

# 基于 GRW 和 FastText 模型的电信用户投诉文本分类应用

赵进，杨小军
（中国电信股份有限公司重庆分公司，重庆 401120）

**摘　要**：随着神经网络的广泛应用，将神经网络应用到自然语言处理文本分类问题中，成为一种有效的解决方法。电信运营商客户服务中心通过多种渠道收集用户投诉信息，为了对投诉文本信息进行自动分类并将其落实到具体责任部门，提升用户感知，提出了一种基于 GRW 模型和 FastText 模型的文本分类方法。首先通过 GRW 模型对投诉文本进行特征选择，提取有效特征词；然后构建基于 FastText 模型的用户投诉文本分类方法；最后在公开数据集和运营商已标注的投诉文本数据集上进行实验。结果表明，基于 GRW 和 FastText 模型的文本分类方法比朴素贝叶斯、双向 LSTM 和 Bert 模型在准确率、Kappa 系数及汉明损失方面的性能有较大提升。

**关键词**：神经网络；文本分类；GRW 模型；FastText 模型
**中图分类号**：TP391.1
**文献标识码**：A
**doi:** 10.11959/j.issn.1000−0801.2021125

# Application on text classification of telecom user complaints based on GRW and FastText model

ZHAO Jin, YANG Xiaojun
Chongqing Branch of China Telecom Co., Ltd., Chongqing 401120, China

**Abstract:** With the widespread application of neural network, the application of neural network to natural language processing text classification problems has become an effective solution. The customer service center of telecom operator collected user complaint information from multiple channels. In order to automatically classify the complaint text information and assign it to the specific responsible department for processing and reply, enhancing customer perception further, a textclassification method based on GRW and FastTextmodel was proposed. Firstly, the GRW model was used to select the features of the complaint text, extract effective feature words, and then a user complaint text classification method based on FastText model was constructed. Experiments on public datasets and a complaint text data by one of telecom company show that the text classification method based on GRW and FastText model is better than naive Bayes, bidirectional LSTM and Bert pre-trained model in accuracy, Kappa coefficient and Hamming loss.

**Key words:** neural network, text classification, GRW model, FastText model

收稿日期：2020−10−15；修回日期：2021−05−25

## 1 引言

大数据和人工智能作为国家提出的新型基础设施，引发了新一轮的科技革命。中国电信作为国内三大电信运营商之一、三化（数字化、信息化、智能化）产业布局者，正着力于大数据和人工智能在其业务和服务上落地生根，实现数字化转型战略目标。中国电信始终将用户的良好体验和用户感知放在首位，对用户的投诉进行业务拆解和快速分析，以求及时准确地反馈用户，规避用户越级投诉的风险和提升用户良好体验和使用感知。但是电信的业务服务类别复杂多样，有效地识别用户投诉类型目前还处于人工分类阶段，在实际工作过程中，业务领域专家的经验起着主导作用[1]，而培养一个业务领域专家需要大量的时间成本和人力成本，因此，需要建立科学合理的文本分类模型。通常用户投诉数据有多种来源，如 10000、微信公众号、线下厅店等，最终都会以非结构化的文本信息记录在系统中，如何对这些用户的投诉信息进行有效、快速、准确的分类成为处理用户投诉、提升用户良好体验和使用感知的关键[2]。

