·人工智能及识别技术· 文章编号：1000-3428(2018)07-0212-07 文献标志码：A 中图分类号：TP391

基于 Co-Training 的微博垃圾评论识别方法

李志欣[1,2]，兰丹媚[1,2]，张灿龙[1,2]，唐素勤[1,2]

（1. 广西师范大学 广西多源信息挖掘与安全重点实验室，广西 桂林 541004；
2. 广西区域多源信息集成与智能处理协同创新中心，广西 桂林 541004）

摘 要： 微博上大量的垃圾评论对个人、社会，甚至是对国家都会造成不良影响。为对微博中的垃圾评论进行识别，提出基于协同训练的微博垃圾评论识别方法。定义一种基于规则的识别方法过滤出显式垃圾评论，剩余的评论归为相关评论，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，判断其是否为垃圾评论，以提高分类精度，节省样本标注工作。实验结果表明，与基于相似度计算的垃圾评论识别方法、基于评论多特征的垃圾评论识别方法相比，该方法具有较好的识别效果。

关键词： 微博垃圾评论；协同训练；同义词词林；支持向量机；相似度计算

中文引用格式： 李志欣，兰丹媚，张灿龙，等. 基于 Co-Training 的微博垃圾评论识别方法[J]. 计算机工程，2018，44(7)：212-218.

英文引用格式： LI Zhixin，LAN Danmei，ZHANG Canlong，et al. Recognition method of microblogging spam comment based on Co-Training[J]. Computer Engineering，2018，44(7)：212-218.

Recognition Method of Microblogging Spam Comment Based on Co-Training

LI Zhixin[1,2]，LAN Danmei[1,2]，ZHANG Canlong[1,2]，TANG Suqin[1,2]

（1. Guangxi Key Lab of Multi-source Information Mining and Security，Guangxi Normal University，Guilin，Guangxi 541004，China；
2. Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing，Guilin，Guangxi 541004，China）

【Abstract】 A large amount of spam comments on microblogging will have an adverse effect on individuals，society，and even the country. In order to identify junk comments in microblogging and reduce junk comments，a microblogging junk comment review method based on collaborative training is proposed. Define a rule-based recognition method to filter out explicit spam comments. The remaining comments are categorized as related comments. The AdaBoost classifier and Support Vector Machine(SVM) classifier are constructed. The Co-Training algorithm is used for collaborative training to determine whether it is a spam comment or not，classification accuracy，saving sample labeling work. Experimental results show that compared with the spam comment recognition method based on similarity calculation and the multi-features comment spam recognition method，this method has a better recognition effect.

【Key words】 microblogging spam comment；collaborative training；synonym word forest；Support Vector Machine(SVM)；similarity computation

DOI： 10.19678/j.issn.1000-3428.0047259

0 概述

微博，即微博客，是一个公众信息传播、分享和获取的平台，能够实时更新 140 字左右的文字信息，保证信息传播和分享的及时性[1]。随着微博平台应用的推广，某些用户为了某种利益或者情绪在该平台发表了许多负面影响的垃圾评论，不但污染了微博环境，而且还浪费了网络资源，有时甚至对读者的心情和情绪产生严重的影响，降低了面向微博评论数据分析和挖掘工作的有效性。因此，对微博中的垃圾评论展开研究和识别显得尤为重要。

垃圾评论是指用户发布的与博文无关的、或没有意义的以及蓄意发表的评论信息，它不但影响读者的情绪和心情，而且还有可能会影响社会的稳定

基金项目： 国家自然科学基金(61663004，61363035，61365009)；广西自然科学基金(2016GXNSFAA380146，2017GXNSFAA198365)；广西多源信息挖掘与安全重点实验室主任基金(16-A-03-02)；广西学位与研究生教育改革专项课题(JGY2015031)。

作者简介： 李志欣(1971—)，男，教授、博士，主研方向为数据挖掘、图像理解、机器学习；兰丹媚，硕士研究生；张灿龙，副教授、博士；唐素勤，教授、博士。

收稿日期： 2017-05-18 **修回日期：** 2017-08-04 **E-mail：** lizx@gxnu.edu.cn

性。各种各样的人想通过制造微博评论宣泄自己的负面情绪、制造舆论、攻击他人和推销产品。他们大多夸大或缩小事实,带有负面性、商业性甚至是政治性,这些垃圾制造者发布了大量的和原微博博文无关的评论。对微博中垃圾评论进行识别不能够只考虑评论和微博之间的相关程度,因为单一的因素考虑会增加垃圾评论的误判率。为此,本文分析和总结微博评论特征,采用协同训练算法构造 2 种不同的分类器进行协同训练,以识别微博垃圾评论。

