Computer Science and Application 计算机科学与应用, 2021, 11(4), 859-866
Published Online April 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.114088

Hans汉斯

基于User-BERT模型的微博谣言检测

缪　鑫

广东工业大学，广东 广州
Email: miaox@mail2.gdut.edu.cn

收稿日期：2021年3月14日；录用日期：2021年4月8日；发布日期：2021年4月15日

摘　要

随社交媒体的快速发展，微博已经成为人们获取信息的主要平台。它给人们生活带来便利的同时，也带来了谣言泛滥的问题。有越来越多研究投入到谣言检测中，从早期的特征工程方法到近期的深度学习方法。但是，目前的工作没有充分利用预训练语言模型与其它特征相结合。因此，文本推出User-BERT模型，使BERT模型能够充分利用文本和用户特征。它使用BERT模型对原文和评论文本进行编码，得到文本表示向量再与用户属性向量结合，最后由深度分类器对其进行解析并预测。在公开微博数据集上，User-BERT取得了当前最好的结果。

关键词

谣言检测，BERT，深度学习，自然语言处理

Microblog Rumor Detection Based on User-BERT

Xin Miao

Guangdong University of Technology, Guangzhou Guangdong
Email: miaox@mail2.gdut.edu.cn

Received: Mar. 14th, 2021; accepted: Apr. 8th, 2021; published: Apr. 15th, 2021

Abstract

With the rapid development of social media, Sina weibo has become the main platform for people to obtain information. While it brings convenience to people's lives, it also brings the problem of spreading rumors. More and more research is devoted to rumor detection, from early feature engineering methods to recent deep learning methods. However, the current work does not make full use of the pre-trained language model combined with other features. Therefore, this work in-

文章引用: 缪鑫. 基于 User-BERT 模型的微博谣言检测[J]. 计算机科学与应用, 2021, 11(4): 859-866.
DOI: 10.12677/csa.2021.114088

troduces the User-BERT model, which enables the BERT model to make full use of text and user characteristics. It uses the BERT model to encode text of source post and comments, obtains the text representation vector and combines it with the user attribute vector, and it finally is parsed by the deep classifier. On the public weibo dataset, User-BERT has achieved the best results currently.

Keywords

Rumor Detection, BERT, Deep Learning, NLP

1. 引言

随因特网和移动设备的普及，微博逐渐成为人们获取和表达信息的主要平台。它给人们生活带来便利的同时，也带来了谣言泛滥的问题。中文社会科学院发布的《新媒体蓝皮书》显示，中国有 59%的网络谣言来自新浪微博[1]。谣言传播恐惧和偏见，它可能会造成(个人、品牌、政府等)被诽谤[2]，甚至导致出现社会信任危机。所以，微博谣言检测是一项十分有必要的研究工作。

在研究早期，谣言检测主要是以特征工程为基础的机器学习方法，例如 Ma 等人[3]用到支持向量机(Support Vector Machine, SVM)。随算力快速提升，以深度学习为基础的方法成为机器学习主流，各类深度学习方法在谣言检测领域取得了长足的进步。Ma 等人[4]和 Wang 等人[5]用到循环神经网络(Recurrent Neural Network, RNN)，Yu 等人[6]用到卷积神经网络(Convolutional Neural Networks, CNN)，Bian 等人[7]用到图卷积网络(Graph Convolutional Network, GCN)，尹鹏博等人[1]和 Geng 等人[8]用到以深度学习为基础的集成学习(Ensemble Learning, EL)。

尽管当前以深度学习为基础的方法取得了不错的效果，但它们没有充分利用预训练语言模型或用户特征。以 BERT(Bidirectional Encoder Representations from Transformers) [9]为代表的预训练语言模型已经在诸多自然语言处理领域取得了领先的效果。其以 Tansformer [10]为基础，经过大规模文本预训练，学到了丰富的语言知识，可将知识迁移至任意下游任务。另外，用户特征已被证明能有效促进谣言检测[11]，通常越是权威的用户发布的消息越真实可靠，反之亦然。因此本文提出 User-BERT 模型，充分将文本特征和用户特征与 BERT 模型相结合，充分发挥 BERT 模型的性能优势。User-BERT 模型将 BERT 模型作为文本编码器对原文和评论文本进行编码，然后将输出的文本表示向量和用户特征向量进行拼接组成新的综合表示向量，最后将综合表示向量输入至全连接层即深度分类器中进行解析并预测结果。实验结果表明，在 Ma 等人[4]提出的公开微博数据集上，User-BERT 取得了当前最好的实验结果。

