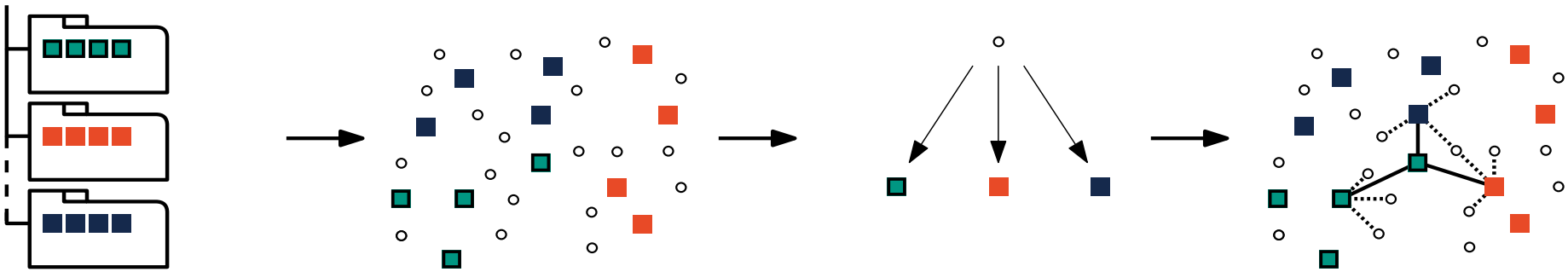
Mining Text Outliers in Document Directories

ICDM 2020: 20th IEEE International Conference on Data Mining

Edouard Fouché*, Y. Meng**, F. Guo**, H. Zhuang**, K. Böhm* & J. Han** | October 19, 2020

* KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT), ** UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



This talk is about...

- Classifying documents into directories
  - For example: emails, news articles, research papers ...

Email 1: Re: Your application...
Email 2: Interview scheduling...
Email 3: Preparing your internship...
Email 4: Self-supervised methods...

Email 5: Fw: Grant Proposal...
Email 6: Support Vector Machines...

Email 7: Re: The best cat memes...
Email 8: ICDM Registration...
Email 9: Get-together, RSVP...

- Documents may be classified wrongly:
  - Type M: Misclassification (wrong folder)
  - Type O: Out-of-distribution (no adequate folder)
- We see those mistakes as semantic "outliers"
  - We present an approach to mine both outliers types simultaneously

This talk is about...

- Classifying documents into directories
  - For example: emails, news articles, research papers ...

Email 1: Re: Your application...
Email 2: Interview scheduling...
Email 3: Preparing your internship...
Email 4: Self-supervised methods...

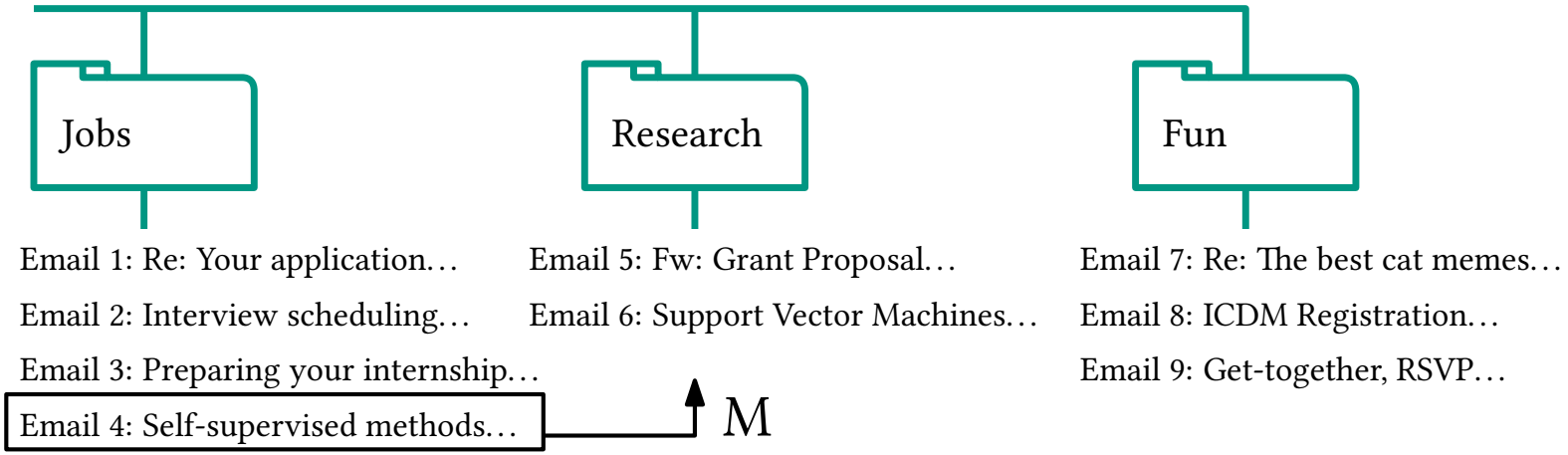
Email 5: Fw: Grant Proposal...
Email 6: Support Vector Machines...

Email 7: Re: The best cat memes...
Email 8: ICDM Registration...
Email 9: Get-together, RSVP...

- Documents may be classified wrongly:
  - Type M: Misclassification (wrong folder)
  - Type O: Out-of-distribution (no adequate folder)
- We see those mistakes as semantic "outliers"
  - We present an approach to mine both outliers types simultaneously

This talk is about...

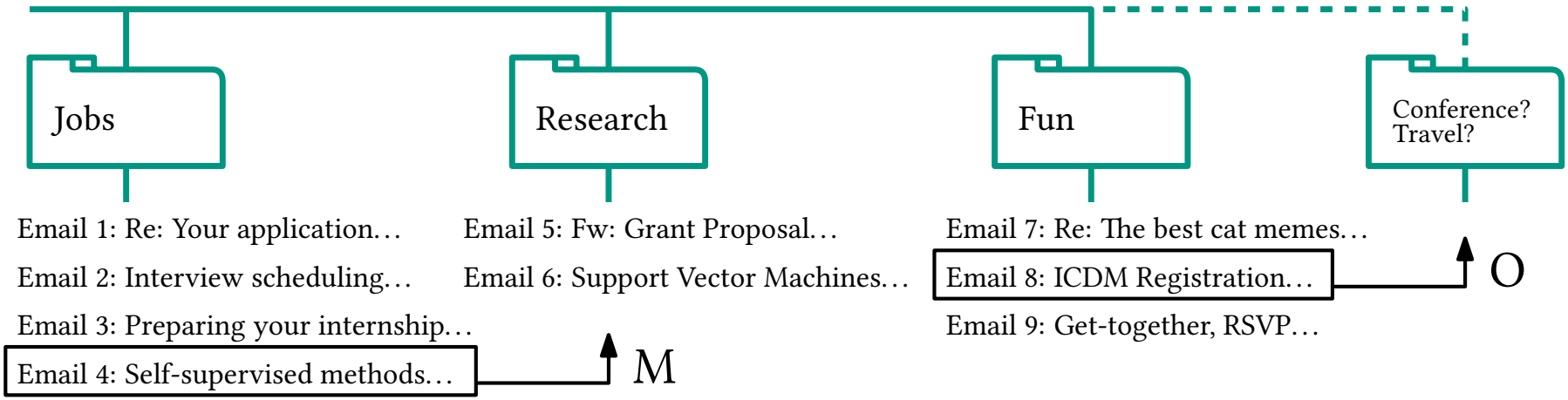
- Classifying documents into directories
  - For example: emails, news articles, research papers ...



- Documents may be classified wrongly:
  - Type M: Misclassification (wrong folder)
  - Type O: Out-of-distribution (no adequate folder)
- We see those mistakes as semantic "outliers"
  - We present an approach to mine both outliers types simultaneously

This talk is about...

