文章编号：2095-2163(2020)10-0105-04 中图分类号：TP391 文献标志码：A

基于协同训练算法的微博垃圾评论识别

曹春萍，杨青林
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：微博上存在大量垃圾评论，这些垃圾评论会带来不良影响，如何识别垃圾评论就成为人们关注的热门。本文针对监督学习框架下大规模标注数据集难以获得和垃圾评论识别不精准的问题，提出基于半监督协同训练算法的微博垃圾评论识别方法。该方法从评论文本和评论用户两个视图构建指标体系，每一个视图用 7 种分类方法挑选出基分类器进行协同训练，以完成对微博垃圾评论的识别。实验结果表明，协同训练算法有更好的识别性能。

关键词：微博垃圾评论；半监督；协同训练；分类器

Microblogging Spam Recognition Based on Co-Training Algorithm

CAO Chunping, YANG Qinglin
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】There are a lot of spam comments on microblog, which will bring adverse effects. How to identify spam comments has become a hot topic. In this paper, a semi supervised collaborative training algorithm is proposed to solve the problem that it is difficult to obtain large-scale annotated data sets under the framework of supervised learning and the identification of spam comments is not accurate. This method constructs an index system from two views of comment text and comment user. Each view uses seven classification methods to select the base classifier for collaborative training to complete the recognition of microblog spam comments. The experimental results show that the cooperative training algorithm has better recognition performance.

【Key words】Microblogging spam comment; Semi-supervised; Co-training; Classifier

0 引 言

微博，微博客（MicroBlog）的简称，通过微博这个平台可以发布、分享和获得信息。微博能够更新 140 字之内的文字信息，实现信息的即时分享。随着微博的流行，微博评论也随之具有研究价值。微博评论有二种不同注解：

（1）针对社会问题、现象直抒胸臆发表的见解、意见，与述评或杂文类似；

（2）对于微博发表的后续评论。

通过这些评论，能够了解民情、民意，以及一些事件的具体情况和后续进展。但是微博评论中存在大量垃圾评论。垃圾评论是指一些没有任何意义或用户带有某些目性质的微博评论，这些评论是由用户随便或者故意发布的不真实的甚至是带有欺骗性质的评论信息，垃圾评论制造者通过发表评论发泄负面情绪、制造舆论或推销产品。这些垃圾评论影响读者情绪，浪费网络资源，还对面向评论的数据挖掘工作造成干扰[1]。因而，识别微博中垃圾评论显得极其重要。识别微博中垃圾评论不能只考虑单一因素，只考虑单一因素会使识别不准确。为此，本文从评论文本和评论用户两个视图选取多种特征，采取协同训练算法，以更好识别微博垃圾评论。

1 相关研究

微博垃圾评论一般分为：

（1）广告评论与超链接评论；

（2）与评论无关的信息；

（3）重复评论；

（4）与微博内容不相关的其他评论；

（5）虚假评论。

国内外学者提出了一系列方法来识别垃圾评论，主要集中在 3 个方面：电子商务、博客和微博。对于微博的评论主要依据内容来识别。例如：黄玲等提取表示微博评论的 8 个特征值向量，这 8 个特征值向量包括相似度、超链接数、评论重复数、情感词数、广告词数、句子长度、名词度、评论的被评论数，通过

基金项目：国家自然科学基金（61803264）。

作者简介：曹春萍（1968- ），女，硕士，副教授，主要研究方向：智能数据处理、个性化服务；杨青林（1994-），男，硕士研究生，主要研究方向：自然语言处理、数据挖掘。

通讯作者：杨青林 Email：1021211663@qq.com

收稿日期：2020-04-23

AdaBoost 算法在这些特征上训练出若干个弱分类器，弱分类器加权集合成强分类器来识别微博垃圾评论[2]，但是对于短小评论和虚假评论识别效果不好；李志欣等提取特殊符号的数量、URL 的数量、情感词的数量、点赞的数量、句子长度、名词比重等 6 个特征构建 AdaBoost 分类器和支持向量机分类器，通过 Co-Training算法进行协同训练，判断其是不是垃圾评论[3]，但对于虚假评论和短小评论识别结果欠好。

