Distilling Knowledge on Text Graph for Social Media Attribute Inference

Quan Li
The Pennsylvania State University
State College, PA, USA
qbl5082@psu.edu

Xiaoting Li
Visa Research
Palo Alto, CA, USA
xiaotili@visa.com

Lingwei Chen*
Wright State University
Dayton, OH, USA
lingwei.chen@wright.edu

Dinghao Wu*
The Pennsylvania State University
State College, PA, USA
dinghao@psu.edu

ABSTRACT

The popularization of social media generates a large amount of user-oriented data, where text data especially attracts researchers and speculators to infer user attributes (e.g., age, gender) for fulfilling their intents. Generally, this line of work casts attribute inference as a text classification problem, and starts to leverage graph neural networks for higher-level text representations. However, these text graphs are constructed on words, suffering from high memory consumption and ineffectiveness on few labeled texts. To address this challenge, we design a text-graph-based few-shot learning model for social media attribute inferences. Our model builds a text graph with texts as nodes and edges learned from current text representations via manifold learning and message passing. To further use unlabeled texts to improve few-shot performance, a knowledge distillation is devised to optimize the problem. This offers a trade-off between expressiveness and complexity. Experiments on social media datasets demonstrate the state-of-the-art performance of our model on attribute inferences with considerably fewer labeled texts.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;
• **Information systems** → *Document representation.*

KEYWORDS

social media, attribute inference, graph neural networks, few-shot learning, knowledge distillation

ACM Reference Format:
Quan Li, Xiaoting Li, Lingwei Chen, and Dinghao Wu. 2022. Distilling Knowledge on Text Graph for Social Media Attribute Inference. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531968

*Corresponding authors. Xiaoting Li's work was done at PSU.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531968

1 INTRODUCTION

Social media allows billions of people to conveniently interact with others, and express personal ideas for social engagements [13]. Such a vibrant environment generates a large amount of user-oriented data. Among them, text data undoubtedly maintains the most basic social media content yet the most important user information, which, more importantly, often embeds intrinsic user attributes, such as age, gender, location, and political view. It has thus attracted different parties to disclose user attributes from their text data and study individual behaviors. For example, researchers leverage user posts for pandemic risk assessment [28], social surveillance [21], and social reaction analysis [17]. Speculators infer users' attributes, especially sensitive and private ones, to deliberately fulfill the economic or political goals, such as promoting advertisements, tracking users, and influencing opinions and votes [13, 16, 30].

While the intents of user attribute inferences on social media vary, the methods used to infer such information from text data are consistent. Information retrieval techniques can be considered as the first attempts, which suggest the personal attributes by searching for words and learning their relevance to attributes [22, 25]. These approaches are immensely limited as users' attributes are usually sparse on words. This naturally leads attribute inferences to text classification problems using either machine learning models over feature engineering (e.g., TF-IDF [6], and LSI [31]), or more advanced natural language processing (NLP) for higher-level text representations (e.g., long short-term memory [10], and transformer [3]). Though with the promising performance, NLP models provide the successful principles to solve the issues raised in feature engineering [9, 14], their inputs are inherently self-contained, and struggle to leverage structural interactions with other texts.

Graph neural networks (GNNs) have recently emerged as one of the most powerful techniques for graph understanding and mining [1, 15, 29]. Therefore, a surge of effective research works utilize GNNs to reveal user attributes on social media [2, 20] or simply perform text classification [5, 12, 18, 26, 27, 32]. For example, Yao et al. customized a GNN model to analyze texts by converting the corpus to a heterogeneous graph with words/documents as nodes and word co-occurrence as edges, which requires high memory consumption yet delivers low expressive power. Huang et al. [12] used global shared word representations to reduce the computational cost, and Ding et al. [5] defined hyperedges to capture high-order interactions between words. Similar refinements can be also found

in this line of work [18, 26, 32]. However, different from siamese or matching networks, these GNN-based models construct text graphs simply using local/global word relations, which may improve text representations to some extent, but barely take effect on application scenarios when labeled texts are few.