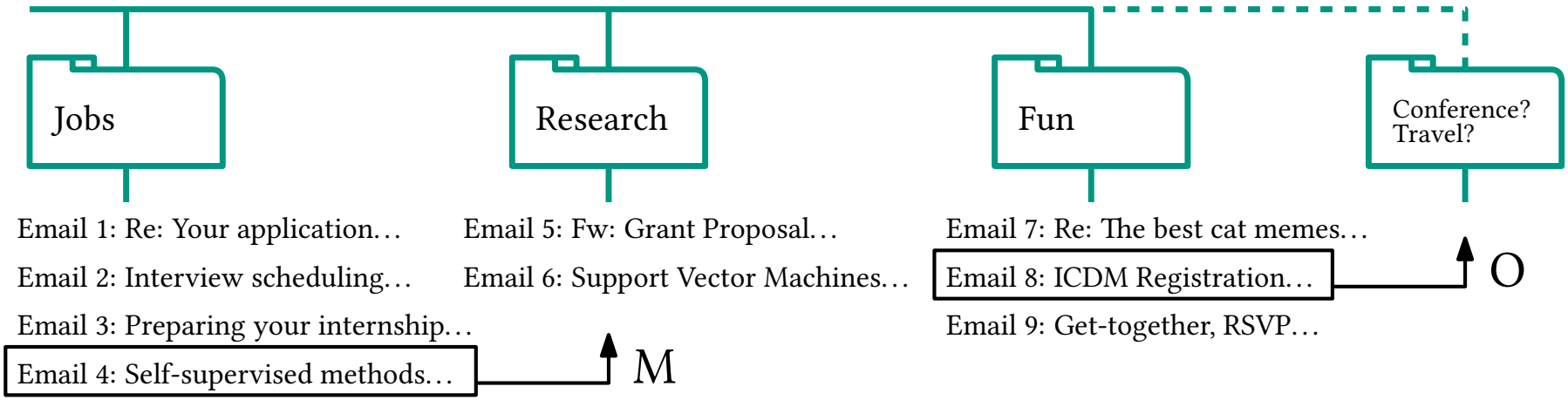
- Classifying documents into directories
  - For example: emails, news articles, research papers ...



- Documents may be classified wrongly:
  - Type M: Misclassification (wrong folder)
  - Type O: Out-of-distribution (no adequate folder)
- We see those mistakes as semantic "outliers"
  - We present an approach to mine both outliers types simultaneously

This talk is about...

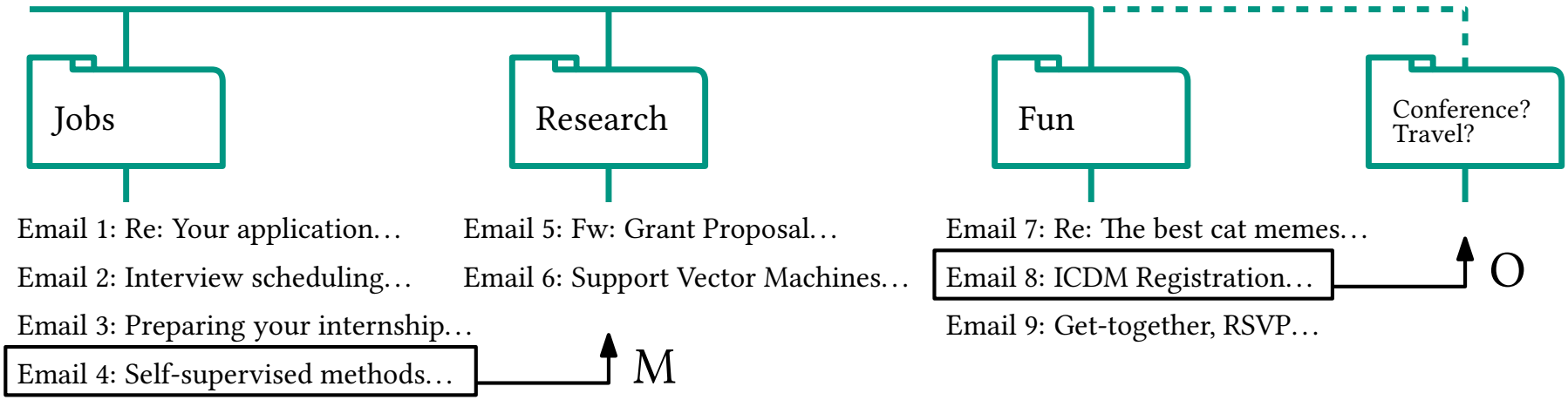
- Classifying documents into directories
  - For example: emails, news articles, research papers ...



- Documents may be classified wrongly:
  - Type M: Misclassification (wrong folder)
  - Type O: Out-of-distribution (no adequate folder)
- We see those mistakes as semantic "outliers"
  - We present an approach to mine both outliers types simultaneously

This talk is about...

- Classifying documents into directories
  - For example: emails, news articles, research papers ...

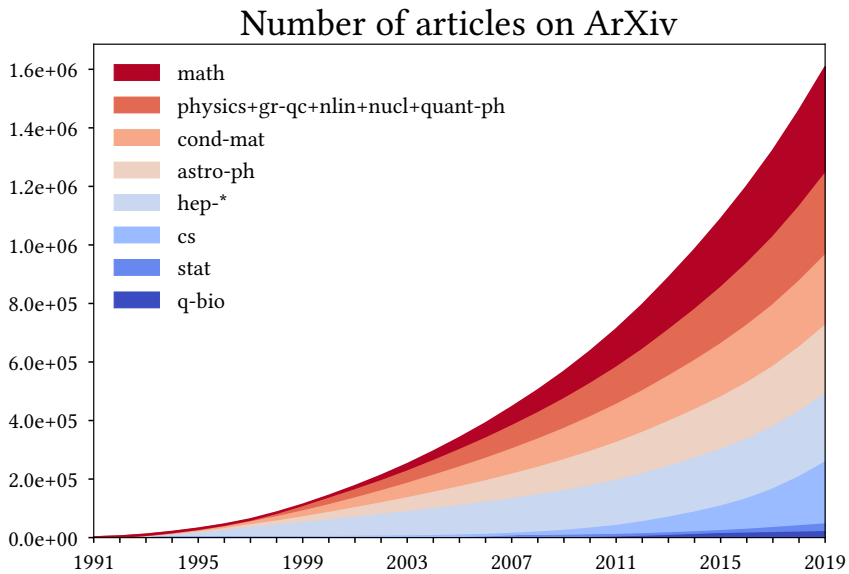


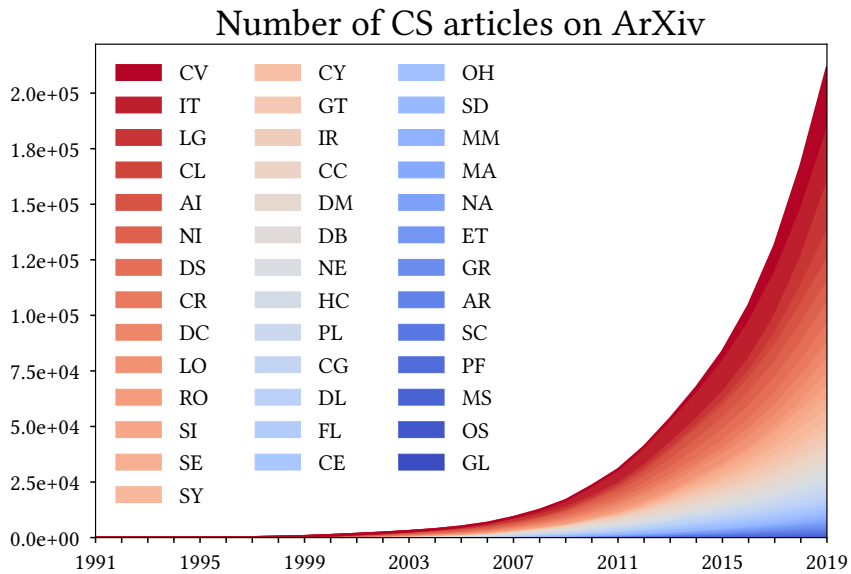
- Documents may be classified wrongly:
  - Type M: Misclassification (wrong folder)
  - Type O: Out-of-distribution (no adequate folder)
- We see those mistakes as semantic "outliers"
  - We present an approach to mine both outliers types simultaneously

Why are there such outliers? (1/2)

We drawning in data

- Document repositories have grown very large
- They tend to be highly multi-modal: many classes, folders





[1]

[1] Data obtained from `https://www.kaggle.com/Cornell-University/arxiv`

Why are there such outliers? (2/2)

Maintenance is difficult

- Human classification: sloppy, unreliable
- Handled by many/different users
- Different user $\rightarrow$ Different classification

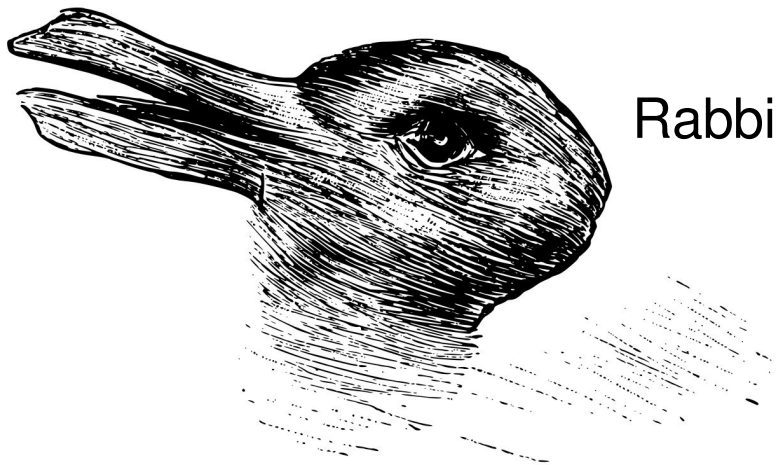
2

[2] The Daily Struggle Meme – Jake Clark – Modified (ML/AI)