1　相关工作

1.1　微博垃圾评论的研究现状

微博垃圾评论是指一些没有任何意义或用户带有某些目的性质的微博评论,这些评论是由用户随意或是蓄意发布的不真实的甚至是带有欺骗性质的评论信息[2]。微博垃圾评论大体可以分为两大类[3-4]:一类是包括广告评论、超链接评论、垃圾字符评论、重复评论等显式垃圾评论;另一类是指其他用户发布的评论与博主发布的微博内容不相关的隐式垃圾评论。

早期识别微博垃圾评论的方法主要两种:一种是基于验证码、链接评论加上 nofollow 标签以及评论加入审核机制等人工识别方法;另一种是基于关键词、链接数量、相关度阈值等方法来识别过滤垃圾评论的自动识别方法等。

目前,研究人员在垃圾评论识别上做了大量的努力,识别垃圾评论的方法已经逐步趋于完善。文献[5]提出垃圾评论检测,随后进行了详细的介绍和分析[6];文献[7]改进相似度公式,并对博客中的垃圾评论进行了 K 轮识别;文献[8]提出一种表示微博评论的特征值向量,通过 AdaBoost 算法在这些特征上训练出强分类器的方法来识别微博垃圾评论;文献[9]提出一个无监督的语言模型(LM),该模型通过计算任意一对评论之间的相似度来检测垃圾评论;文献[10]对于外显型垃圾评论采用基于规则的方法识别,而对于内隐垃圾评论则通过主题特征的方式来识别;文献[11]对于垃圾信息通过 Twitter 中的 Hashtag 特征,首先选用 k-NN 算法过滤掉明显的垃圾信息,然后利用最大期望(EM)算法对剩余的难于识别的垃圾信息进行识别;文献[12]针对给定话题的垃圾微博过滤问题,提出基于朴素贝叶斯分类器和最大期望值算法的半监督中文垃圾微博过滤模型。

1.2　协同训练算法

协同训练是一种应用广泛的半监督机器学习方法,最初由文献[13]于 1998 年提出。协同训练算法的思想是对于给定的样本集,根据样本的 2 个冗余视图构造出 2 个分类器,即要有 2 个相互独立的数据集来解决同一问题,并且两分类器在各自视图上能够独立地学习到一个强分类器。

协同训练算法主要有以下 3 种:

1)Co-Training 算法

Co-Training 最早是由文献[14]应用于一种消除词的歧义的方法上。Co-Training 的命名是 2 个分类器共同协调训练,即 Common-Training,在 2 个分类器的分别迭代训练中,取出自身置信度高的前 k 个无标记样本按分类结果进行标记,并放到对方分类器中进行继续训练,一直到训练结束退出算法[15]。

在 Co-Training 算法中,采用 2 个分类器进行协同训练,算法首先在训练样本的 2 个充分冗余视图上训练出 2 个分类器。然后在每次迭代过程中,轮流选择一个分类器为主分类器,另一个分类器为辅助分类器,辅助分类器选择少量标记置信度高的未标记样本标记后用于主分类器的强化学习。在协同训练算法中,保证 2 个分类器的差异性非常重要,否则算法会退化成 Self-Training 算法。为了保证分类器的差异性,协同训练算法规定训练样本需要 2 个充分冗余视图(Sufficient and Redundant Views),即:特征集能够分成 2 个特征集合;对于每一个子特征集都足以训练出一个较好的分类器;2 个子特征集合是相互独立的。

Co-Training 算法取得成功的关键因素是有 2 个充分冗余数据集,在此基础上构造 2 个完全不同又能相互独立地解决同一问题的分类器,两分类器可以互相将置信度较高的无标记样本进行标记来训练对方,从而可以将两分类器变得越来越强。

2)Tri-Training 算法

由于 Co-Training 算法对数据集有严格的要求,文献[16]提出一种运用 3 个分类器并采用投票的方式解决无标记样本的置信度估计问题,放松了对训练数据集的要求和对协同训练算法的执行条件的算法,即 Tri-Training 算法。该算法将 3 个分类器划分为 1 个主分类器和 2 个辅分类器,并采用投票的方式来选择主分类器再一次训练的数据样本,使分类效果有较大提高。如果 2 个辅分类器对未标记样本预测的结果相同但是是错误的,这样会引入噪声标记。

3)Co-Forest 算法

随机森林(Random Forest)是包含多个决策树的分类器[17]。文献[18-19]借助随机森林提出 Co-Forest 算法,思路是通过随机森林算法构建 M 个初始分类器,在算法迭代过程中,使用 $M-1$ 个分类器为辅分类器,剩下的分类器为主分类器。具体地,使 $M-1$ 个辅分类器对没有标记的样本进行投票,当这个样本的票数高于给定的阈值时,将该样本标记为置信度高的样本,对其他没有参与投票的辅分类器进行进一步的强化训练。为了减少噪声数据的引

