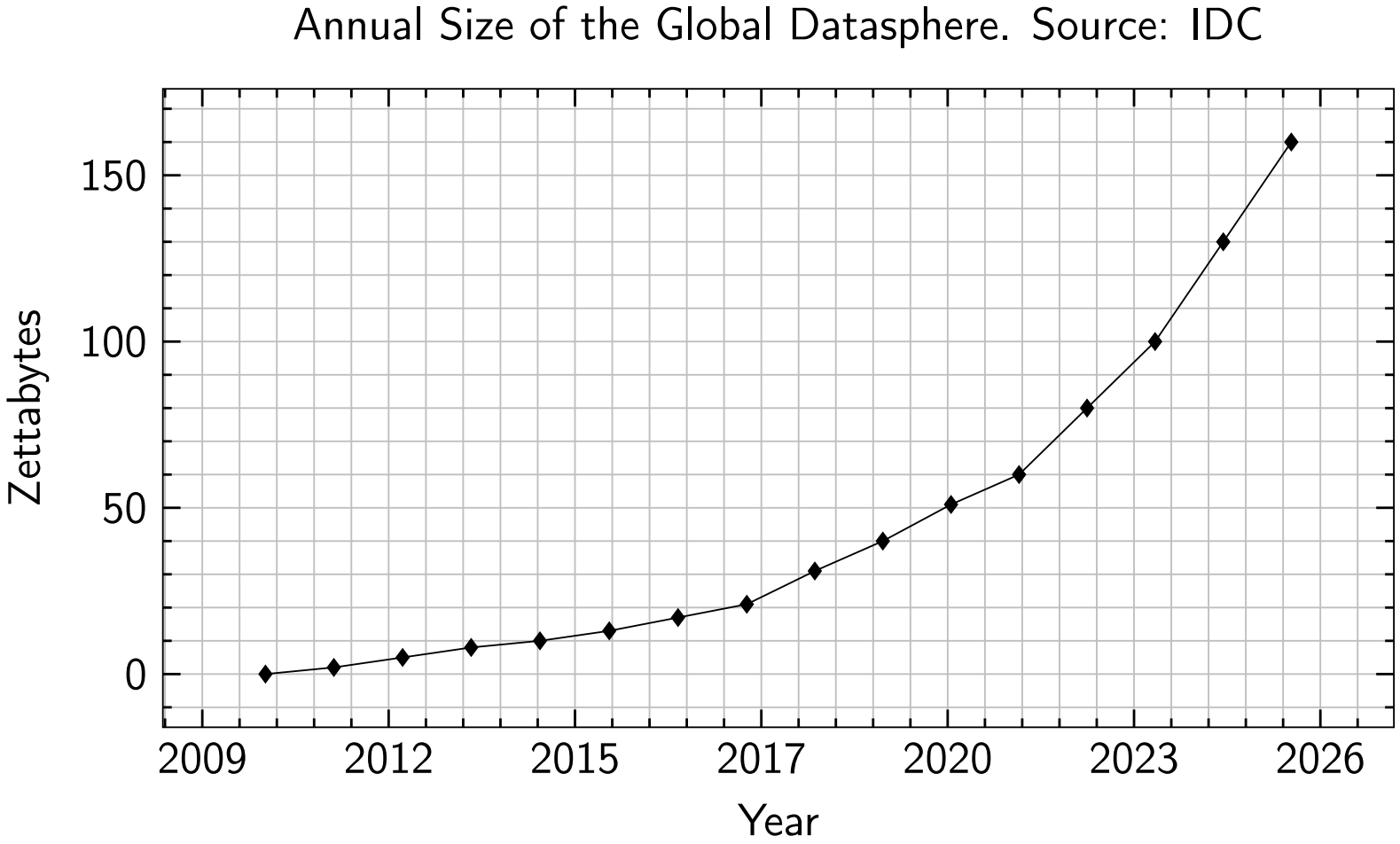
December 4, 2018

Key enabler for Deep Learning: Data growth

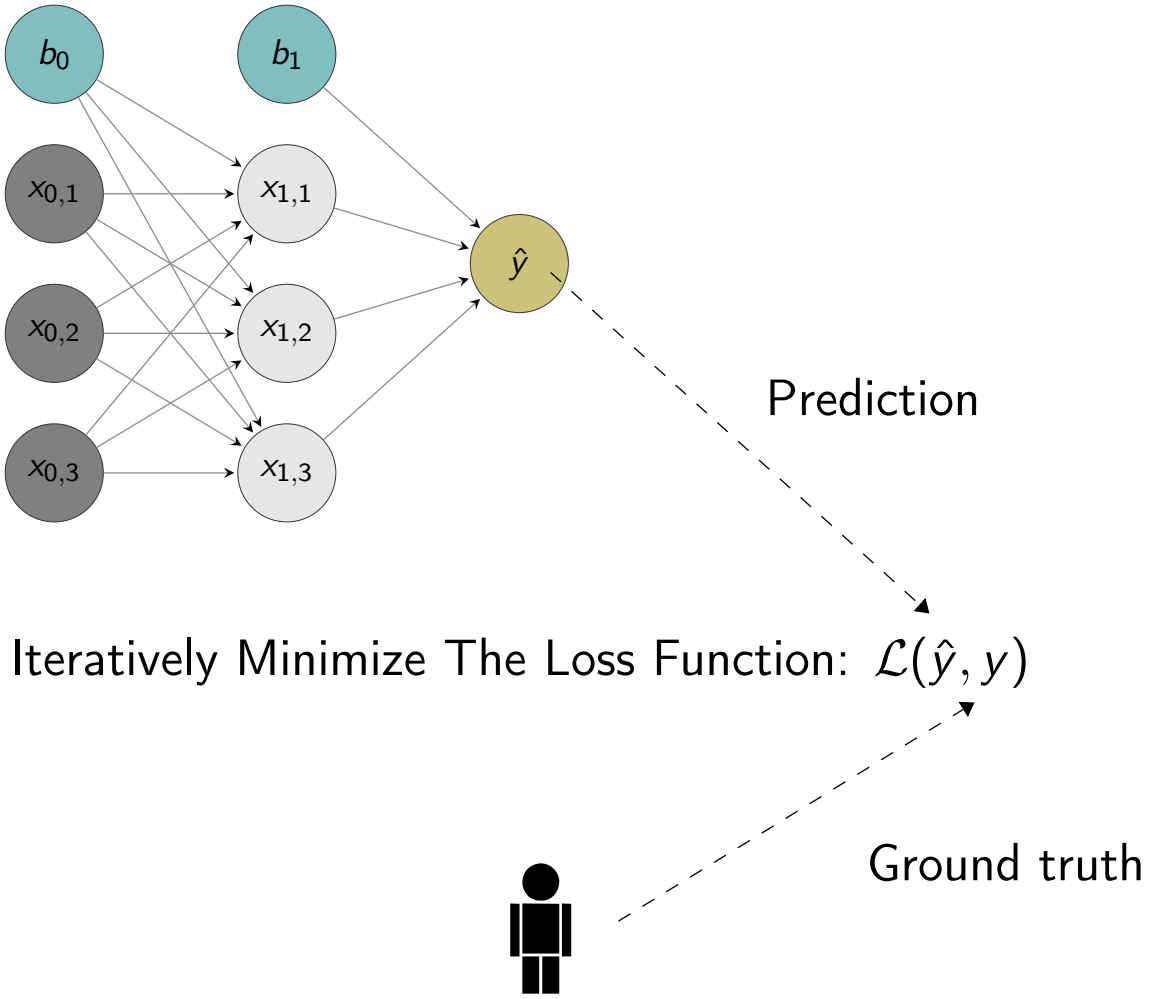
Annual Size of the Global Datasphere. Source: IDC



Key enabler for Deep Learning: Data growth



But what about Labeled Data?

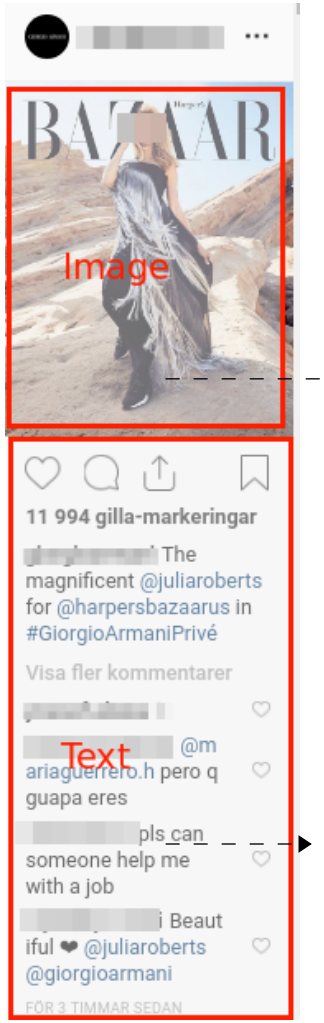