Why are they difficult to find? (1/2)

Ambiguity

- Documents may have complex semantic
- Sometimes, the correct class is unknown yet, e.g., an emerging field
- Folder structures are domain/user-specific $\rightarrow$ unsupervised

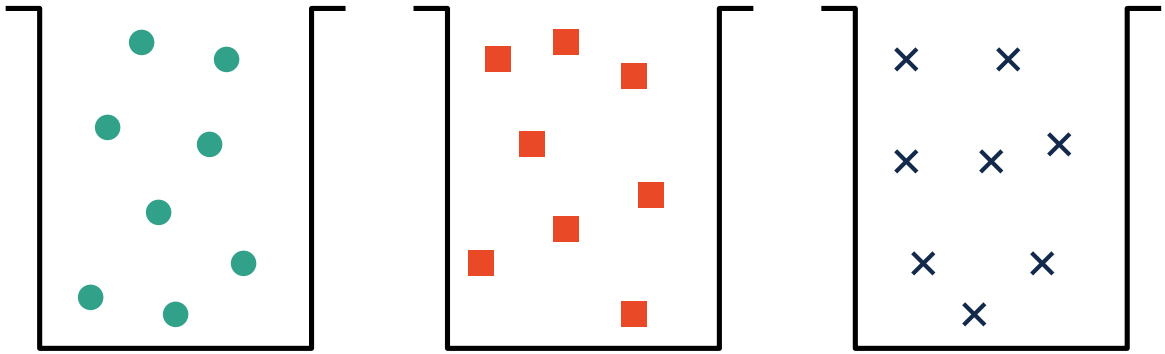


Rabbit, duck, both?[3]

[3] Rabbit and Duck – Fliegende Blätter, 23 October 1892 – Public Domain

Why are they difficult to find? (2/2)

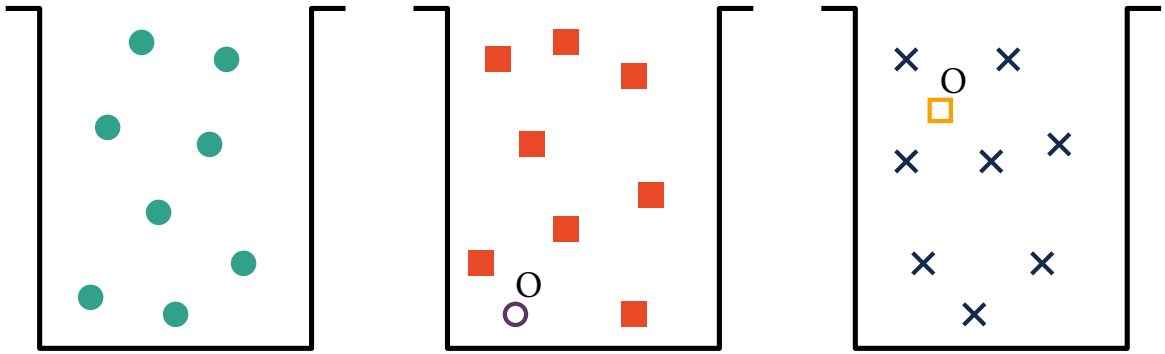
Type O/M must be detected simultaneously



- Type O outliers may be detected as Type M by mistake
- The noise from Type M outliers hinders the detection of Type O

Why are they difficult to find? (2/2)

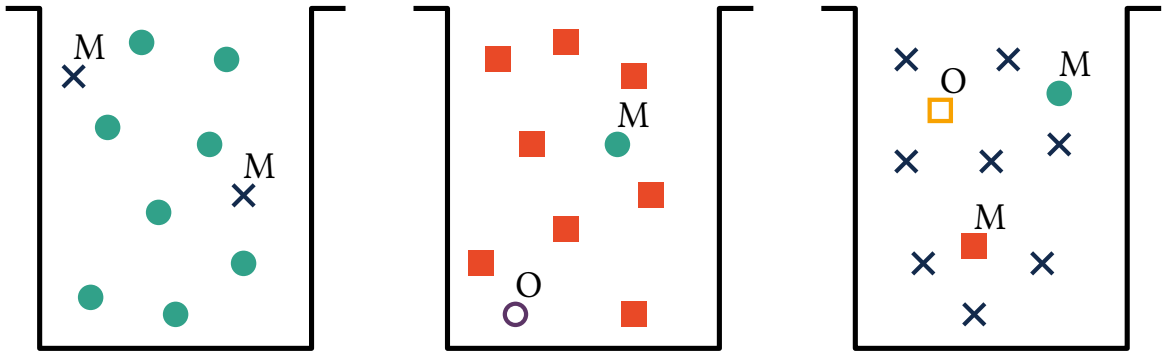
Type O/M must be detected simultaneously



- Type O outliers may be detected as Type M by mistake
- The noise from Type M outliers hinders the detection of Type O

Why are they difficult to find? (2/2)

Type O/M must be detected simultaneously



- Type O outliers may be detected as Type M by mistake
- The noise from Type M outliers hinders the detection of Type O

Our Contributions

We explore the problem of mining text outliers in document directories

- We are first to distinguish between Type O/M outliers

We propose a new approach to detect text outliers

- kj-Nearest Neighbours (kj-NN)
- Exploit similarities between documents/phrases
- Extract semantic labels and similar documents $\rightarrow$ interpretability

We provide an extensive evaluation

- Improve the current state of the art (real-world/synthetic data)
- Interpretable results

Code & data: `https://github.com/edouardfouche/MiningTextOutliers`

Our Contributions

We explore the problem of mining text outliers in document directories

- We are first to distinguish between Type O/M outliers

We propose a new approach to detect text outliers

- kj-Nearest Neighbours (kj-NN)
- Exploit similarities between documents/phrases
- Extract semantic labels and similar documents $\rightarrow$ interpretability

We provide an extensive evaluation

- Improve the current state of the art (real-world/synthetic data)
- Interpretable results

Code & data: `https://github.com/edouardfouche/MiningTextOutliers`

Our Contributions

We explore the problem of mining text outliers in document directories

- We are first to distinguish between Type O/M outliers

We propose a new approach to detect text outliers

- kj-Nearest Neighbours (kj-NN)
- Exploit similarities between documents/phrases
- Extract semantic labels and similar documents $\rightarrow$ interpretability

We provide an extensive evaluation

- Improve the current state of the art (real-world/synthetic data)
- Interpretable results

Code & data: https://github.com/edouardfouche/MiningTextOutliers

Our Contributions

We explore the problem of mining text outliers in document directories

- We are first to distinguish between Type O/M outliers

We propose a new approach to detect text outliers

- kj-Nearest Neighbours (kj-NN)
- Exploit similarities between documents/phrases
- Extract semantic labels and similar documents $\rightarrow$ interpretability

We provide an extensive evaluation

- Improve the current state of the art (real-world/synthetic data)
- Interpretable results

Code & data: `https://github.com/edouardfouche/MiningTextOutliers`

Related Work

Type O (Out-of-distribution)

- "Standard" outlier detectors
  - Distance- [KN98, RRS00], Neighbour- [BKNS00, KSZ08], Probabilistic- [KS12, TB99], Subspace-based [SA18, KMB12]
  - LOF [BKNS00], RS-Hash [SA18]
- Text outlier detectors
  - Von Mises-Fisher mixtures: VMF-Q [ZWT+17]
  - Non-negative Matrix Factorization: TONMF [KWAP17]
  - Context Vector Data Description: CVDD [RZV+19]

Type M (Misclassification)

- Received little attention, while ubiquitous!
- Can be extended from supervised text classification methods
- W-CNN [Kim14], VD-CNN [CSBL17], AT-RNN [ZST+16], RCNN [LXLZ15]

Related Work

Type O (Out-of-distribution)

- "Standard" outlier detectors
  - Distance- [KN98, RRS00], Neighbour- [BKNS00, KSZ08], Probabilistic- [KS12, TB99], Subspace-based [SA18, KMB12]
  - LOF [BKNS00], RS-Hash [SA18]
- Text outlier detectors
  - Von Mises-Fisher mixtures: VMF-Q [ZWT+17]
  - Non-negative Matrix Factorization: TONMF [KWAP17]
  - Context Vector Data Description: CVDD [RZV+19]

Type M (Misclassification)

- Received little attention, while ubiquitous!
- Can be extended from supervised text classification methods
- W-CNN [Kim14], VD-CNN [CSBL17], AT-RNN [ZST+16], RCNN [LXLZ15]