入,该算法还在对未标记数据进行投票前,给每个未标记样本设置一个初始的置信度。

2 基于协同训练的微博垃圾评论识别方法

2.1 系统框架

基于协同训练的微博垃圾评论识别方法系统框架如图 1 所示,实验所涉及的微博数据主要有该条微博的发布者信息、该条微博的博文和该条微博的评论。

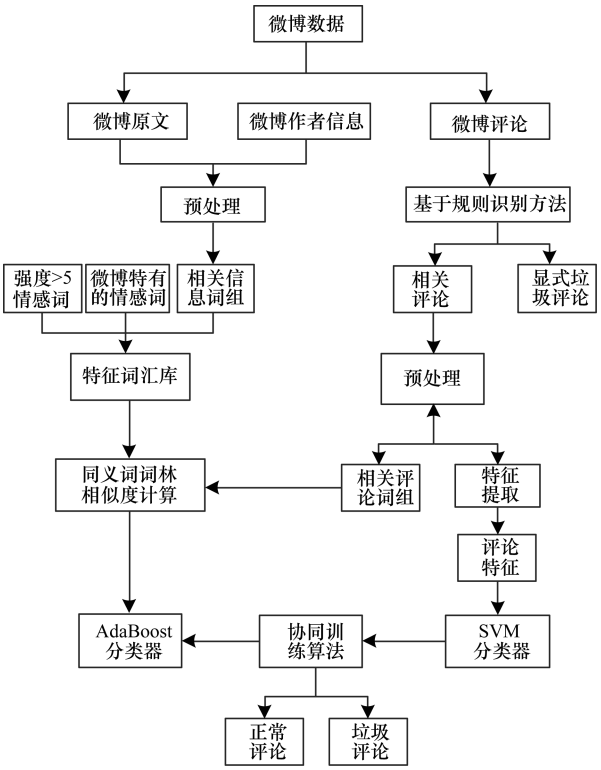


图 1 基于协同训练的微博垃圾评论识别方法框架

对微博作者信息和微博原文进行预处理后,得到文本相关信息词组,即微博作者信息词组和微博原文词组,再和情感词汇本体中情感强度大于5 的情感词及微博特有词汇(“蓝瘦”“香菇”等)共同组成特征词汇库。其中,微博作者信息来自微博发布者首页的微博认证信息,情感词来自大连理工大学信息检索实验室的情感词汇本体。对于微博评论,通过定义的基于规则方法筛选出显示垃圾评论,对于剩下的相关评论,一方面通过预处理并分词后得到的相关评论词组和构建的特征词汇库,通过文献[20]提出的基于同义词词林相似度计算方法计算出结果,送入 AdaBoost 分类器来判断其是否为垃圾评论;另一方面,将微博评论进行特征提取,作为特征向量训练 SVM 分类器,用分类器来判断其是否为垃圾评论;最后将 2 个分类器通过基于微博垃圾评论的 Co-Training 算法进行协调训练,将训练好的模型来判断测试数据集的评论是否为垃圾评论,得到最终的结果。

2.2 基于规则的识别方法

本文认为一条评论中出现的特殊符号、特殊字符、超级链接及随机字符等占该条评论字长的 50% 及以上的评论,就定义为是显式垃圾评论,包括以下几种:用户在评论中含有大量的特殊字符或特殊符号如“ * ●”;用户在评论中正常评论篇幅较多字体很小或不正常,但是垃圾评论字体设置为正常,不仅美观还突出了垃圾评论,如“宝宝不哭,宝宝加油,想做加我 qq:2954683520”;用户的评论中文字较少,含有大量的超级链接等。这些评论不仅给其他用户造成不便,还给很多广告发布者或不法分子可乘之机。

2.3 数据预处理

微博评论数据,通过基于规则的识别方法识别出显示垃圾评论后,对于剩下的相关评论要进行数据预处理,主要包括文本清理、分词去停用词及微博特征提取。

1)微博评论文本清理[21]。对于微博评论数据,要分析其可能包含的噪声数据及清理噪声数据。去除内容有“评论”“回复”“转发”等、@ 及其用户名、评论中的图片、日期等和本文研究无关的东西。

2)中文分词和停用词处理。本文所采用的分词及去停用词的工具是 IKAnalyzer。

3)微博评论特征。本文通过对微博垃圾评评论的分析以及本文的实验需要,引入 6 个特征来表示微博评论特征并进行特征提取:即特殊符号的数量、URL 的数量、情感词的数量、点赞的数量、句子长度、名词比重等。提取出微博的 6 个特征之后,还要建立文档的向量模型,将文本数据转换成计算机可以处理的结构化数据。