Supervised learning: Iteratively Minimize The Loss Function: $\mathcal{L}(\hat{y}, y)$

Labeled Training Data is Still a Bottleneck

Research Problem: Clothing Prediction on Instagram

Instagram Post



Text Model

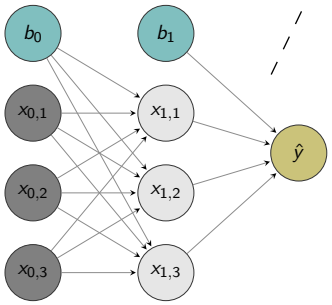
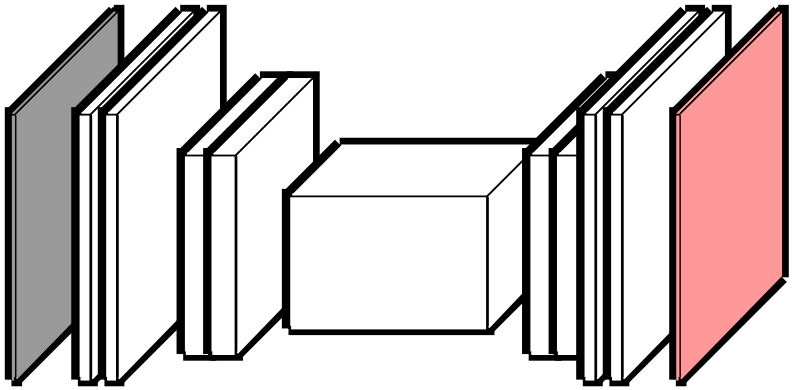