Due to privacy concerns, most social media sites/apps limit the access to some personal information; thus, user attribute labels, especially for those private attributes, may only be available on few texts. In other words, when we reduce user attribute inference on social media to a text classification problem, we face the challenge that our model needs to have the ability to learn from few text examples. To address this challenge, we propose a text-graph-based few-shot learning model that implements attribute inferences on social media text data. Given a text corpus (e.g., tweets, blogs) and an attribute to infer, our model starts by mapping each text to an initial representation; based on these representations, a text graph is constructed where each node is associated with one text, and edges are learned from the current text representations (either initial ones concatenated with one-hot encoding of attribute label at the input, or hidden representations) via manifold learning. This differs from those static text graphs built upon massive words, and offers a better trade-off between expressive power and complexity. The task-driven message passing is then conducted directly between labeled and unlabeled text pairs, which copes better with labeled data scarcity issue as well. To further leverage unlabeled texts to improve few-shot performance, a knowledge distillation operation is devised to optimize our graph-based model for attribute inferences.

2 PROBLEM STATEMENT

The text data posted by users on social media brings attribute inferences to the forefront. In this paper, we put aside the intents of user attribute inferences, and focus on the investigation of how we can generalize the attribute inference model into a more challenging setting with sparse information on words and few labels on texts, which is more realistic for social media environment.

Without loss of generality, we represent social media text data as $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{m} \cup \{x_i\}_{i=1}^{n}$ consisting of $m+n$ sample texts, where m is the number of the labeled texts and n is the number of unlabeled texts. Unlike existing works [5, 12, 18, 26, 27, 32] that use sufficient labeled texts for model training, we practically consider only few of the texts collected from social media have attribute labels. As such, among $\mathcal{X}$, m is much smaller than n. Each text x with label is annotated with a ground truth $y \in \mathcal{Y}$ for a specific attribute. Taking location attribute (main four U.S. regions) as an example: $\mathcal{Y}$ can be specified as $\mathcal{Y} = \{0{:}\text{Northeast}, 1{:}\text{Midwest}, 2{:}\text{South}, 3{:}\text{West}\}$. We deal with discrete text data by mapping each text x into a k-dimensional feature vector $\mathbf{x} = \phi(x)$ where ϕ is a feature representation function $\phi : \mathcal{X} \rightarrow \mathbf{X} \subseteq \mathbb{R}^{(m+n)\times k}$. Resting on text representations, we aim to learn a text classification model $f : \mathbf{X} \rightarrow \mathbf{Y}$ which can take advantage of few labeled texts and large unlabeled texts to perform our attribute inference task. Thus, the attribute label of a given text $\mathbf{x}$ can be inferred using the following formula:

$$y^* = \underset{y \in \mathcal{Y}}{\text{argmax}}\ f_y(\mathbf{x}) \tag{1}$$

where $f_y(\mathbf{x})$ is the confidence score of predicting text $\mathbf{x}$ as attribute label y using the text classification model f.

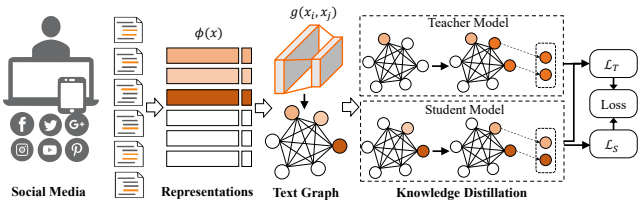


Figure 1: The overview of our proposed model.

3 PROPOSED MODEL

In this section, we present our proposed attribute inference model, the overview of which is illustrated in Figure 1.

