“智慧政务”中的文本挖掘应用

摘要：近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。随着时代的发展与科技的进步，大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。因此，本文基于此挑战和问题，运用大数据和云计算的技术，运用 R 语言编程，通过运用朴素贝叶斯分类算法，建立了关于网络问政平台留言内容的留言标签一级分类模型，并用 F-score 对分类方法进行了评价；首先使用 LDA 获得每个留言主题的主题词分布作为该条评论信息的扩展，然后将留言主题和留言详情一起输入到 word2vec 模型，进行词向量训练，使得留言文本在高维向量空间实现同一主题的聚类，最后使用 Self-Attention 进行动态权重分配并进行分类，选出出现频率前 5 的热点问题，最后基于这 5 个热点问题绘制词云图，从而直观地及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率，与此同时，制作了“热点问题表.xls”和“热点问题留言明细表.xls”以供相关部门核实；最后，通过 G-Caps 模型针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

关键词：大数据　朴素贝叶斯分类算法　F-score　LDA 模型　G-Caps 模型

引言

LDA[5](Latent Dirichlet Allocation)模型在经历了 LSI[6](Latent Semantic Analysis), PLSI[7] (Probabilistic Latent Semantic Analysis)等技术的发展后，被广泛用于文本特征提取。例如吴江等人[8]使用 LDA 模型进行主题特征词选取，组成特征词库进行特征提取。胡勇军等人[9]针对短文本特征稀疏、噪声大的缺点，使用 LDA 模型将概率大于某一阈值的主题词分布的高频词扩展到文本中，以降低短文本分类时噪声和稀疏性的影响。近年来，深度学习算法的快速发展给文本情感分类带来了新的思路。其中基于注意力机制和 LSTM（Long Short Term Memory network）的一类算法成为了主流的文本情感分类算法。其中 LSTM 用于获取文本得上下文依赖关系，注意力机制对 LSTM 编码后得文本进行权重分配，然后进行情感分类[10,11,12]。例如吴小华等[11]针对分词的准确性问题提出了基于字向量的表示方法并使用 Self-Attention 和 Bi-LSTM 进行中文短文本情感分类。陶志勇等[12]使用 Bi-LSTM 网络进行特征提取，将双向长短时记忆网络的两个方向的输入独立输入到注意力机制进行全局权重分配。

基于 LDA 的文本特征提取方法作为一种概率主题模型，虽然能够获得文档之间的关系，然而在建模过程中却忽略了文档的上下文依赖关系，导致了语义信息的丢失。深度学习算法基于序列建模的方法弥补了 LDA 的不足。如吴彦文等[13]使用词嵌入对 LDA 获得的文档特征词进行表示，然后和 LSTM 编码后的文本进行拼接，用于解决数据稀疏问题。张群等人[14]通过拼接相加平均合成的词向量和经过 LDA 特征扩展的短文本向量，利用 kNN 进行分类。

门控循环单元(Gated Recurrent Unit, GRU)[13]方法可以很好地将文本上下文特征的关联进行有效地整合，对情感文本有良好的分类效果。GRU 同时避免了传统 RNN 中出现的梯度消失问题，具有更强的记忆功能。G-Caps 模型不需要人工作业的先验知识也能显示文本属性的强度。为了捕获文本的信息特征[14]，G-Caps 模型将标量信息向量化，使得文本特征的表示更加丰富。实验结果证明将特征向量化的模型会产生更准确的分类效果。

一、基于朴素贝叶斯的多文本分类

（一） 朴素贝叶斯分类模型

贝叶斯分类是以著名的贝叶斯定理为基础的一类算法的总称。下面不加证明的给出贝叶斯公式：

$$P(\mathrm{A}_i \mid \mathrm{B}) = \frac{P(\mathrm{A}_i)P(\mathrm{B} \mid \mathrm{A}_i)}{\sum_{j=1}^{n} P(\mathrm{A}_i)P(\mathrm{B} \mid \mathrm{A}_i)}$$

其中，$P(\mathrm{A})$ 为先验概率，$P(\mathrm{A} \mid \mathrm{B})$ 为后验概率。

朴素贝叶斯分类是一种比较简单的分类算法，这种方法的思想很朴素，朴素贝叶斯的思想基础：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。通俗的讲，就是在先验信息已知的情况下，依据条件概率的大小来选择类别。下面给出贝叶斯分类算法的步骤：

Naive Bayes Algorithm

输入 训练数据 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ ，其中 $x_i = (x_i^{(1)}, \cdots, x_i^{n})^{\mathrm{T}}$ ，$x_i^{(j)}$ 是第 i 个样本的第 j 个特征，$x_i^j \in (\alpha_{j1}, \alpha_{j2}, \cdots, \alpha_{js})$ ，α_{jl} 是第 j 个特征的第 l 个取值。 $j = 1, 2, \cdots, n,\ l = 1, 2, \cdots, s$

输出 x 的分类

（1）计算先验概率和条件概率

$$P(\mathrm{Y} = \mathrm{c}_k) = \frac{\sum_{i=1}^{N} I(\mathrm{y}_i = \mathrm{c}_k)}{N},\ k = 1, 2, \cdots, K$$

$$P(X^j = \alpha_{jl} \mid \mathrm{Y} = c_k) = \frac{\sum_{i=1}^{N} I(X_i^{(j)} = \alpha_{jl}, \mathrm{y}_i = c_k)}{\sum I(y_i = c_k)}$$

$$j = 1, 2, \cdots, n;\ l = 1, 2, \cdots, s;\ k = 12, \cdots, K$$

（2）对于给定的 $x = (x^{(1)}, x^{(2)}, \cdots, x^{(\mathrm{n})})^{\mathrm{T}}$,计算

$$P(\mathrm{Y}=\mathrm{c}_k)\prod_{j=1}^{n}P(X^j=x_j\mid \mathrm{Y}=c_k)$$

（3）确定 x 的类

$$y=\arg\max P(\mathrm{Y}=c_k)\prod_{j=1}^{n}P(X^{(j)}=x_j\mid \mathrm{Y}=c_k)$$

可以看到，整个朴素贝叶斯分类分为三个部分

第一部分，这个部分的任务是为朴素贝叶斯分类做必要的准备，主要工作是根据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。这一阶段的输入是所有待分类数据，输出是特征属性和训练样本。这一阶段是整个朴素贝叶斯分类中唯一需要人工完成的阶段，其质量对整个过程将有重要影响，分类器的质量很大程度上由特征属性、特征属性划分及训练样本质量决定。