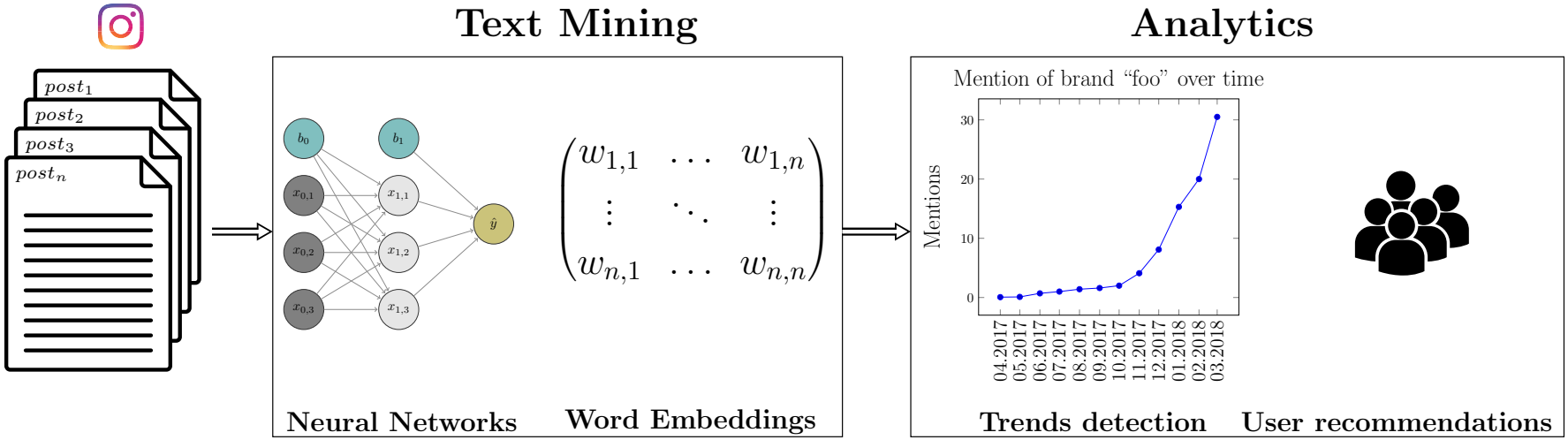
Image Model



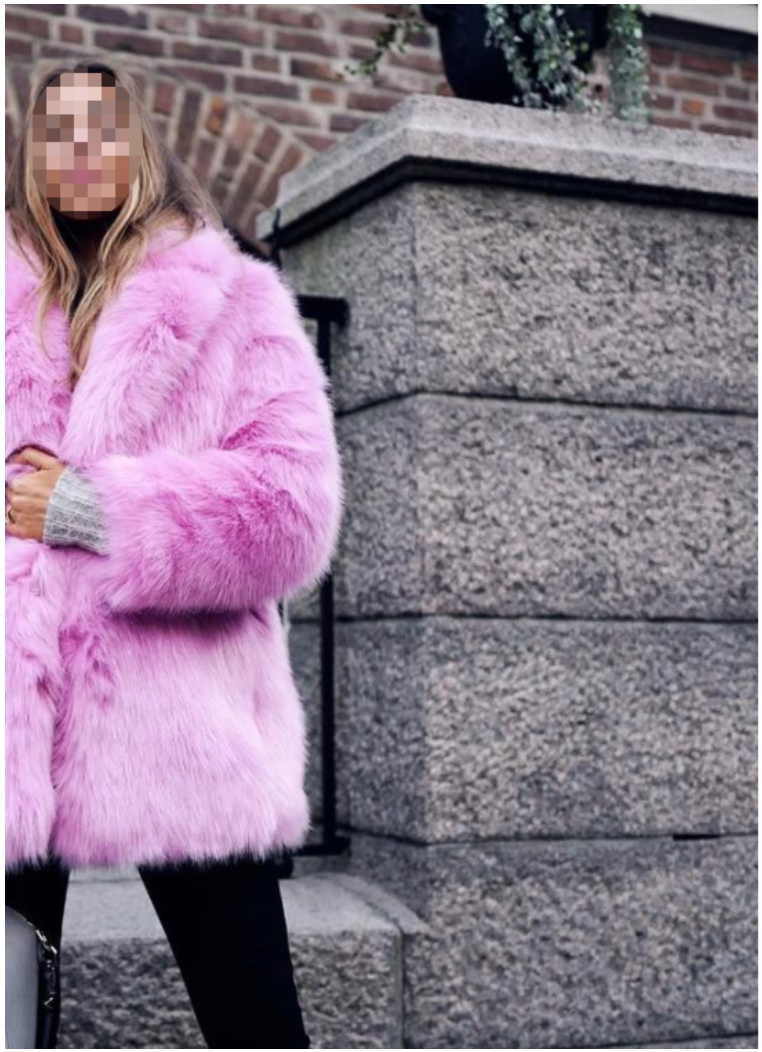
Clothing Prediction

$$\begin{pmatrix} \text{dress} = 0 \\ \text{coat} = 1 \\ \vdots \\ \text{skirt} = 0 \end{pmatrix}$$

