垃圾观点的文档识别

摘　　要

垃圾观点文档已经逐渐影响人们的日常的网络生活。为此本文建立了基于强分类器的 Adaboost-LC 模型、以及结合主成分分析的改进 Adaboost-Ex 模型、最后建立动态变化的 TF-IDF 模型进行商品评论的分类。

问题一，通过对垃圾观点文档的特征进行分析，针对每个特征建立了一个弱分类器，再将每个弱分类器加以训练，形成一个基于强分类器的 Adaboost-LC 模型。对于第一题的评论集合，引入相应词库，利用 ICTCLAS 分词系统将评论进行分词后，将其表示成 8 个特征值组成的特征向量，并编写 java 程序对特征量进行提取，在此基础上，将得到的特征量带入模型就可以对评论进行识别。用该模型对第一问四条评论识别，其正确率达到 75%。

问题二，针对某件产品的评价集合，因各特征量的影响程度不同，在第一问模型的基础上，引入了主成分分析，并将其结果作为 Adaboost-LC 模型训练的初始权重，从而提高了模型的识别率，以及降低了训练样本容量。为了进一步提高模型识别率，引入错误惩罚因子，建立了改进后结合主成分分析 Adaboost-Ex 模型。先选取小米 3 作为样本，建立相应的特征词库，并用四个指标来综合反映模型正确识别垃圾评论的能力，测试结果中能衡量分类器整体分类性能的指标 G 高达 78.04%，较改进前有了很大的改善，并且对垃圾评论中的人身攻击评论判别最好。

问题三，对于一般的产品评价集合，要求有动态的词库，为此建立动态的 TF-IDF 模型，引入词频 TF 和基于分布均匀度 UIDF，根据公式：

$$TF-UIDF=\frac{n_{i,j}}{N*n_i}*\log\left[\frac{N}{1+\{j:t_i\in n_j\}}+\sum_{k=0}^{N}C(s_i)-C(s_k)\right]$$

计算特征词的权重，选择权重值大的自动构成特征词库，并对已有词库进行动态更新。利用从淘宝网站搜集的多种产品的评论，结合优化的 Adaboost-Ex，对模型进行验证。结果得出该模型对一般产品垃圾评论的识别其 G 指标都高于 70%。

最后对于本题改进的 Adaboost-LC 数学模型，不仅可以应用于垃圾观点文档的分类，在实际应用中也有很好的推广。例该模型可应用于人脸的检测。

关键词：特征词库　主成分分析　TF-IDF 模型　Adaboost-Ex

目　　录

摘　　要.....1

一、问题重述.....3

二、问题分析.....3

2.1 概　论.....3

2.2 问题一.....3

2.3 问题二.....3

三、模型假设.....4

四、符号说明.....4

五、模型的建立与求解.....5

5.1 问题一.....5

5.1.1 问题一的分析.....5

5.1.2 基于单阈值简单二值分类器的 AdaBoost-LC 模型的建立.....5

5.1.3 模型求解.....7

5.2 问题二.....9

5.2.1 问题二的分析.....9

5.2.2 结合主成分分析的 AdaBoost-Ex 模型的建立.....10

5.2.2.1、建立主成分分析模型，确定影响程度大的的关键词.....10

5.2.2.2 改进的 AdaBoost-Ex 算法模型.....11

5.2.3 模型求解.....13

5.3 问题三.....19

5.3.1 问题分析.....19

5.3.2 TF-IDF 模型的建立.....20

5.3.3 模型的改进.....21

5.3.4 模型的求解：.....22

六、模型评价.....24

6.1 模型优点.....24

6.2 模型缺点.....24

6.3 模型改进.....24

七、参考文献.....25

附录.....26

一、 问题重述

目前商务网站或博客论坛允许用户发表针对产品或话题的一些评论看法，难免会存在一些虚假或是与产品及话题无关的评论信息，这极大地误导了商家、读者以及观点挖掘系统。因此，垃圾观点文档的识别具有重要研究价值和实用意义。

本任务是对给定的语料集合中，要求参赛系统识别出文档是否为垃圾观点文档。要求完成以下问题：

（1）针对下面介绍的情形，请建立合理的数学模型进行识别，并给出你的算法流程。并通过程序验证，给出你的正确识别率。

（2）请在网络上收集一个更大的关于某件产品的评价集合，建立合理的数学模型和算法进行识别，并给出你的结论。

（3）对一般的产品评价集合，讨论并建立更一般的模型，并谈谈你的该类识别问题的看法。

二、问题分析

2.1 概 论

这是一个垃圾文档分类问题，根据垃圾评论提取特征词，并将其进行量化，对文档分类。问题的特点在于垃圾评论的特征词很多，垃圾评论的种类多，难点在于要随时更新与特证词相匹配的词库，以及对评论中提取的特征量进行适当的量化，并确定合适的阈值进行比较。

2.2 问题一

问题一要求对给定的评论内容进行识别。四条评论中前两条是关于 6plus 的，第三条是关于保时捷的，这三条评论是有效评论，而第四条并不是针对于产品的评论，属于垃圾评论。应该注意到这两种评论的差别在于垃圾评论中并没有评价性词语，而前三条都有。再者评论是很多关键词的组合，必须要对评论进行分词，提取主要关键词，还需建立相应的词典。除此之外还需对关键词进行量化，注意到垃圾评论及有效评论量化后的特征值是有区别的，这就需要建立一个合理的模型区分。并得出模型识别的正确率。

2.3 问题二

与第一问不同的是，对特定的某件产品的评价集合，是对第一问的具体化，需要搜集大量的某件产品的评价参数或者是常出现的关键词，来构建特征词库。对于给定的产品评价集合，其关键词的影响程度不同，对分类结果影响不同，为此可以考虑采用主成分分析法，找到影响大的主要特征量，并结合第一问当中模型对此产品评价集合进行测试，得到识别的效果。

2.4 问题三

与第二问相比不同的是，该题要求对一般的产品评价集合，即对于任意一种产品的垃圾评论进行识别。由于没有明确指明具体产品，所以其特征词库的提取很有困难，可以考虑建立动态的特征词库。再结合第二问的模型加以优化便可以的到最后模型，并总结一、二、三问，得出对这类识别问题的具体看法。

三、模型假设

1、评论文档没有英文评论和图片评论；
2、训练集数据少时，权重的改变不影响模型识别能力；
3、假设每个弱分类器分类能力比随机猜测要好；
4、假设初始每个样本对应的权重是一样的；
5、对于用于训练的样本评论分类事先已经人为分出；
6、假设 Adaboost 模型使用的训练误差就是真正的训练误差。

四、 符号说明

符号	符号说明
$h_j(x)$	样本属于某类别的概率
θ_j	分类特征值
$\mathrm{f}_j(x)$	类结点
$\mathrm{W_{tj}}$	样本
X_i	单词
F	垃圾文档
ε_t	非垃圾文档
T	评论文档
P_1	垃圾单词
R_1、R_2	非垃圾单词
P_1、P_2	平均值
H	密度差

五、模型的建立与求解

5.1 问题一

5.1.1 问题一的分析

垃圾评论的识别与过滤[1]，可以看作是一个文本二值分类问题。通过对评论的特征词进行量化得到相应的特征值，并与每个特征词对应的阈值进行比较，来确定该评论是否为垃圾评论。为此我们建立了基于单阈值简单二值分类器的AdaBoost 模型。

5.1.2 基于单阈值简单二值分类器的 AdaBoost-LC 模型的建立

本题可以先采用线性分类器中的单阈值简单二值分类器对垃圾评论进行简单分类，在机器学习算法中，线性分类器是最简单也很有效的分类器，单阈值简单二值分类器定义如下式：

$$h_j(x)=\begin{cases}1 & \mathrm{p}_j\mathrm{f}_j(x)\geq p_j\theta_j\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

其中，θ_j表示特征的阈值，$\mathrm{f}_j(x)$表示样 X本的第 j 个特征的特征值，p_j为第 j 个特征的偏置，取值为 1。针对评论的单个特征进行训练单阈值简单二值分类器时，需要训练得到阈值。把该特征的最大值和最小值之差分为 100 份，从最小值开始，每次增加一份差值作为阈值，这样只需要在 101 个阈值中确定一个最佳阈值，就可以进行垃圾评论的分类。

计算得出单阈值简单二值分类器的正确识别率过低，所以需要对其进行优化，采用AdaBoost-LC算法。

AdaBoost-LC是一种迭代算法，其基本思想找到若干个分类精度比随机预测略高的弱分类器，再将这些弱分类器集合起来构建成一个高精度的强分类器。AdaBoost-LC 是通过改变数据分布来实现，根据每次训练集中每个样本的分类是否正确，以及上一个弱分类器的分类错误率，来改变每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的弱分类器线性加权集合起来，作为最终的强分类器，弱分类器的错误率越低，其权重越大。

本题将利用背景知识的商品评论的作为训练集，即每条评论都是一个训练样本，附件中的评论作为测试集，步骤如下：

1) 给定n 个训练样本$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$迭代次数$T$，其中$y_i\in\{0,1\}$，1表示正样本(正常评论)，0 表示负样本(垃圾评论);每个样本x_i有k个特征，表示为$\{f_1(x_i),f_2(x_i),\cdots,f_k(x_i)\}$。

2) 初始化权重。设$W_{t,i}$为第t次循环中样本i的误差权重，初始化权重$W_{1,i}$，对于正样本$W_{1,i}=1/2m$，对于负样本$W_{1,i}=1/2l$，其中l，m分别为正样本、负样本个数。

3) 循环For t = 1 to T

a)样本权重归一化，如式(2) 所示:

$$W_{t,i}=W_{t,i}/\sum_{j=1}^{n}W_{t,j} \tag{2}$$

b)对于每个特征j，训练出一个分类器h_j，使得分类器h_j的错误率ε_j最小。

$$\varepsilon_j=\sum_{i=1}^{n}W_{t,i}\left|h_j(x_i)-y_i\right| \tag{3}$$

c)对b) 中每个特征均训练出一个分类器，找出一个具有最小错误率ε_t的分类器h_t并计算其权重，分类器的错误率越小，权重就越大，权重α_t公式如下所示:

$$\alpha_t=\frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t} \tag{4}$$

d) 更新样本权重，分类正确的样本权重减少，分类错误的样本权重增加，更新权重公式如式(5) 所示:

$$W_{t+1,i}=W_{t,i}\times\begin{cases}e^{-\alpha t}, & h_t(x_i)=y_j\\ e^{\alpha t}, & h_t(x_i)\neq y_j\end{cases} \tag{5}$$

e)重复步骤3) ，直到循环T 次结束或者达到边界条件:本次循环训练出的分类器错误率ε_t大于等于0.5时，删除当前的分类器并不再进行循环; 当错误率ε_t等于0时，也不再进行循环，因为这时训练错误率为0，无需继续训练。

4)得到最后的强分类器 $h(x)$ 是 T（实际训练得到的弱分类器个数）个弱分类器的加权平均，如式(6) 所示:

$$h(x)=\begin{cases}0, & \sum_{t=1}^{T}\alpha_t h_t(x)\geq\frac{1}{2}\sum_{t=1}^{T}\alpha_t \\ 1, & \text{其他}\end{cases} \tag{6}$$