第二部分，这个部分的任务就是生成分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。其输入是特征属性和训练样本，输出是分类器。这一阶段是机械性阶段，根据前面讨论的公式可以由程序自动计算完成。

第三部分。这个部分的任务是使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。

（二） 数据预处理

从附件 2 可以明显的看出，该数据集共有 9210 条数据，被人工地分为 7 个一级标签，分别是城乡建设、劳动和社会保障、教育文体、商贸旅游、环境保护、卫生计生、交通运输。首先选择了其中的留言内容和分类标签进行分析，绘制的水平条形图如下：

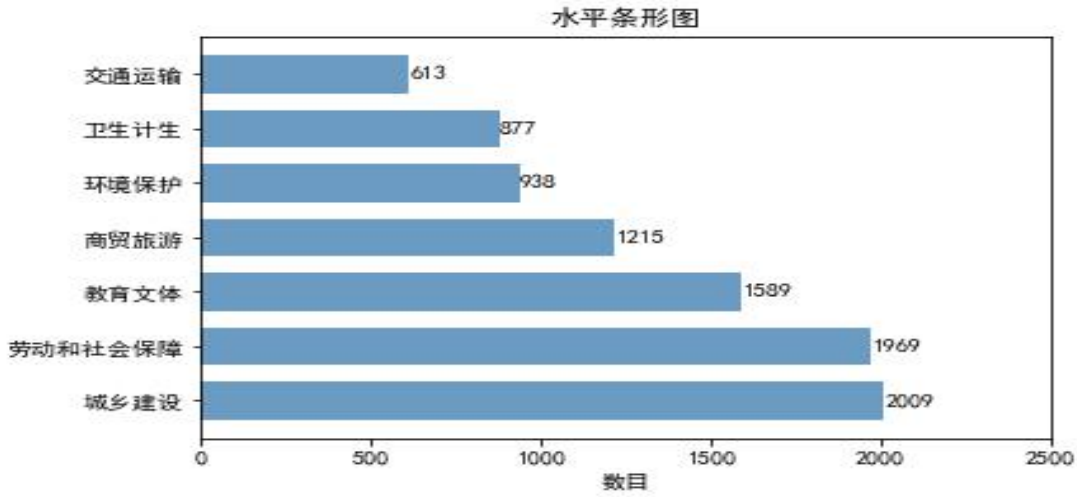


图 1 分类数据的水平条形图

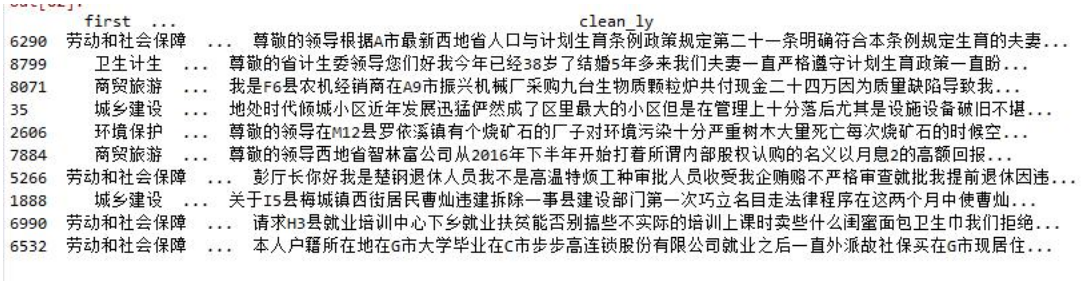
为了方便对数据进行分类处理，这样我们就能通过编程更好地实现分类。首先将 7 个类别分别编为如下表所示

表 1 分类标签

first	*first_id*
城乡建设	0
环境保护	1
交通运输	2
教育文本	3
劳动和社会保障	4
商贸旅游	5
卫生计生	6

为了得到更加“整洁”的数据，接下来将会对数据进行清洗。我们要清洗的

数据是中文的文本数据，首先需要对中文数据进行操作，例如，删除无用的符号、感叹词、特殊符号，还有一些常用的无意义的词语。同时，还有一些高频词无法反映出群众留言内容的主要意思。中文停用词数据来源于 github。我们过滤掉了 ly 中的标点符号和一些特殊符号，并生成了一个新的字段 clean_ly。接下来我们要在 clean_ly 的基础上进行分词,把每个评论内容分成由空格隔开的一个一个单独的词语。.需要注意的是，下图中显示的是进行清洗和分词后的部分结果。



	first	...	clean_ly
6290	劳动和社会保障	...	尊敬的领导根据A市最新西地省人口与计划生育条例政策规定第二十一条明确符合本条例规定生育的夫妻...
8799	卫生计生	...	尊敬的省计生委领导您们好我今年已经38岁了结婚5年多来我们夫妻一直严格遵守计划生育政策一直盼...
8071	商贸旅游	...	我是F6县农机经销商在A9市振兴机械厂采购九台生物质颗粒炉共付现金二十四万因为质里缺陷导致我...
35	城乡建设	...	地处时代倾城小区近年发展迅猛俨然成了区里最大的小区但是在管理上十分落后尤其是设施设备破旧不堪...
2606	环境保护	...	尊敬的领导在M12县罗依溪镇有个烧矿石的厂子对环境污染十分严重树木大量死亡每次烧矿石的时候空...
7884	商贸旅游	...	尊敬的领导西地省智林富公司从2016年下半年开始打着所谓内部股权认购的名义以月息2的高额回报...
5266	劳动和社会保障	...	彭厅长你好我是楚钢退休人员我不是高温特烦工种审批人员收受我企贿赂不严格审查就批我提前退休因违...
1888	城乡建设	...	关于I5县梅城镇西街居民曹灿违建拆除一事县建设部门第一次巧立名目走法律程序在这两个月中使曹灿...
6990	劳动和社会保障	...	请求H3县就业培训中心下乡就业扶贫能否别搞些不实际的培训上课时卖些什么闺蜜面包卫生巾我们拒绝...
6532	劳动和社会保障	...	本人户籍所在地在G市大学毕业在C市步步高连锁股份有限公司就业之后一直外派故社保买在G市现居住...

