Statistics and Application 统计学与应用, 2023, 12(2), 563-568
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122060

Hans汉斯

基于FastText模型的匿名数据文本分类研究

朱美瑶，张寅昊，王宇喆，钟美君

嘉兴南湖学院信息工程学院，浙江 嘉兴

收稿日期：2023年3月27日；录用日期：2023年4月17日；发布日期：2023年4月29日

摘　要

本文主要讨论在数据匿名化情况下，FastText模型相比其它机器学习模型，对文本分类问题是否是更优解。本文对公开新闻数据集的20万条中文文本数据进行匿名化处理，然后分别采用逻辑回归、LGBM、随机森林和FastText模型进行分类，并且针对结果，对FastText提出两方面的改进，通过多个评价指标进行评价后，FastText模型无论在准确率上，还是在运行效率上，均比其它模型更优秀。

关键词

数据匿名化，FastText，TF-IDF，文本分类

Research on Text Classification of Anonymous Data Based on FastText Model

Meiyao Zhu, Yinhao Zhang, Yuzhe Wang, Meijun Zhong

School of Information Engineering, Nanhu University, Jiaxing Zhejiang

Received: Mar. 27th, 2023; accepted: Apr. 17th, 2023; published: Apr. 29th, 2023

Abstract

This paper focuses on whether the FastText model is a better solution to the text classification problem compared to other machine learning models in the case of data anonymization. In this paper, 200,000 Chinese text data from public news datasets are anonymized, and then logistic regression, LGBM, random forest and FastText models are used for classification, and two improvements to FastText are proposed for the results. The FastText model is better than other models in terms of both accuracy and efficiency.

文章引用: 朱美瑶, 张寅昊, 王宇喆, 钟美君. 基于FastText模型的匿名数据文本分类研究[J]. 统计学与应用, 2023, 12(2): 563-568. DOI: 10.12677/sa.2023.122060

Keywords

Data Anonymization, FastText, TF-IDF, Text Classification

1. 引言

在现代经济与科技发展的浪潮中，机器学习无疑是最火热的领域。无论是 alphaGo 战胜了人类世界围棋冠军，还是 ChatGPT 一夜爆火，都是机器学习在某一领域的深度开发。文本分类也是机器学习领域一个比较热门的方向，是工业界能够实际应用的方向之一，比如将文本分类应用于投诉的自动处理、新闻的自动归类等。无论对于中文或是英文的文本分类，为了精确率的目标，都需要对文本进行预处理。但是，有一些敏感或者涉密行业，它们的数据无法对外提供，而这些行业又确实有相关的需求。本文就是对数据匿名化处理后的文本分类进行研究。

文本的匿名化[1] [2]指的是对机器学习中最重要的数据进行处理，将其处理成格式化数据的过程中，消除数据可识别的信息。匿名化的操作，对于个人信息和企业信息的保护和流通都有着非常重要的价值。

2. 相关技术

2.1. FastText

FastText 是一种基于深度学习的字符级别的文本分类算法[3] [4]。它允许迅速建立和训练词向量模型，以解决文本分类和语义分析的问题。它可以处理不同长度的句子，而且还可以处理拼写错误和语义相似性，这使得它成为自然语言处理领域中被广泛使用的框架之一。FastText 将字符串分割为单词，然后将每个单词映射到可以训练的特征向量，最后使用深度神经网络进行分类。FastText 最大的优势是可以处理拼写错误、缩写、数字等文本的不规范形式，比传统的 NLP 方法更快、更容易，并且可以高效处理大型语料库。

国内外对 FastText 的研究进展一直在不断发展。例如，研究者们正在试图利 FastText 提供的技术来更好地理解语义和上下文间的关系[5]。此外，研究者们还在开发新的方法来改善 FastText 的性能和加快其训练过程，以实现更高水平的准确性[6]。另外，一些研究者正在利用 FastText 来处理网络语言分析，对社交媒体，短信和评论进行情感分析[7]，以及使用词向量进行自然语言生成等等。

2.2. 匿名数据

数据匿名化是指在数据传输及处理过程中，将个人身份识别信息(如姓名、身份证号码)经过特殊处理后，改变其原始形态，使得数据持有者无法识别个人身份信息的一种技术手段。数据匿名化可以减少被盗取的个人信息，防止个人信息泄露，有效提高信息保护的安全性[8]。