3.1 Text Representations

To proceed with graph construction in text granularity, the first step is to initialize each text x into k-dimensional feature vector $\mathbf{x}$ with good expressive quality. As BERT [4] provides a context-aware word embedding space and boosts state-of-the-art performance on downstream NLP tasks, we use it to formulate our text representations. More specifically, we leverage SBERT [23] with fine-tuned semantic relations that adds a pooling operation to the output of BERT to derive a fixed-size embedding $\phi_1(x)$ for the input text.

In addition, to facilitate label information propagation among labeled and unlabeled nodes via task-driven message passing, we further map the label of each text into a one-hot encoding $\phi_2(x)$, and concatenate it with SBERT embedding $\phi_1(x)$ as the final text representation at the input of text graph construction:

$$\mathbf{x} = \phi(x) = [\phi_1(x); \phi_2(x)], \mathbf{x} \in \mathbb{R}^k \tag{2}$$

Let $\phi_1(x) \in \mathbb{R}^{k_1}$ and $\phi_2(x) \in \mathbb{R}^{k_2}$ ($k_2 = |\mathcal{Y}|$); then the dimension of our text representation is $k = k_1 + k_2$. For those texts without labels, we replace the one-hot encoding with the uniform distribution over the k_2-simplex, and accordingly get $\phi_2(x) = \mathbf{1}_{k_2}/k_2$. This formulation is helpful for our text-graph-based few-shot learning model to infer the potential attribute similarity between texts.

3.2 Text Graph Construction and Refinement

The goal of our attribute inference model is to learn from few labeled texts and propagate attribute label information from the labeled texts to the unlabeled ones through their relatedness. Recent researches have demonstrated that message passing with graph-based neural networks can effectively work on such label propagation [7, 8, 19]. In this paper, we extend this paradigm to cast attribute inference using task-driven message passing and infer a text's attribute label over the text graph. There are three reasons behind our graph construction over texts rather than word co-occurrences: (1) label propagation can be easily performed as a posterior inference between labeled and unlabeled text pairs, enabling our model to better address labeled data scarcity issue; (2) update on text representations can be immediately used to refine graph structure and improve its expressive power; and (3) the number of graph nodes can be significantly reduced to save the computational cost.

Graph construction via manifold learning. Given a social media text corpus $\mathcal{X}$, we construct a fully-connected graph $G_{\mathcal{X}} = (V, E)$

to associate $\mathcal{X}$, where V denotes the set of texts (both labeled and unlabeled), and $E = V \times V$ denotes the set of edges that connect text pairs. Manifold learning [19] is non-linear dimensionality reduction process which reveals the low-dimensional manifold embedded in the high-dimensional space, which can be feasibly exploited to build up the intrinsic neighborhood among text representations. Thus, we initialize each edge e_{ij} between v_i and v_j in $G_{\mathcal{X}}$ by a layerwise non-linear combination of distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ as

$$e_{ij} = g_{\mathbf{\Theta}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma(\cdots\sigma(|\mathbf{x}_i - \mathbf{x}_j|\mathbf{\Theta}^{(0)})\cdots\mathbf{\Theta}^{(l-1)})\mathbf{\Theta}^{(l)} \quad (3)$$

where $\sigma(\cdot)$ is a non-linear activation function (e.g., ReLU), and $\mathbf{\Theta}$ is learnable weight matrix for each layer. As the constructed structure behaves differently regarding different text representations, the learned edges do not specify a fixed text graph, suggesting the graph can be refined when the neighborhood information is updated.