Related Work

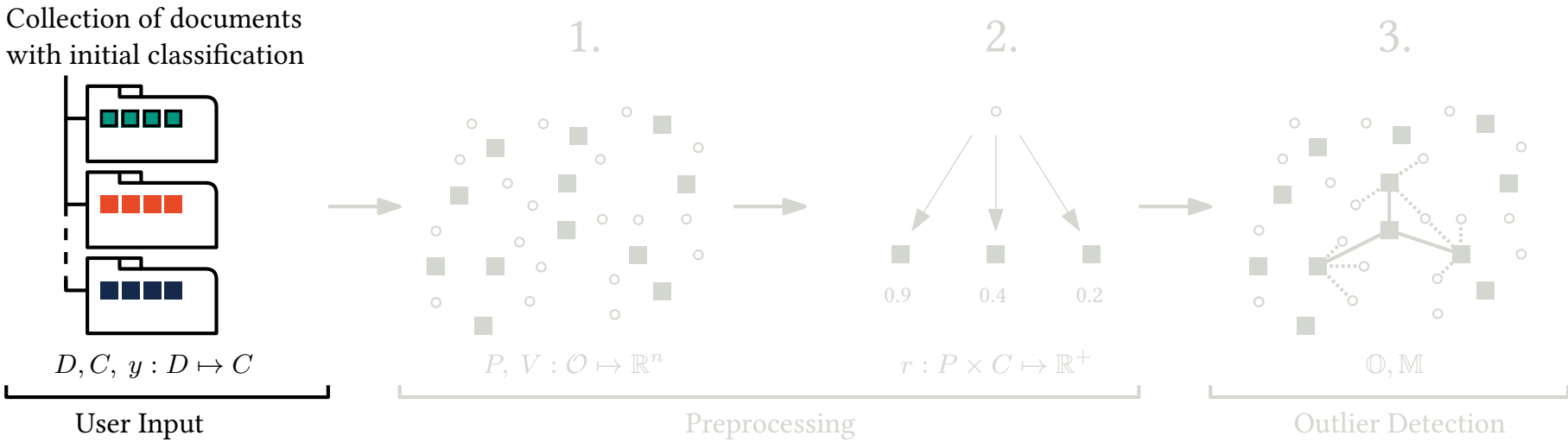
Type O (Out-of-distribution)

- "Standard" outlier detectors
  - Distance- [KN98, RRS00], Neighbour- [BKNS00, KSZ08], Probabilistic- [KS12, TB99], Subspace-based [SA18, KMB12]
  - LOF [BKNS00], RS-Hash [SA18]
- Text outlier detectors
  - Von Mises-Fisher mixtures: VMF-Q [ZWT+17]
  - Non-negative Matrix Factorization: TONMF [KWAP17]
  - Context Vector Data Description: CVDD [RZV+19]

Type M (Misclassification)

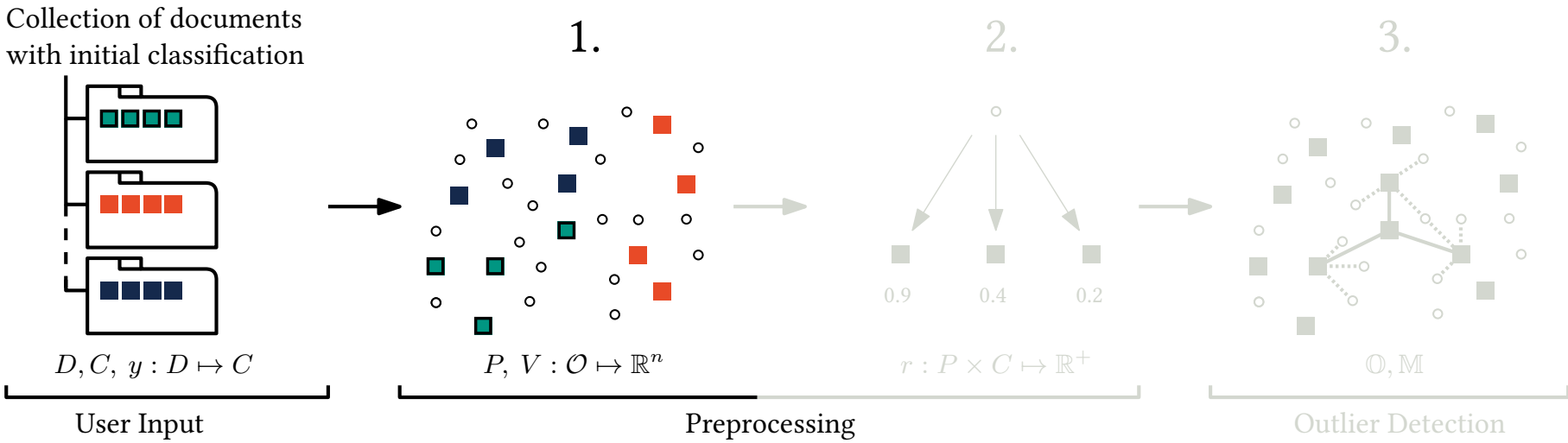
- Received little attention, while ubiquitous!
- Can be extended from supervised text classification methods
- W-CNN [Kim14], VD-CNN [CSBL17], AT-RNN [ZST+16], RCNN [LXLZ15]

Our Framework



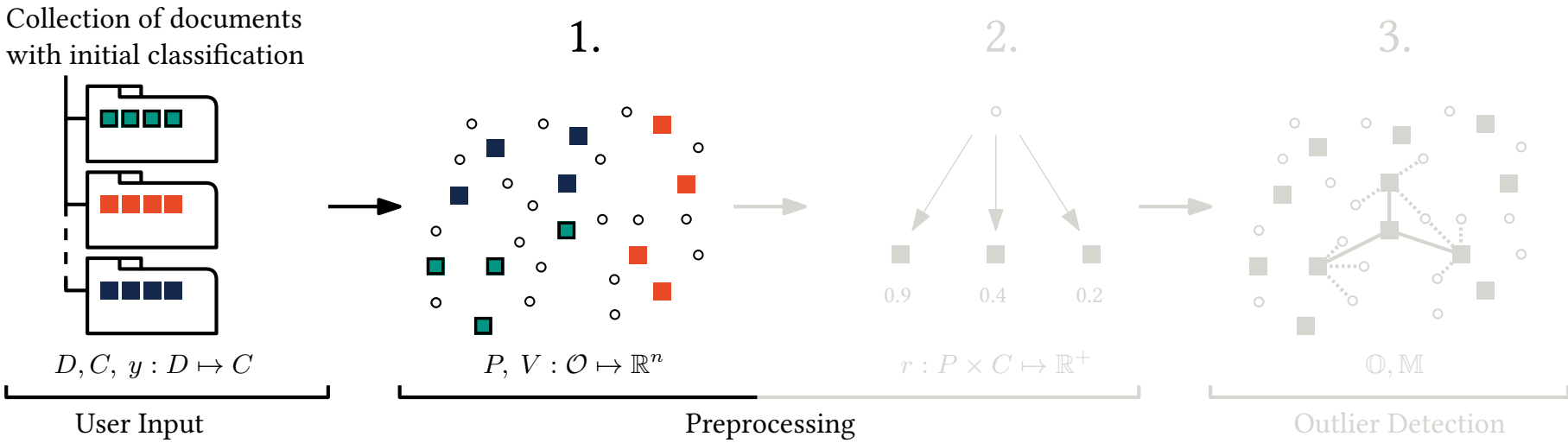
- Given: Documents D, Class C, initial classification $y : D \mapsto C$
- 1a. Extract relevant phrases P (AutoPhrase [SLJ+18]), $\mathcal{O} = D \cup P$
- 1b. Learn joint embedding $V : \mathcal{O} \mapsto \mathbb{R}^n$ (JoSE [MHW+19])
- 2. Mine representativeness $r : P \times C \mapsto \mathbb{R}^+$,
  $r = integrity * popularity * distinctiveness$ (SegPhrase [ZH19])
- 3. Our method: kj-NN $\rightarrow \mathbb{O}$: Type O, $\mathbb{M}$: Type M outlier sets