图 2 分词结果

（三） 朴素贝叶斯分类器

为了训练监督学习的分类器，我们首先将“ly”转变为包含数字的词向量。例如我们前面已经转换好的 tf-idf 的 features。当我们有了词向量以后我们就可以开始训练我们的分类器。分类器训练完成后,就可以对没有见过的 ly 进行预测。朴素贝叶斯分类器最适合用于基于词频的高维数据分类器。这里我们使用的是 sklearn 的朴素贝叶斯分类器 MultinomialNB，我们首先将 ly 转换成词频向量,然后将词频向量再转换成 TF-IDF 向量，还有一种简化的方式是直接使用 TfidfVectorizer 来生成 TF-IDF 向量(正如前面生成 features 的过程)，这里我们还是按照一般的方式将生成 TF-IDF 向量分成两个步骤：1.生成词频向量. 2.生成 TF-IDF 向量。最后我们开始训练我们的 MultinomialNB 分类器。当模型训练完成后,我们让它预测一些自定义的 ly 的分类。我们编写了一个预测函数 predict()。下面是根据某条留言信息预测的结果：

```
In [81]: from sklearn.model_selection import train_test_split
    ...: from sklearn.feature_extraction.text import CountVectorizer
    ...: from sklearn.feature_extraction.text import TfidfTransformer
    ...: from sklearn.naive_bayes import MultinomialNB
    ...:
    ...: X_train, X_test, y_train, y_test = train_test_split(df['cut_ly'], df['first_id'], random_state = 0)
    ...: count_vect = CountVectorizer()
    ...: X_train_counts = count_vect.fit_transform(X_train)
    ...:
    ...: tfidf_transformer = TfidfTransformer()
    ...: X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
    ...:
    ...: clf = MultinomialNB().fit(X_train_tfidf, y_train)
    ...: ##########################预测函数
    ...: def Predict(sec):
    ...:     format_sec=" ".join([w for w in list(jb.cut(remove_punctuation(sec))) if w not in stopwords])
    ...:     pred_first_id=clf.predict(count_vect.transform([format_sec]))
    ...:     print(id_to_first[pred_first_id[0]])
    ...:
    ...: Predict('为什么乡村小学的教室没有电脑，而城镇小学就有')
教育文体
```

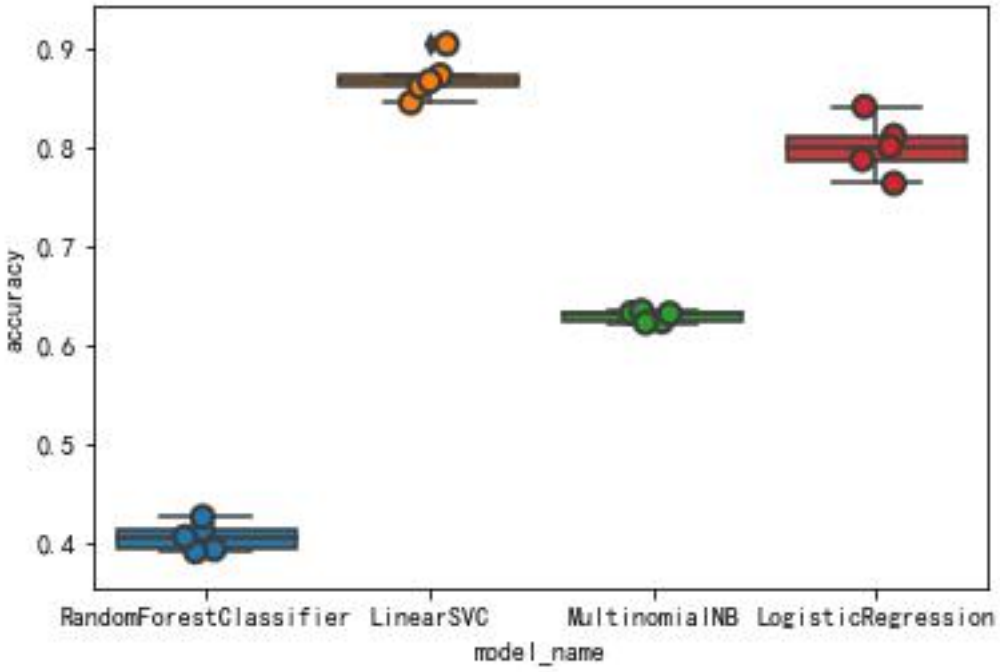
接下来我们尝试不同的机器学习模型,并评估它们的准确率，我们将使用如下四种模型:

（1）Logistic Regression

（2）(Multinomial) Naive Bayes

（3）Linear Support Vector Machine

（4）Random Forest



从箱体图上可以看出随机森林分类器的准确率是最低的，因为随机森林属于集成分类器(有若干个子分类器组合而成)，一般来说集成分类器不适合处理高维数据(如文本数据),因为文本数据有太多的特征值,使得集成分类器难以应付,另外三个分类器的平均准确率都在 60%以上。其中线性支持向量机的准确率最高。

（四）模型评价

多分类模型一般不使用准确率(accuracy)来评估模型的质量,因为 accuracy 不能反应出每一个分类的准确性,因为当训练数据不平衡(有的类数据很多,有的类数据很少)时，accuracy 不能反映出模型的实际预测精度,这时候我们就需要借助于 F1 分数、ROC 等指标来评估模型。，本节采用 $F-Score$ 方法对上述模型进行评价

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

其中，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

下面我们可以得到各个类的 $F1$ 分数。

表 2 $F-Score$ 分数

类别	precision	recall	F1-score	support
城乡建设	0.88	0.89	0.89	2009
环境保护	0.82	0.75	0.85	938
交通运输	0.78	0.7	0.74	613
教育文本	0.84	0.84	0.87	1589
劳动和社会保障	0.86	0.82	0.88	1969
商贸旅游	0.83	0.81	0.84	1215
卫生计生	0.83	0.83	0.77	877
Average	0.83	0.83	0.83	9210