This Paper: Text Classification Without Labeled Data



Example Instagram Post



Following

Wow😻

C'est le même que toi 😍

Where is that coat from??! I have to have it!!

It's soooo fluffy I'm gonna diiiiiie 😍💗

So cozy and noticable 👌💞

H&M! Sold out tho 😞

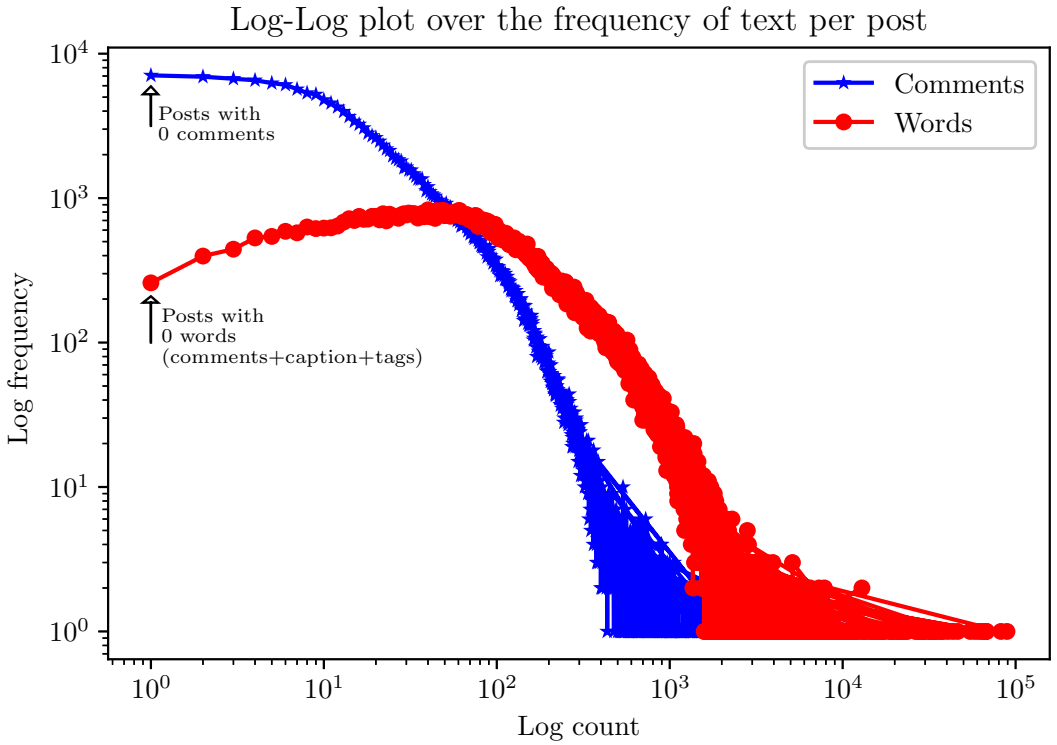
Beauty 💅💅💅💅

Bästa jackan!! 😍 Så glad jag fick tag i en också för ett tag sen 😅 så fruktansvärt varm dock, måste nästan vänta på snö....