文本分类属于自然语言处理（natural language processing，NLP）应用研究范畴中的一个分支[3]，在文本分类方法上，目前已经有了很多有突破性的研究成果，比如基于朴素贝叶斯方法的文本分类方法[4]、基于条件随机场的分类方法[5]等，随着深度学习向文本挖掘领域发展，利用 word2vec、LSTM、CNN、TextRNN、预训练模型 Bert 等深度学习模型对文本进行分类，也取得了比较好的分类效果，并且在部分人工智能产品中落地应用。在传统的文本分类方法中，需要将文本转换成高维向量，这一般会造成向量维度过大，无论计算内存还是计算时间，计算复杂度将会非常大。为了避免高维风险，需要借助特征工程的相关方法或模型对文本的关键特征进行有效的提取，实现降低文本特征向量的维度数。目前文本特征提取有基于词频（TF-IDF）的方法、基于互信息（mutual information，MI）的方法、基于（Chi-square）统计量的方法、基于信息增益（information gain，IG）的方法、基于 WLLR（weighted log likelihood ration，WLLR）的方法和基于 WFO（weighted frequency and odd，WFO）的方法等[6]。在电信用户投诉文本数据中，存在很多与分类结果无关的信息描述，同时电信业务的类别复杂，多达几十种。针对上述问题，本文提出基于 GRW 和 FastText 模型的方法，对用户投诉文本进行多分类，并且基于真实的生产系统数据进行相关验证性实验，验证了本文方法的有效性。

## 2 基于 GRW 模型的特征提取

文本数据的特征提取方式与传统的数值类型特征的提取方式不同。在数据表示上，传统机器学习的数据由一组结构化的数据组成，这些数据的特征一般表现为字符串型的类别性、数值的离散性、数值的连续性、数据的时间性或周期性等，而文本数据属于非结构化数据，由字、词、句组成。因此在对文本数据进行特征提取前，需要对文本进行预处理，将文本拆分成字粒度或词粒度，然后通过不同的方法对文本数据进行数学表示。第一种方法是将文本表示成离散型数据，如 One-hot 编码和 Multi-hot 编码等；第二种方法是将文本表示成分布式型数据，如矩阵表示法、降维表示法、聚类表示法、神经网络表示法。

中文文本数据由字、词、句组成的文本，将中文文本按字、词进行拆分后可以进行一些数学统计，其中 TF-IDF 用来评估一个字或词对一个文本集或语料库中的一条文本的重要程度。当一个字或词在一个文本数据中重复出现时，则认为该字或词具有比较高的重要性，但如果该字或词在所有的文本集或语料库中均出现，且统计其出现

的次数比较高，则该字或词的重要性反而会下降，如一些介词、谓词等。词频（term frequency，TF），即字或词在文本中复现的频率；逆向文本频率指数（inverse document frequency，IDF），即一个字或词在所有的文本中复现的频率，如果一个字或词在很多文本中出现，那么 IDF 值较小，反之，如果一个词在比较少的文本中出现，那么它的 IDF 值较大。TF-IDF 计算式如下：

$$\text{TFIDF}=\text{TF}\times\text{IDF}=\frac{n_{ij}}{\sum_k n_{k_j}}\times\log\frac{|D|}{|\{j:t_i\in d_j\}|} \quad (1)$$

其中，$n_{ij}$ 是该词在文本 $d_j$ 中出现的次数，$\sum^{k} n_{kj}$ 是文本 $d_j$ 中的所有字或词出现的总次数，$|D|$是文本集或语料库中的总的文本数量。$|\{j:t_i\in d_j\}|$表示包含字或词 $t_i$ 的文本数目。

2011 年 Imambi 提出特征加权方案 GRW[7]，赋予特征词 TF-IDF 不同的权重。GRW 模型计算式如下：

$$\text{GRW}(t,C_i)=\text{TFIDF}_j\times\frac{P(T_{ij})}{P(C_i)} \quad (2)$$

其中，$\text{TFIDF}_j$ 是字或词的 T-FIDF 值，$P(T_{ij})$是属于分类 $i$ 的词 $j$ 的概率，$P(C_i)$是属于分类 $i$ 的文档的概率，然后选择字或词的最大 GRW 值确定属于某个分类类别的特征词：

$$\text{GRW}(t)=\max\{\text{GRW}(t,C_i)\} \quad (3)$$

例如，如果 $\text{GRW}(t_1,c_1)=0.3$，$\text{GRW}(t_1,c_2)=0.4$，则词 $t_1$ 被选择为类别 $c_2$ 的特征词。