由上面的分析可以得出，该模型的预测精度 83%，说明我们使用的方法比较适用于该数据集的预测与分类。

二、热点留言的提取

（一）建立模型

1、LDA 模型

基于 LDA 的文本特征提取方法作为一种概率主题模型，虽然能够获得文档之间的关系，然而在建模过程中却忽略了文档的上下文依赖关系，导致了语义信息的丢失。深度学习算法基于序列建模的方法弥补了 LDA 的不足。如吴彦文等[13]使用词嵌入对 LDA 获得的文档特征词进行表示，然后和 LSTM 编码后的文本进行拼接，用于解决数据稀疏问题。张群等人[14]通过拼接相加平均合成的词向量和经过 LDA 特征扩展的短文本向量，利用 kNN 进行分类。

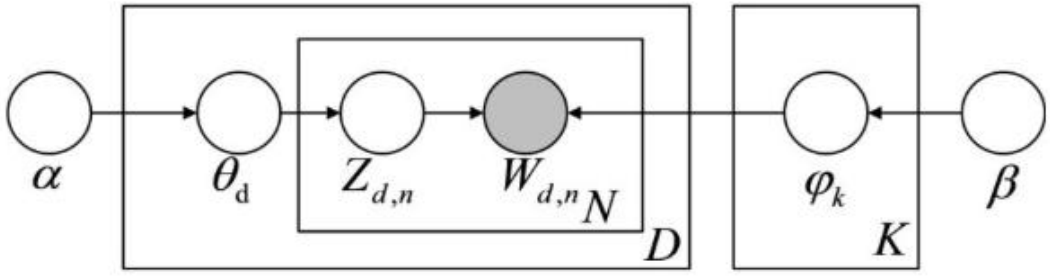


图1 LDA概率图模型

1)文档的主题先验分布服从参数为α的 Dirichlet 分布，其中文档 d 的主题分布为θd=Dirichlet(α)

2)主题中的词的先验分布服从参数β的先验分布，其中主题 k 的词分布为φk=Dirichlet(β)

3)文档 d 中的第 n 个词，从主题分布获得其主题编号分布为 zdn=multi(θd)

4)文档 d 中的第 n 个词分布 wdn 的分布为 wdn=multi(φzdn)

其中 D 是训练数据集的大小，N 是一条训练数据的大小，K 是主题数。

从模型假设可知，已知每个文档的文档主题的 Dirichlet 分布与主题编号的多项式分布满足 Dirichlet-multi 共轭，使用贝叶斯推断的方法得到文档主题的后验分布。同样已知主题词的 Dirichlet 分布与主题编号的多项式分布满足 Dirichlet-multi 共轭，通过贝叶斯推断得到主题词的后验分布。然后通过使用 Gibbs

采样的方法去获得每个文档的主题分布和每个主题的词分布。

2、word2vec 模型

文本信息需要被编码成数字信息才能进行计算处理。传统的模型使用基于 one-hot 编码的方法的 BOW（bagofwords）模型，该方法通过构建词典，统计文本的词频信息，对文本进行编码。然而，one-hot 模型的编码方法孤立了每个词，无法表达出词之间的关系，导致语义信息的丢失。而且，当词的种类过多时，还会带来维度爆炸的问题。因此，提出了词的分布式表示，将经过 one-hot 编码的词，映射到一个低维空间，并保留词之间的语义信息。word2vec 模型是目前主流词分布式表示模型，word2vec 包含两种模型，分别是 CBOW 与 Skip-Gram。CBOW 模型通过输入中心词相关的词的词向量，输出中心词的词向量。Skip-Gram 则相反，通过输入中心词的词向量，输出上下文的词向量[15]。两种模型的结构如图 2 所示：

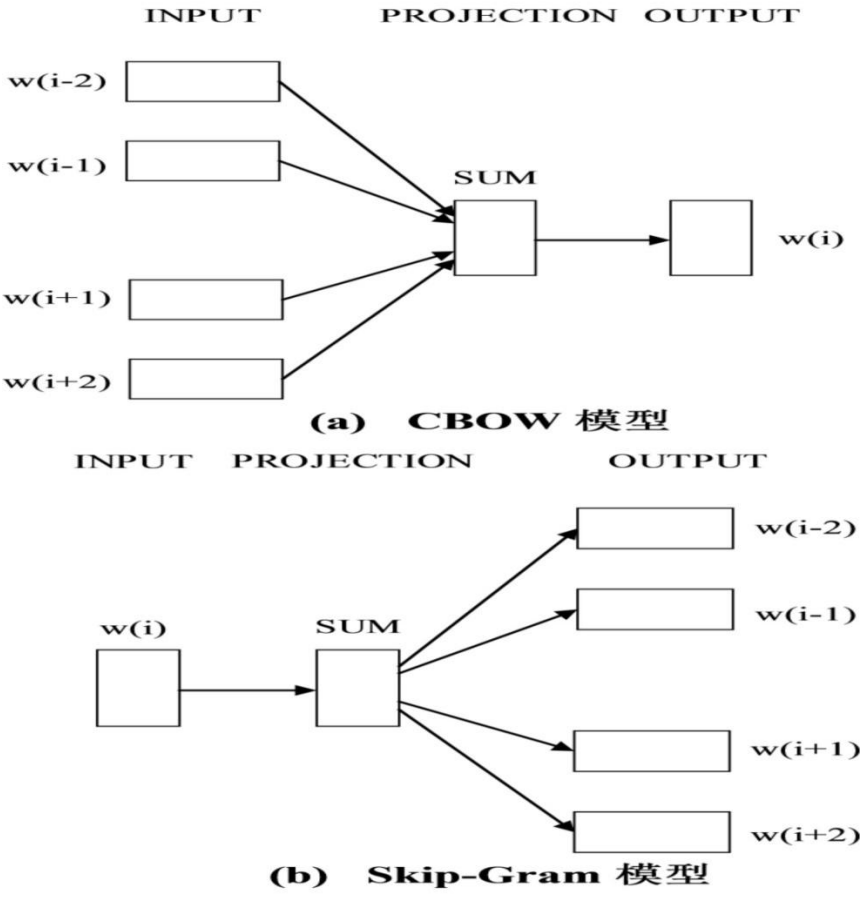


图2 word2vec模型结构

3、attention 模型

注意力机制是一种权重分配机制，通过模仿生物观察行为的过程，将内部经验和外部感觉对齐从而增强观察行为的精细度，在数学模型上表达为通过计算注意力的概率分布来突出某个关键的输入对输出的影响[16,17]。其首先被提出应用于图像特征提取过程，而后被 Bahdanau[10]等人引入到自然语言处理领域。如公式（1）所示，其中 ki(key)与 vi(value) 一一对应，通过计算 qt(query)和各个 ki 的内积，求得与各个 vi 的相似度，然后进行加权求和与归一化。