Our Framework



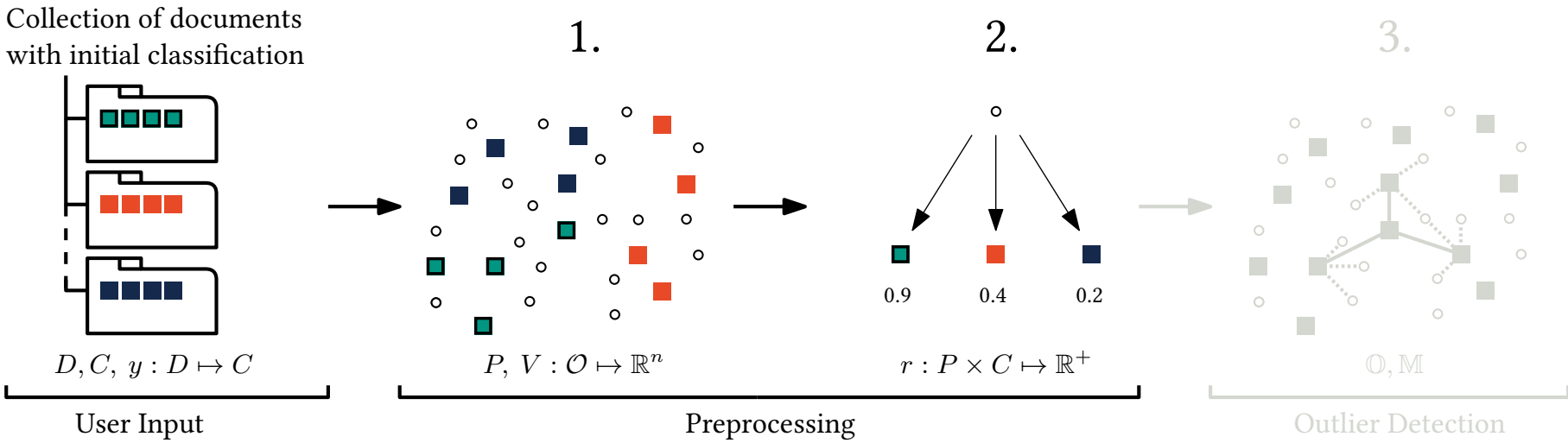
- Given: Documents D, Class C, initial classification $y : D \mapsto C$
- 1a. Extract relevant phrases P (AutoPhrase [SLJ+18]), $\mathcal{O} = D \cup P$
- 1b. Learn joint embedding $V : \mathcal{O} \mapsto \mathbb{R}^n$ (JoSE [MHW+19])
- 2. Mine representativeness $r : P \times C \mapsto \mathbb{R}^+$,
  $r = integrity * popularity * distinctiveness$ (SegPhrase [ZH19])
- 3. Our method: kj-NN $\rightarrow \mathbb{O}$: Type O, $\mathbb{M}$: Type M outlier sets

Our Framework



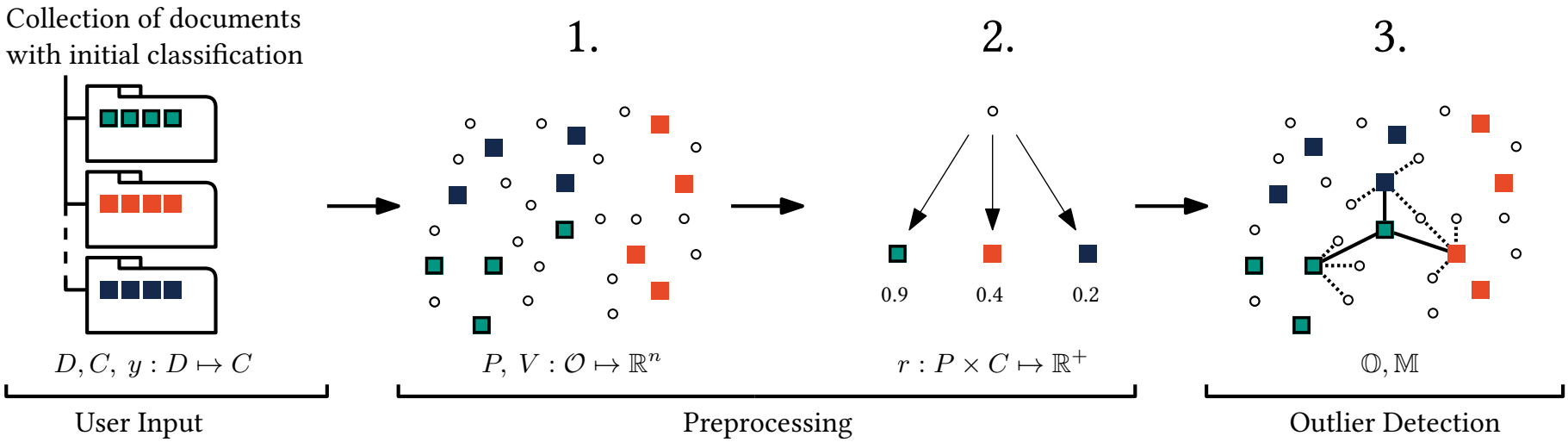
- Given: Documents D, Class C, initial classification $y : D \mapsto C$
- 1a. Extract relevant phrases P (AutoPhrase [SLJ+18]), $\mathcal{O} = D \cup P$
- 1b. Learn joint embedding $V : \mathcal{O} \mapsto \mathbb{R}^n$ (JoSE [MHW+19])
- 2. Mine representativeness $r : P \times C \mapsto \mathbb{R}^+$,
  $r = integrity * popularity * distinctiveness$ (SegPhrase [ZH19])
- 3. Our method: kj-NN $\rightarrow \mathbb{O}$: Type O, $\mathbb{M}$: Type M outlier sets

Our Framework



- Given: Documents D, Class C, initial classification $y : D \mapsto C$
- 1a. Extract relevant phrases P (AutoPhrase [SLJ+18]), $\mathcal{O} = D \cup P$
- 1b. Learn joint embedding $V : \mathcal{O} \mapsto \mathbb{R}^n$ (JoSE [MHW+19])
- 2. Mine representativeness $r : P \times C \mapsto \mathbb{R}^+$,
  $r = integrity * popularity * distinctiveness$ (SegPhrase [ZH19])
- 3. Our method: kj-NN $\rightarrow \mathbb{O}$: Type O, $\mathbb{M}$: Type M outlier sets

Our Framework

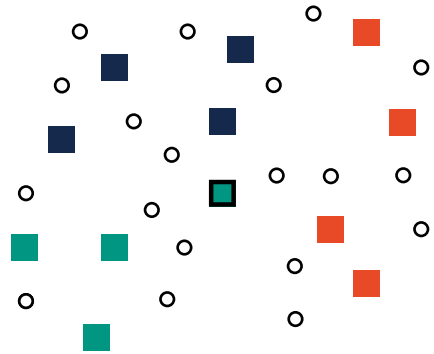


- Given: Documents D, Class C, initial classification $y : D \mapsto C$
- 1a. Extract relevant phrases P (AutoPhrase [SLJ+18]), $\mathcal{O} = D \cup P$
- 1b. Learn joint embedding $V : \mathcal{O} \mapsto \mathbb{R}^n$ (JoSE [MHW+19])
- 2. Mine representativeness $r : P \times C \mapsto \mathbb{R}^+$,
  $r = integrity * popularity * distinctiveness$ (SegPhrase [ZH19])
- 3. Our method: kj-NN $\rightarrow \mathbb{O}$: Type O, $\mathbb{M}$: Type M outlier sets

The kj-Nearest Neighbours (1/2)

Let $\mathcal{K}(d)$ and $\mathcal{J}(d)$ be the k nearest documents and the j nearest phrases of document $d \in D$. For every class $c \in C$, compute:

$$score_{d,c} = \sum_{d' \in \overbrace{\mathcal{K}_c(d)}^{\text{neighbours of class } c}} \underbrace{S(d,d')}_{\text{doc-doc cos sim}} \sum_{p \in \mathcal{J}(d')} \underbrace{S(d',p)}_{\text{doc-phrase cos sim}} \cdot \overbrace{r(p,c)}^{\text{representativeness}},$$



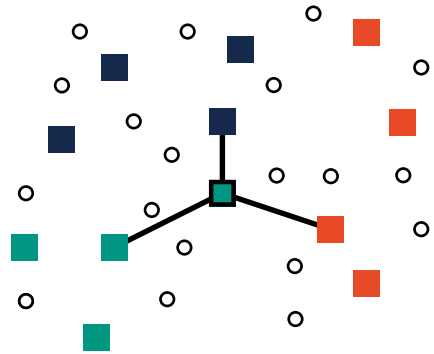
and we predict $\hat{y}(d) = \arg\max_{c \in C} score_{d,c}$

Intuition: *we maximise the posterior* $\Pr(c|d)$*, based on a posterior* $\Pr(c|d')$ *for each nearest documents that is proportional to the representativeness* $r(p,c)$ *of their nearest phrases.* (See our paper)

The kj-Nearest Neighbours (1/2)

Let $\mathcal{K}(d)$ and $\mathcal{J}(d)$ be the k nearest documents and the j nearest phrases of document $d \in D$. For every class $c \in C$, compute:

$$score_{d,c} = \sum_{d' \in \overbrace{\mathcal{K}_c(d)}^{\text{neighbours of class } c}} \underbrace{S(d,d')}_{\text{doc-doc cos sim}} \sum_{p \in \mathcal{J}(d')} \underbrace{S(d',p)}_{\text{doc-phrase cos sim}} \cdot \overbrace{r(p,c)}^{\text{representativeness}},$$



