Sentiment Analysis Method combining Word Embedding with Personalized Text Structure

张兴华，沙灜
中国科学院信息工程研究所，北京
中国科学院大学网络空间安全学院，北京
xing_hua_zhang@126.com shaying@iie.ac.cn

问题

句子/篇章级的情感分析是一项非常重要的自然语言处理任务，目前针对该任务已存在大量出色的研究工作。

由于大部分的文本来源于**社交媒体**，社交网络中的个性化信息实则蕴含着**文本之间的情感关联**。我们的主要任务就是**获取社交文本之间的情感关系，以进一步提升情感分析的性能**。

我们的方法

句子/篇章级的情感分析目前已有的研究工作主要分为两类：①**基于文本内容的情感分析[1]**；②**显式融合社交网络个性化特征的情感分析[2]**。

基于对上述工作以及社交个性化信息（**话题依赖性**、**用户依赖性**和**社交一致性**，如图1所示）的探究，我们提出一种结合**文本语义特征**与**文本之间隐式情感关联特征**的方法——**融合个性化文本结构的词嵌入式情感分析**。

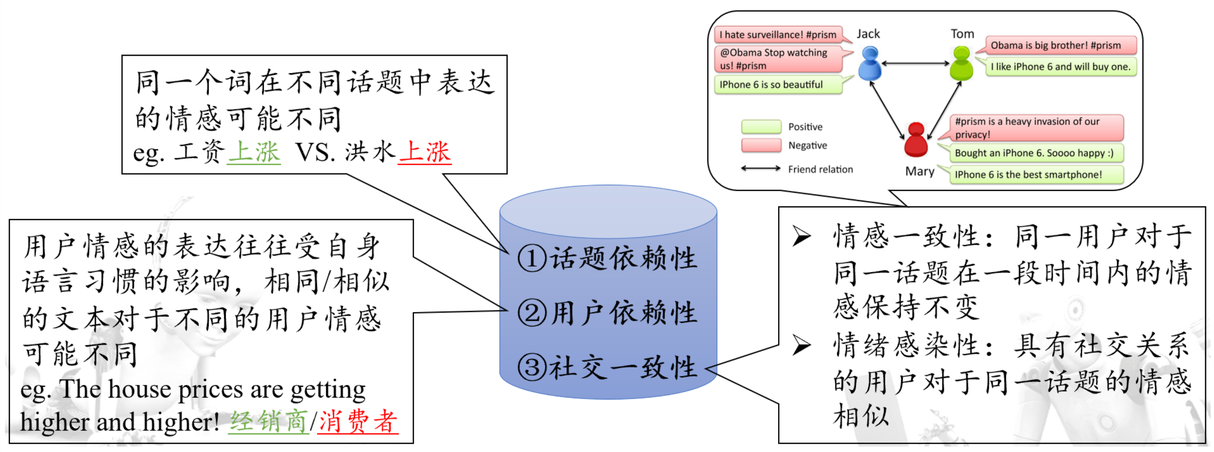


图1 社交个性化信息

融合个性化文本结构的词嵌入式情感分析方法

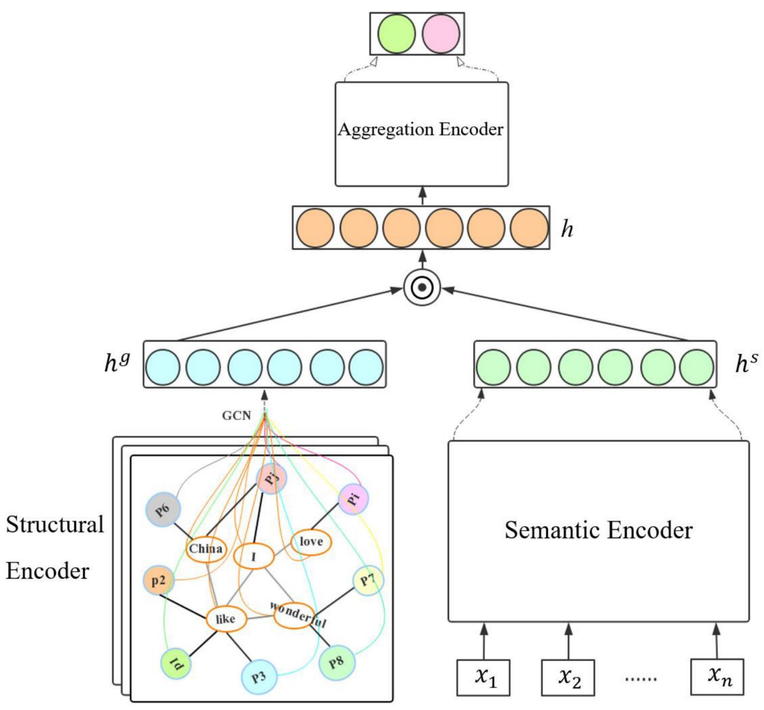


图2 融合个性化文本结构的词嵌入式情感分析框架

融合个性化文本结构的词嵌入式情感分析框架(Word Embedding Sentiment Analysis with Personalized Text Structure, **WESA-PTS**) 采用**图卷积网络**(Graph Convolutional Network, **GCN**)来获取文本之间的隐式情感关联，编码得到个性化文本结构特征，同时结合文本语义特征进行情感分析。

Composition:

- 文本语义编码器 (Semantic Encoder)
- 个性化文本结构编码器 (Structural Encoder)
- 特征集成编码器 (Aggregation Encoder)

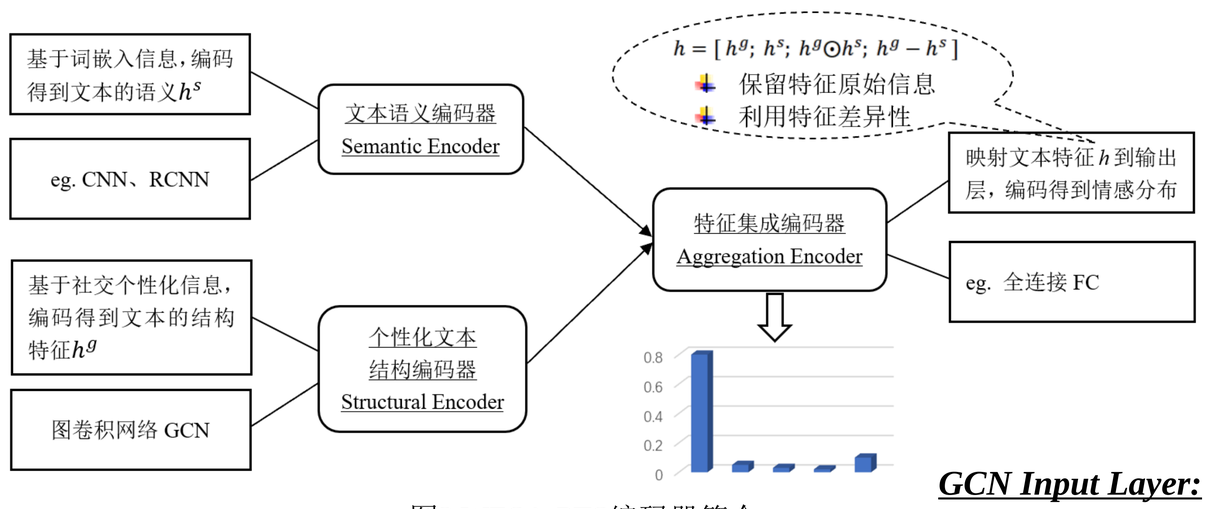
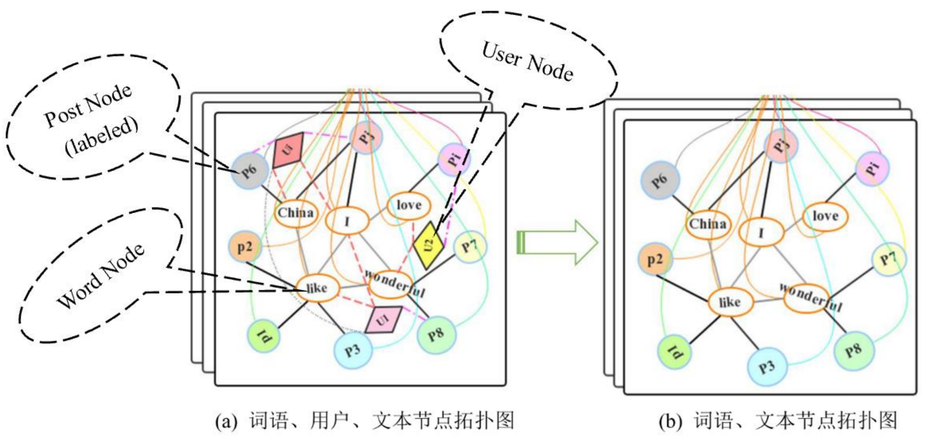


