WOT-Class: Weakly Supervised Open-world Text Classification

Tianle Wang[1,2] , Zihan Wang[2] , Weitang Liu[2] , Jingbo Shang[2]

[1]Shanghai Jiao Tong University

[2]University of California, San Diego

[1]wtl666wtl@sjtu.edu.cn

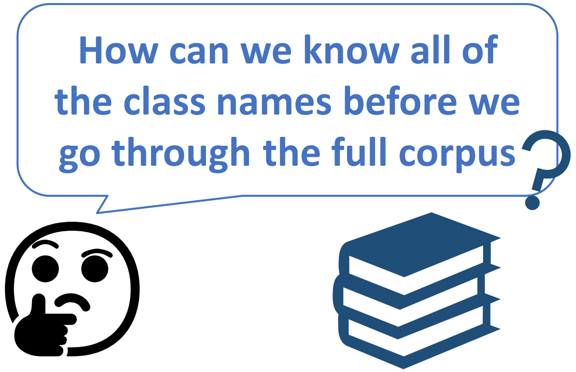
[2]{tiw054, ziw224, wel022, jshang}@ucsd.edu

Agenda

- Motivation
- Main Contribution
- Problem Formulation
- Methodology
- Evaluation
- Conclusion

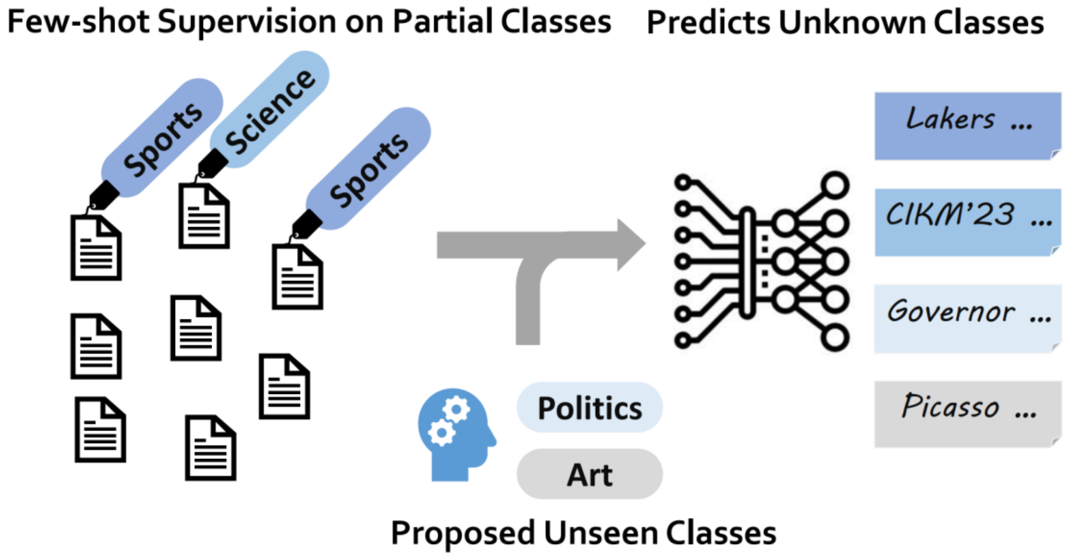
Motivation

- Weakly supervised text classification methods quickly developed these years and significantly reduced the required human supervision
- However, All these methods require human-provided known classes cover all the classes of interest
- Difficult in the real world! E.g., the human expert could be exploring a new, large corpus without a complete picture



Motivation

- How to resolve this problem? Ask machine to find unknown classes!



- The open-world setting here releases the all-class requirement, further reducing the required human effort in weakly supervised text classification.