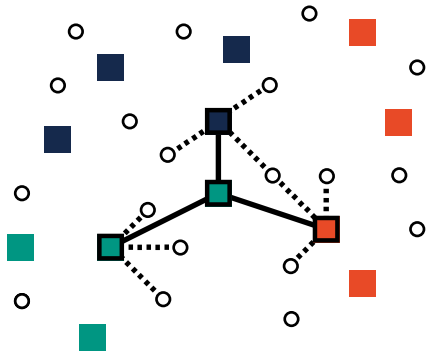
and we predict $\hat{y}(d) = \arg\max_{c \in C} score_{d,c}$

Intuition: *we maximise the posterior* $\Pr(c|d)$*, based on a posterior* $\Pr(c|d')$ *for each nearest documents that is proportional to the representativeness* $r(p,c)$ *of their nearest phrases.* (See our paper)

The kj-Nearest Neighbours (1/2)

Let $\mathcal{K}(d)$ and $\mathcal{J}(d)$ be the k nearest documents and the j nearest phrases of document $d \in D$. For every class $c \in C$, compute:

$$score_{d,c} = \sum_{d' \in \overbrace{\mathcal{K}_c(d)}^{\text{neighbours of class } c}} \underbrace{S(d,d')}_{\text{doc-doc cos sim}} \sum_{p \in \mathcal{J}(d')} \underbrace{S(d',p)}_{\text{doc-phrase cos sim}} \cdot \overbrace{r(p,c)}^{\text{representativeness}},$$



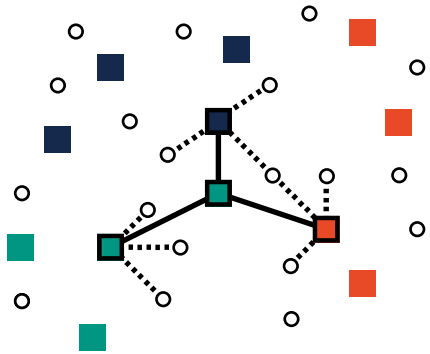
and we predict $\hat{y}(d) = \arg\max_{c \in C} score_{d,c}$

Intuition: *we maximise the posterior $\Pr(c|d)$, based on a posterior $\Pr(c|d')$ for each nearest documents that is proportional to the representativeness $r(p,c)$ of their nearest phrases.* (See our paper)

The kj-Nearest Neighbours (1/2)

Let $\mathcal{K}(d)$ and $\mathcal{J}(d)$ be the k nearest documents and the j nearest phrases of document $d \in D$. For every class $c \in C$, compute:

$$score_{d,c} = \sum_{d' \in \overbrace{\mathcal{K}_c(d)}^{\text{neighbours of class } c}} \underbrace{S(d,d')}_{\text{doc-doc cos sim}} \sum_{p \in \mathcal{J}(d')} \underbrace{S(d',p)}_{\text{doc-phrase cos sim}} \cdot \overbrace{r(p,c)}^{\text{representativeness}},$$



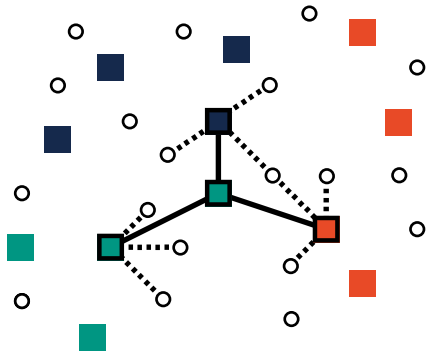
and we predict $\hat{y}(d) = \arg\max_{c \in C} score_{d,c}$

Intuition: *we maximise the posterior $\Pr(c|d)$, based on a posterior $\Pr(c|d')$ for each nearest documents that is proportional to the representativeness $r(p,c)$ of their nearest phrases.* (See our paper)

The kj-Nearest Neighbours (1/2)

Let $\mathcal{K}(d)$ and $\mathcal{J}(d)$ be the k nearest documents and the j nearest phrases of document $d \in D$. For every class $c \in C$, compute:

$$score_{d,c} = \sum_{d' \in \overbrace{\mathcal{K}_c(d)}^{\text{neighbours of class } c}} \underbrace{S(d,d')}_{\text{doc-doc cos sim}} \sum_{p \in \mathcal{J}(d')} \underbrace{S(d',p)}_{\text{doc-phrase cos sim}} \cdot \overbrace{r(p,c)}^{\text{representativeness}},$$



and we predict $\hat{y}(d) = \arg\max_{c \in C} score_{d,c}$

Intuition: *we maximise the posterior* $\Pr(c|d)$*, based on a posterior* $\Pr(c|d')$ *for each nearest documents that is proportional to the representativeness* $r(p,c)$ *of their nearest phrases*. (See our paper)

The kj-Nearest Neighbours (2/2)

Problem: What to do with uncertain predictions? ($score_{d,c}$ are similar $\forall c$)

We compute the entropy of the prediction

$$I(d) = -\sum_{c}^{C} score_{d,c} \cdot \log score_{d,c},$$

and set threshold Γ, based on percentile p^*, i.e., $\frac{|\{d \in D : I(d) < \Gamma\}|}{|D|} = p^*$

- If $I(d) > \Gamma$, then $d \in \mathbb{O}$
- Else if $\hat{y}(d) \neq y(d)$, then $d \in \mathbb{M}$
- Otherwise, d is an inlier

$\rightarrow$ kj-NN returns two lists of outliers $\mathbb{O}$ and $\mathbb{M}$.

The kj-Nearest Neighbours (2/2)

Problem: What to do with uncertain predictions? ($score_{d,c}$ are similar $\forall c$)

We compute the entropy of the prediction

$$I(d) = -\sum_{c}^{C} score_{d,c} \cdot \log score_{d,c},$$

and set threshold Γ, based on percentile p^*, i.e., $\frac{|\{d \in D : I(d) < \Gamma\}|}{|D|} = p^*$

- If $I(d) > \Gamma$, then $d \in \mathbb{O}$
- Else if $\hat{y}(d) \neq y(d)$, then $d \in \mathbb{M}$
- Otherwise, d is an inlier

$\rightarrow$ kj-NN returns two lists of outliers $\mathbb{O}$ and $\mathbb{M}$.

The kj-Nearest Neighbours (2/2)

Problem: What to do with uncertain predictions? ($score_{d,c}$ are similar $\forall c$)

We compute the entropy of the prediction

$$I(d) = -\sum_{c}^{C} score_{d,c} \cdot \log score_{d,c},$$

and set threshold Γ, based on percentile p^*, i.e., $\frac{|\{d \in D : I(d) < \Gamma\}|}{|D|} = p^*$

- If $I(d) > \Gamma$, then $d \in \mathbb{O}$
- Else if $\hat{y}(d) \neq y(d)$, then $d \in \mathbb{M}$
- Otherwise, d is an inlier

$\rightarrow$ kj-NN returns two lists of outliers $\mathbb{O}$ and $\mathbb{M}$.

The kj-Nearest Neighbours (2/2)

Problem: What to do with uncertain predictions? ($score_{d,c}$ are similar $\forall c$)

We compute the entropy of the prediction

$$I(d) = -\sum_{c}^{C} score_{d,c} \cdot \log score_{d,c},$$

and set threshold Γ, based on percentile p^*, i.e., $\frac{|\{d \in D : I(d) < \Gamma\}|}{|D|} = p^*$

- If $I(d) > \Gamma$, then $d \in \mathbb{O}$
- Else if $\hat{y}(d) \neq y(d)$, then $d \in \mathbb{M}$
- Otherwise, d is an inlier

$\rightarrow$ kj-NN returns two lists of outliers $\mathbb{O}$ and $\mathbb{M}$.

The kj-Nearest Neighbours (2/2)

Problem: What to do with uncertain predictions? ($score_{d,c}$ are similar $\forall c$)

We compute the entropy of the prediction

$$I(d) = -\sum_{c}^{C} score_{d,c} \cdot \log score_{d,c},$$

and set threshold Γ, based on percentile p^*, i.e., $\frac{|\{d \in D : I(d) < \Gamma\}|}{|D|} = p^*$

- If $I(d) > \Gamma$, then $d \in \mathbb{O}$
- Else if $\hat{y}(d) \neq y(d)$, then $d \in \mathbb{M}$
- Otherwise, d is an inlier

$\rightarrow$ kj-NN returns two lists of outliers $\mathbb{O}$ and $\mathbb{M}$.

Experiment Setup

Goal: Evaluate kj-NN w.r.t. varying number of classes and outliers.

Two datasets, with many variants:

- NYT: 10,000 articles from the New York Times (5 topics).
  - Inject 1%, 2%, 5% outliers from other 4 topics.
  - Downsample the data by 50%, 20%, 10%.
- ARXIV: 21,467 abstracts published on ArXiv from 10 CS categories.
  - Choose 1 to 5 inlier classes.
  - Inject 1% outliers from the other classes.

We simulate Type M outliers by moving $m\%$ documents to another class.