2.4 协同训练算法

本文提出基于 Co-Training 的协同训练的微博垃圾评论识别算法,一方面通过评论词组与特征词汇库进行相似度计算构造 AdaBoost 分类器进行识别,设为方法 A;另一方面通过微博评论的特征提取构造 SVM 分类器进行识别,设为方法 B;最后设定一种垃圾评论识别的 Co-Training 协同训练算法,将 AdaBoost 和 SVM 两分类器进行协同训练。

方法 A 基于同义词词林相似度计算构造的 AdaBoost 分类器的流程如图 2 所示。

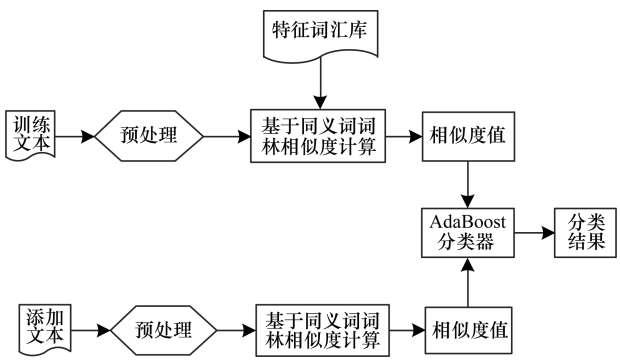


图 2 基于同义词词林相似度计算的 AdaBoost 分类器流程

设训练数据为 X,人工标记垃圾评论为 -1,正常评论为 1。假设每组实验数据的训练数据为 m 个,训练前先给每个数据初始化权重为 $D_1(i)=1/m$。当进行第一次迭代时,将基于同义词词林计算的相似度结果大于等于 0.5 的评论设为正常评论,小于 0.5 的评论设为垃圾评论,将这一分类方法设置为第一次迭代的弱分类器。

AdaBoost 伪代码如下:

Give: $(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)$ where $x_i\in X, y_i\in Y=\{-1,1\}$

初始化 $D_1(i)=1/\mathrm{m}$

输出　$h_{\mathrm{fin}}(x)=\mathrm{sign}\sum_{t=1}^{T}\alpha_t h_t(x)$

For　$\mathrm{t}=1,2,\cdots,\mathrm{T}$:

1. Train weak learner using distribution $\mathrm{D_i}$

2. Get weak hypothesis $\mathrm{h_t}:\mathrm{X}\rightarrow\{-1,1\}$ with error $\varepsilon_t=\mathrm{Pr_{i\sim D_t}}[\mathrm{h_t(x_i)}\neq\mathrm{y_i}]$

3. Choose: $\alpha_t=\frac{1}{2}\ln[(1-\varepsilon_t)/\varepsilon_t]$

4. Update: $\mathrm{D_{t+1}(i)}=\frac{\mathrm{D_t(i)}}{\mathrm{Sum(D_t)}}\times\begin{cases}\mathrm{e^{-\alpha_t}}, \mathrm{h_t(x_i)=y_i}\\ \mathrm{e^{\alpha_t}}, \mathrm{h_t(x_i)1y_i}\end{cases}$

Where $\mathrm{Sum(D_t)}$ is a normalization factor (chosen so that $\mathrm{D_{t+1}}$ will be a distribution)

Output the final hypothesis: $\mathrm{h_{fin}(x)=sign\sum_{t=1}^{T}\alpha_t h_t(x)}$

方法 B 支持向量机(SVM)分类方法的流程如图 3所示,其中 SVM 方法采用文献[22]开发设计的 LibSVM 工具。LibSVM 的参数设置如下:SVM 类型选择 C-SVC;n 交叉验证中设 $n=10$。

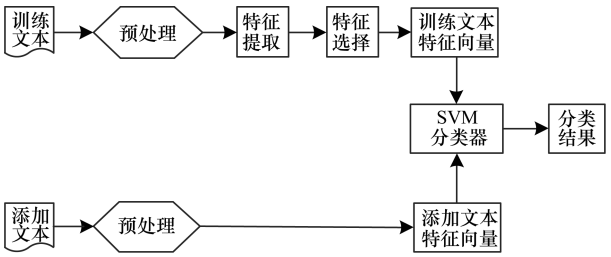


图 3　基于支持向量机分类方法的流程

用训练数据分别训练好 AdaBoost 分类器和 SVM 两分类器后,在基于 Co-Training 协同训练的过程中分别用两分类器预测添加数据为正常评论还是垃圾评论,如图 4 所示。对于获取的每组微博评论中设置的 30% 有标注训练数据 X,通过方法 A,即基于同义词词林相似度计算构造的 AdaBoost 分类器,得到分类器 $\mathrm{C_a}$;同时通过方法 B,得到分类器 $\mathrm{C_b}$,两分类器分别对同一未标注的添加数据进行分类预测(其中添加数据来自实验数据中随机选取 60% 无需标记的协同训练添加数据),取两预测的分类结果中置信度较高的前 h 个结果,由于训练数量越多,训练的分类器性能越好,但选择置信度越高,引入的噪声会更多。实验结果证明,当 $h=50$ 时可取得较好的实验结果,将前 h 个结果标注给该数据,并将该数据添加到有标注训练数据中。