5.1.3 模型求解

本文识别微博垃圾评论的流程主要分为以下 4 步：①数据预处理，对微博和评论的内容进行预处理，为下一步特征值提取做准备；②特征值提取，对评论进行提取本文预定义的几个重要特征的值；③训练，每条评论被表示成具有相同维数的特征值向量，然后加上类标签作为样本集进入 AdaBoost[3]算法流程，直到达到边界条件从而训练出一个强分类器；④识别，对需要识别的每条评论先进行预处理以及提取特征值后用训练出来的强分类器进行判断是否为垃圾评论。本文微博垃圾评论的整体识别流程图如下所示：

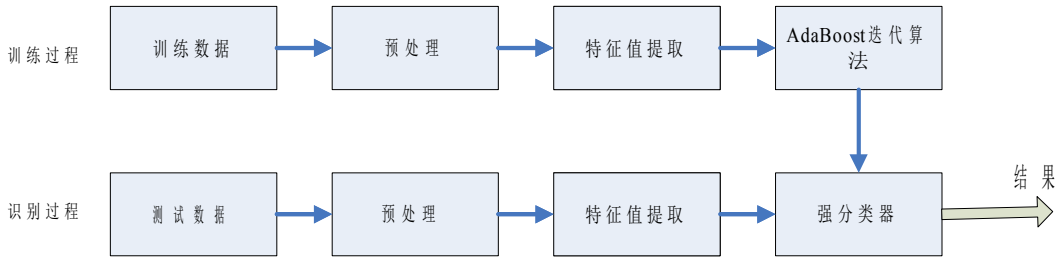

图 1 算法流程图

1、实验的预处理

本文所做的预处理从两个方面考虑，第一方面将采用中国科学院计算技术研究所分词系统 ICTCLAS 对商品评论进行分词，以便下一步特征提取时，计算评论的特征值。另一方面剔除全是数字、字母或者标点符号的垃圾评论，此类垃圾评论很容易过滤处理，本文方法不予考虑。对附件中的商品评论进行系统分词，得到的结果是：

表 1 附件中的商品评论分词结果

	系统分词结果
Doc1	买/的/iphone6plus/轻薄/分辨率/高，电池/耐用，屏幕/够大，散热/比/之前的/有所改进
Doc2	看到/个/视频/iphone6/轻轻/一掰/就/弯/了
Doc3	我/还是/喜欢/保时捷/外观/内饰/都/大气/奢华
Doc4	总是/有人/都/没/用过/买不起/就在哪里/瞎/说说

2、特征量的提取

在本题模型中，特征的选取会直接影响到模型的识别效果。本文从评论、评论者、被评论的商品三个方面进行垃圾商品评论识别特征的选择，特征树状图如图 2 所示。再根据分词以及特征提取的原理，使用 HowNet 情感词典作为种子词典，并加入一些网络新词，去掉一些不常用的词，构建评价词和情感词词典，利用 java 程序进行特征量的提取[4]，求解出特征向量 X(i)，如下图 3:

X1(X2)=评论中出现的正(负)面情感词的次数/评论中出现的所有词的次数

X3=n,(n=0,-1,-2…) (n:评论中广告词的个数，定义为负数)

X4(X5) =评论中正(负) 面评论词的数量/评论中所有词的数量

X6=n,(n=0,-1,-2…) (n:评论中诋毁词以及敏感词的个数，定义为负数)

X7=n，(n = 0，1，2…) (n: 评论中存在该产品的属性的数量)

X8=n(n=0，1)(0: 评论中不存在该产品名称；1: 评论中存在该产品名称)

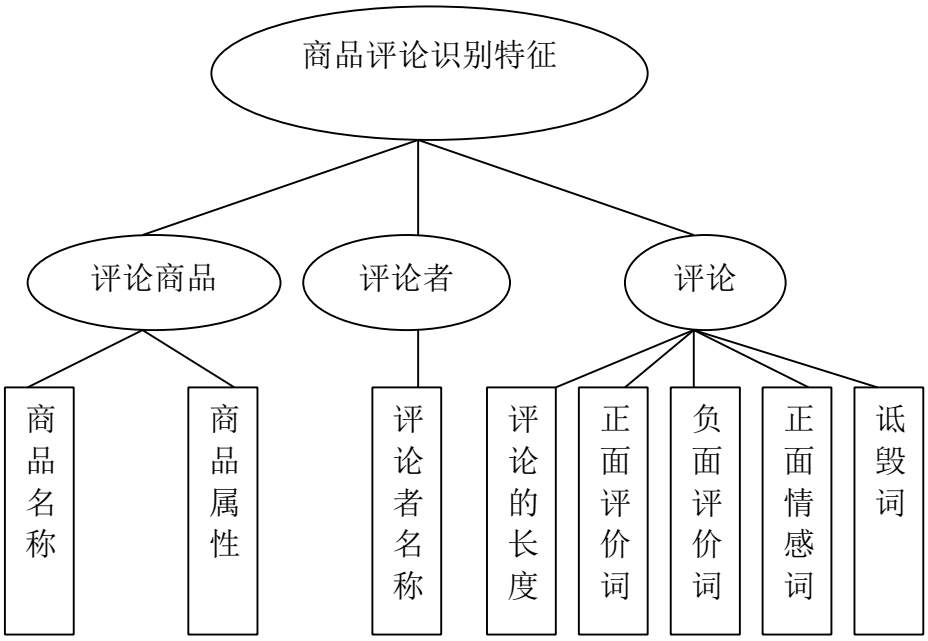

图2 商品评论的识别特征

3、验证模型的正确识别评论的能力

利用建立的AdaBoost算法模型，加权集成弱分类器而得到强分类器，对附件中的评论内容进行分类，最终计算出的正确识别率如下表格：

表2 单阈值简单二值分类器的正确识别率

评价	计算结果	实际结果	正确率
Doc1	0	1	50%
Doc2	1	1	
Doc3	1	1	
Doc4	1	0	

表3 AdaBoost-LC算法模型的正确识别率

评价	计算结果	实际结果	正确率
Doc1	0	1	75%
Doc2	1	1	
Doc3	1	1	
Doc4	0	0	

针对附件已有的评论内容，通过以上两表格的正确识别率的比较，单阈值简单二值分类器的正确识别率为 50%，而 AdaBoost-LC 算法模型的正确识别率为 75%，由此可知该算法模型识别率明显优于单阈值简单二值分类器。但是由于测验的评论比较少，会存在一定误差。

5.2 问题二

5.2.1 问题二的分析

问题二是针对某件产品的评价集合，此时需要构建特征词库，又需考虑到每个特征词所占的权重也是不同的，若直接应用 Adaboost 模型则会导致正确率降低，为此需要做主成分分析，找出影响分类结果较大的特征词，对权重进行重新初始化。由于第一问的模型采用的训练集少，并且模型会因离群点的权重而急剧扩张，AdaBoost-LC 算法出现“退化现象”，为此对其进行优化。

5.2.2 结合主成分分析的 **AdaBoost-Ex** 模型的建立

5.2.2.1、建立主成分分析模型，确定影响程度大的的关键词

1、主成分分析的原理：

主成分分析[10]是设法将原来众多具有一定相关性(比如 P 个指标)，重新组合成一组新的互相无关的综合指标来代替原来的指标。通常数学上的处理就是将原来 P 个指标作线性组合，作为新的综合指标。最经典的做法就是用 F1 的方差来表达，即 Var(F1)越大，表示 F1 包含的信息越多。因此在所有的线性组合中选取的 F1 应该是方差最大的，故称 F1 为第一主成分。如果第一主成分不足以代表原来 P 个指标的信息，再考虑选取 F2 即选第二个线性组合，为了有效地反映原来信息，F1 已有的信息就不需要再出现在 F2 中，用数学语言表达就是要求 Cov(F1, F2)=0，则称 F2 为第二主成分，依此类推可以构造出第三、第四，　　，第 P 个主成分。

2、具体的主成分分析数学模型

$$\begin{cases} F_1 = a_{11}X_1 + a_{21}X_2 + \cdots\cdots + a_{p1}X_p \\ F_2 = a_{12}X_1 + a_{22}X_2 + \cdots\cdots + a_{p2}X_p \\ \qquad\qquad \cdots\cdots \\ F_1 = a_{1\mathrm{m}}X_1 + a_{2\mathrm{m}}X_2 + \cdots\cdots + a_{p\mathrm{m}}X_p \end{cases} \qquad (7)$$

其中 $a_{1i}, a_{2i}, \cdots\cdots, a_{pi}\,(i = 1, \cdots\cdots, m)$ 为 X 的协方差阵 $\sum$ 的特征值多对应的特征向量，$X_1, X_1, \cdots\cdots, X_1$ 是原始变量的值，由于本文中数据无量纲的影响，故没有进行数据的标准化处理。$A = \left(a_{ij}\right)_{p\times m} = \left(a_1, a_2, \cdots\cdots, a_m\right), Ra_i = \lambda_i a_i, R$ 为相关系数矩阵，λ_i 、 a_i 是相应的特征值和单位特征向量，$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ 。

进行主成分分析主要步骤如下:

A、根据研究问题选取指标与数据;

B、进行指标数据标准化(SPSS 软件 Factor 过程自动执行) ;

C、进行指标之间的相关性判定;

D、确定主成分个数 m;

E、确定主成分 Fi 表达式;

F、计算综合主成分值并进行后续步骤。

3、主成分个数的选取

主成分个数提取原则为主成分对应的特征值大于 1 的前 m 个主成分。特征值在某种程度上可以被看成是表示主成分影响力度大小的指标，如果特征值小于 1，说明该主成分的解释力度还不如直接引入一个原变量的平均解释力度大，因此一般可以用特征值大于 1 作为纳入标准。

5.2.2.2 改进的 AdaBoost-Ex 算法模型

由 AdaBoost-LC 模型计算第二问搜集的商品评论集合，得出评论的正确识别率如下表所示：

表 4　第一问的模型求解小米三评论集合的正确识别能力

正常	垃圾	正确识别		错误识别		FN(%)	FS(%)	G(%)	M(%)
		正常	垃圾	正常	垃圾				
50	50	30	26	20	24	57.69%	54.17%	56.04%	48.00%

由表可知，可能由于第一问的训练集样本较少，再加上，模型会因离群点的权重而急剧扩张，AdaBoost-LC 算法出现“退化现象”，导致识别率降低，为此引入错误惩罚因子的概念，在分类过程中，样本被分错次数越多，惩罚因子越大，样本的权重就会相应的降低，这样可以限制样本权重的无节制扩张。因此，当训练弱分类器算法循环步骤中权重更新后再次对权重进行限制：

$$W_{t+1,j} = +W_{t+1,j} \bullet \frac{1}{\log r_i} \qquad \mathrm{r}_i \geq 3 \tag{8}$$

其中，r_i表示样本 i 被分类错误的次数，加入对数使得分错次数 r_i影响减小，不至于矫枉过正，log3 大于 1，因此当错误次数 r_i大于等于 3 时才开始缓慢减小样本的权重，r_i越大，分母越大，样本权重 $W_{t+1,j}$ 受到的限制就会越大，从而限制了样本权重的无节制扩张。而对于 AdaBoost-LC 算法把错误率同等对待的问题，在每轮训练选择弱分类器的时候，不仅以错误率的大小作为评判的标准，而且在错误率相同的情形下选择一个具有最低误判率的弱分类器。