数据匿名化是一种策略，旨在通过减少数据的可识别性来保护用户的隐私。数据匿名化的常用方法有：

1) 数据混淆：通过添加噪声或者编码技术，使得真实数据变得模糊不清，以此来保护数据的隐私[9]。

2) 数据删除：删除不必要的数据，从而减少可识别数据的范围，从而提高信息隐私的安全性。

3) 数据整理：将原始数据转换为只包含统计信息的数据，从而在不影响数据分析的情况下提高信息隐私的安全性。

4) 数据匿名化：使用哈希函数、单向函数或者匿名标识符等技术，取消或替换可识别的数据，从而提高信息隐私的安全性。

2.3. TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency)是一种文本挖掘技术，可以统计一个词语在语料库中出现的次数，以及它在语料库中所占比重，这样从整体上来评估词语的重要性。它的主要思想是，一个词语在一篇文章中出现的次数越多，该词语就越重要，TF-IDF 利用这一点，在每篇文章中都为每个词语计算一个权值，这个权值可以用来评价一个词语是否重要。TF-IDF 是一种基于词频(Term Frequency)和逆文档频率(Inverse Document Frequency)的统计方法，它可以更好地反映一个文档的特征和主题，提高文本搜索和文档分类的准确性[2] [10]。

TF 就是指词语在文件中出现的频率，即某个词语在文件中出现的次数除以文件中的词语总数。IDF 就是指某个词语的逆文档频率，即某个词语在所有文件出现的概率的倒数。TF-IDF 即是 TF 和 IDF 的乘积。

3. 材料与方法

本文采用的数据集是一个公开的中文新闻数据集，数据集共包含 20 万条数据，将其以 8:2 的比例划分成训练集和测试集。所有数据分为科技、股票、体育、娱乐、时政、社会、教育、财经、家居、游戏、房产、时尚、彩票、星座共 14 个类别，每一条数据以字符级别进行匿名处理后，数据如下表 1 所示：

Table 1. Anonymous processing of Chinese text data
表 1. 中文文本数据匿名处理

标签	文本
1	BIFG FGEH CCI B0BA AHED CGCA DACK CADB DADH AEA BCDE CHAA BDAC BACH
...	...