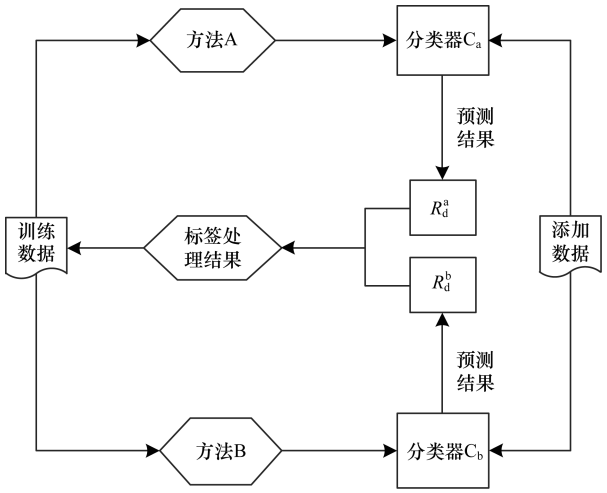


图 4　基于 Co-Training 协同训练的微博垃圾评论识别流程

基于微博垃圾评论的 Co-Training 算法训练过程如下:

输入　有标注样本集 X,无标注样本集 U

输出　样本在分类器上的分类结果 $Y=\{-1,1\}$

迭代下列步骤:

1. 从数据集 U 中取出 K = 500 个数据作为临时集合 U′。

2. 在数据集 X 使用方法 A 得到分类器 $\mathrm{C_a}$。

3. 在数据集 x 使用方法 B 得到分类器 $\mathrm{C_b}$。

4. 分别使用 $\mathrm{C_a}$ 和 $\mathrm{C_b}$ 进行下列步骤:

对于 U′中的数据进行分类预测,分别得到结果 $\mathrm{R_i^a}$ 和 $\mathrm{R_i^b}$;

选置信度较高的前 h = 50 个数据 i 及 i 标注结果添加到数据集中 X,更新 X。

5. 从 U 中补充数据到 U′,直到分类器的 F 测试值收敛或 U′中无数据为止。

基于微博垃圾评论的 Co-Training 算法测试过程如下:

输入　测试样本 S

输出　样本 S 的分类结果 $Y=\{-1,1\}$

1. 对测试数据 $\mathrm{S_i}\in\mathrm{X}$ 使用方法 A 通过分类器 $\mathrm{C_a}$ 得到分类结果 $\mathrm{R_i^a}$。

2. 对测试数据 $\mathrm{S_i}\in\mathrm{X}$ 使用方法 B 通过分类器 $\mathrm{C_b}$ 得到分类结果 $\mathrm{R_i^b}$。

3. 判断 $\mathrm{S_i}$ 分类结果:若 $\mathrm{R_i^a=R_i^b}$,则输出分类结果;若 $\mathrm{R_i^a\neq R_i^b}$,则输出置信度较高的分类结果。

在训练过程中参数设置为:$k=500$,$h=50$;在测试过程中,当两分类器判断结果一致时,输出该结果,当两分类器判断结果不一致时,以置信度高的判断结果为准。

3　实验结果与分析

3.1　实验数据集

本文从新浪微博上获取所需的评论数据集,包括用户名为头条新闻发表的“山东问题疫苗”微博,用户名为王宝强发表的主题为“离婚”微博,用户名为人民日报发表的主题为“高考志愿遭室友篡改”微博,用户名为中国新闻周刊发表的主题为“买

4 700 万衣物再退货赚 49 万”微博,用户名为央视新闻发表的主题为“里皮接任中国国家足球队主教练”微博,分别获取到 6 452、12 601、4 219、7 513、7 033 条,实验数据集如表 1 所示。

表 1 微博评论的实验数据集

编号	主题	获取总评论数	训练数据		添加数据	测试数据	
			正常评论	垃圾评论		正常评论	垃圾评论
D1	山东问题疫苗	6 452	1 104	832	3 871	572	73
D2	王宝强离婚声明	12 601	2 181	1 599	7 561	959	301
D3	室友篡改高考志愿	4 219	759	507	2 531	364	57
D4	买衣物再退货赚钱	7 513	1 350	904	4 508	568	183
D5	里皮任中国足球主教练	7 033	1 248	862	4 220	552	151