Attention(qt,K,V)=∑i=1m1Zexp(dk)vi (1)

其中 Z 是归一化因子，dk 为输入词嵌入向量的维度,起到调节因子的作用，使得内积不至于过大。

（二）模型训练

统计语料集的词频信息建立字典，对文本进行 BOW 编码，输入到 LDA 模型中，获得每条评论的主题分布 d_t=[z1,z2,...,zK]，其中 z 为每个主题编号的概率。然后找到每个主题的词分布 t_w=[w1,w2,...,wN]，其中 w 为字典中每个词的分布概率。则每条评论的主要特征词可以表示为如公式（2）所示。

D_W={z1*w1z1*w2...z2*w1z2*w2.....zK*wN} (2)

通过设置阈值，选取 D_W 中超过阈值的词作为评论文本的主要词特征。

为了更好地实现对评论文本进行聚类，本模型将主题信息融合到评论文本词向量训练的过程。使用 LDA 获得该条评论的主题信息，和原有的评论内容进行拼接，作为评论与主题信息结合后的向量表达。将前述得到的融合主题信息的评论文本作为输入，训练 CBOW 模型。假设词向量的维数为 dk，每条评论文本可以表示为一个行数是词向量的维度 dk，列数是评论文本长度 N 与主题特征词的个数 l 之和的文本矩阵 W=。其中 w 为评论文本的词向量表示，wz 为通过 LDA 获得该评论文本的主题特征的词向量表示。CBOW 模型损失函数如公式（3）、（4）、（5）所示。

L(W)=1N+l∑|c|≤s,c≠0lbP(wi|wi-s,...,wi+s) (3)

其中 wi 为某个中心词，s 为中心词左右窗口大小，P(wi|wi−s,...,wi+s)已知上下文中心词为 wi 的概率大小计算方法如下。

$$P(wi|wi-s,...,wi+s)=exp(w0Twi)\sum w\in dictexp(w0Tw) \quad (4)$$

其中 w0 是 wi 上下文词向量的均值，dict 为字典。

$$w0=12s\sum j=i-s,...,i+s,j\neq iwj \quad (5)$$

通过在评论文本中融合主题的特征信息，从而使得编码后的词向量在高维空间同类信息之间的余弦距离更小，使得相同主题评论文本在向量空间的聚类效果更好。

传统的注意力机制通过计算源端的每个词与目标端的每个词之间的依赖关系来更新训练参数，Self-Attention 机制仅通过关注自身信息更新训练参数，不需要添加额外的信息。将前述通过 CBOW 模型得到的融合主题特征的评论文本向量输入到 Self-Attention 层，通过公式（6）计算权重分布。

$$Self\text{-}Attention=softmax(W\bullet WTdk)W \quad (6)$$

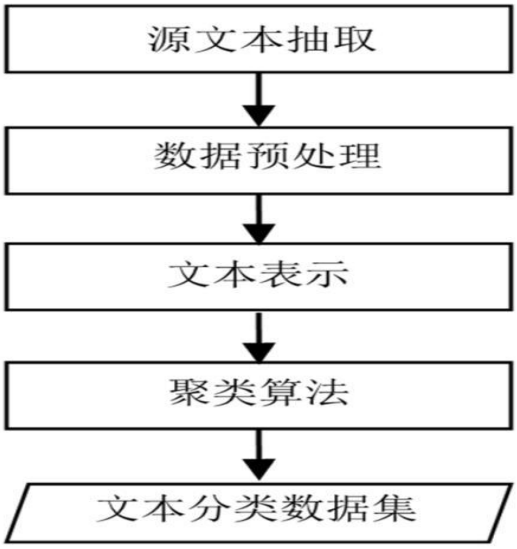
使用交叉熵作为损失函数，利用 Adam 更新网络参数。公式（7）计算评论文本向量 γ x 属于类别 yx 的概率，n_c 为类别的数目。以公式（8）为损失函数，通过迭代更新参数，最小化监督标签 gx 和预测标签之间的交叉熵。

$$yx=exp(\gamma x)\sum q=1n\_cexp(\gamma q) \quad (7)$$

$$L=-\sum n\_cgxlbyx \quad (8)$$

（三）文本数据处理

通过无监督算法(Unsupervised Algorithm)获取热点留言的方法，下图为构建文本分类数据集的流程图。



具体步骤如下：

(1)剔除附件 3 中的其他无关行，保留留言主题和留言详情。

(2)对抽取出的源文本分词、去停用词(Stopwords)、去低频词，避免停用词和低频词对有效信息造成噪声干扰。

(3)应用 K-means 算法对表示后的文本进行聚类操作。

(4)得到类别数量分别为 2、10、20 的三个文本分类数据集，将其称作 Cluster-2、Cluster-10、Cluster-20，为研究不同类别数量的文本分类数据集参与训练对模型生成摘要准确性的影响提供数据集支撑。

(5)绘制关于热点留言详情和留言主题的词云图

（四）实验数据

1、全部留言详情分析

1）、直接分词

查看留言详情的第一条文本分词结果

```
> #查看第一个文本分词
> segword[[1]]
 [1] "座落"   "在"    "市区"   "联丰"   "路米"   "兰春"   "天栋"   "一家"
 [9] "名"     "叫"    "一米"   "阳光"   "婚纱"   "艺术"   "摄影"   "的"
[17] "影楼"   "据"    "说"     "年单"   "这"     "一个"   "工作"   "室营"
[25] "业额"   "就"    "上百"   "万"     "因为"   "地处"   "居民"   "楼"
[33] "内部"   "而且"  "有"     "蛮长"   "的"     "时间"   "了"     "请"
[41] "税务局" "和"    "工商"   "局查"   "一下"   "看看"   "这个"   "一米"
[49] "阳光"   "有"    "没有"   "正常"   "纳税"   "如果"   "没有"   "应"
[57] "该会"   "怎么"  "操作"
```

