X-Class: Text Classification with Extremely Weak Supervision

Zihan Wang [1] Dheeraj Mekala [1] Jingbo Shang [1,2]

[1] Department of Computer Science and Engineering, University of California San Diego, CA, USA
[2] Halıcıoğlu Data Science Institute, University of California San Diego, CA, USA
{ziw224, dmekala, jshang}@ucsd.edu

Preliminary

- Extremely weak supervision refers to the use of only class names to classify documents.
- Unlike previous weak supervision with expert given seed words, extremely weak supervision requires almost no human effort.
- Like previous work, we also assume that class names exist in the corpus. However, X-Class has a much more mild requirement on such existence, and only one occurrence is enough for the model to perform relatively well.

X-Class Model

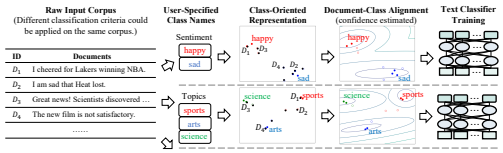
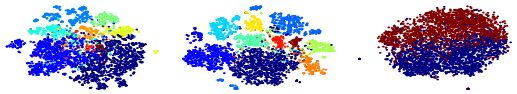


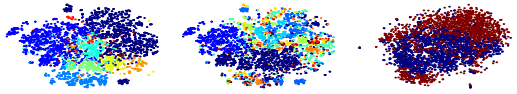
Fig. 1: Overview of X-Class

- First obtain class-oriented representations (Class Representation Estimation → Document Representation).
- Project the representations with PCA and use a Gaussian Mixture Model to cluster the document representations, given priors based on similarity to class representations.
- Based on confidence from the mixtures, select the most confident ones to train a supervised text classifier.

Analysis



(a) Our Class-Oriented Document Representations



(b) Simple Average of BERT Representations

- t-SNE plots with class-oriented document representations and with a simple average document representations [1] on three datasets, NYT-Topic (left), NYT-Location (middle), and Yelp (right).
- Our class-oriented document representations provide much more clear separation of classes than a simple average, and is adaptive to the class names.

Class Representation Estimation

- Start with the class name, iteratively expand class semantics by adding in the most similar word.
- Class representation for each class l is dynamically estimated by

$$\mathbf{x}_l = \frac{\sum_{i=1}^{|\mathcal{K}_l|} 1/i \cdot \mathbf{s}_{\mathcal{K}_{l,i}}}{\sum_{i=1}^{|\mathcal{K}_l|} 1/i}$$

where $\mathbf{s}_{\mathcal{K}_{l,i}}$ is the representation for the words and $|\mathcal{K}_l|$ is the size of the class.

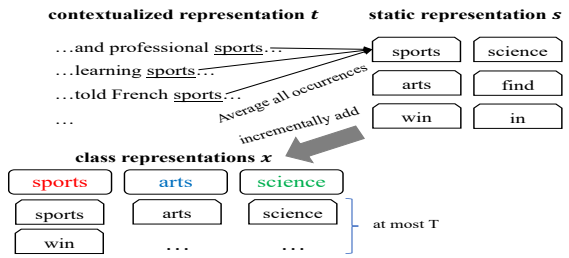


Fig. 2: Overview of Our Class Rep. Estimation.

Document Representation Estimation

- Attend on token representations with class representation.
- Use weighted average of tokens as document representation.

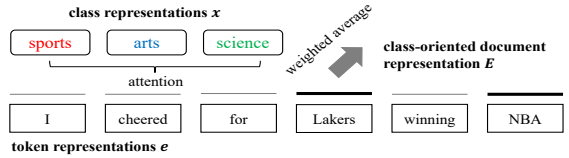


Fig. 3: Overview of Our Document Rep. Estimation.

Dataset

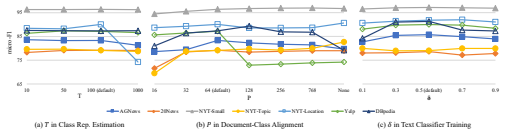
	AGNews	20News	NYT-Small	NYT-Topic	NYT-Location	Yelp	DBpedia
Corpus Domain	News	News	News	News	News	Reviews	Wikipedia
Class Criterion	Topics	Topics	Topics	Topics	Locations	Sentiment	Ontology
# of Classes	4	5	5	9	10	2	14
# of Documents	120,000	17,871	13,081	31,997	31,997	38,000	560,000
Imbalance	1.0	2.02	16.65	27.09	15.84	1.0	1.0

- Seven datasets from various domain, and with different class criterion.

Experiments

Model	AGNews	20News	NYT-Small	NYT-Topic	NYT-Location	Yelp	DBpedia
Supervised	93.99/93.99	96.45/96.42	97.95/95.46	94.29/89.90	95.99/94.99	95.7/95.7	98.96/98.96
WeSTClass†	82.3/82.1	71.28/69.90	91.2/83.7	68.26/57.02	63.15/53.22	81.6/81.6	81.42/81.19
ConWea†	74.6/74.2	75.73/73.26	95.23/90.79	**81.67/71.54**	85.31/83.81	71.4/71.2	N/A
LOTClass	**86.89/86.82**	73.78/72.53	78.12/56.05	67.11/43.58	58.49/58.96	87.75/87.68	86.66/85.98
X-Class	85.74/85.66	**78.62/77.76**	**97.18/94.02**	79.02/68.55	**91.8/91.98**	**90.0/90.0**	**91.32/91.17**
X-Class-Rep	77.86/76.84	75.37/73.7	92.13/83.69	77.06/65.05	86.36/88.1	78.0/77.19	74.05/71.74
X-Class-Align	83.32/83.28	79.19/78.46	96.42/92.32	79.12/67.76	90.09/90.63	87.19/87.13	87.36/87.27

- Micro-/Macro-F_1 scores. † indicates the use of at least 3 seed words per class.
- X-Class performs best on five out of the seven datasets. This is considering models that use expert given seed works.



(a) T in Class Rep. Estimation (b) P in Document-Class Alignment (c) δ in Text Classifier Training

Fig. 4: Ablations

Study on Class Name Frequency

Model	20News		NYT-Small	
	Original	Removed	Original	Removed
X-Class	77.76	74.48	94.02	93.29
LOTClass	72.53	8.82	56.05	29.53

- Has a similar performance even when removing all but one occurrence of a class name in the corpus.

Extension to Hierarchical Classification

Model	Coarse (5 classes)	Fine (26 classes)
WeSTClass	91/84§	50/36§
WeSHClass		87.4/63.2§
ConWea	95.23/90.79	91/79§
X-Class-End	96.67/92.98	86.07/75.30
X-Class-Hier		92.66/80.92

- Works for hierarchical classification by recursively classifying on the hierarchy (X-Class-Hier), but not so well with all the leaves together (X-Class-End).

References

[1] Roee Aharoni and Yoav Goldberg. "Unsupervised domain clusters in pretrained language models". In: *arXiv preprint arXiv:2004.02105* (2020).