3.2 实验环境

本文实验采用的编程语言是 Python,运行环境为 Linux 系统,Python2.7.10,1.6 GHz Intel Core i5,8 GB 1 600 MHz DDR3。

3.3 实验评价指标

本文采用的评价指标是 F-measure 方法,该方法是查全率(Recall)和查准率(Precision)的组合,其值越高表示算法性能越好,是一种非平衡分类数据问题的有效评价准则。建立如表 2 所示的混合矩阵,并根据该矩阵计算出相应的评价指标值。

表 2 微博评论混合矩阵

实验判断	真正为正常评论	真正为垃圾评论
判断为正常评论数	TP	FP
判断为垃圾评论数	FN	TN

在表 2 中,TP 为真正例,TN 为真反例,FP 为假正例,FN 为假反例。

1)查全率也称召回率,测量被正确提取的信息的比例,度量分类器的完整性或灵敏度。

正常评论查全率 R_1 为:

$$R_1 = TP/(TP + FN) \tag{1}$$

垃圾评论的查全率 R_2 为:

$$R_2 = TN/(TN + FP) \tag{2}$$

2)查准率也称准确率,用来测量提取出的信息中有多少是正确的,度量一个分类器的正确性。

正常评论的查准率 P_1 为:

$$P_1 = TP/(TP + FP) \tag{3}$$

垃圾评论的查准率 P_2 为:

$$P_2 = TN/(TN + FN) \tag{4}$$

3)综合评价指标 F-measure。

正常评论 F_1 为:

$$F_1 = (2 \times P_1 \times R_1)/(P_1 + R_1) \tag{5}$$

垃圾评论 F_2 为:

$$F_2 = (2 \times P_2 \times R_2)/(P_2 + R_2) \tag{6}$$

3.4 结果分析

为了验证本文提出的基于 Co-Training 的协同训练的微博垃圾评论识别算法的可行性和有效性,做以下 2 个对比实验:

方法 1 采用文献[4]中利用相似度公式对评论进行 K 轮识别的垃圾评论识别方法;

方法 2 采用文献[7]中基于 LDA 提取主题特征集用 SVM 分类的垃圾评论识别方法。

方法 1 是基于网络常用词和博文主题词的相似度计算的垃圾评论识别方法,方法 2 是基于评论多特征的垃圾评论识别方法。

根据上述方法获取到 5 个评论数据集,每个数据集随机选取 30% 对评论进行人工标记作为训练集,随机选取 60% 无需标记作为协同训练的添加集,剩下的 10% 人工标记作为测试集。不需要进行训练的方法 1 和方法 2 两个对比实验采用同样的测试集。在方法 1 和方法 2 中,实验需要设置阈值或者不良影响因子等参数,因此,本文 2 个对比实验结果都取整体精确率最高的实验结果,方法 1 和方法 2 这 2 个对比实验以及本文的实验结果如表 3 所示。

表 3 3 种方法的实验结果对比

编号	方法 1				方法 2				本文方法			
	正常评论		垃圾评论		正常评论		垃圾评论		正常评论		垃圾评论	
	判断正确	判断错误	判断正确	判断错误	判断正确	判断错误	判断正确	判断错误	判断正确	判断错误	判断正确	判断错误
D1	537	35	50	23	550	22	58	15	567	5	68	5
D2	866	93	174	127	894	65	252	49	917	42	285	16
D3	342	22	45	13	344	20	42	16	352	13	53	5
D4	519	49	130	53	533	35	150	33	542	27	170	13
D5	509	43	101	50	526	26	115	36	535	18	133	18

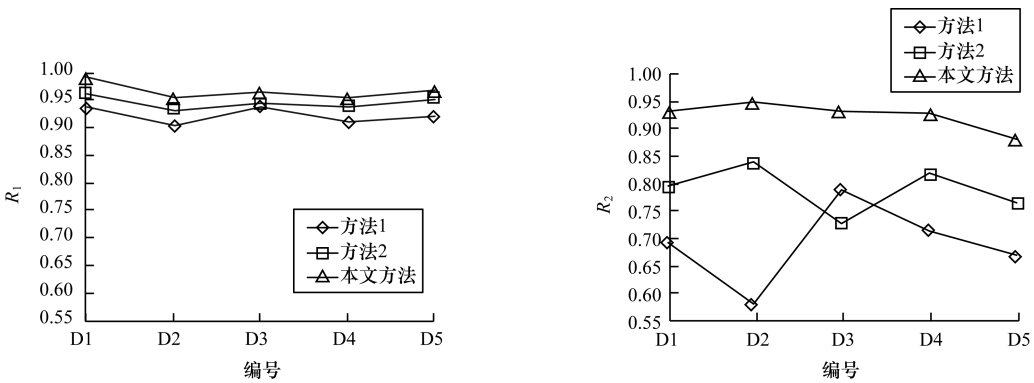
从表 3 可以看出：

方法 1 对 5 个评论数据的测试集进行评论分类，正确识别出了大部分的正常评论和垃圾评论。但是仍有一部分正常评论被识别为垃圾评论，一部分垃圾评论被识别为正常评论，即仍有一部分的正常评论与博文的相似度较低，一部分的垃圾评论与博文的相似度较高。采用基于博文和评论的相似度计算方法容易把此类评论识别错误，从而导致分类器的分类结果不佳。