2. 相关工作

因社交平台的普及，谣言检测引起了更多重视，随机器学习的发展而进步。机器学习的早期主要以特征工程为基础的方法为主流，在谣言检测领域同样如此。Ma 等人[3]从内容、用户、传播三方面共选取了 27 条特征表示数据，用 SVM 模型作为分类器。其它以特征工程为基础的方法大致相同，特征选取是这类方法的关键，发掘有效的特征能给模型带来显著的提升。近些年因深度学习的快速发展，以深度学习为基础的方法已经成为机器学习的主流。深度学习因其自动提取高级特征和更强的拟合数据的能力

得到广泛应用，多种深度学习模型也被应用到谣言检测领域。Ma 等人[4]用词频-逆文本频率指数(Term Frequency-Inverse Document Frequency, TF-IDF)表示原文或评论，然后逐步将特征表示输入到 RNN 模型中，用最后的输出向量预测结果。Wang 等人[5]用 word2vec 对原文或评论作词嵌入，同时结合情感词典加入情感嵌入，分别输入到两个双层 RNN 模型中。Yu 等人[6]用 CNN 模型作为特征提取器构建检测模型。Bian 等人[7]则用到了近来流行的 GCN 模型[12]通过学习谣言的传播特征来判别真假。此外，还有联合多个深度学习模型为基础学习器的深度集成学习模型。尹鹏博等人[1]以 CNN 和 RNN 模型作为基分类器，选取随机森林作为元模型，合并基模型的输出为二次训练集，在元模型上进行二次训练。Geng 等人[8]用三种 RNN 模型作为基分类器，最后以投票的方式整合基分类器的预测结果。

3. 先导

3.1. 问题描述

谣言检测任务可被定义为：设数据集定义为 $D=\{d_1,d_2,\cdots,d_n\}$，$d_i\in D$ 表示一条数据元组，n 表示数据集的大小。且 $d_i=\{s_i,C_i,u_i,l_i\}$，其中 $s_i=\{w_1^i,w_2^i,\cdots,w_m^i\}$ 表示原文，$w_j^i\in s_i$ 表示原文的字，m 为原文的长度；$C_i=\{c_1^i,c_2^i,\cdots,c_o^i\}$ 表示原文 s_i 对应的评论(回复)集，$c_j^i\in C_i$ 表示一条评论，其中每条评论又由若干字组成，o 表示评论个数；u_i 表示发布者的用户画像，用户画像由若干个属性组成；l_i 表示该条数据的标签，具体为{0, 1}，0 表示为真，1 表示为假。总之，对于数据集 D，给定{s, C, u}，需要预测其标签 l。

3.2. Tansformer

BERT 模型以 Transformer 编码器为基础构建，其结构如图 1 所示。Transformer 最早由 Vaswani 等人[10]提出，分为编码器和解码器。BERT 用编码器作为模型的基本单元，Base 版本堆叠 12 层 Tansformer 编码器，Large 版本则堆叠 24 层 Transformer 编码器。Transformer 作为强特征提取器，其成功主要归功于全局的多头自注意力机制，该机制的公式化描述为：

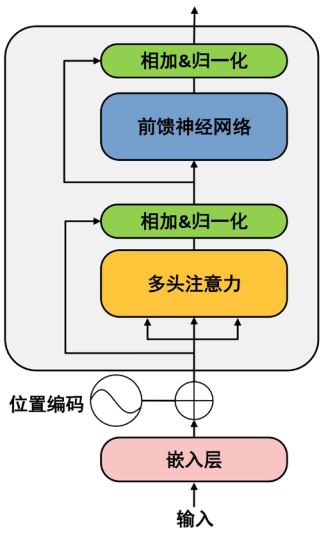

Figure 1. Structure of Transformer encoder
图 1. Transformer 编码器结构图

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{Attention 1},\cdots,\text{Attention } N) \tag{2}$$