SagarImambi 在 5 460 份文档的数据集上，分别采用 GRW、TFIDF、CFS、信息增益比（gain ratio，GR）、Chi-square 和过滤子集（filtered subset，FS）对文档进行特征选择后，在贝叶斯网络、朴素贝叶斯、决策树等分类算法上进行相关实验，实验表明，用 GRW 模型进行特征选择后，具有较高的准确率（准确率比较见表 1），因此文本也将使用 GRW 模型进行特征选择。

表 1　准确率比较[7]

| 特征选择模型 | Cart | Decision-table | Bayes net | Bayes |
|---|---|---|---|---|
| GRW | **99.267 4%** | **99.084%** | **100%** | **70.879 4%** |
| TFIDF | 65.934 1% | 65.934 1% | 65.201 5% | 58.248 4% |
| CFS | 64.652% | 63.736 3% | 65.567% | 66.849 8% |
| 信息增益比 | 64.51% | 63.736 3% | 65.934 1% | 61.172 2% |
| Chi-square | 64.835 2% | 63.736 3% | 65.934 1% | 57.692 9% |
| 过滤子集 | 64.285 7% | 64.285 7% | 65.567 8% | 65.567 8% |

## 3　基于 FastText 模型的文本分类

FastText 是 FAIR（facebook AI research）提出的文本分类模型[8]，该模型只有 3 层：输入层、隐含层和输出层。模型结构如图 1 所示。

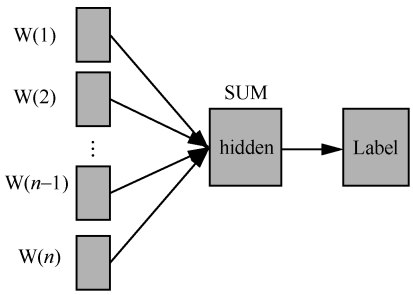


图 1　FastText 模型结构

其中，$W(i)$（$1\leqslant i\leqslant n$）表示文本数据中每个特征词的词嵌入，对每个特征词的词嵌入进行累加后求其均值，用以表示该文本，最后通过 sigmoid 函数得到输出层的标签。在训练该神经网络模型时，通过前向传播算法进行训练：

$$h_{\text{doc}}=\frac{1}{n}\sum_{i=1}^{n} w_i \quad (4)$$

$$z=\text{sigmoid}(W_0 h_{\text{doc}}) \quad (5)$$

其中，$z$ 是一个向量，作为输出层的输入，$W_0$ 是一个矩阵，存放模型隐含层到模型输出层的权重。

通常通过神经网络模型的输出层得到最终的分类类别时，采用 softmax 作为激活函数，将输出层的每个预测值压缩在 0 到 1 之间并归一化。但随着分类类别的增加，softmax 计算量也逐渐增

加，为了减少训练时间，加快模型的训练过程，FastText 使用每个分类类别的权重和模型的参数构建一棵 Huffman 树，形成了层次 softmax，由于具有 Huffman 树的特点，很大程度上提高了模型的训练效率。

FastText 模型的隐含层对每个特征词的词嵌入进行累加后求均值，如果打乱词的顺序，该累加后求均值的方法并不影响隐含层的输出，但失去了语言模型的上下文关系，如：“马上”和“上马”实际含义并不相同。在统计学语言模型中，*N*-gram 算法采用的是滑动窗口法，取一个长度为 *N* 的窗口，按照字或词依次滑过文本，生成片段序列，在 FastText 模型前向传播过程中，把采用 *N*-gram 生成的片段序列表示成向量，也参与到隐含层的累加求均值过程中。然后所有的 *N*-gram 被 Hash 到不同的桶中，但是同一个桶共享嵌入向量，如图 2 所示。

