基于词对主题模型和词嵌入的无数据短文本分类
Dataless Short Text Classification Based on Biterm Topic Model and Word Embeddings

Yi Yang, Hongan Wang, Jiaqi Zhu*, Yunkun Wu, Kailong Jiang, Wenli Guo and Wandong Shi

Published in *the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020), Main Track*, pp. 3969--3975, 2020.

Contact: Jiaqi Zhu, zhujq@ios.ac.cn, 13683257241

Motivation

Problem: How to classify short texts on social media in a dataless manner?

- A huge number of short texts are generated on the Internet and it is crucial to acquire important and interesting information from them.
- Labeling documents is very expensive and time-consuming for domain experts, so labeled data are hard to obtain.
- Only few seed words for each class are available.

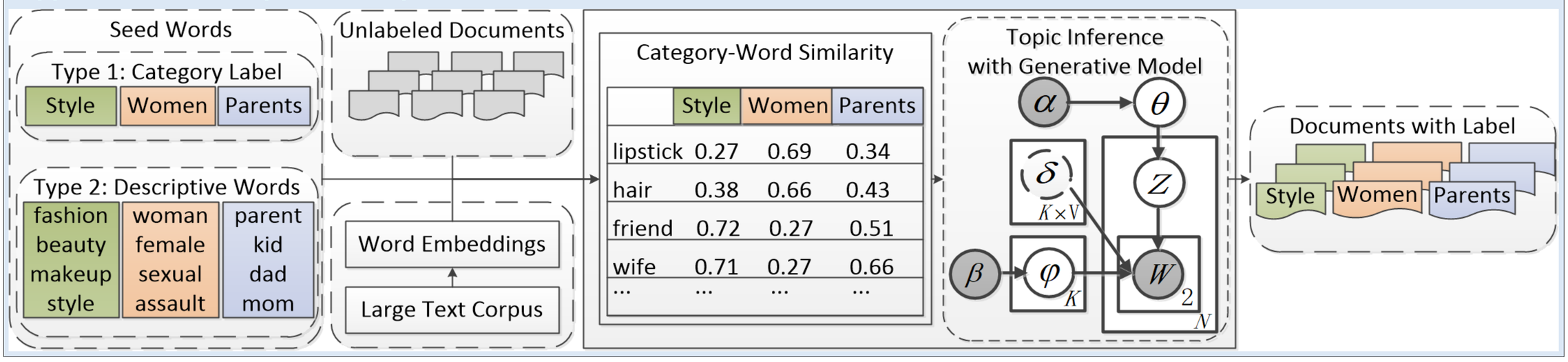
Challenges: The short texts of documents on the Internet are intractable to classify due to the following characteristics.

- They are extremely sparse and only limited word co-occurrence information can be utilized.
- The texts of documents on the Internet is irregular, so keyword-based methods do not work.
- They evolve rapidly so public knowledge bases are not always suitable.

Contribution

- An approach (framework) is presented to solve the task of dataless short text classification with seed words, by combining word co-occurrence information and category-word similarity based on word embeddings.
- Two models SeedBTM and SeedTBTM are respectively proposed through applying our approach on short-text topic models BTM and Twitter-BTM. That indicates our approach is applicable on different topic models by effectively integrating meta information of short texts, such as users (authors).
- Informative experiments are conducted on five real world datasets to show that our models significantly outperform the state-of-the-art baseline methods, especially when the categories are overlapping and interrelated.