Q 代表 Query 矩阵，K 代表 Key 矩阵，V 代表 Value 矩阵，它们从相同的表示矩阵经过线性变换而来，即自注意力机制。d_k 表示向量的维度。公式(1)得到经过自注意力机制计算后的表示矩阵，公式(2)意为将多头注意力表示矩阵做拼接即得到最终表示矩阵。

4. User-BERT 模型

4.1. 整体架构

图 2 展示了 User-BERT 模型的整体架构。架构大致可以分为两部分，本文编码器和深度分类器。文本编码器即为 BERT 模型，其将谣言原文和评论的文本编码成文本表示向量。输出的文本表示向量和用户属性向量进行拼接，形成综合表示向量。深度分类器为全连接层网络(Fully Conneted Network, FCN)，对综合表示向量进行解析并输出最后预测结果。

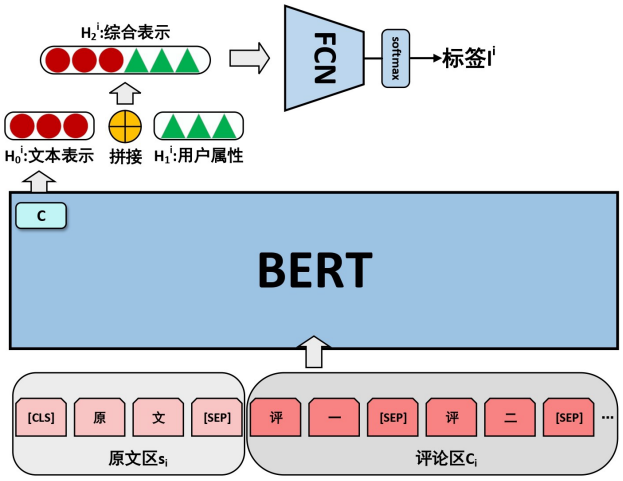

Figure 2. Overall structure of User-BERT
图 2. User-BERT 整体架构

4.2. 文本编码

文本将 BERT 文本编码器的输入分成原文区和评论区两部分，原文区长度限制为 128，评论区长度限制为 384。(BERT 长度限制为 512)原文和评论及评论和评论之间用[SEP]表示符进行分隔。评论按照时间先后顺序排列。本文取 BERT 最后一层的[CLS]表示符对应的表示向量作为整体文本的表示向量。该过程可公式化为：

$$H_0^i=\text{BERT}(s_i,C_i)[0] \tag{3}$$

s_i 表示原文，C_i 表示评论集，H_0^i 表示数据 i 的文本表示向量。文本表示向量不仅包含了原文的特征，还包含了丰富的评论特征。

4.3. 用户特征

本文对用户属性分布进行了详细分析，最终选出六个分布差异明显的用户特征，详见附录。即 u_i = {verified, verified_type, verified_reason_length, followers_count, bi_followers_count, statuses_count}。verified 表示用户是否得到官方认证，verified_type 表示认证类型，verified_reason_length 表示认证原因的文本长

度，followers_count 表示粉丝数量，bi_followers_count 表示双向关注用户数量，statuses_count 表示用户的微博数量。将文本表示向量 H_0^i 和用户属性向量 H_1^i 做拼接，得到综合表示向量 H_2^i：

$$H_2^i = \text{Concatenate}\left(H_0^i, H_1^i\right) \tag{4}$$

4.4. 深度分类器

深度分类器采用 FCN 模型，对输入的综合表示向量 H_2^i 进行解析得到中间结果向量 h_i：

$$h_i = \text{FCN}\left(H_2^i\right) \tag{5}$$

最后通过 softmax [13]激活函数对 h_i 进行归一化处理即可得到数据 d_i 对应标签 l_i 的概率分布：

$$p_j = \frac{\exp\left(h_j^i\right)}{\sum_{k=0}^{n}\exp\left(h_k^i\right)} \tag{6}$$