Graph refinement via message passing. To refine text graph, we apply iterative message passing through neighborhood structure using a graph convolutional network (GCN) [15] to propagate text features and labels along the labeled and unlabeled nodes, and enhance text representations. Specifically, we build the adjacency matrix $A^{(h)}$ by normalizing edge matrix using a softmax at each row, where each e_{ij} is computed on the current text representations:

$$A_{i,j}^{(h)} = \text{softmax}(g_{\mathbf{\Theta}}(\mathbf{x}_i^{(h)}, \mathbf{x}_j^{(h)})) \quad (4)$$

Each message passing iteration can be formalized as multi-layer neighborhood information aggregation, which receives an input $\mathbf{X}^{(h)}$ and produces $\mathbf{X}^{(h+1)}$ as follows:

$$\mathbf{X}^{(h+1)} = \sigma(\widetilde{A}^{(h)}\mathbf{X}^{(h)}\mathbf{W}^{(h)}) \quad (5)$$

where at layer h, $\mathbf{W}$ is weight matrix, $\widetilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}$, $\hat{\mathbf{A}} = \mathbf{A}+\mathbf{I}$, and $\mathbf{D}$ is the diagonal degree matrix defined on $\hat{\mathbf{A}}$, i.e., $\mathbf{D}_{ii} = \sum_{j=1}^{n}\hat{\mathbf{A}}_{ij}$. The text graph $G_{\mathcal{X}}$ is reconstructed after every message passing iteration using the refined text representations, giving our attribute inference model more expressive power.

3.3 Knowledge Distillation

Our constructed and refined text graph can be used directly to perform posterior inference and propagate the attribute labels from few labeled texts to the target texts, and deliver promising attribute inference performance. To further leverage unlabeled texts to improve few-shot learning performance, we devise a knowledge distillation operation over text graph to enrich the optimization problem for attribute inferences. The knowledge distillation technique was designed for model compression, which was then generalized to transfer soft knowledge along teacher neural network to student neural network in a simple way [11].

As such, we divide the labeled texts into two categories: teacher texts $\mathcal{X}_T$ and student texts $\mathcal{X}_S$. A teacher model is first trained on $\mathcal{X}_T$, which is then used to perform attribute inference on $\mathcal{X}_S$. The knowledge distilled by the teacher model can be defined as the inference probability of attribute label for text $\mathbf{x}_S$ in $\mathcal{X}_S$:

$$p(\mathbf{x}_S|\mathcal{X}_T) = \frac{\exp\left(f_y(\mathbf{x}_S/\tau)\right)}{\sum_{y\in\mathcal{Y}}\exp\left(f_y(\mathbf{x}_S/\tau)\right)} \quad (6)$$

where τ is distillation temperature, $\mathbf{x}_S$ is the representation of the text from $\mathcal{X}_S$, and $f_y(\mathbf{x}_S/\tau)$ is the confidence score of predicting text $\mathbf{x}_S$ as attribute label y after iterative message passing over text graph. Similarly, a student model is trained on $\mathcal{X}_S$, which generates inference probability of attribute label for text $\mathbf{x}_S$ as $p(\mathbf{x}_S|\mathcal{X}_S)$. Accordingly, the student model may learn the distilled knowledge from the teacher model by optimizing the cross-entropy loss function:

$$\mathcal{L}_T = -\frac{1}{|\mathcal{X}_S|}\sum_{\mathbf{x}_S\in\mathcal{X}_S} p(\mathbf{x}_S|\mathcal{X}_T)\log p(\mathbf{x}_S|\mathcal{X}_S) \quad (7)$$

Here, $p(\mathbf{x}_S|\mathcal{X}_T)$ is predicted by teacher model on student texts, which are unlabeled data to the teacher. It can be considered soft attribute label with the same distribution as $p(\mathbf{x}_S|\mathcal{X}_S)$ from student model. This formulation significantly advances the model to learn from unlabeled texts.