方法 2 对 5 个评论数据的测试集进行评论分类，先采用基于规则的方法识别出垃圾关键词、重复评论、超链接等显式垃圾评论，然后去除评论长度小于5 字符但不包含特定情感词典中的词或短语的隐式垃圾评论，最后对于剩下的相关评论包含评论、评论者、作者以及博文信息的主题特征集等利用 SVM 在分类特征集上进行评论分类。相对于方法 1，方法 2 不仅考虑了评论与博文的相似度，而且还考虑了垃圾关键词、超链接等显式垃圾评论特征以及博文作者、评论者的信息特征，识别效果有较大提升。

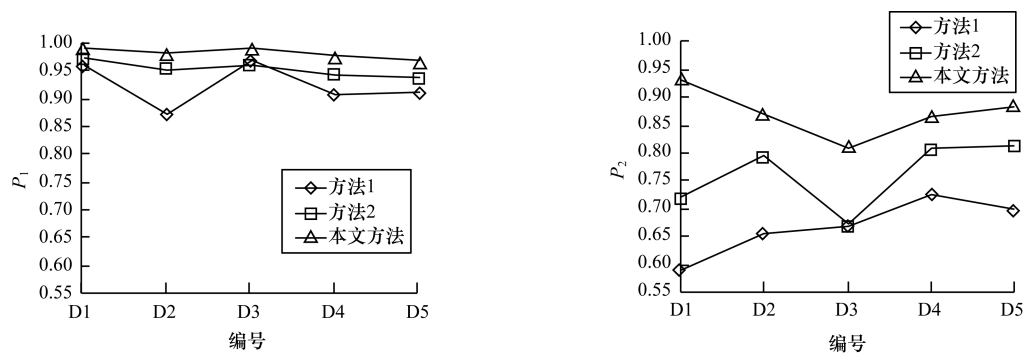
本文方法对 5 个评论数据的测试集进行评论分类，首先通过基于规则识别方法识别出显式垃圾评论，然后对于剩下的相关评论，一方面与博文、情感强度词、微博特有词汇、作者信息词组等组成的特征词汇库进行基于同义词词林的相似度计算用 AdaBoost 分类器进行分类，另一方面进行特征提取，训练 SVM 分类器，再利用添加数据集对两分类器进行 Co-Training 协同训练。本文实验方法不仅有基于相似度计算，还基于多特征进行垃圾评论识别，综合了方法 1 和方法 2 的思想，实验结果证明，识别效果优于方法 1 和方法 2。

本文方法和 2 个对比方法在 5 个测试数据集上的正常评论和垃圾评论的查全率以及查准率如图 5、图 6 所示。可以看出，本文方法的查全率均高于正常评论和垃圾评论的查全率，说明本文分类方法的完整性和灵敏度高于 2 个对比实验；且本文方法的查准率均高于正常评论和垃圾评论的查准率，说明本文分类方法的正确性比 2 个对比实验要高。正常评论和垃圾评论的 F-measure 值如图 7 所示，正常评论的综合评价指标 F_1 值和垃圾评论的综合评价指标 F_2 值都比方法 1 和方法 2 的 F 值要高，说明本文所提出的基于 Co-Training 的协同训练的微博垃圾评论识别模型不仅可行而且是有效的。