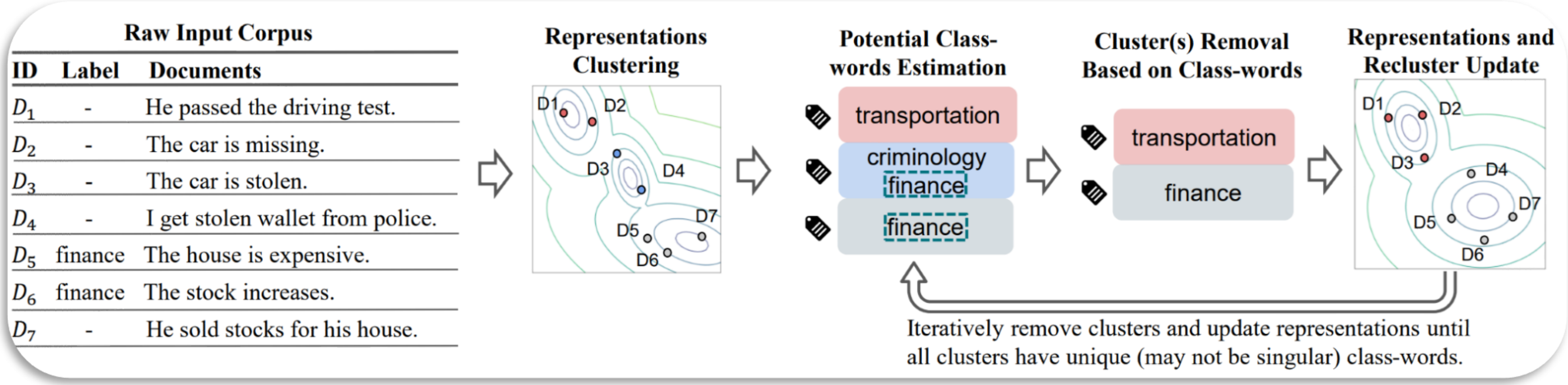
Main Contribution

- We introduce the novel yet important problem of weakly supervised open-world text classification
- We propose a novel, practical framework WOT-Class and extensive experiments demonstrate its power

Problem Formulation

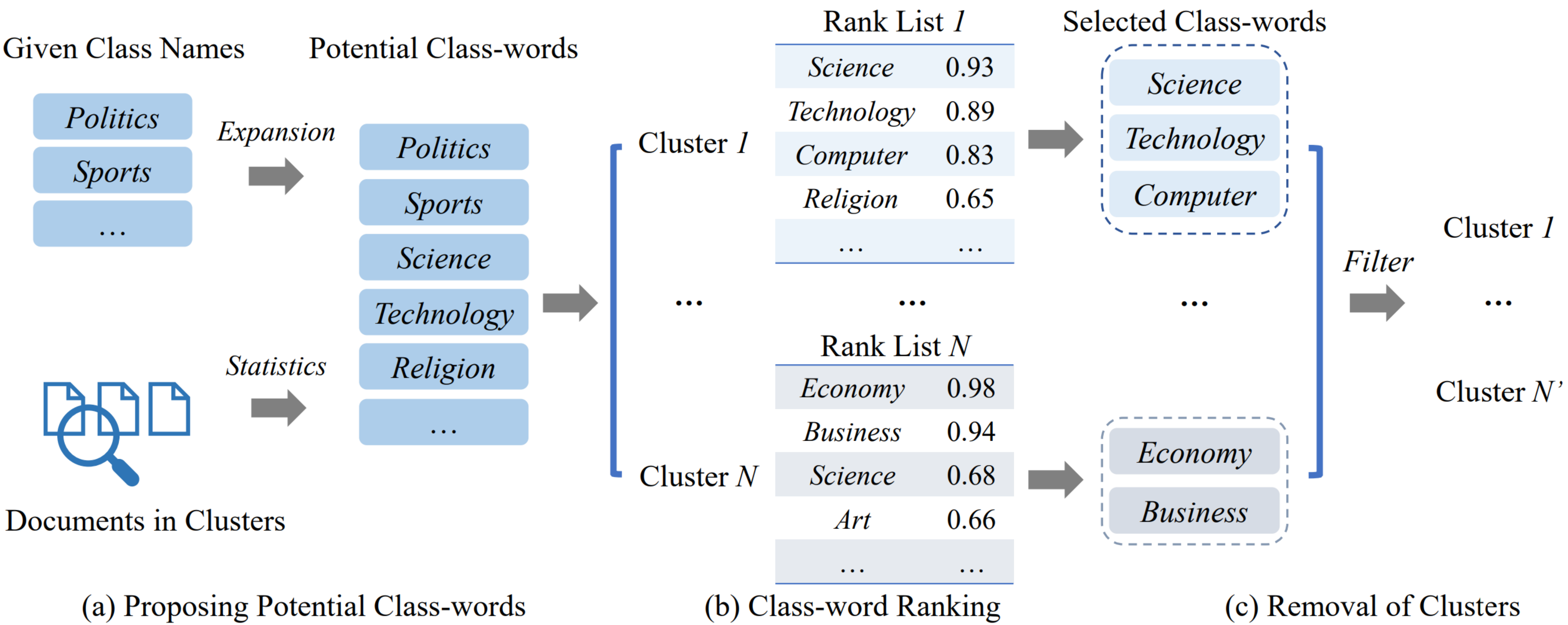
- There exists a not-fully-known set of classes C, which follow the same hyper-concept and a set of documents D, each uniquely assigned to a class.
- A weakly supervised open-world model can observe partial information of C. In this work, the information is given as a labeled few-shot dataset $Ds = \{x_i, y_i\}_{i=1}^{n}$, $y_i \in Cs$, where $Cs \subset C$ is the known subset of classes and n is rather small.
- The goal of the model is to classify the remainder of the dataset, $Du = D \backslash Ds$, where some of the labels in $Cu = C \backslash Cs$ is completely unknown to the model.