因此，训练弱分类器算法循环步骤中计算错误率的同时，还需要计算误判率，在选择弱分类器的时候首先选择错误率最小的弱分类器，如果错误率一样，则选择具有更低误判率的弱分类器。为此引入改进后的 AdaBoost-Ex 算法

第二问，将利用所搜集的全部评论的一部分作为训练集，剩余的作为测试集，具体步骤描述如下：

1) 给定 n 个训练样本$(x_1,x_1),(x_2,x_2),\cdots,(x_n,x_n)$，迭代次数$T$，其中，$i$表示正样本（正常评论），0 表示负样本（垃圾评论）；每个样本x_i有k个特征，表示为$\{f_1(x_i),f_2(x_i),\cdots,f_k(x_i)\}$。

2) 初始化权重

设$W_{t,j}$为第t次循环中样本i 的误差权重，初始化权重$W_{1,j}$，对于正样本$W_{1,j}=1/2m$,对于负样本$W_{l,i}=1/2l$，m，l分别为正样本、负样本个数。

3) 循环 For t=1 to T

a：样本权重归一化，如下式所示：

$$W_{t,j}=\frac{W_{t,j}}{\sum_{j=1}^{n}W_{t,j}} \tag{9}$$

b：对于每个特征j，训练出一个分类器h_j，使得分类器h_j的错误率ε_j最小，如果具有同样错误率的弱分类器，则选择一个具有更低误判率m_j的弱分类器。ε_j、m_j计算公式分别如式下式所示：

$$\varepsilon_j=\sum_{i=1}^{n}W_{t,i}\left|h_j(x_i)-y_i\right| \tag{10}$$

$$m_j=\sum_{i=1}^{n}W_{t,i}\qquad h_j(x_i)=0 \text{ and } y_i=1 \tag{11}$$

其中，误判率m_j表示正常评论被误判为垃圾评论的样本权重值之和。

c：b 中每个特征均训练出一个分类器，从中找出一个具有最小错误率ε_t的分类器h_t，如果存在错误率一样的分类器，则选择具有更低误判率的分类器，并计算其权重，分类器的错误率越小，权重就越大，权重α_t计算公式如式(12)所示：

$$\alpha_t=\frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t} \tag{12}$$

d：更新样本权重，分类正确的样本权重减少，分类错误的样本权重增加，更新权重公式如式(14)所示：

$$W_{t+1,i}=W_{t,i}\times\begin{cases}e^{-\alpha_t} & \mathrm{h}_t(x_i)=y_j\\ e^{\alpha_t} & \mathrm{h}_t(x_i)\neq y_j\end{cases} \tag{13}$$

e：对权重进行限制，限制公式如式（14）所示：

$$W_{t+1,j}=W_{t+1,j}\bullet\frac{1}{\log r_i}\qquad \mathrm{r}_i\geq 3 \tag{14}$$

其中，r_i表示样本i被分类错误的次数，加入对数使得分错次数r影响减小，不至于矫枉过正，log3 大于 1，因此当错误次数r_i大于等于 3 时才缓慢减小样本的权重，r_i越大，分母越大，样本权重的限制就会越大，从而限制了样本权重的无节制扩张。

f：重复步骤 3)，直到循环 T 次结束或者达到边界条件：本次循环训练出的分类器错误率t大于等于 0.5 时，删除当前的分类器并不再进行循环；当错误率t等于 0 时，也不再进行循环，因为这时训练错误率为 0，无需再继续进行训练。

4) 得到最后的强分类器$h(x)$是T (实际训练得到的弱分类器个数)个弱分类器的加权平均，如式（15）所示：

$$h(x)=\begin{cases}0, & \sum_{t=1}^{\mathrm{T}}\alpha_t h_t(x)\geq\frac{1}{2}\sum_{t=1}^{\mathrm{T}}\alpha_t\\ 1, & \text{其他}\end{cases} \tag{15}$$

5.2.3 模型求解

该题要求对于给定的产品评价集合进行垃圾评论，由于其关键词的影响程度不同，对分类结果影响不同，为此，对于特征量进行主成分分析，找到影响程度大的特征量，确定相应权重，将结果应用到改进的 AdaBoost-Ex 模型中去，提高产品垃圾评论的正确判别能力。算法如下图所示：

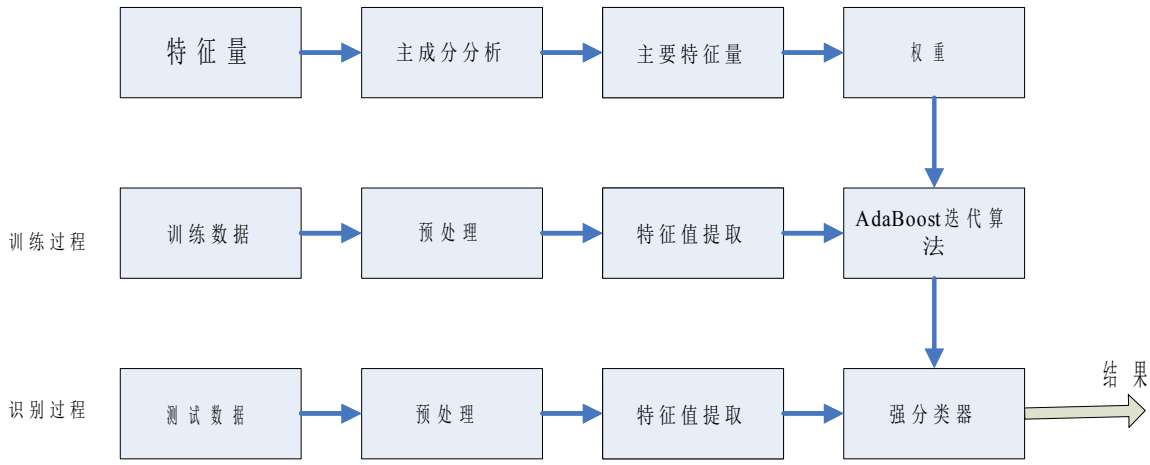

图 3　算法流程图

1、主成分分析求解主要特征量

本题从各大网站上搜取小米 3 的商品评论集合，并对已有的词库进行更新，加入所采集的评论的一些关键词，由于手机的性能因素很多，每个因素所起的作用不同，即特征量的权重也就会相应的发生变化，为此采用 spss 软件[2]对给定的八个特征量进行分析，确定特征量对分类结果的影响程度。

第一问中，根据已有的资料，提取出八个特征量，以下仅为部分评论的数据分析表，现截取部分数据如下：

表 5　　部分评论的各特征量的值

	X1	X2	X3	X4	X5	X6	X7	X8	F
1	0	0	0.0454	0.0909	0	0		0	0.1363
2	0	0	0.3529	0	0	0	0.1764	0	0.5294
3	0	0	0.2083	0.04366	0	0	0	0	0.25
4	0	0	0.0416	0	0	0	0	0	0.0416
5	0	0	0.1818	0	0	0	0	0	0.1818
6	0	0	0.1538	0.1538	0	0	0	0	0.3076
7	0	0	0	0	0	0	0	0	0
8	0	0	0.1304	0	0	0	0	0	0.1304
9	0	0	0.0666	0	0	0	0	0	0.0666
10	0.0526	0	0	0	0	0	0	0	0.0526

通过利用 SPSS 软件进行主成分分析，得各种数据如下：

表 6　相关系数矩阵

Correlation Matrix								
	F1	F2	F3	F4	F5	F6	F7	F8
F1	1.000	-.043	-.191	-.149	.522	.443	-.052	.722
F2	-.043	1.000	-.242	-.190	-.061	.376	-.068	.068
F3	-.191	-.242	1.000	.101	-.087	.127	.875	.261
F4	-.149	-.190	.101	1.000	-.178	.074	.100	.096
F5	.522	-.061	-.087	-.178	1.000	.152	-.074	.153
F6	.443	.376	.127	.074	.153	1.000	.832	.986
F7	-.052	-.068	.875	.100	-.074	.832	1.000	.867
F8	.722	.068	.261	.096	.153	.986	.867	1.000

由上表可以看出：许多变量之间直接的相关性比较强,X3 、X6、X7、X8 有显著关系，可见他们存在信息上的重叠。

表 7 方差分解主成分提取分析表

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative	Total	% of Variance	Cumulative
1	6.320	72.205	72.205	6.320	72.205	72.205
2	1.235	12.346	84.551	1.235	12.346	84.551
3	0.943	8.769	92.224			
4	0.674	5.442	98.654			
5	0.342	0.824	99.223			
6	0.024	0.103	99.971			
7	0.012	0.021	100.000			
8	0.000	0.000	100.000			

通过表 7 可知，提取 2 个主成分，即 m=2, 从表 8 可知 X6、X7、X8 在第一主成分上有较高载荷,X3、X5 在第二主成分上有较高载荷。所以提取两个主成分是可以基本反映全部指标信息，所以决定用两个新变量代替原来的 8 个变量。

表 8 初始因子载荷矩阵

Component		
Component	Component	
	1	2
X1	0.109	0.023
X2	0.112	-0.008
X3	0.943	0.545
X4	0.874	-0.043
X5	0.786	0.735
X6	0.983	0.213
X7	0.978	0.125
X8	0.949	0.221

用表 8 的数据除以主成分相对应的特征值开平方根便得到两个主成分中每个指标所对应的系数。利用 SPSS 进行操作后，就可以得出主成分表达式如下：

$$F_1=0.112X_1+0.042X_2-0.321X_3-0.125X_4+0.004X_5-0.298X_6+0.396X_7+0.313X_8 \quad (16)$$

$$F_2=0.158X_1-0.004X_2-0.742X_3+0.046X_4+0.652X_5-0.175X_6-0.012X_7+0.213X_8 \quad (17)$$

再继续利用 SPSS 进行处理，以每个主成分所对应的特征值占所提取主成分总的特征值之和的比例作为权重计算主成分综合模型，可得到主成分综合模型：

$$F=0.123X_1+0.021X_2-0.323X_3+0.113X_4+0.112X_5-0.266X_6+0.398X_7+0.364X_8 \quad (18)$$

由公式可知 X_3、X_6、X_7、X_8 所对应的四种特征量所占权重比例大，对分类结果会有较大的影响，因此若将此结果作为 Adaboost-LC 模型的初始权重，可大大提升其识别率，并且可以有效的减少训练所需样本容量。

2、改进的 AdaBoost-Ex 算法的求解

根据主成分分析的结果，将其作为 AdaBoost-Ex 模型训练初始权重，进而利用改进的 AdaBoost-Ex 算法进行商品评论的分类。对于影响作用比较大的特征量，在求解 AdaBoost-Ex 模型中给予较高的权重。经由 java 程序提取出评论的特征量，并用正常评论 F_N、垃圾评论 F_S、误判率 M、几何平均 $G-\text{mean}$ 值综合反映模型正确识别垃圾评论的能力。分类的结果可利用混淆矩阵来表示如下图：