5. 实验与分析

5.1. 数据集

本文采用的数据集是 Ma 等人[4]提出的新浪微博数据集 Ma-Weibo。数据集从微博社区管理中心[1]收集而来，其中的谣言数据由人工验证且公开，是来自现实中的真实数据。数据集总共包含 4664 条数据，谣言 2313 条，非谣言 2351 条。公开的数据集[2]带有微博原文、评论、用户属性等信息。

5.2. 实验设置

实验设备情况大致如下，操作系统为 Ubuntu 18.04.4，CPU 型号为 Inter(R) Xeon Silver 4110，显卡型号为 GeForce RTX 2080Ti。采用的 BERT 是谷歌官方[3]提供的中文 Base 版模型，实验环境为 Python3.6、Tensorflow1.14。FCN 模型为三层的全连接层网络，前两层神经元个数为 128，第三层为 2。训练时使用两段式分别训练 User-BERT 的文本编码器和深度分类器，第一阶段使用数据集对 BERT 模型进行调优，第二阶段冻结 BERT 模型参数，对 FCN 模型参数进行训练。此外，训练 BERT 的学习率为 2e-5，epoch 为 8。

5.3. 对比方法

实验充分对比了各种机器学习方法，从特征工程到最新的深度学习方法。对比的方法如下：

- SVM-TS (Ma 等人提出[3])：使用线性 SVM 分类器对人工提取特征进行分类。
- RNN (Ma 等人提出[4])：将原文和评论的表示向量输入到 RNN 网络进行分类。
- CGRU (Wang 等人提出[5])：除原文和评论，加入情感特征到 RNN 网络进行分类。
- CAMI (Yu 等人提出[6])：使用 CNN 模型对谣言进行分类。
- Bi-GCN (Bain 等人提出[7])：使用双向 GCN 网络学习谣言传播特征进行分类。
- RFS-BD (尹鹏博等人提出[1])：使用 CNN 和 RNN 作为基学习器，随机森林为元学习器的集成学习。
- GRU-Ensemble (Geng 等人提出[9])：使用三种 RNN 模型作为基学习器，以投票方式进行分类。

[1] https://service.account.weibo.com/
[2] https://www.dropbox.com/s/46r50ctrfa0ur1o/rumdect.zip?dl=0
[3] https://github.com/google-research/bert

5.4. 结果分析

表 1 展示了所有方法的实验结果，文本提出的 User-BERT 在每个指标都取得了最好的结果。本文取准确率、精确率、召回率及综合考虑精确率和召回率的 F1 值四个指标，充分展示各模型的表现结果。对比方法的结果取其论文和复现实验中最好的结果。通过观察可以得到以下结论：(1) SVM-TS 的结果比其它深度学习方法都要差，说明深度学习方法比特征工程方法确实有明显的性能优势。(2) User-BERT 的结果好于 RNN 和 CGRU，显然是因为 Tansformer 比 RNN 单元有更强的特征提取能力，RNN 单元随着序列长度增加会遗忘部分信息，而 Transformer 则能从全局注意力中学习。(3) User-BERT 的结果明显好于 CAMI，说明 Tansorformer 比卷积神经网络更适合用于处理文本信息。(4) User-BERT 的结果好于 Bi-GCN，说明即使不使用谣言的传播信息也可以达到更好的结果。(5) User-BERT 比两个集成模型即 RFS-BD 和 GRU-Ensemble 的表现明显更加优异，展示了 User-BERT 单模型的强大性能。

Table 1. Comparison of experimental results
表 1. 实验结果对比

方法	准确率	精确率	召回率	F1 值
SVM-TS	0.846	0.845	0.845	0.845
RNN	0.910	0.914	0.910	0.910
CGRU	0.963	0.963	0.963	0.963
CAMI	0.933	0.933	0.933	0.933
Bi-GCN	0.961	0.962	0.963	0.961
RFS-BD	0.929	0.927	0.923	0.925
GRU-Ensemble	0.956	0.956	0.957	0.956
User-BERT(Ours)	**0.968**	**0.969**	**0.968**	**0.968**

5.5. 消融实验

为了更好理解 User-BERT 每个部分的作用，文本对 User-BERT 模型进行了消融实验，实验结果如表 2 所示。User-BERT/User 表示只去除用户属性，User-BERT/Comment 表示只去除评论信息，User-BERT/User/Comment 表示去除用户属性和评论信息。可以看到，去除用户属性或评论信息都会造成性能下降，说明它们对谣言检测都有帮助。而从下降幅度来看，评论信息比用户属性更加重要，这可能是因为参与用户能有效反映原文的情感和观点[14]，给原文补充更多有效特征。