Overview of WOT-Class



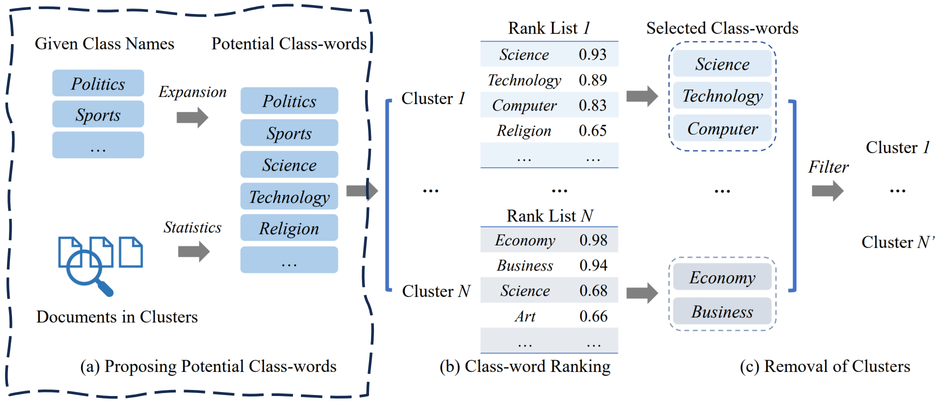
- Overestimate the number of classes and give a rough clustering
- Find class-indicative words for each cluster and delete redundant ones
  - Propose potential class-words
  - Rank class-words in each cluster
  - Remove redundant clusters based on class-words
- Utilize WS-TC method for keyword-based text classification (clustering)

Methodology



Propose Potential Class-words

- Class-words are words that are related to or highly indicative of the class (cluster):
  - **Semantics**: Entity expansion algorithms (e.g., CGExpan) generate words under the same semantic hyper-concept of known class names
  - **Statistics**: TF-IDF liked methods find statistically representative words within each cluster
- Merged as the set of potential class-words