表 9 混淆矩阵

	正常评论	垃圾评论
识别的正常评论	m_1	n_1
识别的垃圾评论	m_2	n_2

定义一：正常评论召回率 R_1，垃圾评论召回率 R_2，

$$R_1 = \frac{m_1}{m_1 + n_1}, R_2 = \frac{n_2}{m_2 + n_2} \tag{19}$$

定义二： 正常评论精确率 P_1，垃圾评论精确率 P_2

$$P_1 = \frac{m_1}{m_1 + m_2}, P_2 = \frac{n_2}{n_1 + n_2} \tag{20}$$

定义三： 正常评论 F_N 值，垃圾评论 F_S 值，

$$F_N = \frac{2 * R_1 * P_1}{R_1 + P_1}, F_S = \frac{2 * R_2 * P_2}{R_2 + P_2} \tag{21}$$

定义四：几何平均 $G-mean$

$$G-mean = \sqrt{\frac{m_1}{m_1 + m_2} \bullet \frac{n_2}{n_1 + n_2}} \tag{21}$$

几何平均 $G-mean$[7]是正常评论的精确率与垃圾评论的精确率乘积的平方根，如果两个精确率中有一个较低，则会使得 G 值较低，只有当两个精确率都较高时,G 值才会较高，因此 G 值能比较合理的衡量分类器对于非平衡数据集的整体分类性能，也是常用的应用于非平衡数据集的评价指标。

定义五： 误判率 M

$$M = \frac{n_1}{n_1 + n_2} = 1 - P_2 \tag{22}$$

在对垃圾评论识别效果评判标准中，误判率 M 也是一大重要指标。误判率是指被分类器判别为垃圾评论的评论中正常评论的比例。

通过四个指标来反映模型的识别能力，能够充分地说明模型的特点，因此，本文选择正常评论值 F_N，垃圾评论 F_S 值，几何平均 $G-mean$ 以及误判率 M 四个指标来评价分类器的分类性能。

3、结论

将模型的判别结果利用 excel 做出如下表格：

表 10 三种方法分别得出的正确识别率

方法	正常	垃圾	正确识别		错误识别		FN(%)	FS(%)	G(%)	M(%)
			正常	垃圾	正常	垃圾				
P	50	50	30	26	20	24	57.69%	54.17%	56.04%	48.00%
Q	50	50	37	37	13	13	74.00%	74.00%	74.00%	26.00%
O	50	50	40	38	10	12	78.43%	77.55%	78.04%	24.00%

其中： P 代表 AdaBoost-LC 算法；

Q 代表主成分分析+AdaBoost-LC；

O 代表主成分分析+优化的 AdaBoost-Ex 模型

由表可知，在 P、Q、O 三种方法的测试结果中，G 值依次增大，由于 G 值能比较合理的衡量分类器对于非平衡数据集的整体分类性能，所以三种方法的整体分类性能依次增加。结合正常评论 F_N 值，垃圾评论 F_S 值，以及误判率 M 三种指标就可以看出：对于小米 3 的商品评论判别，随着模型的不断优化，以及主成分分析的加入，垃圾评论的识别率依次提高，误判率逐渐降低，正确评论值逐渐升高，错误评论值逐渐升高（垃圾评论识别为垃圾评论正确的比例）。

最后根据表格中的四个判别正确识别率的指标，做出柱状图，进行对比，更好的说明模型优化的效果。

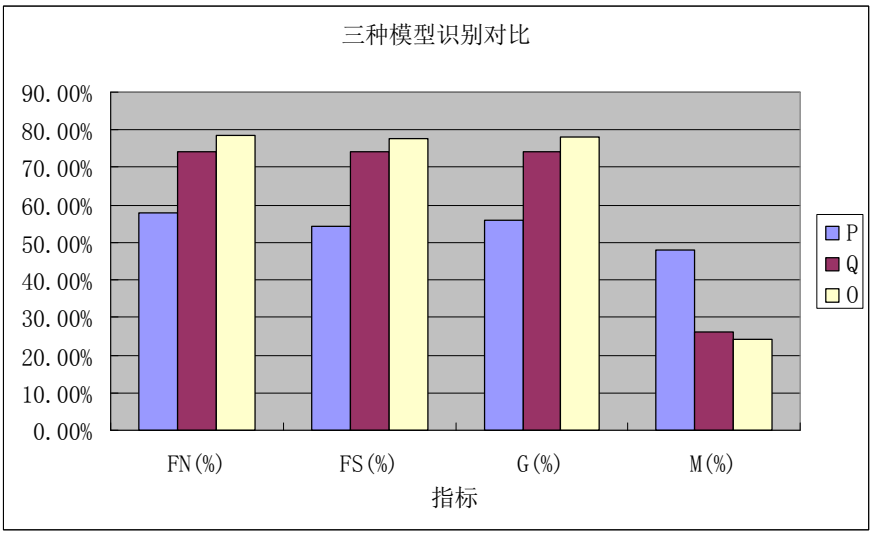

图 3 三种模型应用某件产品评论判别结果对比

由上述柱状图可看出，结合主成分分析的 Adaboost-Ex 模型对垃圾评论的识别率有明显提高，误判率逐渐降低，正确评论值逐渐升高，错误评论值逐渐升高。

表 11 改进的模型对于不同种垃圾评论的识别效果

垃圾评论归类	总数	正确识别	错误识别	识别率
评错产品评论	8	4	4	50.0%
广告评论	15	14	1	93.3%
人身攻击评论	10	8	2	80.0%
其他评论	17	10	7	58.8%

由于无用评论可分为很多种，此处分为广告评论，评错产品评论，人身攻击评论、其他评论，结合主成分分析的 Adaboost-Ex 模型对于垃圾评论当中的人身攻击评论识别最好，对于广告评论识别较好，对于评错产品评论识别最差，无用评论当中的其他评论识别中等。对于选定的小米 3 商品评论集合，由表 9 可得出结合主成分分析的 Adaboost-Ex 模型判别垃圾评论的能力，可表示为：正确评论值为 78.43%，错误评论值为 77.55%，几何平均 $G-mean$ 为 78.04%，误判率为 24%。

5.3 问题三

5.3.1 问题分析

由于第二问确定了对特定产品的评论进行识别，因此可以轻易获得其特征词库，但是该题要求对一般的产品评价集合，即对于任意一种产品的评论，其特征词库的获得非常困难，这时如果直接用 Adaboost-Ex 模型就会导致识别率大大降低，因此建立 TF-IDF 模型来动态提取特征词库，结合 Adaboost-Ex 模型对垃圾评论分类。与第一、第二问相比，不同之处在于建立动态的特征词库，为此需要随时改变词频和分布均匀度，利用建立的 TF-IDF 数学模型进行求特征词的权重，将权重大的特征词纳入词库，与其他词库共同构成动态词库。收集不同产品的评论文档，再结合第二问优化的模型，对 TF-IDF 数学模型进行检验。

5.3.2 TF-IDF 模型的建立

由于上述模型中对 F7（产品属性）、F8（产品名称）权重非常大，那么模型的正确率非常依赖于特征词库的选取，因此上述模型不能直接应用于一般产品的分类，为了解决此问题，建立 TF-IDF 算法模型进行扩展。

TF-IDF实际上是：TF *IDF，TF词频，IDF反文档频率。

TF表示词条在评论库中出现的频率（某一个给定的词在库中出现的次数）。

IDF指出：在一组文档中，刻画某一文档特征的特征项(词)可以根据其在这组文档中出现的频率赋予相应的权重，只在少数文档中出现的较特殊的词，权重要比在多篇文档中出现的词的权重要高，并给出如下权重计算公式:

$$\log_2 N - \log_2 n + 1 \qquad (23)$$

其中N代表总文档数，n指包含特征项的文档数。如果特征项在所有文档中出现的频率越高，则它包含的信息熵就越少；如果特征项的出现较为集中，只在少量文档中有较高的出现频率，则它拥有较高的信息熵。因此IDF可以理解为在一个特定条件下关键词的概率分布的交叉熵。

TF-IDF主要体现了以下思想：一个词在特定的评论库中出现的频率越高,说明它在区分该评论内容属性方面的能力越强(TF)；一个词在文档中出现的范围越广，说明它区分文档内容的属性越低(IDF)。

传统的TF-IDF模型只需要考虑字词在当前文本集中出现的次数和在其它文本集中出现的次数。假设有类别 Sj，则字词 Wi的 TF-IDF 权重计算公式如下：

$$TFID = TF_{i,j} * IDF_i \qquad (24)$$

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, IDF_i = \log \frac{N}{1 + \{j : t_i \in n_j\}} \qquad (25)$$

其中：

1) $n_{i,j}$： W_i在当前类别 s_j 中出现的次数；

2) $\sum_k n_{k,j}$：样本训练集中所有词的总数之和；

3) N：样本训练集包含的所有类别数量之和；

4) $\{j : t_i \in n_j\}$：除当前类别外，包含字词 t_i 的类别的数目；

但是该模型存在以下两个重大缺陷：

一：在实际的垃圾评论识别应用中，仅仅考虑字词在某个评论库中出现的次数是远远不够的，会引起较大的权重计算偏差。

二：在字词的传统 TF-IDF 权重计算公式中，对TF的计算，并没有考虑不同评论库下的字词总数数量级的差别，对于数量级差别大的评论库，其 TF-IDF 权重计算的结果可能会有较大的偏差。

5.3.3 模型的改进

针对于缺陷一：引入“分布均匀度”的 UIDF 公式，即在 IDF 计算公式中引入一个可以表示“分布均匀度”的值，也就是一个词在评论库 i 中的数量远远大于在其他评论库中的数量，则该字词对于评论库 i 具有较高的标识能力；如

果一个字词在当前评论库中的数量与其它评论库中的数量分布是均匀的，则该字词对当前评论库的标识能力较低。

基于分布均匀度的 IDF 计算公式称为 UIDF 公式，如下：

$$UIDF_i = \log\left[\frac{N}{1+\left\{j:t_i \in n_j\right\}} + \sum_{k=0}^{N} C(s_i) - C(s_k)\right] \tag{26}$$

在公式（26）中：

1) N：评论训练集包含的评论数量之和；

2) $C(s_i) - C(s_k)$：当前评论 s_i 与评论 s_k 中包含 t_i 的数量差；

3) $C(s_i) - C(s_k) \le 0$：差值小于等于 0 时取 0；

针对于缺陷二：引入构建平衡数据集[8]的 TF 公式。针对在传统 TF 计算公式存在的问题，可以通过在计算中对数量级较小按比例进行放大，从而平衡各评论库下的字词数量方式进行解决，根据按比例放大的思想，本文构造出新的 TF 计算公式：

$$TF_{i,j} = \frac{n_{i,j} * \max\left[S_i\right] / n_i}{N * \max\left[S_i\right]} \tag{27}$$