Challenge: Noisy Text and No Labels

A case study of a corpora with 143 fashion accounts, 200K posts, 9M comments

Challenge 1: Noisy Text with a Long-Tail Distribution

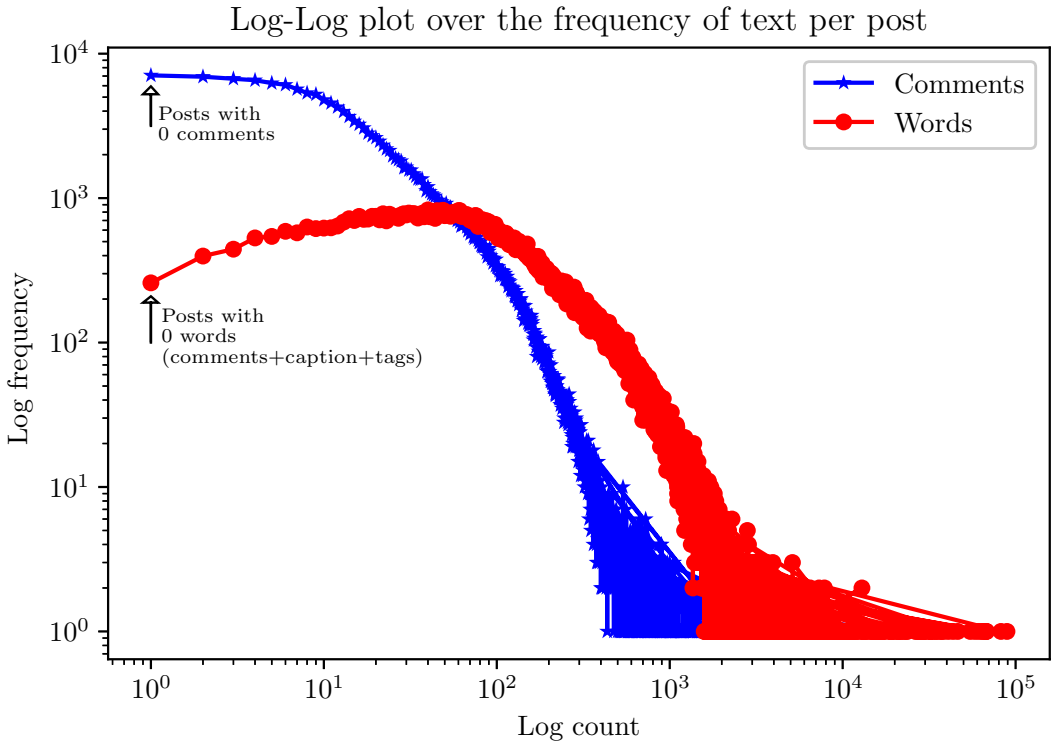


Text Statistic	*Fraction of corpora size*	*Average/post*
Emojis	0.15	48.63
Hashtags	0.03	9.14
User-handles	0.06	18.62
Google-OOV words	0.46	145.02
Aspell-OOV words	0.47	147.61

Challenge: Noisy Text and No Labels

A case study of a corpora with 143 fashion accounts, 200K posts, 9M comments

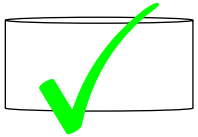
Challenge 1: Noisy Text with a Long-Tail Distribution



Text Statistic	*Fraction of corpora size*	*Average/post*
Emojis	0.15	48.63
Hashtags	0.03	9.14
User-handles	0.06	18.62
Google-OOV words	0.46	145.02
Aspell-OOV words	0.47	147.61