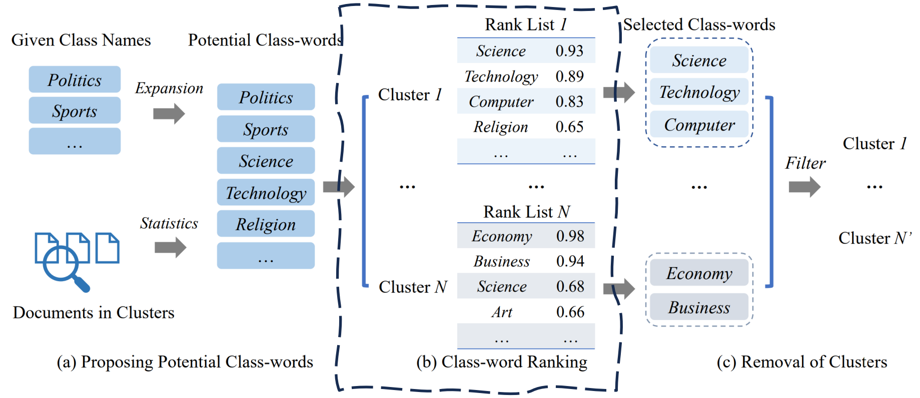
Class-word Ranking

- We construct a MLP to compute the similarity score $p(w,i)$ between a cluster i and a potential class-word w:
  - **Feature Design**: Select a large list of keyword in cluster, compute cos similarity and distance between potential words and these words as features
  - **Training Dataset**: Utilize the few-shot supervision to build virtual clusters, positive sample is given class name, negative sample is the furthest word
- Further design a penalty coefficient $\mu(w,i)$ to penalize generic words:

$$\mu(w,i) = \log\left(\frac{M\ \{rank_j(w) \mid 1 \le j\ \le C\}}{1 + rank_i(w)}\right)$$

- Final indicativeness ranking is based on:

