基于 web 的股评观点挖掘系统①

莫 倩, 姜 越, 胡航丽

(北京工商大学 计算机与信息工程学院, 北京 100037)

摘 要: 互联网已经逐渐成为散户投资者获得投资信息的主要渠道. “大盘走势”是散户投资股市主要考虑的因素. 这里基于股评文章的特征设计实现了股评观点挖掘系统. 该系统利用基于模式的倾向性分析股评的方法, 识别并提取预测性观点句并通过倾向性分析最终获得股评的分类. 实验表明, 基于该方法的观点挖掘系统, 查准率达到了 91.7%.

关键词: 观点挖掘; 模式; 倾向性分析; 股评; 观点挖掘系统

An Opinion Mining System of Stock Recommendations Based on Web

MO Qian, JIANG Yue, HU Hang-Li

(Department of Computer Science and Technology, Beijing Technology and Business University, Beijing 100037, China)

Abstract: Internet has become the main channel of information on investment for individual investors. “market trend” is a major consideration for individual investors to investment market. Here to try to design the opinion mining system of stock, the system uses model-based method of tendentious analysis of stock analysts, to identify and extract the predictable view of statement classification and tendentious analysis of the final stock analysts. The experiment results show that the use of the approach makes the opinion mining system’s precision rate arrive at 91.7%.

Key words: opinion mining; mode; orientation analysis; stocks comments; opinion mining system

1 引言

根据《2010 年中国散户投资者报告》, 互联网已经成为绝大多数单户获得投资信息的主要渠道. 80.93%的单户表示其投资决策信息主要源自互联网. 在散户投资理念及操作策略方面, “大盘走势”是散户投资股市主要考虑的因素, 68.77%的散户是根据走势来决定是否要投资股市的. 而互联网上的股评铺天盖地, 大部分股民没有足够的时间和精力来阅读和总结股评, 他们关心的只是股票未来是涨还是跌. 所以对股评文章进行倾向性分类, 为股民提供总体的股市的走势是很有意义的一项工作.

目前, 在国内观点挖掘的研究主要集中在汽车评论、银行服务评论、电影评论、旅游目的地评论、电子产品评论等. 已有观点挖掘系统的一些实现, 如姚天昉等人的用于汉语汽车评论的意见挖掘系统[1], 该研究借助于极性词词典, 构建汽车本体有效处理句子中的实体、特征、极性词之间多对多关系、处理缺省主题的文本, 可以有效处理代词和成分缺省的文本. 王忠辉等人的中文网络评价信息挖掘系统(COMP)实现了在特征层次上对评价对象进行评价挖掘.

国外也有很多成型的观点挖掘系统, 最早的情感分析工具 ReviewSeer[2]; Yi 等人开发的 WebFountain 系统[3]; Wilson 等人开发的可以自动识别主观性句子以及句子中的其它主观性成分 OpinionFinder 系统[4]; 以及 Liu 等人开发的处理在线用户评论的 Opinion Observer[5]等; 还有 Wei Jin 等人在设计的 Opinion Minner 系统[6], 它提出了一种建立在词汇化 HMMs 架构下的新型机器学习方法.

本文在证券领域设计了一种股评观点挖掘系统, 旨在对互联网上大量的股评文章进行倾向性分类, 为

① 基金项目:国家自然科学基金(61170112);北京市教委科技创新平台建设项目(PXM2011_014213_113631)
收稿时间:2012-04-13;收到修改稿时间:2012-05-12

投资者(尤其是散户投资者)提供更为精简、明确的分类的股评信息，更好的帮助他们做出或买入或卖出或持有的决定.

2 系统框架

整体的互联网股评观点挖掘系统的架构图如图 1 所示.