Measures: ROC AUC, Average Precision (AP), Recall/Precision at 1, 2, 5%

Experiment Setup

Goal: Evaluate kj-NN w.r.t. varying number of classes and outliers.

Two datasets, with many variants:

- NYT: 10,000 articles from the New York Times (5 topics).
  - Inject 1%, 2%, 5% outliers from other 4 topics.
  - Downsample the data by 50%, 20%, 10%.
- ARXIV: 21,467 abstracts published on ArXiv from 10 CS categories.
  - Choose 1 to 5 inlier classes.
  - Inject 1% outliers from the other classes.

We simulate Type M outliers by moving $m\%$ documents to another class.

Measures: ROC AUC, Average Precision (AP), Recall/Precision at 1, 2, 5%

Experiment Setup

Goal: Evaluate kj-NN w.r.t. varying number of classes and outliers.

Two datasets, with many variants:

- NYT: 10,000 articles from the New York Times (5 topics).
  - Inject 1%, 2%, 5% outliers from other 4 topics.
  - Downsample the data by 50%, 20%, 10%.
- ARXIV: 21,467 abstracts published on ArXiv from 10 CS categories.
  - Choose 1 to 5 inlier classes.
  - Inject 1% outliers from the other classes.

We simulate Type M outliers by moving $m\%$ documents to another class.

Measures: ROC AUC, Average Precision (AP), Recall/Precision at 1, 2, 5%

Experiment Setup

Goal: Evaluate kj-NN w.r.t. varying number of classes and outliers.

Two datasets, with many variants:

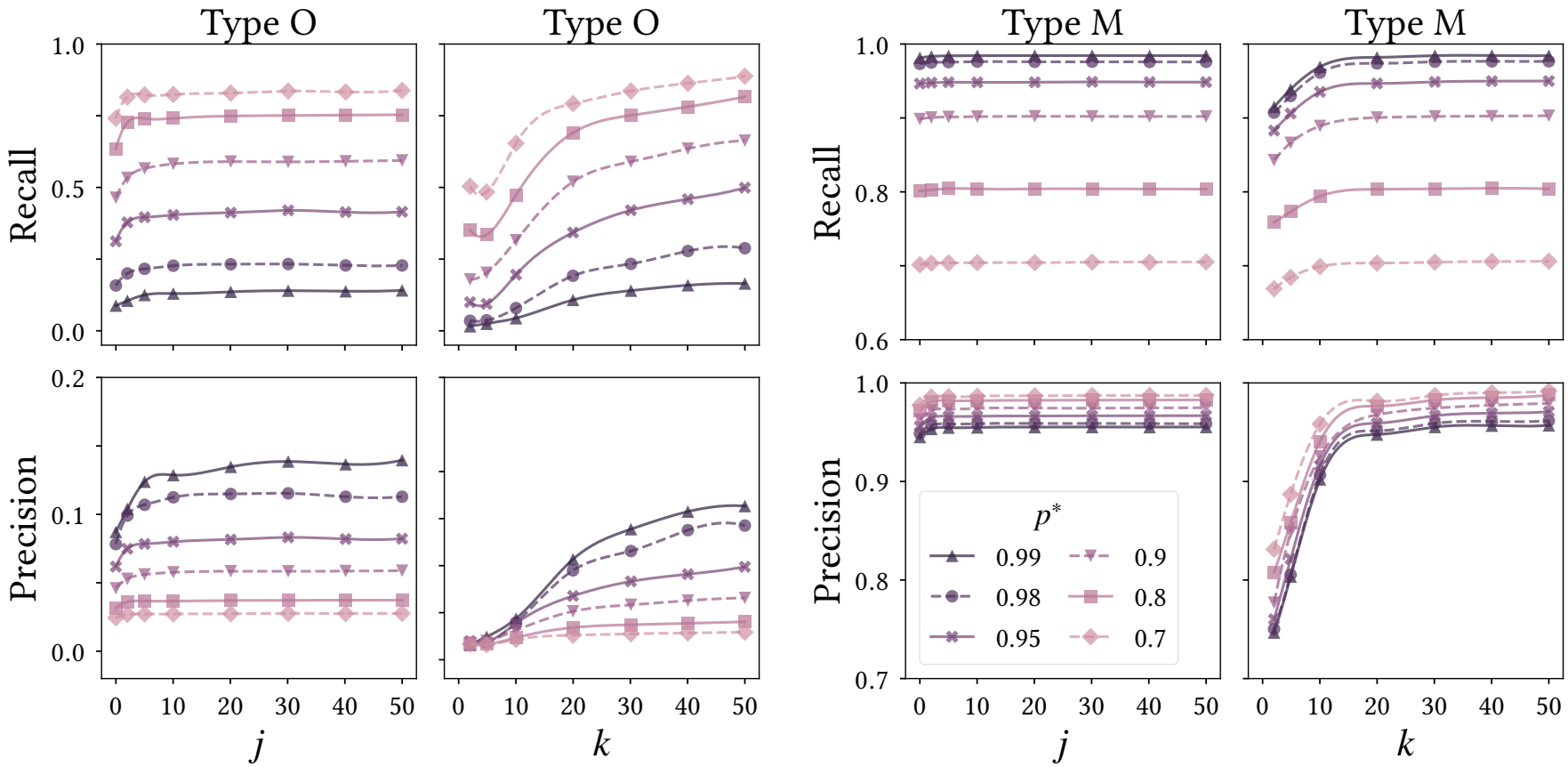
- NYT: 10,000 articles from the New York Times (5 topics).
  - Inject 1%, 2%, 5% outliers from other 4 topics.
  - Downsample the data by 50%, 20%, 10%.
- ARXIV: 21,467 abstracts published on ArXiv from 10 CS categories.
  - Choose 1 to 5 inlier classes.
  - Inject 1% outliers from the other classes.

We simulate Type M outliers by moving $m\%$ documents to another class.

Measures: ROC AUC, Average Precision (AP), Recall/Precision at 1, 2, 5%

Parameter Sensitivity

Influence of k, j, p^*



Outlier Detection (Type O)

Table 1: Comparison w.r.t. our competitors (Type O, NYT).

		AUC	AP	R1	R2	R5		AUC	AP	R1	R2	R5
LOF	NYT-1	66.45	1.62	3.00	3.00	9.00	NYT-50	60.91	1.77	2.00	6.00	12.00
RS-Hash		46.62	0.87	0.00	1.00	1.00		46.14	0.90	0.00	0.00	2.00
ANCS		66.63	2.57	6.00	10.00	21.00		63.94	4.35	14.00	16.00	20.00
k-ANCS		83.89	3.65	3.00	7.00	19.00		81.08	4.43	12.00	14.00	20.00
TONMF		58.66	7.30	0.00	2.00	9.00		61.42	**31.01**	0.90	0.90	4.90
VMF-Q		76.34	2.21	2.00	3.00	11.00		85.76	4.00	6.00	8.00	22.00
CVDD		78.47	7.75	11.7	18.09	22.34		76.29	15.89	21.62	27.03	29.73
kj-NN		**92.51**	**17.57**	**25.00**	**39.20**	**61.40**		**93.33**	27.76	**37.20**	**46.80**	**62.80**
LOF	NYT-2	60.66	2.45	2.00	2.50	6.00	NYT-20	74.76	4.12	5.00	5.00	15.00
RS-Hash		48.05	1.87	0.50	1.00	5.50		42.91	0.89	0.00	0.00	0.00
ANCS		67.59	5.19	8.00	12.00	18.00		78.52	5.60	10.00	15.00	**40.00**
k-ANCS		82.51	5.94	3.00	5.50	14.50		88.56	6.59	15.00	20.00	30.00
TONMF		54.61	1.78	2.50	3.00	9.00		64.94	6.35	0.00	0.00	15.00
VMF-Q		83.67	6.87	4.00	11.00	20.00		83.98	4.59	5.00	10.00	20.00
CVDD		73.10	10.00	11.58	14.74	22.11		88.83	**24.00**	**25.00**	**25.00**	25.00
kj-NN		**94.51**	**42.64**	**30.40**	**44.50**	**64.50**		**91.42**	6.57	3.00	11.00	38.00
LOF	NYT-5	52.28	5.44	1.80	3.40	7.00	NYT-10	77.93	2.77	0.00	0.00	20.00
RS-Hash		48.76	4.58	0.80	1.40	2.80		56.94	1.76	0.00	0.00	10.00
ANCS		67.67	11.13	6.60	9.80	19.20		83.89	13.65	30.00	40.00	40.00
k-ANCS		75.56	10.45	3.40	6.40	11.40		91.37	10.93	30.00	30.00	30.00
TONMF		52.95	1.50	1.40	2.40	6.40		71.13	29.92	0.00	0.00	0.00
VMF-Q		77.11	12.92	4.60	7.40	15.60		63.39	2.55	0.00	10.00	20.00
CVDD		72.81	18.69	9.04	13.05	21.49		85.13	**44.99**	**42.86**	**42.86**	**42.86**
kj-NN		**97.04**	**71.69**	**19.12**	**36.96**	**68.28**		**91.52**	8.45	10.00	16.00	38.00

Outlier Detection (Type M)

Table 2: Comparison w.r.t. our competitors (Type M, NYT).