图3 WESA-PTS编码器简介



(a) 词语、用户、文本节点拓扑图 (b) 词语、文本节点拓扑图

图4 情感分析拓扑图*A*

GCN Input Layer:

- **社交个性化信息**主要体现在 **Word** 、 **User** Level
- 构建图4所示的(a)、(b)两种拓扑图，邻接矩阵计算如公式(1)
- 图4(a)显式地引入了User关联关系

$$A_{i,j} = \begin{cases} \text{TF-IDF}(i,j) & i,j \text{ 分别为文本（用户）节点和词语节点} \\ \text{PMI}(i,j) & i,j \text{ 均为词语节点，PMI}(i,j)>0\text{，否则等于 0} \\ \text{P}(i,j) & i,j \text{ 分别为用户节点和文本节点} \\ \text{F}(i,j) & i,j \text{ 均为用户节点} \\ 1 & i=j \\ 0 & \text{其它} \end{cases} \quad (1)$$

注：

1. TF-IDF(i, j): 计算节点i和j之间的词频-逆向文档频率。
2. PMI(i, j): 计算节点i和j之间点互信息量。
3. P(i, j): 文本发表的所属关系，值为0或1。
4. F(i, j): 社交好友关系，值为0或1 。

Experiments Setting:

Encoder / WESA-PTS	Semantic Encoder	Structural Encoder	Aggregation Encoder
G2CNN	CNN	GCN	FC
GRCNN	RCNN		

实验结果

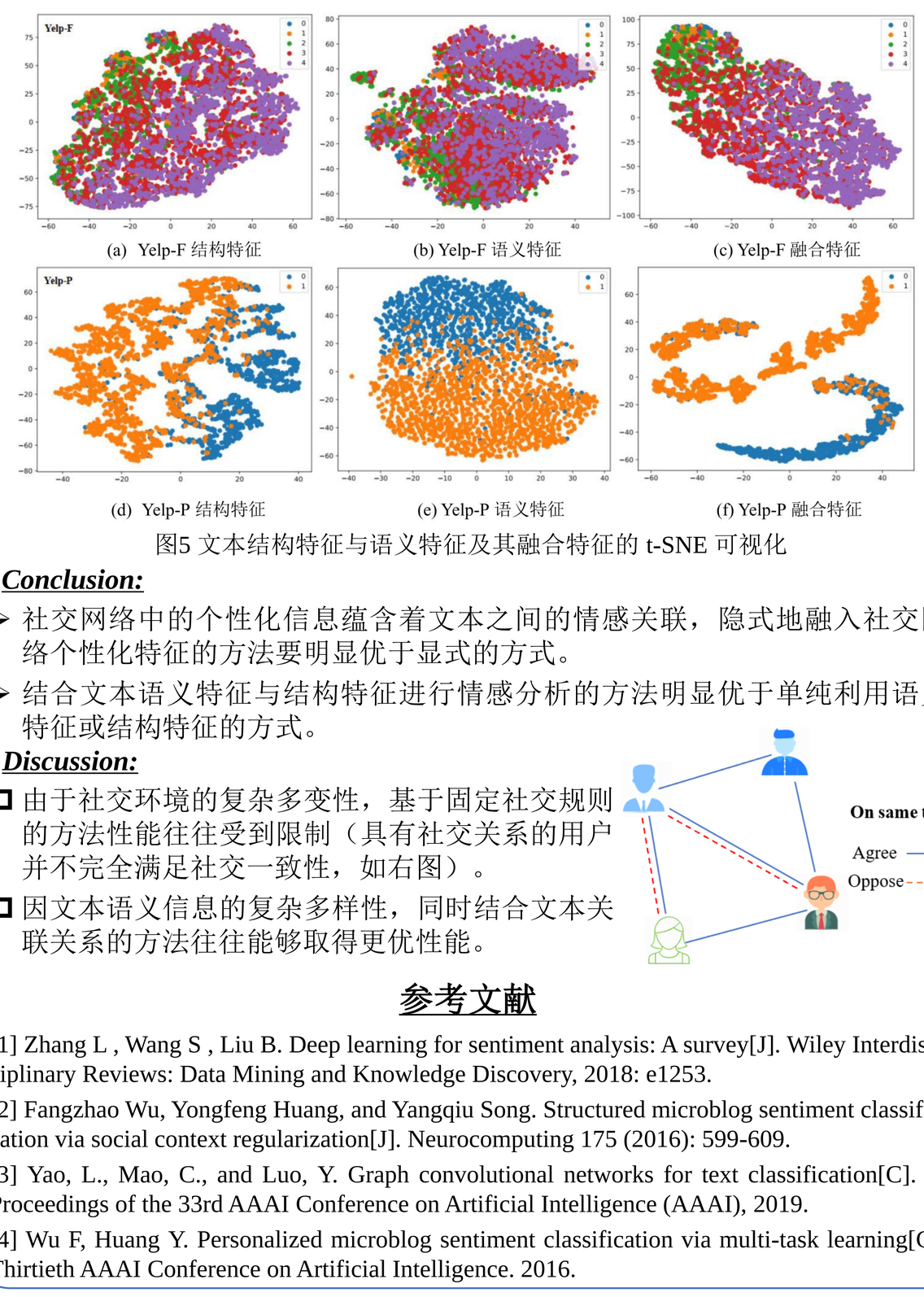
实验结果准确率Accuracy ± Standard Deviation (%)