Challenge 2: Lack of Expensive Labeled Training Data

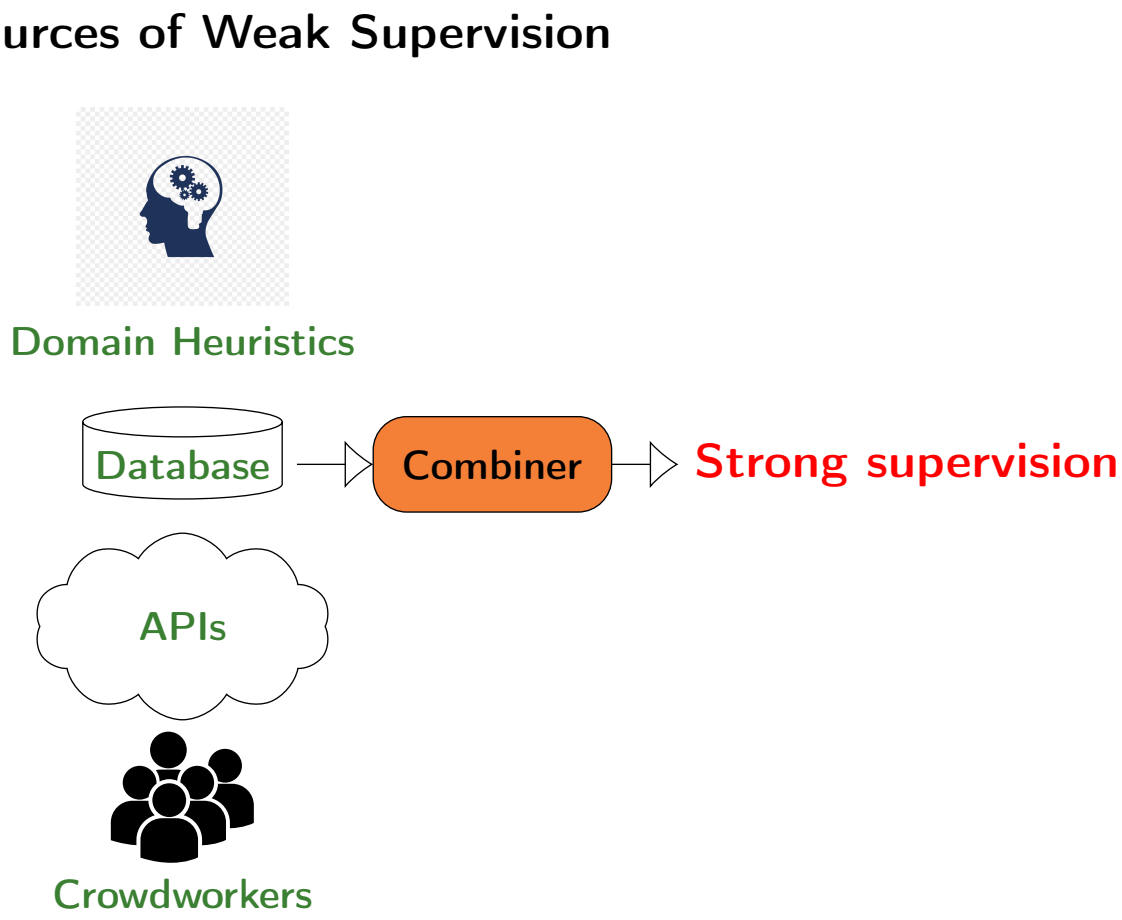
Raw Instagram Text



Human Annotations

Alternative Sources of Supervision That Are Cheap but Weak

- **Strong supervision:** Manual annotation by expert
- **Weak supervision:** A signal that does not have full coverage/perfect accuracy



Weak Supervision in the Fashion Domain

- Open APIs:

clarifai Google Cloud Platform deepomatic

[1] https://github.com/jolibrain/deepdetect

Weak Supervision in the Fashion Domain

- Open APIs:

clarifai Google Cloud Platform deepomatic

- Pre-trained Clothing Classificiation Models:

DeepDetect[1]

[1] https://github.com/jolibrain/deepdetect

Weak Supervision in the Fashion Domain

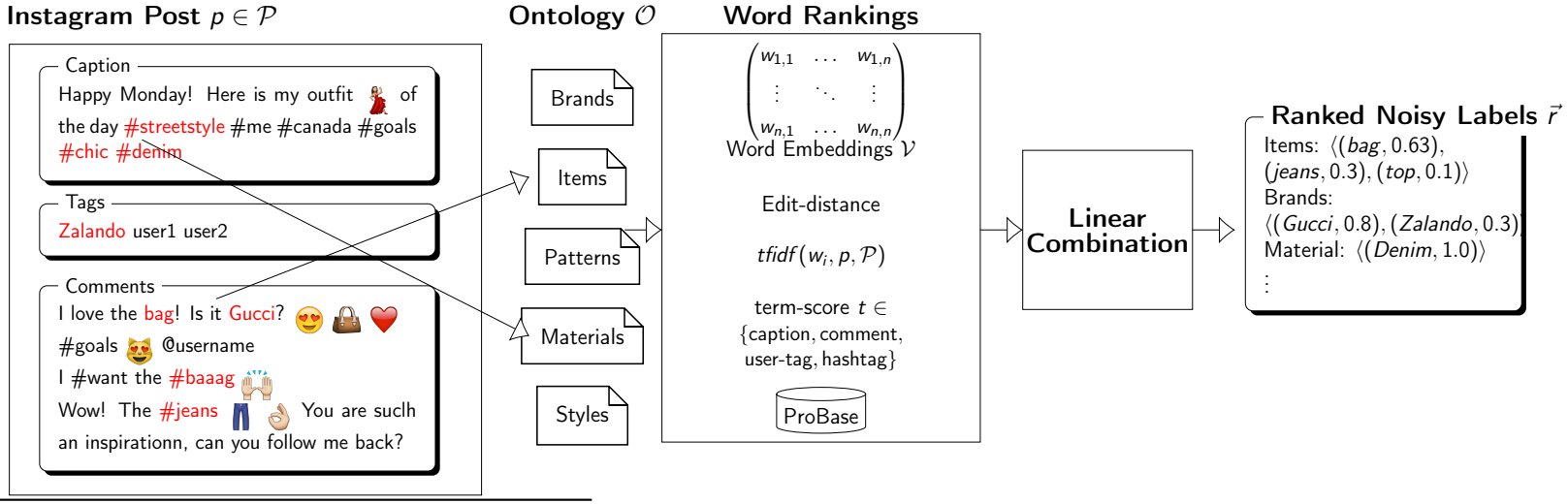
- Open APIs:

clarifai Google Cloud Platform deepomatic

- Pre-trained Clothing Classificiation Models:

DeepDetect[1]

Text mining system based on a fashion ontology and word embeddings:



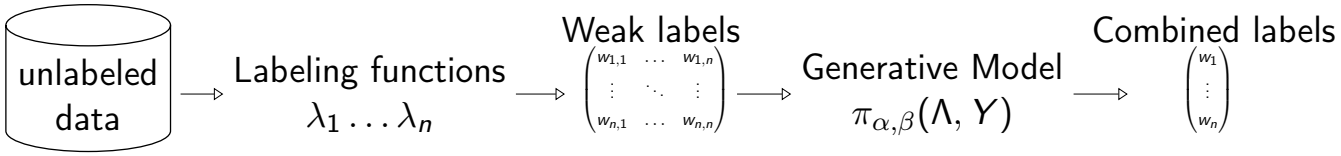
[1] https://github.com/jolibrain/deepdetect

How To Combine Several Sources Of Weak Supervision?