在公式（27）中：

1) $n_{i,j}$：字词 W_i 在当前评论库 S_j 中出现的次数；

2) $\max\left[S_i\right]$：包含字词数量最多评论库的总词数；

3) n_i：当前评论库中的总词数；

4) N：评论训练集包含的评论数量之和；

把公式（27）化简，可得：

$$TF_{i,j} = \frac{n_{i,j}}{N * n_i} \tag{28}$$

综合以上两点就可以得到一个比较完善的TF-UIDF模型

$$TFUIDF = \frac{n_{i,j}}{N * n_i} * \log\left[\frac{N}{1+\left\{j:t_i \in n_j\right\}} + \sum_{k=0}^{N} C(s_i) - C(s_k)\right] \tag{29}$$

改进的 TF-UIDF 公式很好地弥补了传统 TF-IDF 计算公式的两个缺陷，并且对不同评论库下总词数数量级差别较大而影响特征权重计算结果的问题进行了解决。对比传统的 TF-IDF 算法，在实际应用中，TF-UIDF 具备更好的适应性和更贴合实际的权重计算结果。

实际上，如果一个词条在一个类的评论库中频繁出现，则说明该词条能够很好代表这个类的文本的特征，选来作为该类文本的特征词以区别与其它类文档。将权重值最大的词特征词库里。这样就可以得到任意产品的特证词库。有了特证词库之后，结合hownet的情感词库，以及诋毁词库、敏感词库等组成一个词库系统，结合Adaboost-Ex模型就可以得出一个优化的针对于一般产品评价集合的Adaboost-Ex模型（简称为一般Adaboost-Ex模型）。

5.3.4 模型的求解：

从淘宝网上分别找出苹果iPad Air 2、魅族 MX4、真彩迪士尼书包、惠普g14-a003TX四件产品的评价集合，每件产品评论都选取了50条垃圾评论和50条有效评论，作为评论库，首先利用TF-UIDF模型自动提取各产品的特征词库，利用一般Adaboost-Ex模型进行垃圾评论的筛选。为了确定垃圾评论在总评论中所占比例对结果的影响，将每种产品的100组数据分为两组进行测试，第一组由40条正确评论及10条垃圾评论组成。第二组由40条垃圾评论及10条正常评论组成。第一组测试结果如下表：

表12 一般Adaboost-Ex模型对不同产品评论的第一组识别结果

产品	正常	垃圾	正确识别		错误识别		F_N(%)	F_S(%)	G(%)	M(%)
			正常	垃圾	正常	垃圾				
苹果 iPad Air 2	40	10	34	8	6	2	89.47%	66.67%	73.46%	20.00%
魅族 MX4	40	10	36	7	4	3	91.14%	66.67%	76.64%	30.00%
真彩迪士尼书包	40	10	36	7	4	3	91.14%	66.67%	76.64%	30.00%
惠普 g14-a003TX	40	10	36	8	4	2	92.31%	72.73%	79.47%	20.00%

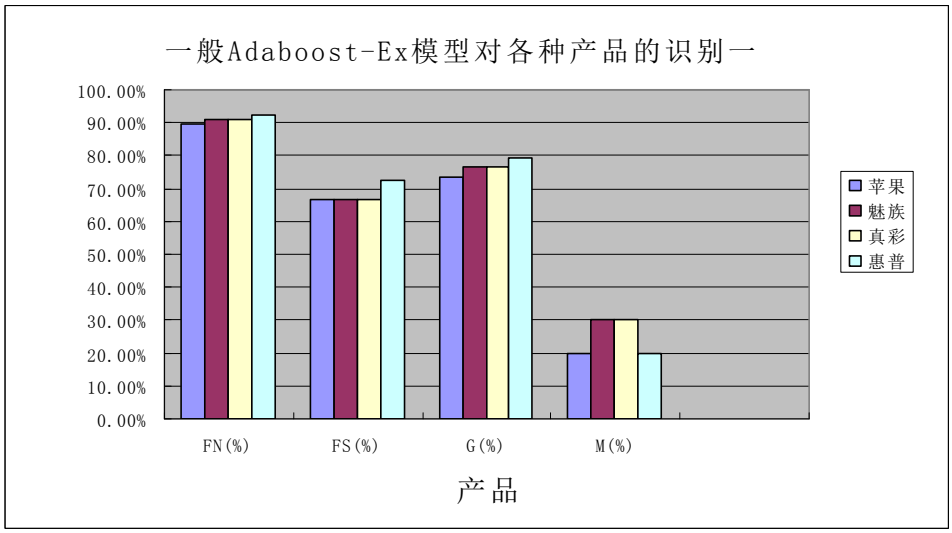

图4 一般Adaboost-Ex模型对不同产品评论的第一组识别结果

由图4可以看出用一般Adaboost-Ex模型对四种产品评论的识别效果都是很好的，其中衡量分类器对于非平衡数据集的整体分类性能的G指标在四种产品中都可以达到70%以上，这就说明一般Adaboost-Ex模型有很好的普遍适用性。

第一组测试结果如下表：

表13 一般Adaboost-Ex模型对不同产品评论的第二组识别结果

产品	正常	垃圾	正确识别		错误识别		FN(%)	FS(%)	G(%)	M(%)
			正常	垃圾	正常	垃圾				
苹果 iPad Air 2	10	40	7	35	3	5	63.64%	89.74%	73.30%	12.50%
魅族 MX4	10	40	8	35	2	5	69.57%	90.91%	76.30%	12.50%
真彩迪士尼书包	10	40	9	33	1	7	69.23%	89.19%	73.89%	17.50%
惠普 g14-a003TX	10	40	8	36	2	4	72.73%	92.31%	79.47%	10.00%

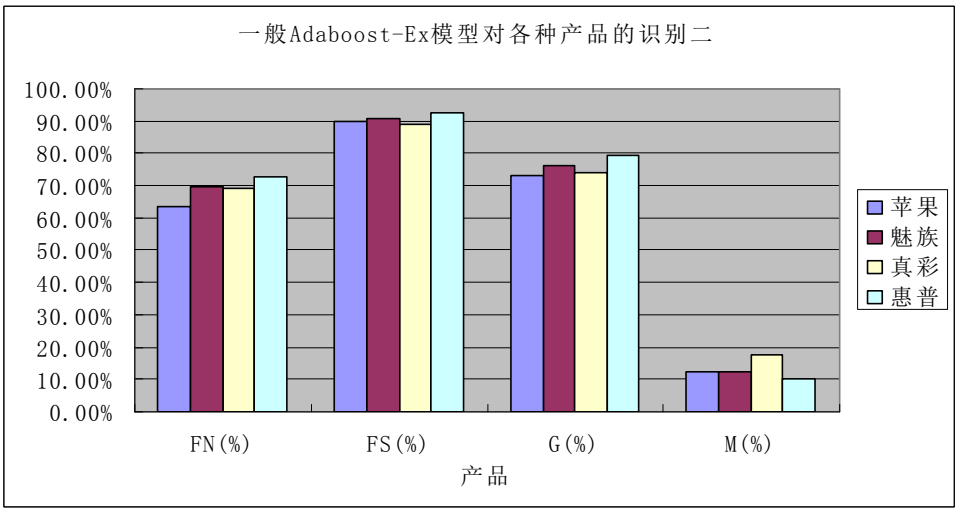

图5 一般Adaboost-Ex模型对不同产品评论的第二组识别结果

由此可以看出对于对于第二组测试其G指标也高于70%，综合对比一、二两组测试可知：一般Adaboost-Ex模型无论垃圾评论在占总评论的多少比重，都可以对垃圾评论由较好的识别效果，但是当垃圾评论占比例较小时，此模型的误判率会相对较高。

总的来说一般Adaboost-Ex模型在对各类产品垃圾评论的识别中，无论垃圾评论在占总评论的多少比重，各项指标总处于一个比较可观的水平。

六、模型评价

6.1 模型优点

1）考虑每条评论多个特征的值，表示成具有相同维数的特征值向量。有效利用了评论的多个重要特征，从而提高了垃圾评论的识别率。

2）考虑到通过线性加权的方式集成弱分类器，构建具有高精度的强分类器，有效提高弱分类器对于微博垃圾评论的识别率

3）该模型考虑了随着时间的流逝，现有评论的特征不足以清晰的区分正常评论与垃圾评论时，只需要增加新的评论特征即可，方便简单，有良好的扩展性。

4）考虑产品的评论集合的广泛性，从而利用动态词库调整与 Adaboost 模型相结合的想法，实用性更加广泛。

6.2 模型缺点

1）根据分词系统得到的分词，对于一些的、了、之类的词会加大作业量，没有进行相应的筛选。

2）对于某些只有几个字符的短小评论识别效果不好，此类评论的特征值几乎均为 0，导致很多短小评论拥有同样的特征值向量却分属于不同的类别。

6.3 模型改进

本文采用单阈值简单二值分类器作为弱分类器，在每个特征上进行训练时得到的阈值，不一定能很好的对正常评论和垃圾评论进行区分，为此需要寻找识别效果更好的弱分类器，例如双阈值简单二值分类器。从而来提高识别垃圾评论的能力。

七、 参考文献

[1] 李霄, 丁晟春, 垃圾商品评论信息的识别研究, 南京理工大学信息管理系, 高等教育出版社, 现代图书情报技术, 第 1 期, 2013 年.

[2] 张文霖, 主成分分析在 SPSS 中的操作应用, 慧聪国际行业研究院广州分公司.

[3] 黄玲, 李学明, 基于 AdaBoost 的微博垃圾评论识别方法, 重庆大学 计算机学院, 计算机应用, 33（12）,2013.

[4] 何海江, 凌云, 由 Logistic 回归识别 Web 社区的垃圾评论, 长沙学院 计算机中心, 计算机工程与应用, 45（23）,2009.

[5] 邓冰娜, 王煜, 刘宇, 一种应用于博客的垃圾评论识别方法, 河北大学 数学与计算机系, 43（1）,2011.

[6] 王庆福, 常广炎, 基于 TF-IDF 算法在文本分类中的应用研究, 辽宁行政学院, 电脑编程技巧与维护, 2014.

[7] 陈昀, 基于数据挖掘技术的产品垃圾评论识别研究, 河北大学, 硕士学位论文.

[8] 游贵荣, 吴为, 钱沄涛, 电子商务中垃圾评论检测的特征提取方法, 浙江大学计算机科学与技术学院, 251（10）,2014.

[9] 许少岩, 钟敏娟, 互联网产品信息评论中垃圾评论的识别方法浅析, 江西财经大学,2014.