其中，标签代表这条数据的类别，文本是一篇新闻的具体数据。字符集匿名化处理的规则如下：

1) 将所有文本以 n_gram 为 1 的滑动窗口进行分词。分词后形成了以字符为单位的一个集合，将集合中的字符按统计频率从低到高排序。

2) 将所有字符进行从 0 开始的编号，获得一个自然数集合。

3) 按顺序获取用户输入的 10 个英文字符，使用英文字符替换数字中的每一位，获得一个英文字符组合的集合

4) 将英文字符的集合替换原文本中的中文字符，形成了新的训练和测试数据集。

5) 将数据的标签类别

本文的训练集的数据分布如图 1 所示：

其中，类别 0 也就是匿名化前的科技类别数据最多，有 38918 条数据。类别 13 数据最少，只有 908 条数据。

通过对匿名化后的字符编码进行统计，可知所有训练数据中，总共包含不同类型的字符 7550 个。通

过对所有字符进行统计，可得出现率最高的前 3 个中文字符为：DHFA、JAA、GEI，初步猜测可能是逗号、句号等标点符号。

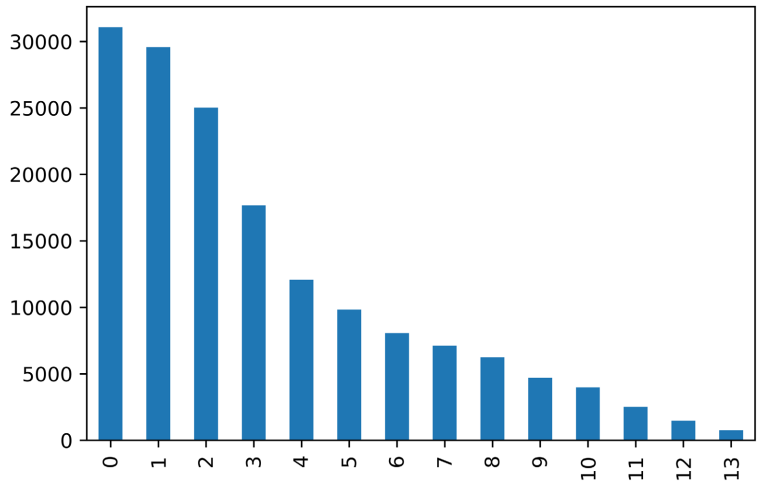

Figure 1. Training set data category distribution
图 1. 训练集数据类别分布

4. 实验结果与分析

本文通过对数据进行匿名化后，使用 FastText 模型与其它模型进行对比实验。在相同的训练集和测试集下，其它模型为了提高其分类的准确性，均在训练前对数据进行 TF-IDF 的预处理。在四种模型准确率的表现上，FastText 模型 93.62%，逻辑回归模型 90.68%，随机森林模型 91.03%，LGBM 模型 94.42%，结果如图 2 所示。

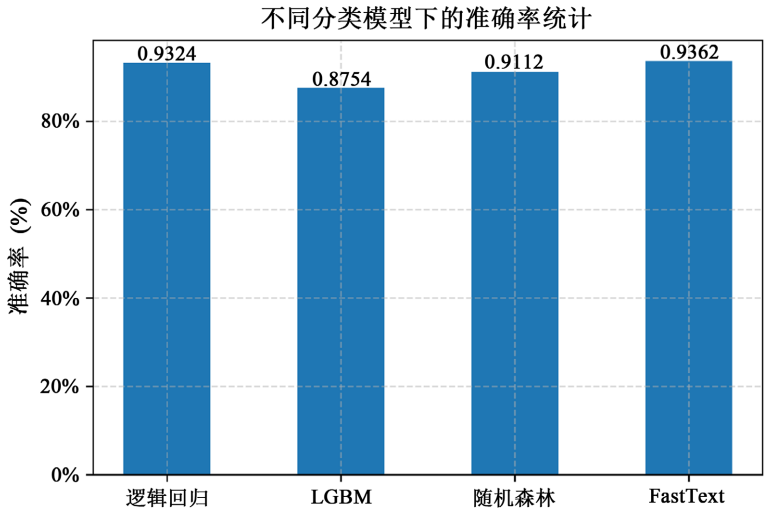

Figure 2. The accuracy of four classification models
图 2. 四种分类模型的准确率

除了准确率以外，本文还通过精确率、召回率及 F1 值来评价不同分类模型的表现，如表 2 所示。根据实验结果，LGBM 算法在各项指标上均优于其它模型。逻辑回归数据表现较差，并且其耗时远高于其它模型。从准确率来看，FastText 算法仅次于 LGBM，在运行速度上高于其它模型。

Table 2. The results of different models under four evaluation indicators
表 2. 不同模型在四种评价指标下的结果

模型	精确率	准确率	召回率	F1 值
逻辑回归	0.8836	0.9068	0.8849	0.8841
LGBM	0.9384	0.9442	0.9232	0.9306
随机森林	0.9310	0.9103	0.8407	0.8787
FastTest	0.9153	0.9362	0.9136	0.9144

5. 对 FastText 模型的改进 TT-FastText

在中文文本分类中，存在特征稀疏、分词效果不佳、无法处理语义信息、无法处理长文本等。因此，本文基于新闻数据的特征，使用两种方法改进 FastText 的分类效果：

预先使用 TF-IDF 对文本进行特征值处理。

将每个文本拆分成标题和内容，分别进行 embedding 后，提高标题的权重，加权拼接以获得更好的分类效果。

TT-FastText 的算法步骤如下：

1) 将文本拆分成标题和内容
2) 分别对标题和内容计算每个词的 TF-IDF 值
3) 分别对标题和内容计算每个词的 FastText 向量表示
4) 将每个词的 TF-IDF 和 FastText 向量进行加权平均
5) 将标题和内容进行加权平均