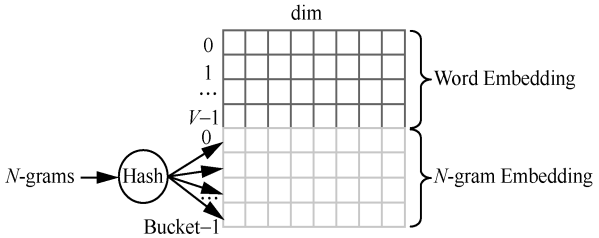


图 2　Embedding 矩阵

Embedding（嵌入）矩阵中的每一行代表了一个字、词或 *N*-gram 的嵌入向量，其中，前 0 到 *V*−1 行是词嵌入，后 0 到 Bucket−1 行是 *N*-gram 嵌入。经过大小为 *N* 的滑动窗口生成的片段序列被哈希函数作用后，被分散到 0 到 Bucket−1 的位置，得到每个片段序列的嵌入向量。

## 4　实验结果及分析

### 4.1　数据集预处理

本实验在公开数据集和私有数据集上进行。其中，公开数据集采用新浪新闻从 2005 年到 2011 年的历史数据，一共覆盖 10 类新闻，每类新闻包括 65 000 条样本。私有数据来源于某运营商已进行人工标注的 29 084 条投诉文本信息，这些投诉文本信息中包括了一些业务部门规定的基础的模板信息（如投诉时间、投诉事件、业务号码、投诉原因、用户要求、联系时间等），在对数据集进行特征提取前，需要对这些模板信息（如删除固定的符号、删除规定的规范用词等）进行初步的清理，投诉内容的原文以及对应的清理后的投诉内容见表 2，这有利于初步降低特征词向量的维度、过滤无用符号等。

表 2　用户投诉文本信息

| 投诉内容（原文） | 投诉内容（清理后） |
| --- | --- |
| 用户来电反映：【用户称 XX 时间已向 XX 分公司 XX 营业厅反映过】，【宽带号码：XX】在【月份：2019-12-01 到 2020-02-29】产生的【费用项目：基本套餐费包含违约金】，【用户原因：称疫情期间无法缴费，自己使用的也是老年手机，无法线上缴费，现对 2019 年 12 月到 2020 年 2 月的违约金不认可，并提及工业和信息化部投诉】，查 BSN【月份：2019-12-01 到 2020-02-29 费用项目：基本套餐费包含违约金金额：623 元】有此费用，已给用户解释，用户拒不认可，现在【用户要求：尽快核实处理】，请相关单位查实处理后回复用户及 10000，谢谢。【联系时间：随时】 | 用户来电反映：宽带号码：XX 在 2019−12−01 到 2020−02−29 产生的基本套餐费包含违约金，称疫情期间无法缴费，自己使用的也是老年手机，无法线上缴费，现对 2019 年 12 月到 2020 年 2 月的违约金不认可，并提及工业和信息化部投诉，查 BSN2019−12−01 到 2020−02−29，费用项目：基本套餐费，包含违约金金额：623 元有此费用，已给用户解释，用户拒不认可，现在，请相关单位查实处理后回复用户及 10000，谢谢。 |

由于投诉文本以最终分类到具体的业务类别为标准，进而供业务单位处理。某运营商在业务类别分类上采用多级标识，以“>”作为上下级分割符号，且各业务类别均表示得比较独立，因此将此投诉文本分类问题，定义为标签多分类问题而不是多标签分类问题。本实验样本已标记的数据分类类别共有 89 类，标签均为文本，需要对类别进行编码，部分类别编码后见表 3。

表 3　业务类别编码

| 业务类别 | 业务编码 |
| --- | --- |
| 移动业务>移动数据网络>用户端原因>用户使用不当或操作设置错误 | 12 |
| 信息安全及专项>通信诈骗>电话类>关停涉嫌诈骗电话>数据模型分析及呼出超频关停 | 23 |
| 移动业务>渠道>自有营业厅>业务不熟练/受理差错>其他 | 35 |
| 固话业务>规则政策>业务办理规则>拆机/退网规则 | 3 |
| 固话业务>渠道>合作渠道>代理服务点>业务不熟练/受理差错 | 44 |
| 移动业务>终端及卡>终端与 UIM 卡故障>UIM 卡故障 | 5 |