附录

附录一：AdaBoost-LC 模型求解小米三评论集合的各个特征量的部分特征值：

评论	X1	X2	X3	X4	X5	X6	X7	X8	F
1	0.0000	0.0000	0.0454	0.0909	0.0000	0.0000	0.0000	0.0000	0.1363
2	0.0000	0.0000	0.3529	0.0000	0.0000	0.0000	0.1764	0.0000	0.5294
3	0.0000	0.0000	0.2083	0.0437	0.0000	0.0000	0.0000	0.0000	0.2500
4	0.0000	0.0000	0.0416	0.0000	0.0000	0.0000	0.0000	0.0000	0.0416
5	0.0000	0.0000	0.1818	0.0000	0.0000	0.0000	0.0000	0.0000	0.1818
6	0.0000	0.0000	0.1538	0.1538	0.0000	0.0000	0.0000	0.0000	0.3076
7	0.0000	0.0554	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0554
8	0.0000	0.0000	0.1304	0.0000	0.0000	0.0000	0.0000	0.2018	0.1304
9	0.0000	0.0000	0.0666	0.0000	0.0000	0.0000	0.0000	0.0000	0.0666
10	0.0526	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0526
11	0.0000	0.0000	0.1509	0.1509	0.0188	0.0000	0.0000	0.0000	0.3207
12	0.0000	0.0000	0.4166	0.0000	0.0000	0.0000	0.0000	0.0000	0.4166
13	0.0000	0.0000	0.1600	0.0800	0.0000	0.0000	0.0000	0.0000	0.2400
14	0.0000	0.0000	0.0769	0.0000	0.0000	0.0000	0.0000	0.0000	0.0769
15	0.0000	0.0000	0.0769	0.0000	0.0000	0.0000	0.0000	0.1222	0.0769
16	0.0000	0.0000	0.0645	0.1290	0.0000	0.0000	0.0967	0.0000	0.2903
17	0.0000	0.0000	0.0666	0.1333	0.0000	0.0000	0.0000	0.0000	0.2000
18	0.0000	0.0000	0.1153	0.1153	0.0000	0.0000	0.0000	0.0000	0.2307
19	0.0000	0.0000	0.4285	0.1428	0.0000	0.0000	0.1428	0.0000	0.7142
20	0.0000	0.0000	0.1111	0.0741	0.0741	0.0000	0.0000	0.0000	0.2592
21	0.0000	0.0000	0.2195	0.0975	0.0000	0.0000	0.0000	0.0000	0.3170
22	0.0112	0.0000	0.0779	0.0224	0.0000	0.0000	0.0000	0.1625	0.1982
23	0.0000	0.0000	0.0909	0.0151	0.0151	0.0000	0.0000	0.0000	0.1212
24	0.0000	0.0000	0.0000	0.2000	0.0000	-1.0000	0.0000	0.0000	-0.9120
25	0.0000	0.0000	0.0833	0.0000	0.0000	0.0000	0.0000	0.0000	0.0833
26	0.0000	0.0000	0.1818	0.1818	0.0000	0.0000	0.0000	0.0000	0.3636
27	0.0000	0.0000	0.1428	0.0000	0.2857	0.0000	0.0000	0.0000	0.4286
28	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0554
29	0.0000	0.0000	0.0666	0.0000	0.0000	0.0000	0.0000	0.0000	0.0666
30	0.0000	0.0000	0.1428	0.1428	0.0000	0.0000	0.0000	0.0000	0.2857
31	0.0000	0.0000	0.3333	0.0000	0.0000	0.0000	0.0000	0.0000	0.3333
32	0.0000	0.0000	0.0000	0.0000	0.1818	-1.0000	0.0000	0.0000	-0.8182
33	0.0000	0.0000	0.2352	0.0000	0.0000	0.0000	0.0000	0.0000	0.2352
34	0.0000	0.0000	0.2857	0.1428	0.0000	0.0000	0.0000	0.0000	0.4285
35	0.2000	0.0000	0.0000	0.0000	0.2000	0.0000	0.0000	0.0000	0.4000
36	0.0000	0.0698	0.0000	0.0000	0.0000	0.0000	0.0000	0.1235	0.1155
36	0.0151	0.0000	0.0606	0.0606	0.0303	0.0000	0.0000	0.0000	0.1666

附录二：关于小米 3 的 100 条评论集合以及人工与系统判定结果：

评论	内容	事实结果	模型结果
1	它是我喜爱的， 帮朋友买的。	0	1
2	不知道是不是话筒坏了，而且有时候接电话的时候，向右滑动滑动不了	1	1
3	很快很好。。。。。。	1	1
4	一次很不满意的购物！想不要了一个星期才到账？	0	0
5	用了 2 个月开不了机了 冲电手机发热 找售后没回应	1	1
6	带独立包装，非常好，绝对值得收藏	0	0
7	暂时没问题，还好用，下载速度很快	1	0
8	刚买几天的米 3 手机就有问题，屏幕不受人控（自动滚动屏幕），自动拨打电话，自动发信息，为了不冤枉扣通信费及流量只有关机，不知道怎么处理	0	0
9	手机大容易受热，看电视，玩游戏几分钟就碰不得了，充电也比较慢，苹果半个钟还看着电视就充满了	1	1
10	用着不错 看了一下 是真的	1	1
11	外观漂亮	1	1
12	帮同事买的，我也不知道哪好，他非的要，就买了，给个五星吧	0	0
13	还可以吧，一直在用，	1	1
14	充电不行 坏了 破手机	1	1
15	东西靠谱，价格符合商品！十个字	1	1
16	米 3 没的说。好用。尤其 GPS 定位快。	1	1
17	好好好好好好好好好好好好好	1	1
18	宝贝到手?	0	0
19	丈母娘喜欢，点名要小米。	0	1
20	很好，支持国产，支持小米	1	0
21	系统自带的系统不知道哪个版本的，还好我自己会刷，不然能给他砸了，大型游戏带不起来！	0	0
22	挺好用的前段时间有买了个小米 4 挺好的	0	0
23	说实话颜色不太好看，其他都还不错！不得不承认京东就是快……	1	1
24	商品非常好我很满意…	1	1
25	还可以，还没有插卡使用呢	1	1
26	好好用啊啊啊啊啊啊啊啊啊啊	1	1

27	机子超级好用，我喜欢，。	1	1
28	很好！京东值得信赖！	1	1
29	还行，，，，，　，　，	1	1
30	我认为这个还是可以的	1	1
31	还行，就是电池不太耐用，目前还没发现什么问题	1	1
32	写那么多字干什么呢？？？？？	0	0
33	买个女朋友老爸用的	0	0
34	还没有用呢，先来占个位置	0	0
35	生产日期是 2014 年 3 月的	0	0
36	太抠他们芋头陈军军模拟	0	0
37	才用一星期就被扒手偷走了	0	0
38	正在使用	0	0
39	呵呵，，，快递员竟然没给我找一块钱。	0	1
40	速度快，物流很给力！！！	0	0
41	…….....	0	0
42	好吧……YY h MM GBA	0	1
43	.*^o^*好空哦哦吐了咯哦哦	0	0
44	？？？？？？？？？？？	0	0
45	给老爸^ω^买的	0	1
46	京东白条，分期 3 个月买的，现在 第二个月，钱没还完，手机坏了，这心情。。	1	1
47	Xdufovotivigvfxertuficticuf	0	0
48	.。。。。。。。。。。。。。。。。。。。。	0	0
49	小米啊……降价真是快……哎~	0	0
50	公司搞活动，抽奖用的	0	0
51	小米 3 联通 64G 星空灰 联通 3G 手机	0	0
52	给堂哥买的，不知道怎么样	0	0
53	还不错吧，和小米 4 比起来味道不错	0	1
54	还行，为京豆儿评价，还行，为京豆儿评价	0	0
55	快递小哥不错	0	0
56	我对世界上的人充满爱，但给你的与众不同	0	0
57	用了这么久还没问题！	1	1
58	……～～……～…～…～～～	0	0
59	客服服务态度很好、物流配送很快	0	0
60	好不错就去可以的没问题	0	1
61	这会 3 降价了，赶紧给老妈搞一个，老人家用够了	0	0
62	耳机都没有，一千多块就配一个充电器	0	0
63	用了一个多月可是手机总是死机黑屏。求解是否可以换货，退货？	0	0
64	还可以哦？？？？？？？	0	1
65	暂时没有问题，不晓得质量能过关不？	0	

66	總體不錯就是有些次手沒碰到屏幕就會打開應 用程序	0	0
67	1.11111E+31	0	1
68	为什么比之前贵了 100 元	0	0
69	十个字十个字十个字十个字	0	0
70	帮他们买的，放了好多天都没用	0	0
71	!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!	0	0
72	买后居然涨了 100。。。。。。。。。。。。。。。 。。。。。。。。。。。。。。。。	0	1
73	人地说万句不如自己整部玩玩	0	0
74	入手米 3 的最好时刻，京东商城的服务越来越 好	0	0
75	长度在 5-200 个字之间 填写您对此商品的使用心得， 例如 该商品或某功能为您带来的帮助，或使用过程中遇到 的问 题等。最多可输入 200 字	0	0
76	龙记进攻龙年大吉恩公农行	0	
77	ai1234567889	0	1
78	支持国货。。。。。。。。。。。。。。。	0	0
79	太轻了 没有小米二好	0	0
80	京东实在实太给力了	0	1
81	但是当时的数据库上单多少	0	0
82	就是比官网上贵了一百	0	0
83	为什么非要填写啊！我是懒人。。。。。	0	1
84	hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh hhhhhhh	0	0
85	不咋滴当初直接上米 4 就好多了	0	0
86	昨天订的今天就到了给力	0	0
87	京东购物正规，令人放心。快递及时，快递人员服务 周到、热情。想不给五星都找不出理由。	0	0
88	轻薄，小巧，大屏，待机长，OTG 功能不错	1	1
89	给人家那个。。。。。。	0	0
90	非常好。。。。。。。。。。。。	1	0
91	拿回来还有 50gb 还是相当不错	1	1
92	感觉还可以，	1	1
93	商品不错 物流也快 一个字 爽	1	1
94	不错不错不错不错不错	1	1
95	还蛮好的，妈妈蛮满意的	1	0
96	用着还行，比三星用着流畅	1	1
97	手机非常好，京东的送货员也很友好，很幽默	1	1

98	老公买的手机，上一个被摔坏了，又买的同款	1	0
99	内存不小，但是真如所说为发烧而生那个烫不是一般烫	1	1
100	挺好的 给力给力给力	1	0

附件三：关于魅族 MX4 的 100 条评论集合以及人工与系统判定结果：

评论	内容	事实结果	模型结果
1	其实早就到了。一直没评价。	0	1
2	还是支持国产	0	0
3	心疼老大，希望给力。	0	1
4	我只想说 你们千万不要 Root	0	0
5	x4 很流畅 外观漂亮 感觉不错	1	1
6	值得一提的是快递是真快，从下订单到收货 14 小时给个赞	0	1
7	发货速度好快，半天到了	0	0
8	快递小哥真真是风雨无阻 习惯好评	0	0
9	山东分公司担任公司个人我让他玩儿他给我	0	1
10	加油京东我相信京东	0	0
11	flyme 做的越来越好了 但是还有很长的路要走 手机还行吧 没有过大的惊喜 就这样吧 支持国产 等用段时间再追评 感觉还是要华为的荣耀 6 好点	1	0
12	虽然退掉了，可服务绝对好的没话说	0	
13	物流人员超给力，下次再买	0	0
14	魅族 MX4 16GB 银白色 移动 4G 手机	0	1
15	感觉这手机没有之前用的 mx3 好用，有点卡	1	
16	支持国产机	1	1
17	……！……	0	0
18	垃圾手机.刚送来外观严重缺陷	1	1
19	5 星给快递员的，为我一个件，晚上 7 点多了从顺德大良送到伦教，谢谢。	0	1
20	ghiyDFgghughjgbju	0	0
21	Freeson 纳米防爆系列钢化玻璃手机保护贴膜 适用于魅族 MX4	0	1
22	就比 mx3 大一点点	1	1
23	小米 4，三星 s4,魅族 mx4 最后选了魅族，果然不失所望！	1	0