目前微博垃圾评论识别，仅仅从评论自身特征出发，而忽略评论者的一些特性，识别效果差。

在电子商务和博客领域，识别垃圾评论很多时候是从内容和评论者特征两方面着手。如：在电子商务领域，Jindal 等人提出垃圾评论检测，将产品评论中垃圾评论分为 3 类：虚假评论即只针对品牌的评论以及非评论；从评论内容即评论者以及被评论的产品 3 方面提取特征，采取构造二类分类器的方法对产品评论进行分类[4]；在博客领域，将垃圾评论分为二大类，对博客中显示垃圾评论用基于规则的方法识别，对隐式垃圾评论采取基于主题的特征选取和基于主题的检索模型二种方法来识别，从评论、评论者、作者、博文 4 方面出发构建特征集。

本文提出基于协同训练的微博垃圾评论识别方法，在特征提取时从评论文本和评论用户二个视图选取多种特征，从 7 种分类方法中选出合适的基分类器，通过协同训练完成对微博垃圾评论的识别。

2 基于协同训练的微博垃圾评论识别方法

2.1 识别流程

微博垃圾评论识别流程如图 1 所示。

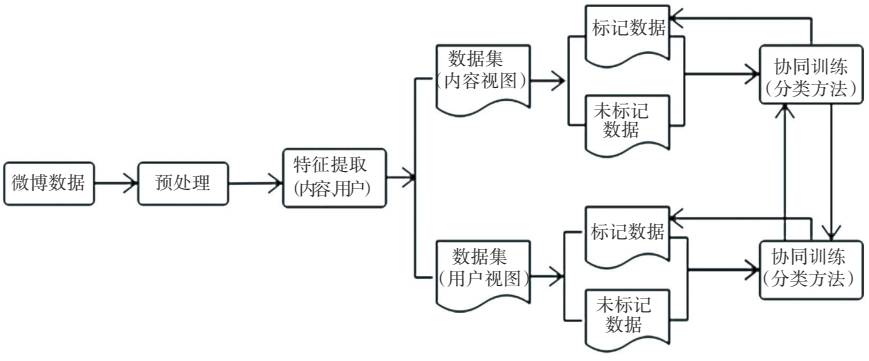

图 1 微博垃圾评论识别流程

Fig. 1 Microblog spam comment recognition process

在识别过程中，首先对数据预处理，提取评论内容和评论用户的特征，选择出合适的基分类器，在每个视图中，利用有标记数据和部分未标记数据训练分类器，然后进行垃圾评论识别。

2.2 数据预处理

数据预处理主要包括二方面：

(1)微博评论文本清理，清理微博评论中的噪声数据包括评论、回复、转发、@ 及其用户名、评论中的图片、日期等；

(2)采用 IKAnalyzer 工具对微博原文和评论分词，以便于下一步的特征提取。

2.3 特征提取

特征提取是微博垃圾评论识别流程中的重要步骤，本文从评论内容和评论者二个方面对特征做细化，以便协同训练，具体内容见表 1。

表 1 特征指标集合

Tab. 1 Characteristic index set

一级指标	二级指标	指标描述
评论内容特征	相似度	度量评论与微博的相似性。采用同义词词林计算
	超链接数	正常评论中很少包含超链接
	点赞的数量	点赞数越多，属于正常评论可能性越大
	特殊符号的数量	正常评论中很少包含特殊符号
	评论重复数	评论重复数越多，越有可能是垃圾评论
	句子长度	垃圾评论中往往句子长度较长
	名词比重	评论中名词数量/评论中词语数量×100%
评论用户特征	评论用户信用等级	评估用户的可信程度
	评论用户认证情况	用户是否为认证用户
	评论用户影响力	用户粉丝数与粉丝数关注数之和的比值
	评论用户发布微博数目	用户的微博数量

2.4　协同训练算法

协同训练的基本步骤：首先从数据集中选出部分评论，对这些评论进行标注，将数据集分为有标注数据集 L 和无标注数据集 U，从 U 中随机选取 $2p + 2n$ 的数据放入缓冲池 U_1 中；在迭代过程中，利用 L 的两个子集 L_1 和 L_2 分别训练得到分类器 h_1 和 h_2，h_1 和 h_2 分别挑选置信度最高的 $(p + n)$ 个正反例给对方，以便训练更新；将 $2p + 2n$ 个标记好的数据加入到 L 中，再次从 U 中随机选出 $2p + 2n$ 个数据放入缓冲池 U_1 中。协同训练迭代算法步骤描述如下：