从图 3 投诉数据集分类统计可以看出，本次实验数据集分布不均匀，业务类别记录数从 51 到 3 546 条，属于类别不均衡问题。此时，准确率不能作为最后分类模型效果评估的唯一标准，本文采用准确率（accuracy）、Kappa 系数以及汉明损失（Hamming loss）作为评价模型好坏的指标。

将实验样本数随机打乱顺序后，按 7:3 比例进行拆分，70%的数据作为训练集，用于模型训练；30%的数据作为测试集，用于模型验证。

### 4.2　特征提取

对中文文本数据进行特征提取前，需要对文本进行分词，本实验在采用 jieba 分词工具的精确模式对文本进行分词的同时，使用电信行业专用业务词汇作为用户词典和中文标准停用词，以提高分词的准确性。

使用 GRW 模型对用户投诉文本进行特征提取，提取出每个类别（标签）的特征词组，部分特征词组见表 4。

表 4　分类类别及对应特征词组

| 业务编码（标签） | 特征词组 |
| --- | --- |
| 0 | “处理”“营销”“升级包”“多扣”“自动”“免费”，“业务受理”“调帐”“流量畅享”“流量达量”… |
| 1 | “移动”“流量”“产生”“费用”“金额”“流量包”“收取”“20 元”“数据通信”“流量费用”“退费”… |
| 2 | “爱游戏”“点错”“异议”“300 元”“小额退赔”，“预付费”“天冀”“阅读”“天阅文化”… |

### 4.3　评价指标

二分类问题的评价指标通常为准确率、精确率、召回率、F1-score、AUC、ROC、P-R 曲线等，多分类问题的模型评估方式有两种：第一种是将多分类的问题通过某种方式转化为 $N$ 个二分类的