Table 2. Comparison of ablation experiment results
表 2. 消融实验结果对比

方法	准确率	精确率	召回率	F1 值
User-BERT	**0.968**	**0.969**	**0.968**	**0.968**
User-BERT/User	0.960	0.961	0.960	0.960
User-BERT/Comment	0.946	0.946	0.947	0.946
User-BERT/User/Comment	0.935	0.934	0.936	0.935

6. 总结与展望

本文提出了 User-BERT 模型，充分利用预训练语言模型 BERT 和文本信息及用户属性相结合，并在

公开微博数据集 Ma-Weibo 上取得了最好的结果。通过实验还证明了评论信息和用户属性对谣言检测的作用。未来将考虑把更多可能的特征融入到预训练语言模型中，比如知识被证明对谣言检测很有帮助[15]，传播路径也被证明有帮助[7]。此外，还将会使用更多其它的预训练语言模型进行实验。

参考文献

[1] 尹鹏博, 彭成, 潘伟民. 基于集成学习的微博谣言早期检测[J]. 微电子学与计算机, 2021, 38(1): 83-88.

[2] Liu, X., Nourbakhsh, A., Li, Q., *et al*. (2015) Real-Time Rumor Debunking on Twitter. *Proceedings of the* 24*th ACM International on Conference on Information and Knowledge Management*, Melbourne, 19-23 October 2015, 1867-1870. https://doi.org/10.1145/2806416.2806651

[3] Ma, J., Gao, W., Wei, Z., *et al*. (2015) Detect Rumors Using Time Series of Social Context Information on Microblogging Websites. *Proceedings of the* 24*th ACM International on Conference on Information and Knowledge Management*, Melbourne, 19-23 October 2015, 1751-1754. https://doi.org/10.1145/2806416.2806607

[4] Ma, J., Gao, W., Mitra, P., *et al*. (2016) Detecting Rumors from Microblogs with Recurrent Neural Networks.

[5] Wang, Z. and Guo, Y. (2020) Rumor Events Detection Enhanced by Encoding Sentimental Information into Time Series Division and Word Representations. *Neurocomputing*, **397**, 224-243. https://doi.org/10.1016/j.neucom.2020.01.095

[6] Yu, F., Liu, Q., Wu, S., *et al*. (2017) A Convolutional Approach for Misinformation Identification. *Proceedings of the* 26*th International Joint Conference on Artificial Intelligence*, Melbourne, 19-25 August 2017, 3901-3907. https://doi.org/10.24963/ijcai.2017/545

[7] Bian, T., Xiao, X., Xu, T., *et al*. (2020) Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 549-556. https://doi.org/10.1609/aaai.v34i01.5393

[8] Geng, Y., Lin, Z., Fu, P., *et al*. (2019) Rumor Detection on Social Media: A Multi-View Model Using Self-Attention Mechanism. In: *International Conference on Computational Science*, Springer, Cham, 339-352. https://doi.org/10.1007/978-3-030-22734-0_25

[9] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[10] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need.

[11] Li, Q., Zhang, Q. and Si, L. (2019) Rumor Detection by Exploiting User Credibility Information, Attention and Multi-Task Learning. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 1173-1179. https://doi.org/10.18653/v1/P19-1113

[12] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks.

[13] Liu, W., Wen, Y., Yu, Z., *et al*. (2016) Large-Margin Softmax Loss for Convolutional Neural Networks. *The* 33*rd International Conference on Machine Learning* (*ICML* 2016), New York, 19-24 June 2016, 7.