实验方法		Yelp-P	Amazon-P	Yelp-F	Amazon-F
LSTM		89.47±0.23	86.46±0.26	48.43±0.45	62.92±0.14
LSTM+attention		87.68±0.17	85.54±0.13	52.32±0.33	64.90±0.17
Bi-LSTM		89.72±0.31	86.63±0.26	49.01±0.48	63.62±0.29
Bi-LSTM+attention		90.92±0.06	88.05±0.07	**56.22±0.17**	66.27±0.09
CNN_static		88.67±0.07	86.56±0.07	52.33±0.07	65.02±0.05
CNN_non_static		89.70±0.08	88.29±0.05	53.20±0.14	65.70±0.07
RCNN		90.91±0.08	88.79±0.11	55.70±0.29	66.36±0.07
GCN+user		88.03±0.00	88.18±0.00	53.92±0.02	66.01±0.00
GCN		90.28±0.00	88.61±0.00	54.54±0.00	66.26±0.00
WESA-PTS(Ours)	G2CNN	90.50±0.18	89.31±0.08	55.56±0.34	66.32±0.04
	GRCNN	**91.27±0.08**	**89.58±0.15**	55.88±0.21	**66.64±0.11**

数据集

- Yelp-P: 2分类
  Train 7216
  Test 2264
- Amazon-P: 2分类
  Train 7003
  Test 2335
- Yelp-F: 5分类
  Train 10000
  Test 5000
- Amazon-F: 5分类
  Train 16624
  Test 5542

注：训练集的10%作为验证集



(a) Yelp-F 结构特征 (b) Yelp-F 语义特征 (c) Yelp-F 融合特征

(d) Yelp-P 结构特征 (e) Yelp-P 语义特征 (f) Yelp-P 融合特征

图5 文本结构特征与语义特征及其融合特征的 t-SNE 可视化

Conclusion:

- 社交网络中的个性化信息蕴含着文本之间的情感关联，隐式地融入社交网络个性化特征的方法要明显优于显式的方式。
- 结合文本语义特征与结构特征进行情感分析的方法明显优于单纯利用语义特征或结构特征的方式。

Discussion:

- 由于社交环境的复杂多变性，基于固定社交规则的方法性能往往受到限制（具有社交关系的用户并不完全满足社交一致性，如右图）。
- 因文本语义信息的复杂多样性，同时结合文本关联关系的方法往往能够取得更优性能。

参考文献

[1] Zhang L , Wang S , Liu B. Deep learning for sentiment analysis: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018: e1253.

[2] Fangzhao Wu, Yongfeng Huang, and Yangqiu Song. Structured microblog sentiment classification via social context regularization[J]. Neurocomputing 175 (2016): 599-609.

[3] Yao, L., Mao, C., and Luo, Y. Graph convolutional networks for text classification[C]. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI), 2019.

[4] Wu F, Huang Y. Personalized microblog sentiment classification via multi-task learning[C]. Thirtieth AAAI Conference on Artificial Intelligence. 2016.