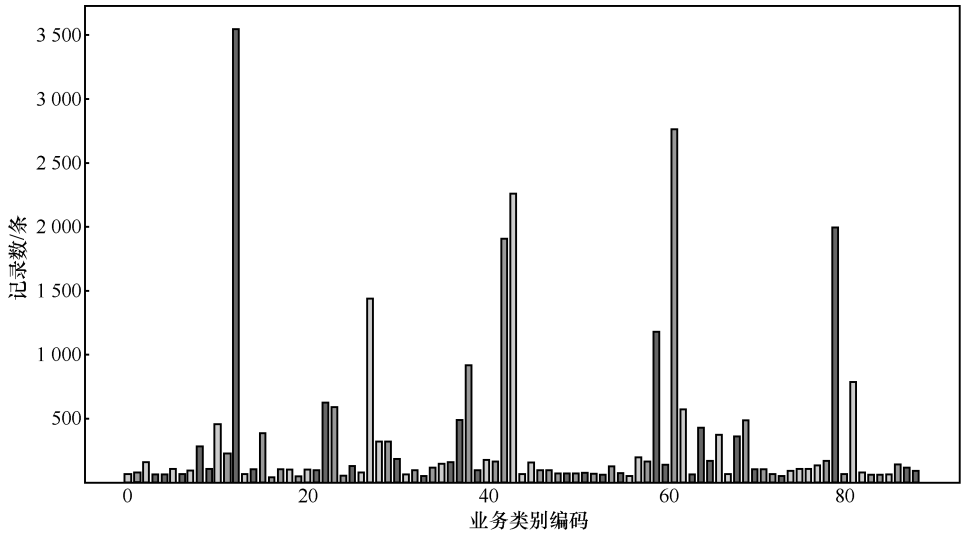


图 3　投诉数据集分类统计

问题；第二种是采用多分类指标。常见多分类问题评价指标如下。

（1）准确率

一般分类问题采用混淆矩阵来判断分类的效果，混淆矩阵如图 4 所示。

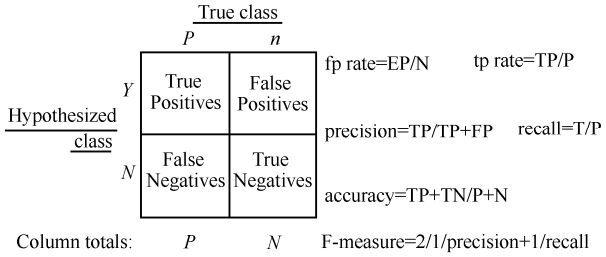


图 4　混淆矩阵

其中，TP（true positives）表示预测对，实际也对；FP（false positives）：表示预测为对，实际为错；TN（true negatives）：预测为错，实际为错；FN（false negatives）：预测为错，实际为对。准确率（accuracy）计算式如下：

$$\text{accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{6}$$

（2）Kappa 系数

Kappa 系数用于衡量多项独立指标的一致性（可靠性），表示具有有限范围的任何统计量的重新定标，形式为（S-SR）/ max（S−SR），其中 SR 是在随机性假设下的 S 值。Kappa 系数取值范围为−1 到+1，值越高，代表模型实现的分类准确度越高。

$$K=\frac{P_{\mathrm{A}}-P_{\mathrm{C}}}{1-P_{\mathrm{C}}} \tag{7}$$

其中，$P_{\mathrm{A}}$ 是总体分类精度，$P_{\mathrm{C}}=P^2+P'^2$，$P$ 是所有分类的比例，和为 1。Kappa 系数及一致性等级对应关系见表 5。

（3）汉明损失[10]

汉明损失考查实例标签被错误分类的情况。对于给定的向量 $f$ 和预测函数 $F$，汉明损失函数计算式如下：

$$L_{\text{hamloss}}(F(f(X),Y)=\frac{1}{Q}\sum_{i=1}^{Q}I[\hat{y}_i\neq y_i] \tag{8}$$

其中，

$$\hat{Y}=F(f(X))=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_Q) \tag{9}$$

汉明损失计算可描述为相关的分类标签不在已预测的分类标签集合中，或者描述为无关的分类标签在已预测的分类标签集合中，因此，汉明损失指标值与模型的分类能力呈现正相关性。

### 4.4　结果分析

本实验分别采用词袋模型和朴素贝叶斯模型、Embedding 和双向 LSTM 模型、预训练 Bert 模型、FastText 模型作为基础模型在公开数据集和私有数据集上进行训练，再用 GRW 模型对特征进行提取后，结合 FastText 模型在训练集上进行训练，最后在同样的测试集上进行验证，结果见表 6、表 7。

**表 6　模型对比（公开数据集）**

| 模型 | 准确率 | Kappa 系数 | 汉明损失 |
|---|---|---|---|
| Bag of Words+朴素贝叶斯 | 0.893 52 | 0.808 93 | 0.137 74 |
| Embedding+双向 LSTM | 0.916 45 | 0.856 74 | 0.109 95 |
| Bert | 0.959 36 | 0.911 63 | 0.051 93 |
| FastText | 0.980 91 | 0.937 14 | 0.031 56 |
| GRW+FastText | 0.986 13 | 0.946 92 | 0.014 69 |

对实验结果进行比较分析可以看出，本文提出的基于 GRW 模型和 FastText 模型的分类方法在准确率、Kappa 系数、汉明损失等指标上，比其他几个模型有所提升，主要在于 GRW 模型在特征词的提取上，进一步优化了特征词的重要性。

**表 5　Kappa 系数及一致性等级对应关系**

| Kappa 系数 | 0～0.2 | 0.2～0.4 | 0.4～0.6 | 0.6～0.8 | 0.8~1.0 |
|---|---|---|---|---|---|
| 一致性等级 | 极低 | 一般 | 中等 | 高度 | 几乎完全一致 |