Table 1: Comparing statistics of the two datasets

Dataset	Attribute	#Post	#Class	#Vocabulary
Twitter	Gender	13,926	2	21k
Blog	Gender, Age	25,176	2	30k

3.4 Loss Generation for Transductive Training

The student model itself computes training loss between predictions and ground truth (hard attribute label), which is defined as:

$$\mathcal{L}_S = -\frac{1}{|\mathcal{X}_S|}\sum_{\mathbf{x}_S\in\mathcal{X}_S} y\log p(\mathbf{x}_S|\mathcal{X}_S) \quad (8)$$

In this respect, the final objective loss function of our learning model can be formalized as:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_S + \lambda\mathcal{L}_T \quad (9)$$

where λ is a balance parameter to trade off $\mathcal{L}_S$ and $\mathcal{L}_T$. We train our text-graph-based few-shot learning model in a transductive (or semi-supervised) manner, where all texts (labeled and unlabeled) are accessible during training.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Setup

Datasets. We test our model on two real-world social media datasets: Twitter dataset[1] and Blog dataset [24]. The Twitter dataset contains 13,926 tweets for gender attribute inference, and the blog dataset contains 25,176 blogs for two attribute inference settings: gender and age. The statistics of the dataset are shown in the Table 1.

Baselines. In our study, we choose five state-of-the-art GNN-based models on texts classification tasks to be our baselines:

- TL-GNN [12]: It constructs graph by using global shared word representations and uses message passing for text classification.
- HyperGAT [5]: It defines hyperedge to connect all words and designs dual-attention method for text classification.
- TextGCN [27]: It builds graph with word nodes and document nodes, and uses GCN for text classification.
- TextING [32]: It builds individual graph for each text and utilizes gated GNN to learn embedding of word nodes.

[1] https://www.kaggle.com/crowdflower/twitter-user-gender-classification

Table 2: Comparisons of different graph-based baselines (2×15)

Inference	TL-GNN		HyperGAT		TextGCN		TextING		HGAT		Our model	
	ACC(%)	F1	ACC(%)	F1	ACC(%)	F1	ACC(%)	F1	ACC(%)	F1	ACC(%)	F1
Twitter-gender	50.49	0.3616	50.76	0.4755	49.36	0.4314	51.36	0.4898	52.41	0.3439	**61.16**	**0.5816**
Blog-gender	51.26	0.3636	51.88	0.3487	53.43	0.5302	52.76	0.5110	51.69	0.3407	**59.80**	**0.5655**
Blog-age	56.10	0.4282	47.43	0.4666	52.30	0.5209	58.28	0.5696	58.56	0.4770	**69.06**	**0.6845**

- HGAT [18]: It constructs heterogeneous information network (HIN) for texts and applies dual-level attention mechanism to learn importance of nodes and update the representations.

Parameter setting. We select 15 labeled instances per class as training data and randomly select 20% instances from the remaining as test data for each inference task. We set the knowledge distillation temperature $\tau = 3$ and the balance parameter $\lambda = 0.3$ for the training loss. We also evaluate the impacts of different training sizes and distillation temperatures in Section 4.2.

4.2 Evaluation of Our Model

Effectiveness. In this section, we evaluate the effectiveness of our model over three inference settings under different parameters. In particular, we test the inference accuracy of our model with training size $m \in \{2 \times 1, 2 \times 5, 2 \times 10, 2 \times 15, 2 \times 20\}$ respectively, while the knowledge distillation temperatures $\tau \in \{2, 3, 5, 7, 10\}$ when $m = 2 \times 15$. The experimental results are shown in Figure 2. As we can see, though different parameters contribute to different test results, which will be discussed later, our model achieves the state-of-the-art results of inferring attributes on social media texts when only few labeled texts are available. When "1-shot" (2×1) is set, the inference accuracy is 55.40%, 53.19%, and 60.70% for Twitter-gender, Blog-gender, and Blog-age respectively, which are much better than most of the baselines trained on (2×15); averagely, their inference accuracies are 59.63%, 57.93%, and 66.51%.