24	帮别人买的啊啊啊啊啊啊啊	0	0
25	手机收到了。手机很好玩 哈哈哈	0	1
26	fuvjfHGFF???	0	0
27	为什么要凑足十个字了？	0	0
28	国产最好。。。。。。。	0	0
29	关注这个挺久了，虽然没有降价，但是比较放心	0	1
30	怎么我买的时候，没有送钢化膜 和这个保护套呢？	0	0
31	京东正品，送货速度也快！一直 信赖京东！		
32	连个垃圾耳机都没的送！！！	0	0
33	j?s?j?s?j?d?j?g?h?s?k?s?k	0	0
34	不是本人使用的,暂且不便于评论.	0	0
35	早上下单，下午就送到了。好快的速度。京东直 营的速度好快	0	0
36	不能耍糖果传奇 还经常黑屏 不 好用	1	1
37	好大，超乎想像，手都没那么大。	1	1
38	哈哈哈哈哈	0	0
39	aaaaaaaaaaaaaaaaaaaa	0	1
40	就多个膜跟耳机就贵 160 元，比粗 粮还粗粮	0	0
41	帅，送货快，京东爱死你了…	0	1
42	不知道说什么好。。。	0	0
43	点解可以甘好用噶....	0	1
44	墙高.....	0	
45	碉堡了，就是没想象中那么好看	0	1
46	货品已收到,一天送达,京东物流不错,服务好, 热情周到,送货快,不错,赞！	0	0
47	期望已久的银翼，京东天狗确实比某猫要好，送 货死快，我能说早上 10 点半付，下午 3 点就送到 了吗？	0	0
48	评价大于 100 元的订单可以获得 20 个京豆（订 单完成 3 个月内有效）	0	1
49	京东购物向来都是快准好。服 务到位。	0	0
50	手机还是白色的好看	0	0
51	支持国产。。。。。。。。。。。	0	1
52	。。。。。。？。。。。	0	0
53	只能买套餐，有点贵， 别的都可以，京东发 货快	0	0
54	配送人员的服务满意度	0	1

55	第一次买到，虽然是套餐。	0	0
56	套餐,你懂的.....	0	0
57	国产神器，抢的人太多	0	0
58	11 月 11 号 0 点下单，2 分钟完成订单，4 分一刷新，就没货了，下午 3 点收到手机 ，非常不错做工好 ，高档大气。。。。。。。。。。。	1	0
59	支持国产机支持国产机	0	0
60	棒棒的擦擦擦擦擦擦擦擦	1	1
61	性价比高、中国 iphone 精品，支持	1	0
62	套装没有耳机让我很伤心。。。89 块钱啊也能漏了发？快递竟然把我的手机放在地上。。。？	0	0
63	我擦，这是愚人节的节奏吗	0	0
64	物流超快，第二天中午就到了。	0	0
65	今天很冷，送货员大哥辛苦了	0	0
66	京东送货很快，双十一抢到，下午到货	0	0
67	漂亮完美的四妹，现在已经爱不释手	1	0
68	膜挺好的。就是不知道耐刮么	0	0
69	可能双十一，货到的比较慢	0	0
70	不错过任何时候	0	0
71	贴上钢化膜之后像果六一样，但外观又比果六好看，比 iPhone6+好看很多哦，正面超赞	0	0
72	魅族 EP-21 耳机 适用于魅族 MX2/MX3/MX4 手机	0	0
73	帮妹妹买的，银色很漂亮，适合女孩子用	1	1
74	这个我也是看得着摸不着	0	0
75	太惊喜了，十点四十多秒抢到的，第二天就到啦，手机超酷	0	0
76	梦寐以求的，终于抢到手了	0	0
77	刚到手，物流很慢，也许是双十一缘故吧	0	0
78	在某猫抢金色的没抢上，来京东抢个白的也不错。支持国产。呵呵。。。咱也能晒单了你们就羡慕嫉妒恨吧？！鲁大师跑分，很牛逼。。	0	0
79	完美真的很喜欢，我的银翼我最爱	1	1
80	不错不错不错 已经在京东买了十来部手机了	1	1
81	很好，第二次用魅族了	1	1

82	非常好，很满意。。。。。	1	1
83	很好用……	1	1
84	手机外观好看，做工也好，但是手机放音乐外放时有很大的咂咂声，让我觉得很不满意	1	1
85	非常完美！喜欢！	1	1
86	一级棒！！！！！	1	1
87	超赞的！mx4 比我那三妹强不少啊！拍照太强悍	1	1
88	魅族这款手机绝对是卖的最好的一种	1	1
89	间歇性的跳屏+屏幕失灵，一般关闭屏幕重点亮一下就可以了，偶尔要重启。可是今天竟然重启时点关机或重启键都没反应。只能强制 15 秒重启。在考虑是否要退货了	1	1
90	手机还好啦，暂时买什么问题	1	1
91	魅族手机不错，系统还待提高！	1	1
92	手机很给力 物流有点慢 好评	1	1
93	支持魅族 支持国产 真心不错	1	1
94	质量好，系统流畅，像素高，比某米强	1	1
95	好用，好看，除去不支持内存卡，其它都不错	1	1
96	手机不错。。。。。。。。。。	1	1
97	买白色要买套餐而且算起来配件贵很多，京东好卑鄙	0	0
98	外观很好，系统流畅，大魅族还不错，先体验几天再说，哇咔咔	1	1
99	好的吧不错不错不错不错	1	1
100	试用中！好评！好评！好评论	1	1

附件四：关于真彩（TrueColor） M606156 迪士尼书包 卡通拉杆书包的 100 条评论集合以及人工与系统判定结果：

评论	评论内容	人工评判	模型评判
1	水很足水很足水很足水很足	1	0
2	就那样吧，还能咋的	1	1
3	为啥买 5 个就 6 折，买一个起码打个 8 折吗。	0	1
4	孩子挺喜欢的，一买回来就立刻把书装好了	1	0
5	质量不错的书包，质量挺好的	1	1

6	从 uiojk 网站刷出来的，历史低价。	0	1
7	活动时候买的，价格非常给力啊	0	0
8	.....	0	0
9	凑字凑字。。。。。	0	1
10	送货及时 值得购买	1	0
11	活动的时候买的，很划算，宝宝 很喜欢！	1	1
12	京东的东西比较让人容易接受 ，最好是自营的	0	0
13	古古怪怪方法刚刚不会吧韩国哈哈哈哈哈哈急	0	0
14	大概呵呵恶魔的问题是什么时候我知道的东西 噶车程侧扎的地方的时候就后悔否决策应	0	0
15	信赖京东 11111111111111	0	1
16	试了试感觉还行，质量有待考证， 用用看再说	0	0
17	还好？？？	1	0
18	送人的。。。。。。。。	0	0
19	快递师傅非常给力，认真负责， 谢谢！	0	1
20	小孩用用，挺好的 。	0	0
21	看上去还可以，牢不牢再说	1	1
22	稳定性要用段时间才知道，至于信号，我只能说不如我之前买的思科 CVR100W。同放在书房，且信道一个是 11，一个是 6，没有重叠扰，在厨房角落，思科的信号和网件的信号都只剩一个点，但思科可以流畅打开网页，网件就不行了，会提示网络连接有问题。	0	0
23	真彩的东西质量一般吧	0	1
24	京东非常棒，继续努力，加油	0	0
25	垃圾货……！	1	1
26	希望活动给力 继续支持	0	0
27	团购价格比较划算，感谢淘缤果网站推荐值 得购买	1	1
28	携带方便逢考必备日常使用使用 舒适	0	1
29	信赖京东。足不出户就能买到想 要的东西。	0	0
30	物流很快，京东值得信任	0	0
31	~~~~~~~~~~~~~~~~~~~~~~~~~~	0	0

32	买来送人的，赶上活动太值了	0	1
33	送小朋友必备，挺好的	0	0
34	客服还算好，帮我从别的仓库拿货	0	0
35	错过 600-500，悔啊，现在是 129 入的，唉	0	1
36	。。。。。。。。。。。。。。。。。。。。。。	0	0
37	物流非常的快，昨天晚上买的，第二天就到了	0	0
38	京东做活动优势没话说，希望多做点	0	0
39	和超市做活动时的价格一样	0	0
40	适合小学生携带，就是不知道学校让不让带	0	0
41	理解并学习与人交往和情绪控制的 EQ 技能。	0	0
42	再看看使用情况吧	0	0
43	价格再便宜就很好了。。。。	0	0
44	搞活动买的，便宜啊！	0	0
45	价格很划算喔 满 200 省 100 有点小味道 放在阳台上吹吹	0	1
46	活动时买的真便宜 比商场便宜多了	0	0
47	现在小孩上学都得拎这个拖着的，坑啊~~~贵	0	0
48	很感谢快递大哥在我下班前先给我送来	0	0
49	搞活动买的，买给马上上小学的女儿	0	0
50	东西不错，参加活动购买的，非常划算	0	1
51	等孩子上了学就可以用了。其实出去旅行也可以用的	0	0
52	400-300 买了两个，囤货，宝宝只有 2 岁。。。。是不是太早了。质量还可以，折下来每个 50 多	0	1
53	给我即将上小学的小妹妹买的，高兴得不得了	1	1
54	活动价买的，非常划算，满意。	1	1
55	好~~~~~~~~~~~~~~~~~~~~~~	0	0
56	送人合适，价格还行！！	0	0
57	活动价超值！	0	1

58	给我侄女买的礼物	0	0
59	呵呵给同事带的孩子要上小学了	0	0
60	特价买入，囤货中。。。。。	0	0
61	haixing~~~~~~~~~~~~~~`	1	1
62	满 600-500，很划算的哦	0	0
63	挺好，活动买的，相当的值哦	0	0
64	京东购物放心，很好，正品	1	1
65	很实惠，活动买的	0	0
66	价格实惠～～～～～～～～～～	0	0
67	对于刚上一年级的孩子来说，书包有些大，估计 3 年级以后使用会好些	1	1
68	轮子不是发光底轮，与图片不符 ，其他还 ok	1	1
69	拉杆书包好贵哦，不过活动下来 就很好了	0	0
70	看起来做工还可以，希望侄女喜欢	0	1
71	一折买的，超级划算	0	0
72	神价格不解释，期待下一次！！	0	0
73	家常必备 老妈满意就好 家常必备 老妈满意 就好	0	0
74	600-500 买的，你懂的	0	0
75	五花马千金裘呼儿将出换美酒	0	0
76	折下来 50 多吧...神价不解释，整体硬质的，小孩子用太大...可能到时春游什么的能用，现在就当行李箱了	1	1
77	贵了点，做活动后也要 100 多到手	0	0
78	送人的，没看到东西。。	0	0
79	京东购迪斯尼真彩丈青拉杆书包物美价廉送货快	0	0
80	东西可以 送货速度也很到位 就是送货员有些二 不付款不能看货 刷卡机没有信号自己在那发脾气 要他再试试就在那烦躁 要我付现金	0	0
81	给力非常给力，真的很给力。给 我好大力量	1	1
82	还不错，好评	1	1
83	不错，速度很快@！！！！！	1	1
84	呵呵，也得刚上小学的小朋友的必备用具。很轻松，装的东西也很多。	1	1