$$I(w,i) = p(w,i) \times \mu(w,i).$$

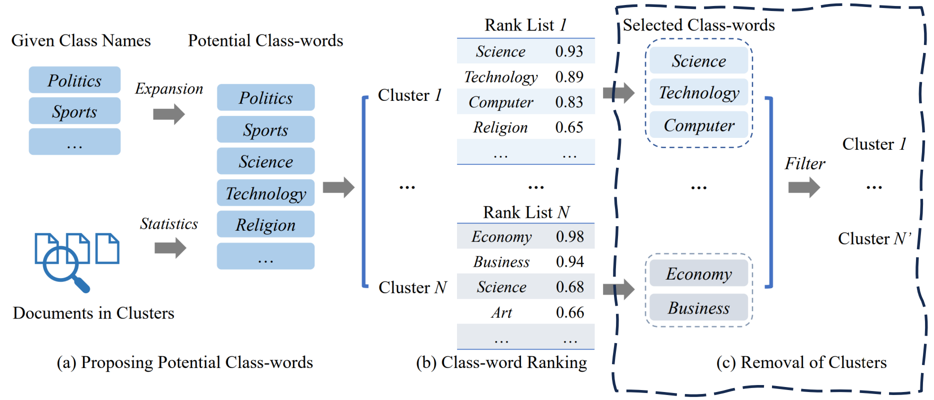


Removal of Clusters

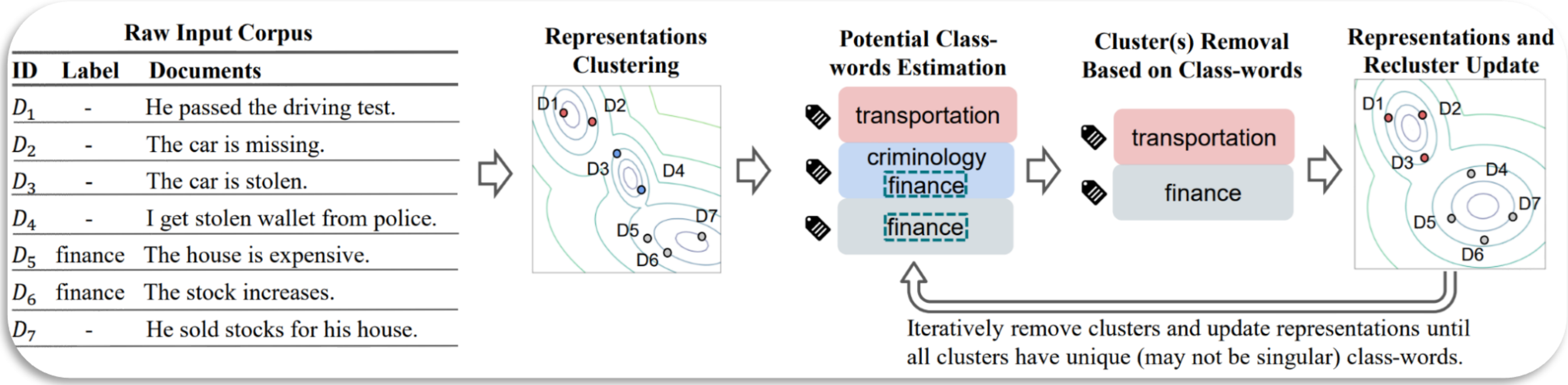
- In simple terms, we remove clusters that have non-empty intersections in the class-words:
  - After ranking, we pick the T highest ranked class-words for a cluster to compare, where T is the number of iterations in the removal process
  - We introduce a cutoff threshold β such that we do not pick words that have a low ratio of score to the highest score in the cluster
  - When overlapping, we remove the cluster with a low coherent η

$$\eta = \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} \cos(r, \overline{\mathbf{R}})$$

- Re-rank the class words and continue the process until no clusters require removal



Iterative Framework



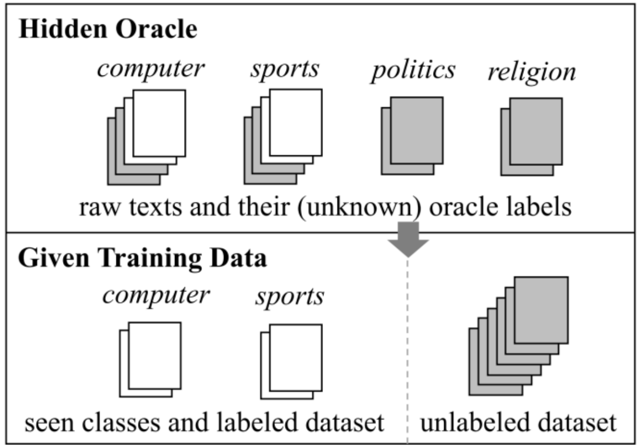
- After identifying unique class-words for each cluster, we apply the seed-word driven text classification method (e.g., X-Class) to update the clusters
- The whole iterative framework ends until it can no longer remove clusters. We then train a final text classifier based on the pseudo-labels assigned to each text

Experimental Setting

- **Dataset**
  - 7 popular datasets
  - from different textual sources and criteria of classes
- **Compared Methods**
  - 3 open-world methods adapted from image field: Rankstats+, ORCA, GCD
  - 2 solid baseline in text field: BERT+Clustering (GMM/SVM)

	AGNews	20News	NYT-Small	NYT-Topics	NYT-Locations	Yahoo	DBpedia
Corpus Domain	News	News	News	News	News	QA	Wikipedia
Class Criterion	Topics	Topics	Topics	Topics	Locations	Topics	Ontology
# of Classes	4	5	5	9	10	10	14
# of Documents	12,000	17,871	13,081	31,997	31,997	18,000	22,400
Imbalance	1.00	2.02	16.65	27.09	15.84	1.00	1.00

Experimental Setting



- Make the most infrequent half of classes as unseen
- Among the seen classes, we give 10-shot supervision
- Since all compared methods require the total number of classes as input, we evaluate them in two ways:
  - Our Estimation (OE)
  - Baselines' Estimation
- Evaluation: calculate the F1 Score after maximum matching

Overall Performance

- WOT-Class performs noticeably better than SOTA general methods and BERT+Clustering baselines
- Gains a 23.33% greater average absolute macro-F1 over the current best method across all datasets