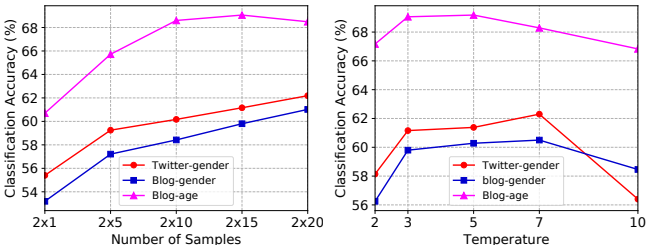
Impact of training size and distillation temperature. As illustrated in Figure 2(a), when "higher-shot" is applied in training, the performance of our model generally continues to improve, but the improvements of few-shot learning in $[2 \times 10, 2 \times 20]$ are less significant (or more stable) than that of $[2 \times 1, 2 \times 10]$. With the training size increases, the advantage of our few-shot model narrows since more labeled texts are used and the inference performance is closer to the upper bound. As for the distillation temperature, Figure 2(b) indicates that when we enlarge τ, the attribute inference accuracy first significantly increases, rises to a stable high level at $\tau \in [3, 7]$, and then drastically drops when τ changes from 7 to 10. It's not difficult to understand this trend: when τ is relatively small, the soft attribute label probabilities distilled from teacher model are informative and helpful to facilitate optimizing student model; when τ is large, the distilled knowledge is ambiguous, which may in turn smooth the student model's inference ability.

4.3 Comparisons with Baselines

In this section, we compare our model with five baselines that work on text classification over graph structure, including TL-GNN [12], HyperGAT [5], TextGCN [27], TextING [32], and HGAT [18]. The comparative results are illustrated in Table 2 with $m = 2 \times 15$. We can observe that among baselines, HGAT, TextGCN, and TextING slightly take the lead in Twitter-gender, Blog-gender, and Blog-age respectively with respect to accuracy and F1-score. Obviously, our model completely outperforms baselines with a large margin in lower-shot (i.e., the improvement margin of accuracy is $(6.3, 21.7)$%, and the improvement margin of F1-score is $(0.04, 0.26)$). Another observation from Table 2 and Figure 2(a) is that our model with only 1-shot is either outperforming or comparable to baselines with 15-shot. This confirms that (1) graphs built upon word co-occurrence can improve text representations, but hardly learn from few labeled texts; (2) the text-level graph with neighborhood refinement contributes better to few-shot learning than the word-level graph, and (3) our model offers a better trade-off between expressive power and complexity in terms of node number, and thus provides a better solution for social media attribute inferences.

Table 3: Evaluation on model components (accuracy %)

Method	Twitter-gender	Blog-gender	Blog-age
SBERT	51.20	50.93	54.59
SBERT+Graph	58.04	55.35	64.64
SBERT+Graph+KD	**61.16**	**59.80**	**69.06**



(a) Sizes of training samples **(b) Distillation temperatures**

Figure 2: Evaluation on different model parameters.

4.4 Ablation Study

In this section, we conduct the ablation study to further investigate how different components contribute to the performance of our model. Our model proceeds with text representations, graph construction and refinement, and knowledge distillation. We gradually add these components one by one and formulate three attribute inference models: (1) SBERT: directly feed SBERT representations to fully-connected and softmax layers for text classification; (2) SBERT+Graph: construct and refine a text graph using SBERT representations and perform posterior inference by transductive learning; (3) SBERT+Graph+KD: the complete design of our model. The results are reported in table 3.

From the results, we can see that SBERT representations provide good expressive quality for texts, which deliver comparable performances to some baselines over world-level graphs. The constructed and refined text graph has the greatest contribution to our model, which significantly improves the inference results by $(4.0, 11.0)\%$ of accuracy. Knowledge distillation is able to further advance the state-of-the-art performance to a higher level, which implies that this operation yields an additional advantage for few-shot learning. These observations reaffirm the effectiveness of our design to infer user attributes on social media when labeled texts are few.