表 7　模型对比（私有数据集）

| 模型 | 准确率 | Kappa 系数 | 汉明损失 |
|---|---|---|---|
| Bag of Words+朴素贝叶斯 | 0.493 18 | 0.451 23 | 0.506 82 |
| Embedding+双向 LSTM | 0.640 23 | 0.620 94 | 0.359 77 |
| Bert | 0.673 7 | 0.653 73 | 0.326 3 |
| FastText | 0.702 35 | 0.685 51 | 0.297 65 |
| GRW+FastText | 0.823 91 | 0.713 85 | 0.201 73 |

### 4.5　应用效果

2020 年 10 月在某运营商采用本文提出的模型对电信用户投诉文本进行分类，分别从“5G”“携号转网”和“橙分期”3 个专题投诉分析过程中验证模型。10 月“5G”大类投诉共 656 件，“携号转网”大类投诉共 508 件，“橙分期”大类投诉 140 件，采用人工对这 3 类进行投诉类别细分，平均耗时 1 min/件，准确率为 90%。采用本文提出的模型进行自动分类后，总共耗时不到 2 s，准确率达 86%，接近人工分类准确率，提高了业务人员工作效率，缩短了投诉处理时长。

## 5　结束语

本文提出了一种基于 GRW 模型和 FastText 模型的文本分类方法，并应用于电信用户的投诉文本分类问题中，在实验结果对比上，基于 GRW 模型和 FastText 模型的方法比 Bag of Words+朴素贝叶斯模型、Embedding+双向 LSTM、Bert 和 Fast 模型在准确率、Kappa 系数、汉明损失上有所提升。GRW 模型是加权的 TF-IDF 模型，在数据分析过程中有较好的准确性和解释性，使用 FastText 模型进行投诉文本分类具有较快的分类速度和较好的准确率。综上，使用本文提出的方法，有利于业务人员快速识别投诉业务类型，进行针对性的业务受理和用户服务，以提升用户感知。

## 参考文献：

[1] FAED A, CHANG E, SABERI M, et al. Intelligent customer complaint handling utilising principal component and data envelopment analysis (PDA)[J]. Applied Soft Computing, 2016, 47: 614-630.

[2] 梁昕露, 李美娟. 电信业投诉分类方法及其应用研究[J]. 中国管理科学, 2015, 23(S1): 188-192.
LIANG X L, LI M J. Text categorization of complain in telecommunication industry and its applied research[J]. Chinese Journal of Management Science, 2015, 23(S1): 188-192.

[3] Handbook of natural language processing[Z]. 2010.

[4] 李丹. 基于朴素贝叶斯方法的中文文本分类研究[D]. 保定: 河北大学, 2011.
LI D. The study of Chinese text categorization based on Naive Bayes[D]. Baoding: Hebei University, 2011.

[5] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001). [S.l.:s.n.], 2001.

[6] KALAIVANI K S, KUPPUSWAMI S. Exploring the use of syntactic dependency features for document-level sentiment classification[J]. Bulletin of the Polish Academy of Sciences. Technical Sciences, 2019, 67(2).

[7] IMAMBI S S, SUDHA T. A novel feature selection method for classification of medical documents from pubmed[J]. International Journal of Computer Applications, 2011, 26(9): 29-33.

[8] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. [S.l.:s.n.], 2017: 427-431.

[9] KRAEMER H C. Kappa coefficient[J]. Wiley StatsRef: Statistics Reference Online, 2014: 1-4.

[10] GAO W, ZHOU Z H. On the consistency of multi-label learning[J]. Journal of Machine Learning Research, 2011(19): 341-358.

[作者简介]

**赵进**（1989− ），男，现就职于中国电信股份有限公司重庆分公司，主要研究方向为大数据、人工智能、开源软件等。

**杨小军**（1970− ），男，现就职于中国电信股份有限公司重庆分公司，主要研究方向为云计算、大数据、人工智能、开源软件等。