Method	Extra Info	AGNews	20News	NYT-S	NYT-Top	NYT-Loc	Yahoo	DBpedia	Average
Rankstats+		39.53/28.55	24.94/13.88	52.01/23.13	42.23/19.98	39.68/23.13	29.66/20.44	48.20/39.15	39.47/24.04
ORCA	✗	72.44/72.27	48.92/39.83	74.34/42.22	62.23/39.02	58.71/44.81	35.57/32.71	69.27/67.92	60.21/48.40
GCD		66.37/66.51	51.75/42.96	82.59/63.35	66.36/39.69	70.25/53.41	36.73/35.39	75.81/72.97	64.27/53.47
WOT-Class		**79.42/79.75**	**79.07/79.29**	**94.78/88.46**	**78.67/69.48**	**80.94/79.55**	**54.46/56.23**	**85.15/84.87**	**78.93/76.80**
		+5.52%	+36.33%	+25.11%	+29.79%	+26.14%	+20.84%	+11.90%	+23.33%
Rankstats+ (OE)		61.44/57.50	53.65/38.12	40.82/31.67	19.93/15.07	21.96/16.81	32.79/26.94	50.03/44.31	40.09/32.92
ORCA (OE)		64.38/64.50	51.85/40.04	70.44/46.21	59.42/38.29	42.99/33.08	43.87/41.43	82.54/81.30	59.35/49.26
GCD (OE)	# of Classes	65.42/65.44	61.27/56.42	78.82/56.59	70.51/42.44	55.37/44.86	39.01/37.58	84.14/83.60	64.93/55.28
BERT+GMM (OE)		38.25/37.14	29.32/25.21	58.79/24.79	26.88/14.08	11.64/9.47	14.11/13.64	14.74/14.20	27.68/19.79
BERT+SVM (OE)		45.20/44.15	39.07/34.96	51.97/22.34	24.95/12.83	13.91/7.45	15.25/13.39	16.28/14.41	29.52/21.36

Imbalance Tolerance

- We construct 3 imbalanced DBpedia datasets with different degrees of imbalance

	Low	Medium	High
Δ	2%	4%	6%
# of Documents	19,480	16,565	13,652
Imbalance	1.35	2.09	4.56

- WOT-Class reaches the lowest performance drop compared with ORCA and GCD's Pareto optimal w/ & w/o extra info

Method	DBpedia			DBpedia-Low			DBpedia-Medium			DBpedia-High		
	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen
ORCA	81.30	95.19	67.42	76.07	95.51	56.64	72.21	97.20	47.23	69.99 -11.31%	97.63	42.34
GCD	83.60	93.48	73.71	82.92	94.32	71.51	81.78	94.42	69.15	75.57 -8.03%	92.53	58.61
WOT-Class	84.87	87.16	82.59	85.81	91.82	79.97	84.97	93.31	76.63	79.35 -5.52%	88.81	69.90

Prediction of # of Classes

Method	AGNews	20News	NYT-S	NYT-Top	NYT-Loc	Yahoo	DBpedia	Average Offset
Rankstats+	2.00_{0}	2.00_{0}	$2.33_{0.58}$	$4.33_{0.58}$	5.00_{0}	5.00_{0}	$10.33_{3.06}$	3.71
ORCA	$69.00_{6.08}$	$53.67_{45.65}$	$39.33_{32.35}$	$96.33_{2.89}$	$94.67_{3.21}$	$66.67_{55.16}$	$66.00_{3.61}$	62.48
GCD	$23.33_{2.89}$	$16.00_{19.92}$	$59.33_{17.01}$	$29.67_{26.63}$	$27.33_{9.24}$	$20.00_{16.64}$	$14.00_{3.46}$	19.81
WOT-Class	$19.67_{1.15}$	$20.67_{0.58}$	$27.00_{2.89}$	$18.67_{3.61}$	$11.67_{0.58}$	$21.00_{2.00}$	$17.67_{2.89}$	11.33
Ground Truth	4	5	5	9	10	10	14	-