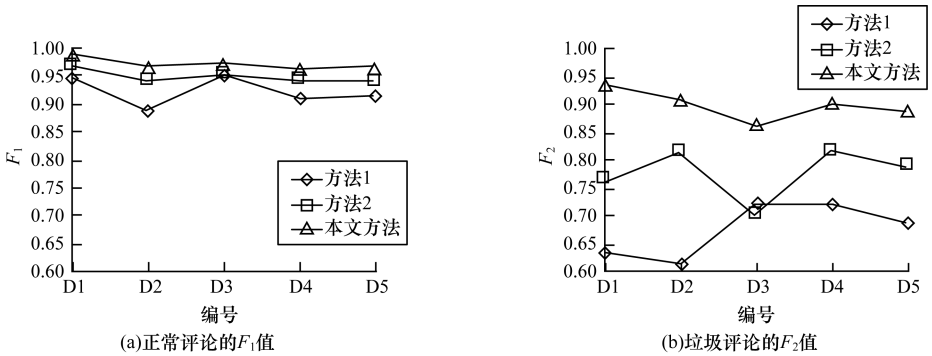
(a)正常评论的查全率　　(b)垃圾评论的查全率

图 5　正常评论和垃圾评论的查全率



(a)正常评论的查准率　　(b)垃圾评论的查准率

图 6　正常评论和垃圾评论的查准率



(a)正常评论的F_1值 (b)垃圾评论的F_2值

图 7 正常评论和垃圾评论的综合评价指标 F-measure 值

4 结束语

本文提出一种基于 Co-Training 的协同训练微博垃圾评论识别方法。运用该方法进行 2 个对比实验,即基于评论和博文的相似度来识别垃圾评论与对 5 个评论数据的测试集进行评论分类。将本文方法和对比实验进行比较分析,结果表明了本文方法的可行性、有效性以及优越性。下一步将考虑博文和评论中带有图像语音及视频与本文方法相结合对微博垃圾评论进行识别,以及如何对短小评论进行识别,提高短小评论的识别效果。

参考文献

[1] 丁兆云,贾 焰,周 斌. 微博数据挖掘研究综述[J]. 计算机研究与发展,2014,51(4):691-706.

[2] LIU B. Web data mining:exploring hyperlinks,contents,and usage data [M]. Berlin,Germany:Springer,2009.

[3] 杨 亮,许 侃,林鸿飞,等. 博客作者声誉度分析[J]. 计算机科学与探索,2013,7(9):838-847.

[4] 杨风雷,黎建辉. 用户生成内容中的垃圾意见研究综述[J]. 计算机应用研究,2011,28(10):3601-3605.

[5] JINDAL N,LIU B. Review spam detection [C]// Proceedings of IEEE International Conference on World Wide Web. Washington D. C.,USA:IEEE Press,2007:1189-1190.

[6] JINDAL N,LIU B. Opinion spam and analysis[C]// Proceedings of IEEE International Conference on Web Search and Data Mining. Washington D. C.,USA:IEEE Press,2008:219-230.

[7] 邓冰娜,王 煜,刘 宇. 一种应用于博客的垃圾评论识别方法[J]. 郑州大学学报(理学版),2011,43(1):65-69.

[8] 黄 铃,李学明. 基于 AdaBoost 的微博垃圾评论识别方法[J]. 计算机应用,2013,33(12):3563-3566.

[9] LAI C L,XU K Q,LAU R Y K,et al. High-order concept associations mining and inferential language modeling for online review spam detection [C]// Proceedings of IEEE International Conference on Data Mining Workshops. Washington D. C.,USA:IEEE Press,2010:1120-1127.

[10] 刁宇峰,杨 亮,林鸿飞. 基于 LDA 模型的博客垃圾评论发现[J]. 中文信息学报,2011,25(1):41-47.

[11] SURENDRA S,AIXIN S. Hspam14:a collection of 14 million tweets for hashtag-oriented spam research[C]// Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York,USA:ACM Press,2015:223-232.

[12] 姚子瑜,屠守中,黄民烈,等. 一种半监督的中文垃圾微博过滤方法[J]. 中文信息学报,2016,30(5):176-186.

[13] FREUND Y,SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[C]//Proceedings of European Conference on Computational Learning Theory. Berlin,Germany:Springer,1995:23-27.

[14] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods[C]//Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics,Washington D. C.,USA:IEEE Press,1995:189-196.

[15] NIGAM K,MCCALLUM A K,THRUN S,et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning,2000,39(2):103-134.

[16] ZHOU Z H,LI M. Tri-training:exploiting unlabeled data using three classifiers [J]. IEEE Transactions on Knowledge & Data Engineering,2005,17(11):1529-1541.

[17] BREIMAN L. Random forests[J]. Machine Learning,2001,45(1):5-32.

[18] ZHOU Z H,ZHAN D C,YANG Q. Semi-supervised learning with very few labeled training examples[C]// Proceedings of AAAI Conference on Artificial Intelligence.[S.1.]:AAAI Press,2007:675-680.

[19] LI M,ZHOU Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems,Man,and Cybernetics,Part A,2007,37(6):1088-1098.

[20] 田久乐,赵 蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版),2010,28(6):602-608.

[21] 张剑峰,夏云庆,姚建民. 微博文本处理研究综述[J]. 中文信息学报,2012,26(4):21-27.

[22] CHANG C C,LIN C J. LIBSVM:A library for support vector machines[J]. ACM Transactions on Intelligent Systems & Technology,2007,2(3):27-33.

编辑 索书志