85	很不错的东西哦~很不错的东西哦~	1	1
86	100 元买了三个书包，还有无数文具，送货员送了个 26 寸拉杆箱那么大的货物，京东真彩的活动真心给力啊，良心电商	0	0
87	这件宝贝确实不错，和描述的一样好，卖家也很好，快递、服务都没的说，全 5 分！宝贝比想像中的要好,只是后悔买得太少.很不错 价格便宜	1	1
88	不错，送外甥！比较满意	1	1
89	对于幼儿园的宝贝来讲 太大	1	1
90	很不错的包包 喜欢	1	1
91	有个拉杆再也不怕书包太重了	0	0
92	给外甥买的，可惜活动好东西都被抢光了。。	0	0
93	买来送朋友家小朋友的，质量很不错	1	1
94	质量不错，图案很可爱，好评	1	1
95	很漂亮，价格超给力了	1	1
96	下单订好的货没了也不赔偿，说是工作人员发现货物有瑕疵，懒得扯，不满意	0	0
97	活动买的，再加上什么值得买的神卷，价格真的没话说	0	0
98	不错 就是有点贵哦，送小朋友的	1	1
99	一直想给孩子买个拉杆书包，价格很美貌哦	0	0
100	东西不错，好值得啊！！！	1	1

附件五：关于苹果 iPad Air 2 的 100 条评论集合以及人工与系统判定结果：

评论	内容	事实结果	模型结果
1	上午 9 点半下的定单，下午 6 点前就收到了’速度好快啊’必须?	0	1
2	长度在 5-200 个字之间 填写您对此商品的使用心得，例如该商品或某功能为您带来的帮助,或使用过程中遇到的问题等。最多可输入 200 字	0	0
3	hhhhhhhhhhhhjh	0	1
4	帮家人买的，小孩子玩游戏太疯狂	0	0
5	配送速度值得赞许，且给苹果客服点赞	1	1
6	?????????	0	1

7	包装很好，应用还在慢慢熟练过程中	0	0
8	给公司买的，应该还可以吧。	1	1
9	给老人买一个，外出携带看看东西，用着不错	0	0
10	我竟然只买了 16G 的 有点不够我用	0	0
11	这没啥说的，谁用谁知道	0	0
12	菲拉格慕（Ferragamo）仲夏之梦淡香水 30ml 菲拉格慕（Ferragamo）仲夏之梦淡香水 30ml	0	0
13	京东一直很值得信赖！	0	0
14	信赖京东，快，很好。	0	1
15	送货很快，用过后来评价	0	0
16	还没怎么用了 ……	0	0
17	出门旅行带着很实用	0	1
18	苹果的东西，大家都懂得。	1	1
19	。。。。。。。。。。。。。	0	0
20	想买就买不解释！想买就买不解释！	0	0
21	力量力量力量力量力量	0	0
22	还没有激活了，用了再来追加哦	0	0
23	二次选择夜间送货都没在夜间送过	0	0
24	一次性买了几十台。作为员工奖励	0	1
25	中午到手了，非常喜欢，早准备拍了，犹豫了好长时间	1	1
26	可以 现在正在琢磨…	1	1
27	满意京东服务	0	0
28	比起来实体店便宜很多呢哦哈	0	0
29	正品，京东购物放心，买	1	1
30	给外甥女的新年礼物，我自己是不喜欢这玩意的	0	0
31	帮朋友买的 快递很快	0	0
32	是正品！第一次来京东买东西！	1	1
33	买来送人的,购物方便	0	0
34	哈哈哈哈哈哈哈丰富的	0	0
35	评价晚了，非常满意京东的货物和服务	1	1
36	还把哈哈哈哈哎哎哎查查哈哈拆才拆才 c	0	0
37	发票呢？咋没有呢	0	1

38	送给老妈的大屏幕，再也不用看 小小的手机了	0	0
39	整体不错，看新闻用！	1	1
40	儿子的礼物，比玩手机强多了，强大的游戏机! ！呵呵	0	0
41	快递给力，次日达挺好。	0	0
42	送货很快。没什么问题	0	0
43	至少来说确实比安卓系统的要好。但是相同价位的就没有三星的那么美观了	1	1
44	给孩子买的，非常清晰	1	1
45	评价得京豆，评价得京豆	0	0
46	家里大大小小的东西基本都在京 东买的，放心	0	0
47	帮别人买的，昨天晚上下的单，今天就到了，赞一个，正在使用中	0	0
48	京东的物流太给力了！！	0	0
49	京东的快递太奇葩了，货直接扔家门口，也没有电话联系。东西掉了算你的吗？以后在京东买贵重物品要注意了。	0	0
50	携带方便.....	1	1
51	大小合适，携带方便，很符合个人的喜爱，大大的赞！就是京东的包装实在是寒碜啊，啥时候可以改善改善呢？！	1	1
52	在观察使用中，以后再补充	0	0
53	昨天晚上定的货，中午收到了，包装还行，还没 试试机，应该不会差	0	0
54	相信京东、放心京东、	0	0
55	买给朋友的，话说也没有给我 说什么不满意的话	0	1
56	在京东上购买就是方便，节省 不少时间	0	0
57	才过了 10 天今天一看 又降了 100 这都不给退京 东券？	0	0
58	东京的物流速度真是快，买的本本是国行的，回 来用了下，挺好的	0	0
59	没发票呢，发票什么时候到。	0	0
60	送了十五天才送到，真是服了，客户只会说等着	0	0

	把，也不跟你说货的状态		
61	这个还没用 看包装应该不错	0	0
62	差点就在别家买了，转到京东来瞅了一眼，真便宜，果断京东。	0	0
63	多岁的萨实打实的 多岁的萨实打实的 多岁的萨 实打实的 多岁的萨实打实的 多岁的萨实打实的 多岁的萨实打实的 多岁的萨实打实的 多岁的萨实打实的	0	1
64	范德萨范德萨范德萨范德萨发	0	0
65	送给嘟嘟滴礼物，希望他喜欢	0	0
66	不能任意拷文件进去 你懂的	1	1
67	京东当天购买当天收货，满意。	0	0
68	Dfsyhvdhncdhbv	0	1
69	yuhe86621yuhe86621yuhe86621 yuhe86621	0	0
70	jixin43jixin43jixin43	0	0
71	chongqian5224chongqian5224c hongqian5224	0	1
72	居然说我评价不规范 什么鬼 那我只能机智的截 图然后添加到晒单图片了 doge	0	0
73	啦咯啦咯来咯五天来咯哦理论陌 路加 Q 速龙	0	0
74	降价太快了，5 个月降了 700 多 ，也是醉啦	0	1
75	京东发货快，服务态度好，售 后有保障。	0	0
76	白条免分期 2488 入手 超值 没 拆封~不错不错	0	0
77	可以，就是速度慢了	0	0
78	送爸妈的结婚纪念日礼物	0	0
79	还好 ，，，挺不错的机子，，，	1	1
80	ipad 没什么好说的，速度很快	1	0
81	是正品，用起来就是爽	1	1
82	特价买的，性价比爆棚了，买对 时候了	1	1
83	很好用.....	1	0
84	老妈喜欢得不得了！天天抱着玩	1	1
85	满意，超级满意，京东保障！	1	1
86	送货及时，价格合理，商品适用	1	1

87	刚买完就降价*块，以后下单一 定要慎重了	0	1
88	还可以还可以还可以还可以	1	0
89	平板很赞，屏幕清晰，运行速度快	1	1
90	用了几天，感觉还可以，很强大	1	1
91	只能说游戏还是选择 32g 的。。16g 的只剩 10 个了 。。强烈建议	0	0
92	还行，为京豆儿评价，还行，为 京豆儿评价	0	0
93	考虑很久，买了。不后悔，很喜欢	1	1
94	不错！用几几天没发现问题！	1	1
95	这是第二次买这个 iPad，之前用的还不错，就 是 有时候有的卡，总体而言还不错	1	1
96	超好用，送货速度快，就是不能用 zhifubao 付 款 不是很方便	1	1
97	感觉就是不一样啊。好好好	1	0
98	好东西，比国产强多了。	1	1
99	不错不错不错不错不错不错	1	1
100	非常好，一点也不卡。速度快。	1	0

附件七：特征值提取 java 程序：

```
import java.io.*;

public class StringMatcher implements FilenameFilter
 {
     static String comment_words[];
     static int comment_times[];

     public static void main(String args[])
      {
            double sum;
            double persent[] = new double[8];//persent 为每个特征的特征值
            double bad;
            double good;
              try
           {
            File dir = new File("E:/program");//词库所在位置
            StringMatcher matcher = new StringMatcher();
            File file[] = dir.listFiles(matcher);
                  comment_times = new int[file.length];
```

```
            File comment_file = new File("E:/program/words/comment.txt");//分词后的文件所在位置
            FileReader in = new FileReader(comment_file);
            BufferedReader reader = new BufferedReader(in);

                    for(int m=0;m<100;m++)
                {
                    String comment = reader.readLine();
                    System.out.println(comment.trim());
                            comment_words = comment.trim().split(",");
                            sum = 0;
                            for(int i=0;i<file.length;i=i+1)
                                {
                                double    ck= 0;
                                comment_times[i] = 0;
                        for(int j=0;j<comment_words.length;j=j+1)
                                 {
                                    boolean                exist                =
lookUp(comment_words[j],file[i]);
                                    if(exist)
                                     {
                                                comment_times[i]                =
comment_times[i] + 1;
                                                }
                                    //System.out.println(file[i].getName());
                                    }
//                          System.out.println(comment_times[i]);
                                }
                        bad = comment_times[1] + comment_times[5];
                        good = comment_words.length - bad;
                        persent[0] = comment_times[0]/(good);
                        if(bad == 0)
                            {
                                persent[1] = 0;
                                persent[5] = 0;
                            }
                        else
                            {
                                persent[1] = comment_times[1]/bad;
                                persent[5] = comment_times[5]/bad;
                            }
                        persent[2] = comment_times[2]/(good);
                        persent[3] = comment_times[3]/(good);
                        persent[4] = comment_times[4]/(good);
```

```
                    persent[6] = comment_times[6]*3/(good);
                    persent[7] = comment_times[7]*3/(good);
                    System.out.println(comment_words.length);
                    for(int i=0;i<file.length;i++)
                     {
                              sum = sum + persent[i];
                              System.out.println(persent[i]);
                        }
                    System.out.println();
                    System.out.println(sum);//在初始权重下得出的评论的特
征值
                }
                }
        catch(Exception e) {}

    }

    public static boolean lookUp(String key,File file)
     {

        try
         {
            FileReader read = new FileReader(file);
            BufferedReader buffer = new BufferedReader(read);

                String str = buffer.readLine().trim();
             while(str!=null)
                    {

                if(str.equals(key))
                        {
                        return true;
                        }
                str = buffer.readLine().trim();
                    }
          }
        catch(Exception e) {}
        return false;
     }
    public boolean accept(File dir,String name)
     {
         if(name.endsWith(".txt"))
          {
              return true;
```

```
            }

        return false;
    }
}
```