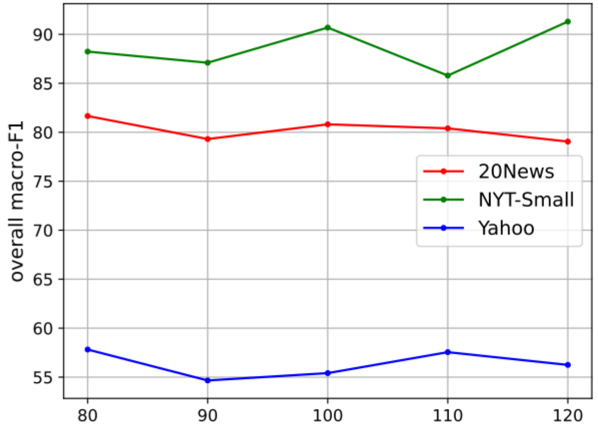
- Rankstats+, ORCA, and GCD’s ability to estimate the number of classes in the few-shot setting is more unreliable

- Examples of class-words we find:

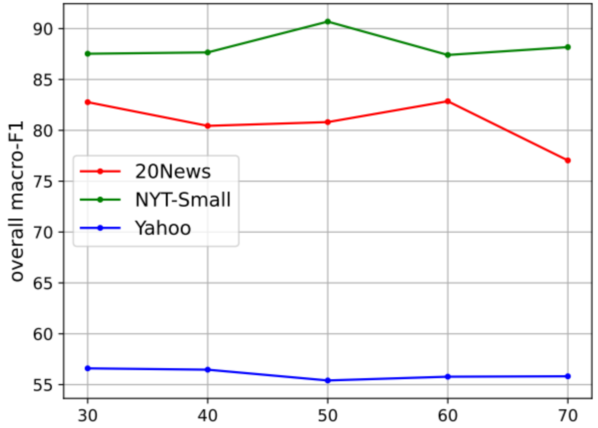
Dataset	Ground Truth	WOT-Class
NYT-Loc	Russia	[Ukraine, Russia]
	Germany	[Germany]
	Canada	[Canada]
	France	[France]
	Italy	[Italy]
DBpedia	athlete	[footballer], [Olympics]
	artist	[painting, painter, art], [tv, theatre, television]
	company	[retail, company, business]
	school	[school, education, academic]
	politics	[politician]
	transportation	[aircraft, locomotive]
	building	[architecture, tower, church]

Hyper-parameter Sensitivity

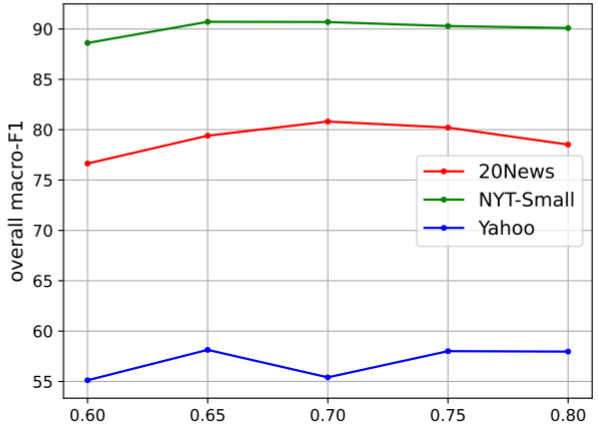
- Conduct our study on 20News, NYT-Small and Yahoo using a fixed random seed (42)
- The performance fluctuations remain within reasonable margins, basically under 5%



(a) K



(b) W



(c) β

Conclusion

- We first introduce the challenging yet promising weakly supervised open-world classification task into text domain
- We have identified the key challenges and unique opportunities of this task and proposed WOT-Class that achieves quite decent performance with minimal human effort

Future Work

- Maybe open-world text classification can be conducted with even less manual annotation. For example, by only requiring user-provided hyper-concept (e.g., Topics, Locations) or custom instructions
- Open-world text classification is an emerging field, demanding more algorithms, datasets, evaluation metrics … to truly unleash its potential

Thanks for your listening!

tiw054@ucsd.edu