2）、删除停用词

将已经分好词的所有词条中的如还、不、是等无意义的词语用自定义停用词进行删除，查看删除停用词之后的 4326 条文本分词结果。

```
> segword2[[1]]
 [1] "座落"   "市区"   "联丰"   "兰春"   "天栋"   "一家"   "一米"   "阳光"
 [9] "婚纱"   "艺术"   "摄影"   "影楼"   "年单"   "一个"   "工作"   "室营"
[17] "业额"   "上百"   "万"     "地处"   "居民"   "楼"     "内部"   "蛮长"
[25] "时间"   "请"     "税务局" "工商"   "局查"   "一下"   "看看"   "这个"
[33] "一米"   "阳光"   "没有"   "正常"   "纳税"   "如果"   "没有"   "操作"
```

3）、创建语料库

将已分完词的列表导入语料库，并进一步加工处理语料库。从结果可以看出，

语料库中存放了 4326 条留言详情的分词结果。

```
> text_corpus
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 4326
```

4）、创建文字词条

从文字词条结果图中可以看出，留言文档词条矩阵包含 4326 行和 56592 列，行代表 4326 条留言，列代表 56592 个词；而且该矩阵实际上为稀疏矩阵，其中矩阵中的非 0 元素有 476552 个，而 0 元素有 244340440 个，稀疏率达到 100%；最后，这 56592 个词中，最频繁的一个词出现在了 5 条留言中。由于稀疏矩阵的稀疏率过高，这里将剔除一些出现频率极低的词语。

```
> dtm
<<DocumentTermMatrix (documents: 4326, terms: 56592)>>
Non-/sparse entries: 476552/244340440
Sparsity           : 100%
Maximal term length: 5
Weighting          : term frequency (tf)
```

5）、降低稀疏率

保留稀疏矩阵中稀疏率在 90%以下的词条，去除稀疏矩阵中稀疏率达到 90%以上的词条。通过去稀疏化矩阵的稀疏率降低到 81%并且其中的列大幅度减少，当前矩阵只包含 130 列，即 130 个词语。之后，为了便于进一步的统计建模，需要将矩阵转化为数据框格式。

```
> dtm <- removeSparseTerms(x = dtm, sparse = 0.9)
> dtm
<<DocumentTermMatrix (documents: 4326, terms: 130)>>
Non-/sparse entries: 106953/455427
Sparsity           : 81%
Maximal term length: 3
Weighting          : term frequency (tf)
```

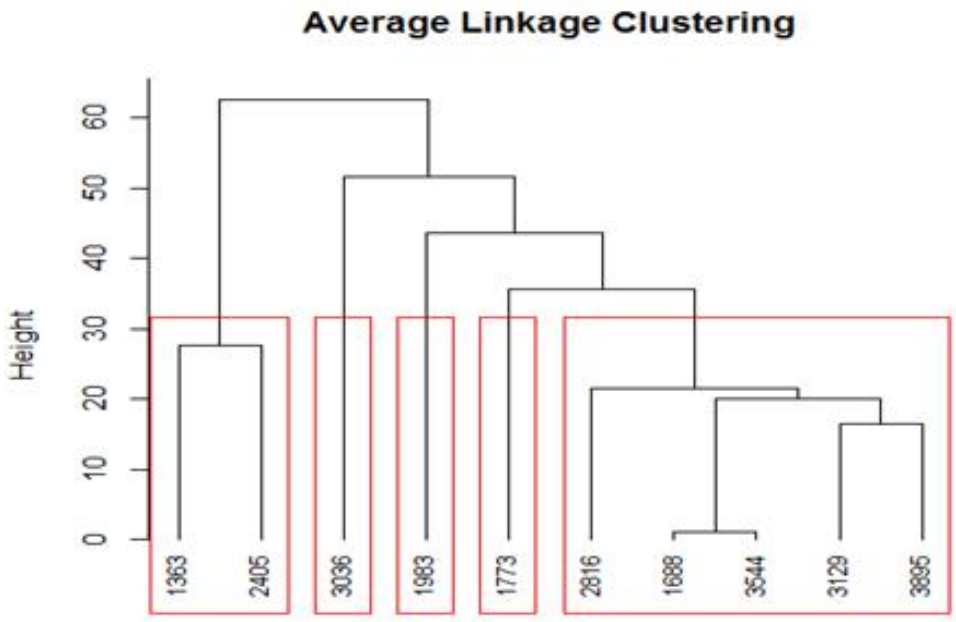
6）、创建数据框

将矩阵转化为数据框格式，结果如下，稀疏率达到 84%，矩阵的列为 113 列即词条降低到 113 个。

```
> df <- as.data.frame(inspect(dtm))
<<DocumentTermMatrix (documents: 4326, terms: 113)>>
Non-/sparse entries: 79738/409100
Sparsity           : 84%
Maximal term length: 3
Weighting          : term frequency (tf)
Sample             :
      Terms
Docs   部门 公司 领导 没有 人 问题 小区 业主 一个 政府
  1363    4    2    0    4 66   15    3    6    0    5
  1688    4    8    1   13  3    9    2   32    1    6
  1773    8    2    4    6  2    3    1    0    2   10
  1983    6   39    1   13 15    5    0    0    3   16
  2405    1    0    2    0 45    0    2    0    0    0
  2816    9   19    7    2  3   12    1   21    1    7
  3036    2    1    2    1  2    1   48   23    2    4
  3129    4   15    0    2  1    3   11   29    1    1
  3544    4    8    1   12  3    9    2   32    1    6
  3895    6   22    1    6 12    1   11   37    1    4
```