		P	R	F1	R10	R20		P	R	F1	R10	R20
W-CNN	NYT-1	54.38	86.04	66.64	27.28	53.51	NYT-50	47.38	87.61	61.50	22.28	47.62
VD-CNN		90.71	69.22	78.52	15.04	30.18		**98.58**	55.44	70.97	49.31	54.95
AT-RNN		67.12	51.88	58.52	32.92	51.34		89.75	69.22	78.16	14.83	29.92
RCNN		**96.02**	9.65	17.54	9.55	9.55		57.46	87.31	69.31	27.23	56.93
kj-NN		95.93	**90.02**	**92.88**	**50.19**	**90.02**		95.78	**90.36**	**92.99**	**50.22**	**90.36**
W-CNN	NYT-2	54.16	88.02	67.06	27.30	54.56	NYT-20	39.34	**91.56**	55.03	20.54	40.59
VD-CNN		88.99	69.69	78.17	14.71	29.61		93.50	81.80	87.26	15.20	30.81
AT-RNN		80.36	49.03	60.90	40.29	48.14		50.44	85.11	63.34	25.25	52.23
RCNN		89.60	5.59	10.52	5.49	5.49		39.63	**91.56**	55.32	20.54	40.59
kj-NN		**94.63**	**91.15**	**92.86**	**50.74**	**91.15**		**93.80**	90.64	**92.19**	**50.52**	**90.64**
W-CNN	NYT-5	51.12	89.07	64.96	25.14	50.38	NYT-10	36.12	88.94	51.38	16.83	35.15
VD-CNN		58.21	82.39	68.22	8.98	18.46		86.20	69.40	76.89	14.10	27.26
AT-RNN		61.71	70.49	65.81	31.38	61.62		36.12	88.94	51.38	16.83	35.15
RCNN		90.82	44.93	60.12	42.86	42.86		36.12	88.94	51.38	16.83	35.15
kj-NN		**92.43**	**93.44**	**92.93**	**52.27**	**93.44**		**88.57**	**90.77**	**89.65**	**50.50**	**89.97**

→ See our paper for results w.r.t. the ARXIV benchmark and ablation analysis

Interpretable results (1/2)

O (Education): NYC will build a new home for one of its premier high schools, Stuyvesant, [...] under a schedule that seeks to show that its public schools can be built fast and well, Mayor Koch and Governor Cuomo said yesterday. The new school, incorporating the latest in modern laboratory equipment, fiber optic systems and an olympic size swimming pool will be built [...] in lower manhattan, with work to begin at the end of next year...

1st-NN (Business): Hong Kong on the first floor of a hulking residential building, at the end of a dimly lighted corridor, a narrow door opens up into Hong Kong's economic underbelly [...]. Hong Kong's housing situation is now one of the reasons the government of Leung Chun Ying, who took the helm of the city 's administration last year, is deeply unpopular...

2nd-NN (Politics): Praising the work of young scientists and inventors [...], President Obama on monday announced a broad plan to create and expand [...] initiatives designed to encourage children to study science, technology, engineering and mathematics. [...] Obama said he was committed to giving students the resources they need to pursue education...

3rd-NN (Politics): After [...] intense political pressure, schools chancellor Rudy Crew [...] said he would accept the candidate. Dr. crew had provoked harsh criticism last month when [...] he used his new veto power [...] to reject Claire Mcintee, an elementary school principal who was unanimously selected [...] to be the district's top administrator...

Top phrases: City, state, program, buildings, education, office, schools, year, project, company...

Interpretable results (2/2)

Type O: Review of : Brigitte Le Roux and Henry Rouanet, geometric data analysis, from correspondence analysis to structured data analysis, ...

→ **Top phrases:** Data, paper, challenges, learning, ...

Type O: The paper has been withdrawn due to an error in Lemma 1.

→ **Top phrases:** Problem, work, error, conjecture, ...

Type M (cs.AI → cs.CL): Open-text (or open-domain) semantic parsers are designed to interpret any statement in natural language by inferring a corresponding meaning representation (MR). Unfortunately, large scale systems cannot be easily machine-learned, due to lack of directly supervised data. We propose here a method that learns to assign MRs to a wide range of text (using a dictionary of more than 70,000 words, which are mapped to more than 40,000 entities) thanks to a training scheme that combines learning from WordNet and ConceptNet with learning from raw text. The model learns structured embeddings of words, entities and MRs via a multi-task training process operating on these diverse sources of data [...]. This work ends up combining methods for knowledge acquisition, semantic parsing, and word-sense disambiguation ...

→ **Top phrases:** Representations, word, semantic, model, embeddings, information, word embeddings, ...

Conclusions

- Mining text outliers is difficult: manifold, domain-specific, unsupervised
  - Outliers fall into two types: Out-of-distribution (O), Misclassification (M)

- Our approach, kj-NN, is the first one to detect both types
  - Exploit document/phrase similarities
  - Improved performance, compared to existing work
  - Results are interpretable

- Future work: There are many possible extensions
  - Other domains: Multivariate time series
  - Other settings: Streams, multi-class, few shots...

Conclusions

- Mining text outliers is difficult: manifold, domain-specific, unsupervised
  - Outliers fall into two types: Out-of-distribution (O), Misclassification (M)
- Our approach, kj-NN, is the first one to detect both types
  - Exploit document/phrase similarities
  - Improved performance, compared to existing work
  - Results are interpretable
- Future work: There are many possible extensions
  - Other domains: Multivariate time series
  - Other settings: Streams, multi-class, few shots...

Conclusions

- Mining text outliers is difficult: manifold, domain-specific, unsupervised
  - Outliers fall into two types: Out-of-distribution (O), Misclassification (M)
- Our approach, kj-NN, is the first one to detect both types
  - Exploit document/phrase similarities
  - Improved performance, compared to existing work
  - Results are interpretable
- Future work: There are many possible extensions
  - Other domains: Multivariate time series
  - Other settings: Streams, multi-class, few shots…

References I

[BKNS00] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying Density-Based Local Outliers. In *SIGMOD*, pages 93–104, 2000.

[CSBL17] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann LeCun. Very Deep Convolutional Networks for Text Classification. In *EACL (1)*, pages 1107–1116, 2017.

[Kim14] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *EMNLP*, pages 1746–1751, 2014.

[KMB12] Fabian Keller, Emmanuel Müller, and Klemens Böhm. HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. In *ICDE*, pages 1037–1048, 2012.

[KN98] Edwin M. Knorr and Raymond T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *VLDB*, pages 392–403, 1998.

[KS12] JooSeuk Kim and Clayton D. Scott. Robust Kernel Density Estimation. *J. Mach. Learn. Res.*, 13:2529–2565, 2012.

[KSZ08] Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-Based Outlier Detection in High-dimensional Data. In *KDD*, pages 444–452, 2008.

[KWAP17] Ramakrishnan Kannan, Hyenkyun Woo, Charu C. Aggarwal, and Haesun Park. Outlier Detection for Text Data. In *SDM*, pages 489–497, 2017.

[LXLZ15] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent Convolutional Neural Networks for Text Classification. In *AAAI*, pages 2267–2273, 2015.

[MHW+19] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance M. Kaplan, and Jiawei Han. Spherical Text Embedding. In *NeurIPS*, pages 8206–8215, 2019.

[RRS00] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. In *SIGMOD*, pages 427–438, 2000.

[RZV+19] Lukas Ruff, Yury Zemlyanskiy, Robert A. Vandermeulen, Thomas Schnake, and Marius Kloft. Self-Attentive, Multi-Context One-Class Classification for Unsupervised Anomaly Detection on Text. In *ACL (1)*, pages 4061–4071, 2019.

References II

[SA18] Saket Sathe and Charu C. Aggarwal. Subspace histograms for outlier detection in linear time. *Knowl. Inf. Syst.*, 56(3):691–715, 2018.

[SLJ+18] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. Automated Phrase Mining from Massive Text Corpora. *IEEE Trans. Knowl. Data Eng.*, 30(10):1825–1837, 2018.

[TB99] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[ZH19] Chao Zhang and Jiawei Han. *Multidimensional Mining of Massive Text Data*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2019.

[ZST+16] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *ACL (2)*, pages 207–212, 2016.

[ZWT+17] Honglei Zhuang, Chi Wang, Fangbo Tao, Lance M. Kaplan, and Jiawei Han. Identifying Semantically Deviating Outlier Documents. In *EMNLP*, pages 2748–2757, 2017.