- Simplest way to combine many weak signals: Majority Vote
- Recent research on combination of weak signals: Data Programming[2]

[2] Alexander J Ratner et al. "Data Programming: Creating Large Training Sets, Quickly". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 3567–3575. URL: http://papers.nips.cc/paper/6523-data-programming-creating-large-training-sets-quickly.pdf.

Model Weak Supervision With Generative Model



- **Model weak supervision as labeling functions λ_i**
  - λ_i(unlabeled data) $\rightarrow$ label
- **Learn Generative Model $\pi_{\alpha,\beta}(\Lambda, Y)$ over the labeling process.**
  - Based on conflicts between labeling functions assign the functions an estimated accuracy α_i.
  - Based on empirical coverage of labeling functions assign the functions a coverage β_i.
- **Given α and β for each labeling function, it can be used to combine labels into a single probabilistic label**
  - Give more weight to high-accuracy functions
  - If there is a lot of disagreement$\rightarrow$ low probability label
  - If all labeling functions agree $\rightarrow$ high probability label

Data Programming Intuition

Low accuracy labeling functions

"it is not a coat"

High accuracy labeling functions

"it is a coat"

Probabilistic Label: 0.6 probability that it is a coat

Majority Vote: 1.0 probability that it is not a coat

Extension of Data Programming to Multi-Label Classification

- Problem: **Data programming only defined for binary classification in original paper**
- **To make it work for multi-class setting:** model labeling function as $\lambda_i \rightarrow k_i \in \{0, \ldots, N\}$ instead of $\lambda_i \rightarrow k_i \in \{-1, 0, 1\}$.
- **Idea 1 for multi-label:** model labeling function as $\lambda_i \rightarrow \vec{k_i} = \{v_0, \ldots, v_n\} \wedge v_j \in \{-1, 0, 1\}$
- **Idea 2 for multi-label:** learn a separate generative model for each class, and let each labeling function give binary output for each class $\lambda_{i,j} \rightarrow k_{i,j} \in \{-1, 0, 1\}$.

Trained Generative Models: Labeling Functions' Accuracy Differ Between Classes

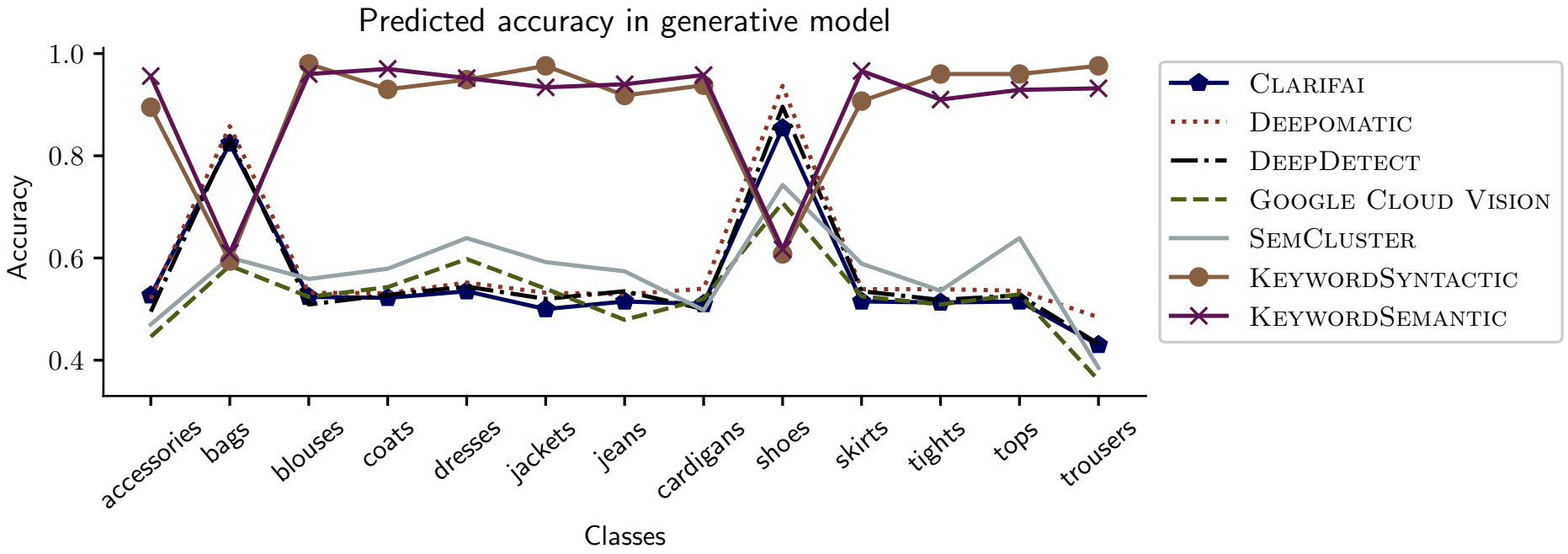


Figure: Multiple generative models can capture a different accuracy for labeling functions for different classes.