5 CONCLUSION

In this paper, we generalize attribute inferences over social media text data into the more challenging yet more realistic setting with sparse information on words and few labels on texts, and design a text-graph few-shot learning model to address this challenge. To evaluate the performance of our model, we conduct extensive experiments over two real-world social media datasets and three inference settings. The state-of-the-art results validate its attribute inference effectiveness, its superiority to baselines, and its significance and helpfulness to cope with few-shot learning in practice.

ACKNOWLEDGMENTS

The work was supported in part by a seed grant from the Penn State Center for Security Research and Education (CSRE).

REFERENCES

[1] Lingwei Chen, Xiaoting Li, and Dinghao Wu. 2020. Enhancing robustness of graph convolutional networks via dropping graph connections. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 412–428.

[2] Weijian Chen, Yulong Gu, Zhaochun Ren, Xiangnan He, Hongtao Xie, Tong Guo, Dawei Yin, and Yongdong Zhang. 2019. Semi-supervised User Profiling with Heterogeneous Graph Attention Networks.. In *IJCAI*, Vol. 19. 2116–2122.

[3] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[5] Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. Be more with less: Hypergraph attention networks for inductive text classification. *arXiv preprint arXiv:2011.00387* (2020).

[6] George Forman. 2008. BNS feature scaling: an improved representation over tf-idf for svm text classification. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*. 263–270.

[7] Victor Garcia and Joan Bruna. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043* (2017).

[8] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*. 1263–1272.

[9] Neil Zhenqiang Gong and Bin Liu. 2018. Attribute inference attacks in online social networks. *ACM Transactions on Privacy and Security (TOPS)* 21, 1 (2018), 1–30.

[10] Alex Graves. 2012. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*. 37–45.

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[12] Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. *arXiv preprint arXiv:1910.02356* (2019).

[13] Jinyuan Jia and Neil Zhenqiang Gong. 2018. Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In *27th USENIX Security Symposium (USENIX Security 18)*. 513–529.

[14] Jinyuan Jia, Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. 2017. Attriinfer: Inferring user attributes in online social networks using markov random fields. In *Proceedings of the 26th International Conference on World Wide Web*. 1561–1569.

[15] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[16] Xiaoting Li, Lingwei Chen, and Dinghao Wu. 2021. Turning Attacks into Protection: Social Media Privacy Protection Using Adversarial Attacks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 208–216.

[17] Chung-Ying Lin. 2020. Social reaction toward the 2019 novel coronavirus (COVID-19). *Social Health and Behavior* 3, 1 (2020), 1.

[18] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4821–4830.

[19] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. 2018. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002* (2018).

[20] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. 2020. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *CVPR*. 14424–14432.

[21] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science* (2020).

[22] Jay M Ponte and W Bruce Croft. 2017. A language modeling approach to information retrieval. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 202–208.

[23] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[24] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging.. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 6. 199–205.

[25] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.

[26] Yaqing Wang, Song Wang, Quanming Yao, and Dejing Dou. 2021. Hierarchical Heterogeneous Graph Representation Learning for Short Text Classification. *arXiv preprint arXiv:2111.00180* (2021).

[27] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7370–7377.

[28] Yanfang Ye, Shifu Hou, Yujie Fan, Yiyue Qian, Yiming Zhang, Shiyu Sun, Qian Peng, and Kenneth Laparo. 2020. α-Satellite: An AI-driven System and Benchmark Datasets for Hierarchical Community-level Risk Assessment to Help Combat COVID-19. *arXiv preprint arXiv:2003.12232* (2020).

[29] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *SIGKDD*. 974–983.

[30] Sixie Yu, Yevgeniy Vorobeychik, and Scott Alfeld. 2018. Adversarial classification on social networks. In *International Conference on Autonomous Agents and MultiAgent Systems*. 211–219.

[31] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of TF-IDF, LSI and multi-words for text classification. *Expert Systems with Applications* 38, 3 (2011), 2758–2765.

[32] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every document owns its structure: Inductive text classification via graph neural networks. *arXiv preprint arXiv:2004.13826* (2020).