Approach Overview (Framework)



Model SeedBTM

Estimating Category-Word Similarity: We assume that word similarity scores based on word embeddings could represent the semantic correlations between words. For a category seed word *s* and a corpus word *w*, we get the word vectors v_s and v_w through word embeddings, and calculate the word similarity sim(*s*, *w*). Then, we compute the category-word similarity $\delta_{z,w}$ as the maximum similarity between each seed word and *w*.

$$\mathrm{sim}(s,w) = \max(\cos(v_s, v_w), \epsilon) \qquad \delta_{z,w} = \max_i(\mathrm{sim}(s_{z,i}, w))$$

Seeded Biterm Topic Model: With the similarity scores, we extend the base model BTM and proposed SeedBTM. Given a sampled topic *z*, we think the generation of a word *w* in SeedBTM is influenced by both prior category word similarity δ_z and the topic-word distribution φ_z. The Gibbs Sampling becomes:

$$P(z_b|\mathbf{z}_{\neg b}, B, \boldsymbol{\delta}) \propto \delta_{z,w_i} \cdot \delta_{z,w_j} \cdot (n_z + \alpha) \cdot \frac{(n_{w_i|z}+\beta)(n_{w_j|z}+\beta)}{\sum_w (n_{w|z}+1+M\beta)(\sum_w n_{w|z}+M\beta)}$$

SeedBTM treats the expectation of the topic proportions of biterms in a document as the topic proportions of the document *P*(*z*|*d*). *P*(*b*|*d*) is estimated based on the relative frequency of *b* in *d* and *P*(*z*|*b*) can be calculated via Gibbs sampling. Finally, for a document *d*, the category label z_d can be predicted as the topic with the highest probability

$$P(z|d) = \sum_b P(z|b)P(b|d) \qquad z_d = \arg\max_i P(z_i|d)$$

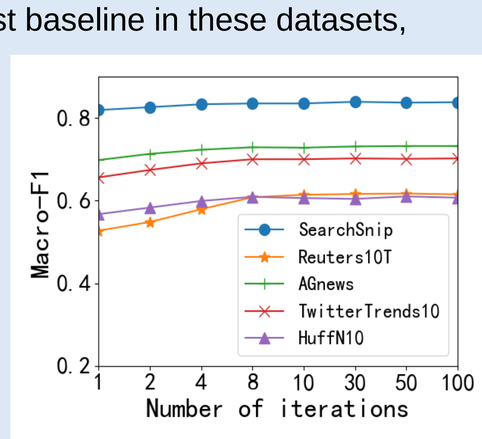
Experiments

We show the model effectiveness on five real short-text datasets, against five baseline methods with Macro-F1 value.

Dataset	SeedBTM		WeSTClass		WeSTClass*		STM		SSCF	DescLDA
	S^L	S^D	S^L	S^D	S^L	S^D	S^L	S^D	S^D	S^D
SearchSnip	67.6	**83.4**	13.7	15.0	65.7	80.3	64.9	80.0	80.2	70.0
Reuters10T	39.2	**61.5**	22.5	16.9	33.4	38.1	37.6	54.9	57.9	59.1
AGnews	61.8	73.1	76.6	**76.7**	75.8	74.7	66.3	69.5	71.4	58.9
TwitterTrends10	43.0	**71.5**	9.3	68.5	40.6	49.8	9.1	62.9	67.8	57.3
HuffN10	54.4	**60.8**	37.1	58.0	39.4	38.3	10.1	53.9	53.3	37.6
HuffN4-SWPB	70.8	**71.5**	32.6	34.9	65.2	70.3	26.8	14.1	60.6	46.8
HuffN4-ECMC	51.7	**61.9**	30.0	35.4	55.6	54.9	41.3	53.2	54.5	39.0

We can observe SeedBTM performs significantly better than baseline models on 6 of 7 datasets except AGnews. Specifically, the Macro-F1 value increases about 1.2~6.3 percent compared to the best baseline in these datasets, especially for difficult and confusing tasks.

We vary the iteration number in the range of [1,100], and the results show that SeedBTM can achieve good classification performances when the number is 10. The fast convergence should give credit to the regulating effect of the prior knowledge from word embeddings.



Applications

Public Opinion Analysis: For emerging hot events on social media such as Twitter and Weibo, classify the opinions of related posts without labeled data and discover the target content the government or companies concern.

Case Investigation: From chat messages, discover suspicious users and identify the criminal behaviors with some prior knowledge.