7）、进行文本分类

类平均法进行文本分类，如图可知，文本被分为 5 类，第 1363 与 2405 条留言为第一类，第 3036 条为第二类、第 1983 条为第三类、第 1773 条为第四类、第 2816、1688、3544、3129 和 3995 这 5 条留言为第五类。



8）、绘制留言详情词云图

从词云图中可以看出，关键词的热点排序依次是小区、业主、政府、公司、领导、部门、物业、问题等词。可以看出留言所关心的问题围绕小区、政府、部

门、公司这些方面张开。具体问题我们根据热点留言进行分析。

2、热度前五的留言分析

1）、热点留言指标的选取

通过附件 4 给出的所有留言信息，我们可以知道所有留言的总体点赞数与反对数，而这两个指标恰恰是热度评价指标最重要的两个指标，点赞数越高，证明留言所反映的问题是大家所关注所赞同的；反对数越高，证明留言所反映的问题是大家所关注所反对的事情。因此，通过以上分析，我们可以通过加权平均的办法，对点赞数加权 60%，反对数加权 40%，来构建热度指数，然后按热度指数从大到小选出热度排名前五的留言信息详情，进行下一步的分析。

2）、热点留言主题分析

根据热点问题留言明细表中的信息，绘制留言主题的星星词云图，从词云图中可以直观得看出，热点排名前五的词云关注的主要问题是车贷案、金毛湾、市区、配套入学、诈骗案等一系列问题，其中受关注度最高的问题是车贷案问题；其中还有一些关键词是警官、集资、保护伞等，证明可能存在监察机关以公谋私、为违法犯罪人员提供便利的行为；而市区、汇金路等关键词可知某市区小区可能存在物业管理不善等一系列问题；通过市、政府、书记、金水湾、配套入学等关键词可知，可能反应金水湾小区配套入学政策可能存在一些问题，相关政府部门应多多关注。

3）、热点留言详情分析

根据热点问题留言明细表中的信息，绘制留言详情的五星词云图，从词云图中可以直观得看出，关键词出现的次数从多到少依次是小区、金毛湾、业主、受害人、没有、物业、教育局、配套入学、车贷、涉案人员等。可知，在热点留言详情中大家最关注的还是小区的物业管理不善问题，其次是车贷问题、最后是配套入学的问题。

（五）总结

通过上述分析，我们可以看出，热度指标选取比较合理，因为热度前五的留言所包含的关键词如小区、业主、物业等与所有留言所包含的关键词信息基本一致，证明热度指标选取效果良好。而且通过热点留言词云分析可以总结出，大家最关心的热点问题有三个，第一是某高档小区物业管理不善问题；第二个是 58 车贷诈骗案问题；第三个是金毛湾小区配套入学的问题。

