X-Class: Text Classification with Extremely Weak Supervision

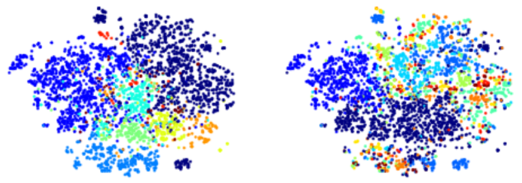
NAACL2021

Zihan Wang, Dheeraj Mekala, Jingbo Shang

UCSD

Overview

- Task
  - Text classification with extremely weak supervision, i.e.,only relying on the surface text of class names.



(a) NYT-Topics (b) NYT-Locations

Figure 1: Visualizations of News using Average BERT Representations. Colors denote different classes.

- Key insights
  - ideal document representations should lead to very close results between clustering and the desired classification
    - i.e., doc embeddings should reflect class info in clustering

Overview - three modules

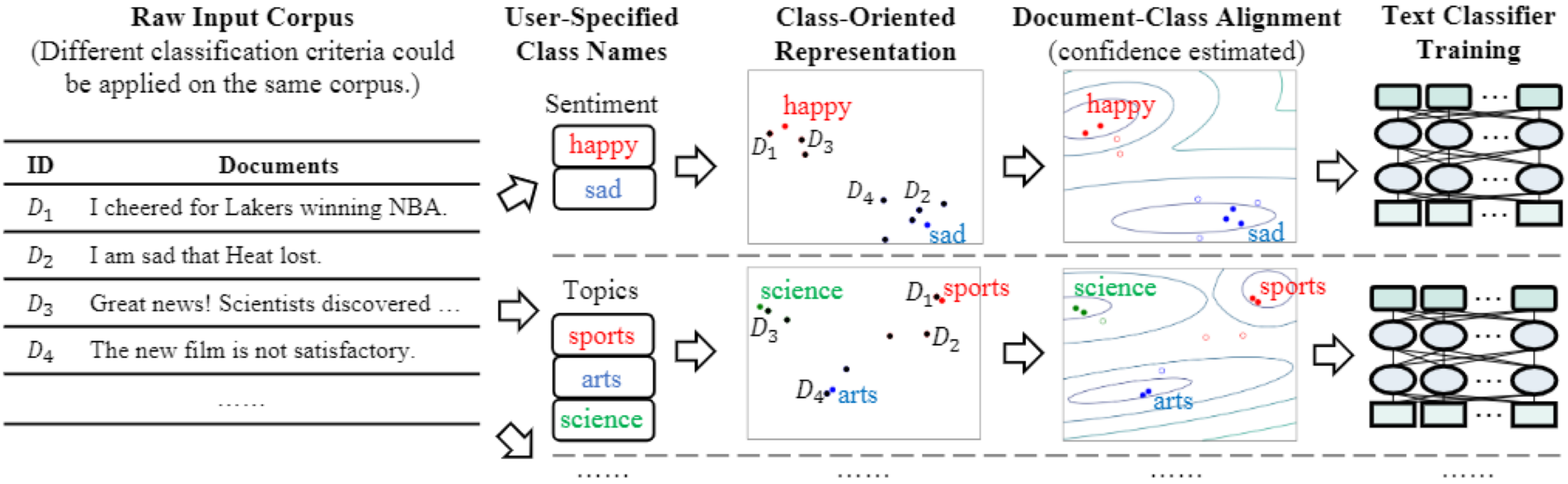


Figure 2: An overview of our X-Class. Given a raw input corpus and user-specified class names, we first estimate a class-oriented representation for each document. And then, we align documents to classes with confidence scores by clustering. Finally, we train a supervised model (e.g., BERT) on the confident document-class pairs.

M1: Class-oriented Document Representation

- Class Representation Estimation
  - Weighted average representation based on a ranked list of keywords
  - Incrementally add new keywords to list by ranking similarities of out-of-list words
    - Top-ranked keywords are expected to have more similar **static representations** to the class representation
  - Stop condition
    - New class rep. **Changed** the current list OR reach max T
- Document Representation Estimation
  - 4 ways to compute attention weight
    - 2 token rep. + 2 attention mechanisms
  - A unified list of geometric mean of the 4 ranks
    - Assign a weight of 1/r to a token ranked at r-th position

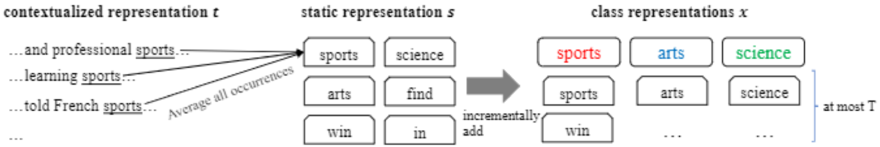


Figure 3: Overview of Our Class Rep. Estimation.