[14] Lu, Y.J. and Li, C.T. (2020) GCAN: Graph-Aware Co-Attention Networks for Explainable Fake News Detection on Social Media. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, July 2020, 505-514. https://doi.org/10.18653/v1/2020.acl-main.48

[15] Li, Q., Zhang, Q., Si, L., *et al*. (2019) Rumor Detection on Social Media: Datasets, Methods and Opportunities. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom*: *Censorship, Disinformation, and Propaganda*, Hong Kong, November 2019, 66-75. https://doi.org/10.18653/v1/D19-5008

附录

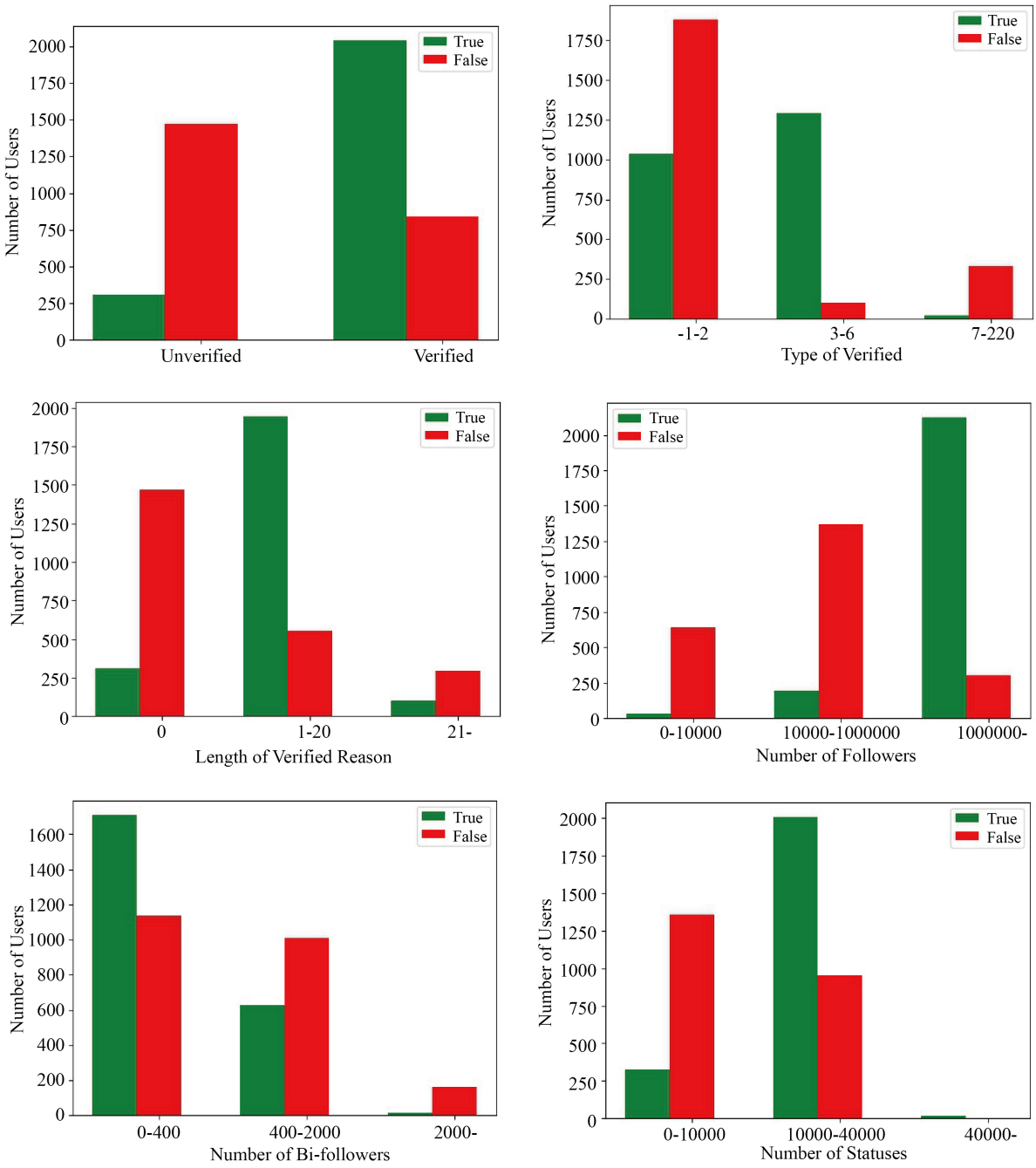

Figure A1. Distribution of user characteristic
图 A1. 用户特征分布