三、答复意见评价方案

（一）建立模型

本文提出的模型结构如图 1 所示，它由三层网络结构组成：GRU 层、初始胶囊层和主胶囊层。这一章节会对这三层网络的关键部分进行详细地介绍。

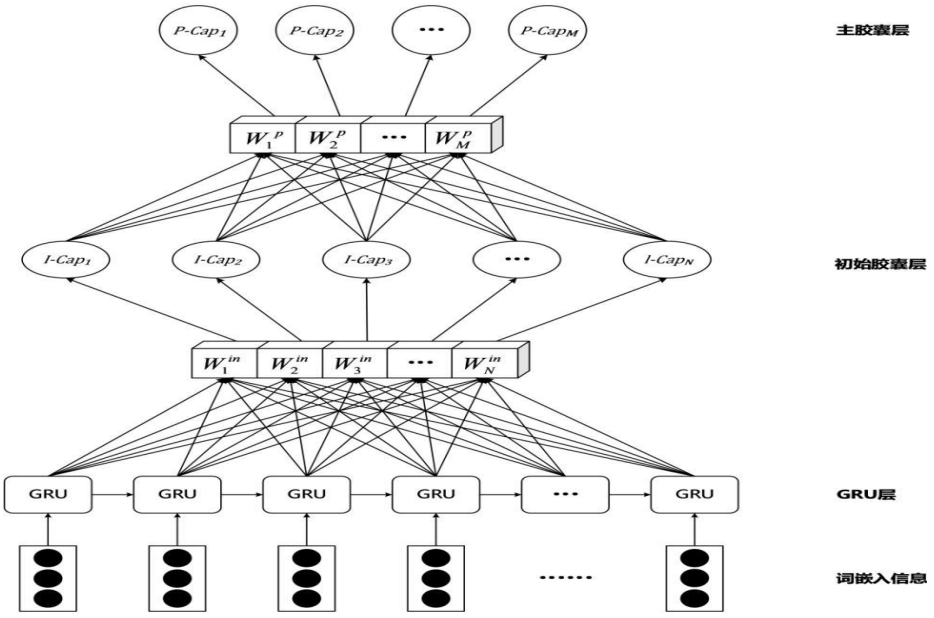


图1G-Caps模型结构图

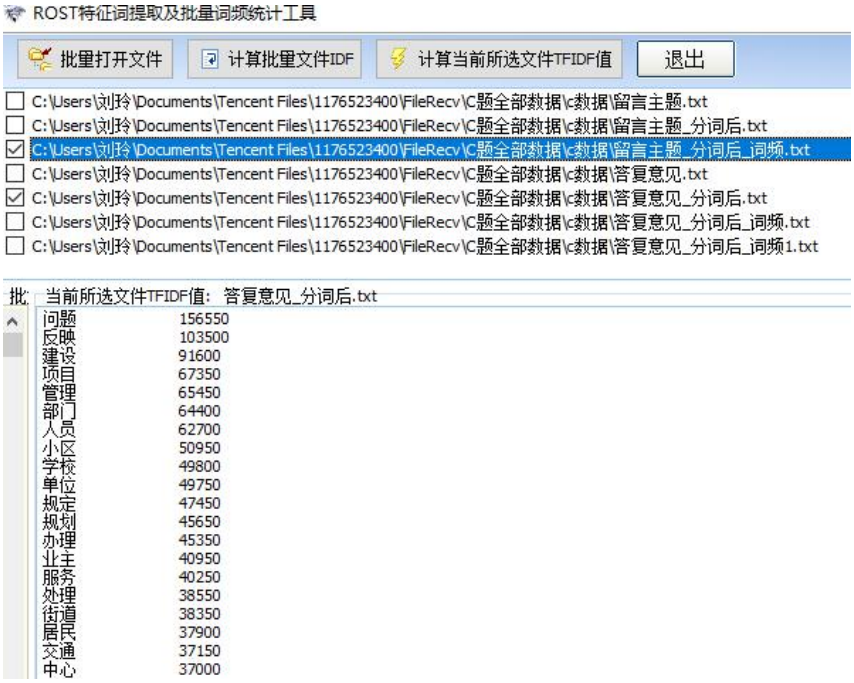
Fig. 1Structure of G-Caps model

（二）实验数据

1、分词；选取附件一中的三级分类作为自定义词表，对数据进行分词；基于 ROSTCM6 中的过滤词库对分词后的数据进行过滤，统计词频，并绘制答复意见作图如下。从图一答复意见词云图可以看出问题、反映、网友、城乡、住房、现场、教育局、方案等是答复意见中的高频词。

图一答复意见词云图

2、我们选取了附件 4 中的留言主题和答复意见数据，用 TFIDF 对选取的数据进行关键词提取，并绘制答复意见关键词的词云图如下，从中可以看出关键词出现的次数从多到少依次是问题、工作、进行、情况、网友、反映、相关、反复、我们、回复、建设、管理、部门、项目、公司等。



3、对选取的留言主题和答复意见数据的关键词，用 R 计算留言主题和答复意见之间的混淆矩阵，计算两者之间的查准率和查全率，两者之间的查准率为 0.5098039，达到了 50%以上，表明答复的相关性还不错，查全率为 0.962963。F1 度量的值达到了 66%，表明答复意见的可解释性较好。

```
> Pb
[1] 0.5098039
> Rb
[1] 0.962963
> F1
[1] 0.6666667
```

（三）总结

对选取的留言主题和答复意见数据的关键词，用 R 计算留言主题和答复意见之间的混淆矩阵，计算两者之间的查准率和查全率，两者之间的查准率为 0.5098039，达到了 50%以上，表明答复的相关性还不错，查全率为 0.962963。F1 度量的值达到了 66%，表明答复意见的可解释性较好。

四、结语

通过运用朴素贝叶斯分类算法，建立了关于网络问政平台留言内容的留言标签一级分类模型，得出 F-score 的得分为 0.83，证明我们使用的方法比较适用于该数据集的预测与分类。首先使用 LDA 获得每个留言主题的主题词分布作为该条评论信息的扩展，然后将留言主题和留言详情一起输入到 word2vec 模型，进行词向量训练，使得留言文本在高维向量空间实现同一主题的聚类，最后使用 Self-Attention 进行动态权重分配并进行分类，选出出现频率前 5 的热点问题，最后基于这 5 个热点问题绘制词云图，通过研究我们发现，热度指标选取比较合理，因为热度前五的留言所包含的关键词如小区、业主、物业等与所有留言所包含的关键词信息基本一致，证明热度指标选取效果良好。而且通过热点留言词云分析可以总结出，大家最关心的热点问题有三个，第一是某高档小区物业管理不善问题；第二个是 58 车贷诈骗案问题；第三个是金毛湾小区配套入学的问题。相关部门应该进行有针对性地处理，提升服务效率，与此同时，制作了“热点问题表.xls”和“热点问题留言明细表.xls”以供相关部门核实；最后，通过 G-Caps 模型针对相关部门对留言的答复意见，对选取的留言主题和答复意见数据的关键词，用 R 计算留言主题和答复意见之间的混淆矩阵，计算两者之间的查准率和查全率，两者之间的查准率为 0.5098039，达到了 50%以上，表明答复的相关性还不错，查全率为 0.962963。F1 度量的值达到了 66%，表明答复意见的可解释性较好。

参考文献

[1]余传明,王曼怡,林虹君,朱星宇,黄婷婷,安璐.基于深度学习的词汇表示模型对比研究[J/OL].数据分析与知识发现:1-19[2020-05-08].
[2]陈欢,黄勃,朱翌民,俞雷,余宇新.结合 LDA 与 Self-Attention 的短文本情感分类方法[J/OL].计算机工程与应用:1-8[2020-05-08].
[3]张洋,胡燕.基于多通道深度学习网络的混合语言短文本情感分类方法[J/OL].计算机应用研究:1-7[2020-05-08].
[4]廖胜兰,吉建民,俞畅,陈小平.基于 BERT 模型与知识蒸馏的意图分类方法[J/OL].计算机工程:1-8[2020-05-08].
[5]杨云龙,孙建强,宋国超.基于 GRU 和胶囊特征融合的文本情感分析[J/OL].计算机应用:1-6[2020-05-08].
[6]曾凡锋,李玉珂,肖珂.基于卷积神经网络的语句级新闻分类算法[J].计算机工程与设计,2020,41(04):978-982.
[7]徐建国,肖海峰,赵华.基于多示例学习框架的文本分类算法[J].计算机工程与设计,2020,41(04):1017-1023.
[8]薛兴荣,靳其兵.基于词典的文本极性计算及分类研究[J] ,2020(04):57-61.
[9]郭梁,王佳斌,马迎杰,朱新龙.基于模型融合的搜索引擎用户画像技术[J].科技与创新,2020(07):17-19+22.
[10]方炯焜,陈平华,廖文雄.结合 GloVe 和 GRU 的文本分类模型[J/OL].计算机工程与应用:1-9[2020-05-08].
[11]彭俊利,谷雨,张震,耿小航.融合改进型 TC 与 word2vec 的文档表示方法[J/OL].计算机工程:1-7[2020-05-08].
[12]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学报,2020,35(01):64-69.
[13]杨锐,陈伟,何涛,张敏,李蕊伶,岳芳.融合主题信息的卷积神经网络文本分类方法研究[J].现代情报,2020,40(04):42-49.
[14]杨锋.基于线性支持向量机的文本分类应用研究[J].信息技术信息化,2020(03):146-148.
[15]姚佳奇,徐正国,燕继坤,熊钢,李智翔.基于标签语义相似的动态多标签文本分类算法[J/OL].计算机工程与应用:1-8[2020-05-08].