输入　有标注数据集 L，无标注数据集 U；

输出　分类器 h_1 和 h_2

(1)根据有标注数据集 L 得到基于 2 个视图的已标注数据 L_1 和 L_2；

(2)从未标记数据 U 中随机选取 u 个示例放入缓冲池 U_1 中；

(3)使用训练集 L_1 训练出分类器 h_1；

(4)使用训练集 L_2 训练出分类器 h_2；

(5)h_1 在 U_1 中挑选置信度最高的 p 个正例和 n 个反例，加入到 L_2 中；

(6)h_2 在 U_1 中挑选置信度最高的 p 个正例和 n 个反例，加入到 L_1 中；

(7)将以上 $2p + 2n$ 条评论从缓冲池 U_1 移除；

(8)从 U 中随机产生 $2p + 2n$ 个未标记样本放入缓冲池 U_1；

(9)U 为空或分类器不发生改变或迭代次数达到最大值，停止迭代。

3　实验结果与分析

3.1　数据集

本文从新浪微博上抓取评论得到评论数据集，包括用户名为头条新闻发表的“雪乡宰客”微博，用户名为奢车志发表的“奔驰漏油事件”微博，用户名为央视新闻发表的“巴黎圣母院火灾”微博，用户名为腾讯体育发表的“周琦发球失误”微博。

3.2　实验评价标准

本文参考相关研究，采用 $F - measure$ 方法作为评价指标，评价标准包括召回率 R、查准率 P、准确率 $Accuracy$ 以及综合评价指标 $F - measure$ 值。建立混合矩阵，见表 2，并计算相应的评价指标值。

(1)召回率。测量被正确提取的信息的比例，公式(1)：

$$R = \frac{TP}{TP + FN}. \tag{1}$$

(2)查准率。测量提取出的信息中有多少是正确的，公式(2)：

$$P = \frac{TP}{TP + FP}. \tag{2}$$

(3)准确率。整体的正确率，公式(3)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}. \tag{3}$$

(4)综合评价指标 $F - measure$，评论 F_1，公式(4)：

$$F_1 = \frac{2PR}{P + R}. \tag{4}$$

表 2　混合矩阵

Tab. 2　Mixed matrix

类型	分类结果	
	垃圾评论	正常评论
垃圾评论	TP	FN
正常评论	FP	TN

3.3　基分类器的选择

使用相同训练集，在不同视图特征上，分别测试了随机森林算法(RF)、朴素贝叶斯算法(NB)、K 近邻分类算法(KNN)、逻辑回归算法(LR)、决策树算法(DT)、支持向量机算法(SVM)、梯度提升决策树算法(GBDT)的分类性能，以构造基分类器，结果为图 2 所示。

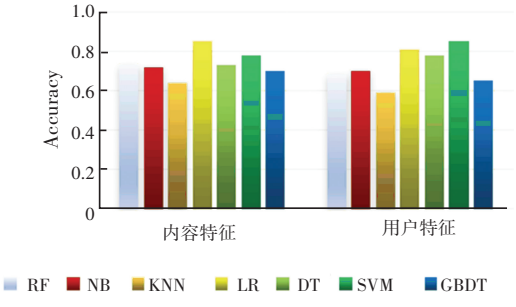

图 2　不同视图特征上分类器的总体准确率

Fig. 2　The overall accuracy of classifiers on different view features

由图 2 可知，LR 与 SVM 的性能要比剩余几种分类模型要好，因此实验中选择 LR 与 SVM 作为协同训练的二个基分类器。测试 LR 与 SVM 互相结合形成的 4 种协同方式的分类性能，结果见表 3，其中 h_1 为评论内容特征视图上的分类器，h_2 为评论用户视图上的分类器，组合 1 的分类性能最佳。因此选取组合 1 作为本文方法所选用的基分类器组合。

3.4　实验结果及分析

为了验证本文方法的有效性，设计了几组对比试验方法：

(1)采用 LR 和 DT 作为基分类器的协同训练算法；

(2)采用从相似度和其它内容特征分为两视图的协同训练算法[5]。

(下转第 111 页)