Putting Everything Together

1. **Apply weak supervision** to unlabeled data (open APIs, pre-trained models, domain heuristics etc.)

Putting Everything Together

1. **Apply weak supervision** to unlabeled data (open APIs, pre-trained models, domain heuristics etc.)
2. **Combine labels** using majority voting or generative modelling (data programming)

Putting Everything Together

1. **Apply weak supervision** to unlabeled data (open APIs, pre-trained models, domain heuristics etc.)
2. **Combine labels** using majority voting or generative modelling (data programming)
3. **Use the combined labels for training** a discriminative model using supevised machine learning.

Pipeline for Weakly Supervised Classification in Instagram

Problem: A Multi-class Multi-label classification problem with 13 output classes (dresses, coats, blouses, jeans, ...)

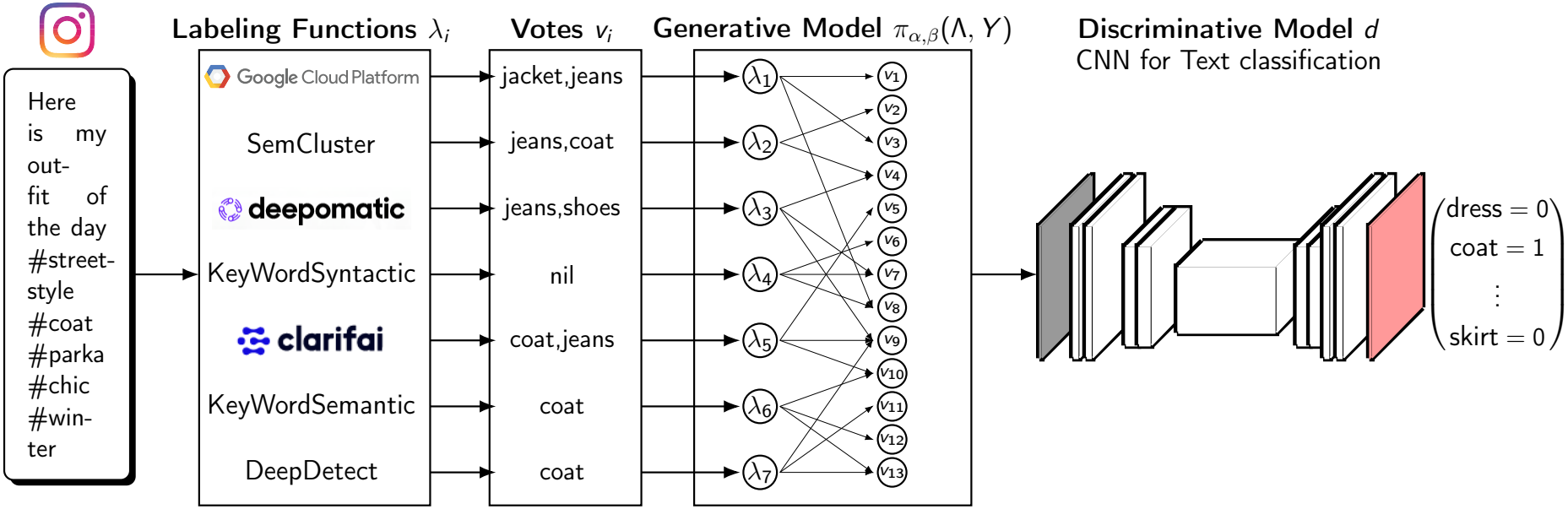


Figure: A pipeline for weakly supervised text classification of Instagram posts.

Data Programming Beats Majority Voting

Results

- **Data programming gives 6 F_1 points improvement over majority vote[3], achieving an F_1 score of 0.61 (On level with human performance)**

Model	*Accuracy*	*Precision*	*Recall*	*Micro-F_1*	*Macro-F_1*	*Hamming Loss*
CNN-DataProgramming	**0.797** $\pm$ 0.01	**0.566** $\pm$ 0.05	0.678 $\pm$ 0.04	**0.616** $\pm$ 0.02	**0.535** $\pm$ 0.01	**0.195** $\pm$ 0.02
CNN-MajorityVote	0.739 $\pm$ 0.02	0.470 $\pm$ 0.06	**0.686** $\pm$ 0.05	0.555 $\pm$ 0.03	0.465 $\pm$ 0.05	0.261 $\pm$ 0.03
DomainExpert	**0.807**	**0.704**	0.529	0.604	0.534	**0.184**

- **Main cause of error: data sparsity (can not extract clothing items from the text if it is never mentioned in the text)**

[3] A smaller, hand-labeled dataset by experts was used for evaluation

Conclusion

- Instagram text is jus as noisy as Twitter, has a long-tail distribution, and is multi-lingual
- In shifting data domains where accurate labeled data is a rarity, like social media, weak supervision is a viable alternative.
- Combining weak labels with generative modeling beats majority voting.
- To extend Data programming to the multi-label scenario, a collection of generative models can be used to incorporate per-class accuracy.

Thank you

- All code and most of the data is open source: `https://github.com/shatha2014/FashionRec`
- Questions?