经过实验，TT-FastText 和 FastText 准确率如图 3 所示。

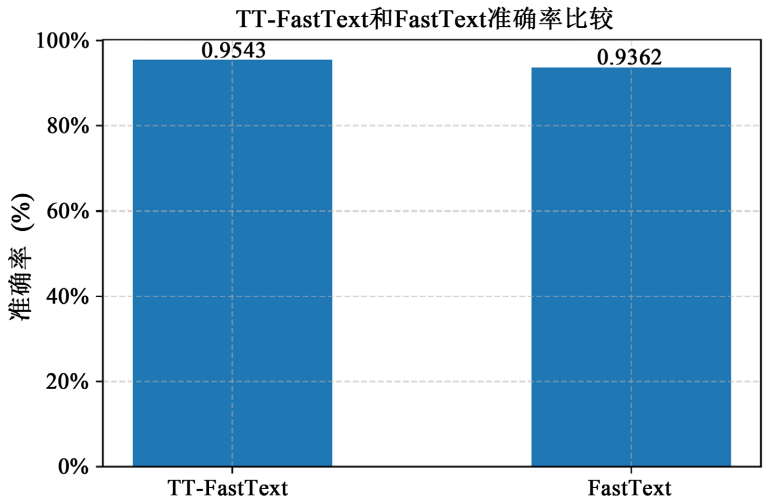

Figure 3. Compared with TT-FastText and FastText accuracy after improvement
图 3. 改进后的 TT-FastText 与 FastText 准确率比较

TT-FastText 与 FastText 各项指标如表 3 所示。

针对以上两方面的改进，TT-FastText 模型在各项评价指标上均有明显提升，相比逻辑回归、LGBM、随机森林等模型，有更好的分类效果和运行效率。

Table 3. TT-FastText and FastText various indicators
表 3. TT-FastText 与 FastText 各项指标

模型	精确率	准确率	召回率	F1 值
TT-FastTest	0.9416	0.9543	0.9394	0.9401
FastTest	0.9153	0.9362	0.9136	0.9144

6. 结束语

本文提出了 FastText 对匿名化数据分类应用的研究意义，说明了匿名化数据的需求与使用场景，然后介绍了目前的技术基础及采用的实验方法。通过处理实验数据，将其完全匿名化，然后使用 FastText 与逻辑回归、LGBM 和随机森林模型进行对比，采用精确率、准确率、召回率和 F1 值进行评价，最后通过对 FastText 进行两方面的改进，突显 FastText 模型对匿名化文本分类问题的性能和效率。FastText 模型亦可通过模型融合等方法，进一步提高其准确率。对于非全匿名化的数据，在数据预处理阶段，可通过人工标注等方式，进一步提高其效率。

参考文献

[1] 孙广中, 魏燊, 谢幸. 大数据时代中的去匿名化技术及应用[J]. 信息通信技术, 2013, 7(6): 52-57.

[2] 李媛. 大数据时代个人信息保护研究[D]: [博士学位论文]. 重庆: 西南政法大学, 2016.

[3] 代令令, 蒋侃. 基于 fastText 的中文文本分类[J]. 计算机与现代化, 2018(5): 35-40+85.

[4] 冯勇, 屈渤浩, 徐红艳, 王嵘冰, 张永刚. 融合 TF-IDF 和 LDA 的中文 FastText 短文本分类方法[J]. 应用科学学报, 2019, 37(3): 378-388.

[5] 霍光煜, 张勇, 孙艳丰, 尹宝才. 基于语义的档案数据智能分类方法研究[J]. 计算机工程与应用, 2021, 57(6): 247-253.

[6] 阴爱英, 吴运兵, 郑一江, 余小燕. 基于 fastText 模型的词向量表示改进算法[J]. 福州大学学报(自然科学版), 2019, 47(3): 314-319.

[7] 范昊, 李鹏飞. 基于FastText字向量与双向GRU循环神经网络的短文本情感分析研究——以微博评论文本为例[J]. 情报科学, 2021, 39(4): 15-22. https://doi.org/10.13833/j.issn.1007-7634.2021.04.003

[8] Britton, K.E. and Britton-Colonnese, J.D. (2017) Privacy and Security Issues Surrounding the Protection of Data Generated by Continuous Glucose Monitors. *Journal of Diabetes Science and Technology*, **11**, 216-219. https://doi.org/10.1177/1932296816681585

[9] 蔡婷. 基于数据混淆的隐私保护机制研究[D]: [硕士学位论文]. 西安: 西安建筑科技大学, 2016. https://doi.org/10.27393/d.cnki.gxazu.2016.000150

[10] 叶雪梅, 毛雪岷, 夏锦春, 王波. 文本分类 TF-IDF 算法的改进研究[J]. 计算机工程与应用, 2019, 55(2): 104-109+161.