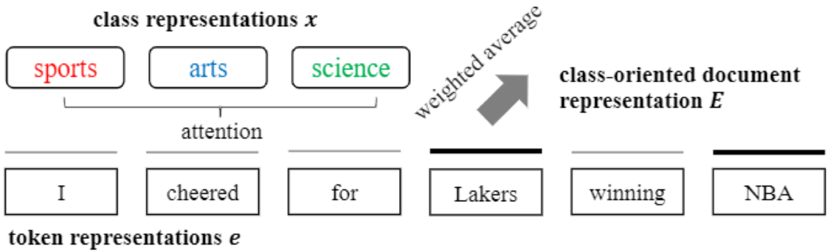


Figure 4: Overview of Our Document Rep. Estimation.

M1: Class-oriented Document Representation

- Class Representation Estimation
  - Weighted average representation based on a ranked list of keywords
  - Incrementally add new keywords to list by ranking similarities of out-of-list words
    - Top-ranked keywords are expected to have more similar **static representations** to the class representation
  - Stop condition
    - New class rep. **Changed** the current list OR reach max T
- Document Representation Estimation
  - 4 ways to compute attention weight
    - 2 token rep. + 2 attention mechanisms
  - A unified list of geometric mean of the 4 ranks
    - Assign a weight of 1/r to a token ranked at r-th position

Algorithm 1: Class-Oriented Document Representation Estimation

Input: n documents D_i, k class names c_j, max number of iterations T, and attention mechanism set $\mathcal{M}$

Output: Document representations $\mathbf{E}_i$.

Compute $\mathbf{t}_{i,j}$ (contextualized token rep.)

Compute $\mathbf{s}_w$ for all words (Eq. 1)

`// class rep. estimation`

for $j = 1 \dots k$ **do**

  $\mathcal{K}_j \leftarrow \langle\, c_j \,\rangle$

  for $i = 2 \dots T$ **do**

    Compute $\mathbf{x}_j$ based on $\mathcal{K}_j$ (Eq. 2)

    $w = \arg\max_{w \notin \mathcal{K}_j} sim(\mathbf{s}_w, \mathbf{x}_j)$

    Compute $\mathbf{x}'_j$ based on $\mathcal{K}_j \oplus \langle w \rangle$

    `// consistency check`

    if $\mathbf{x}'_j$ *changes the words in* $\mathcal{K}_j$ **then**

      break

    else

      $\mathcal{K}_j \leftarrow \mathcal{K}_j \oplus \langle w \rangle$

`// document rep. estimation`

for $i = 1 \dots n$ **do**

  for *attention mechanism* $m \in \mathcal{M}$ **do**

    Rank $D_{i,j}$ according to m

    $r_{m,j} \leftarrow$ the rank of $D_{i,j}$

  Rank $D_{i,j}$ according to $\prod_m r_{m,j}$

  $r_j \leftarrow$ the final rank $a_j \leftarrow 1/r_j$

  $\mathbf{E}_i \leftarrow \frac{\sum_j a_j \cdot \mathbf{t}_{i,j}}{\sum_j a_j}$

M2: Document-Class Alignment & M3: Text Classifier Train

- M2: Document-Class Alignment
  - each document is assigned to its nearest class $L_i = \arg\max_c cos(\mathbf{E}_i, \mathbf{x}_c)$
  - Gaussian Mixture Model (GMM) clustering

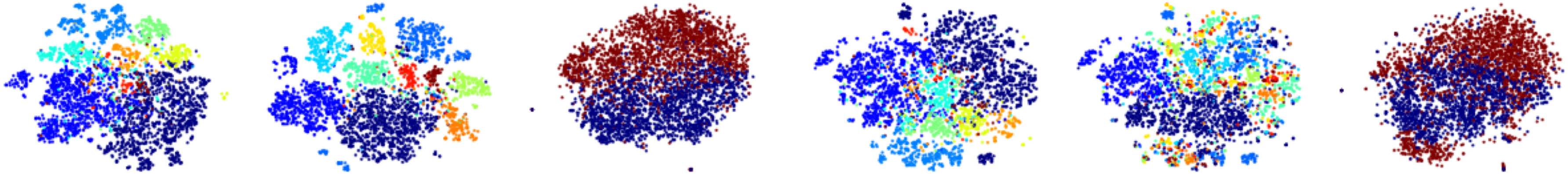
- M3: Text Classifier Training
  - select most confident samples to train a text classifier (BERT) using the pseudo labels

Experiments

Table 2: Evaluations of Compared Methods and X-Class. Both micro-/macro-F_1 scores are reported. WeSTClass and ConWea consume at least 3 seed words per class. Supervised provides a kind of upper bound. We are not able to re-run WeSTClass and ConWea on DBpedia due to the large size.

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	81.1/ N/A
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	84.8/84.65	**81.36/80.6**	**96.67/92.98**	80.6/69.92	**90.5/89.81**	**88.36/88.32**	**91.33/91.14**
X-Class-Rep	77.92/77.03	75.14/73.24	92.13/83.94	77.85/65.38	86.7/87.36	77.87/77.05	74.06/71.75
X-Class-Align	83.1/83.05	79.28/78.62	96.34/92.08	79.64/67.85	88.58/88.02	87.16/87.1	87.37/87.28

Experiments



(a) Our Class-Oriented Document Representations (b) Simple Average of BERT Representations

Figure 5: T-SNE Visualizations of Representations. From left to right: NYT-Topics, NYT-Locations, Yelp.