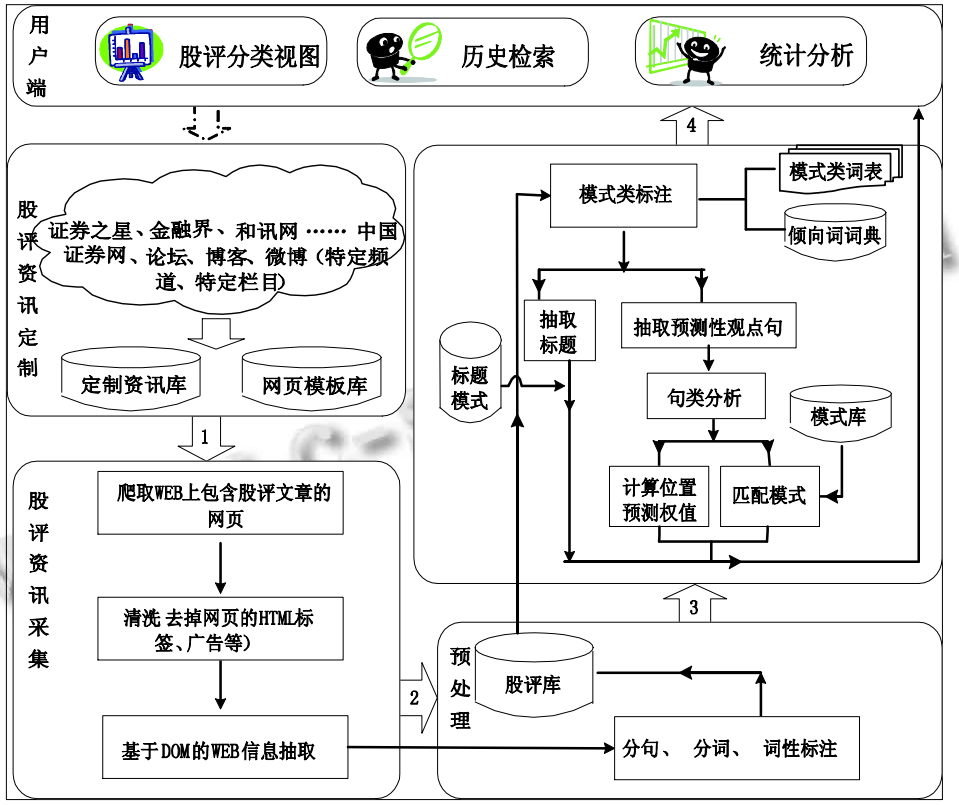

图 1 互联网股评观点挖掘系统框架图

互联网股评观点挖掘系统主要实现的功能，是采集互联网上的股评文章并对网页进行处理，对互联网上的某一时间段内的股评进行倾向性分析，并把结果展示给股民，为股民提供检索、统计分析等服务，使他们更为快捷地获得直观的“大盘走势”. 根据互联网股评观点挖掘系统的主要功能，该系统可以分为四个主要模块：股评资讯定制，股评资讯采集，预处理，混合倾向性分析，系统服务.

1) 股评资讯定制. 股评资讯定制模块为股民提供个性化的股评资讯定制服务. 该模块中的网页模板库是一些主要的证券网站、微博、论坛的网页模板.

2) 股评资讯采集. 股评资讯采集模块的功能是爬取互联网上包含股评的文章，并对网页进行进一步的清洗处理——去掉网页中的 HTML 标签、广告等，然后利用基于 DOM 的 WEB 信息抽取的方法抽取网页中的股评文章.

3) 预处理. 预处理模块功能是对股评文章进行分词、断句等处理，最后将处理后的股评文章存入股评数据库，为下一个模块的股评倾向性分析做好数据准备. 本文使用的是哈工大的分词程序对股评文章进行分词、词性标注处理.

4) 股评倾向性分析. 该模块是整个系统的核心模块，通过对股评进行构建倾向性词词典、构建模式类词表、获取模式，循环计算预测句子集中每个句子的倾向性，最后，利用预测句的倾向度和句子预测权值、位置权值计算出文档级别的倾向性.

5) 用户端. 用户端可以为用户提供股评分类视图显示、历史检索、趋势分析等服务. 股评分类视图展示旨在为股民提供友好、清晰的股评分类的结果视图，即将股评混合倾向性分析模块产生的结果可视化出来. 历史检索，即用户可以输入时间范围从而可以检索出这个时间段内所有采集到的股评文章，还可以检索出该时间段的股评的分类视图. 统计分析服务，可以通过某时间段的股评分类情况，预测未来股市的趋势.

3 关键技术

该股评观点挖掘系统中，最核心的模块是股评倾向性分析. 该模块中倾向性分析有两大子任务: 主观句的选择(识别并抽取表示预测观点的语句); 股评倾向性分类.

3.1 识别并提取预测性观点句

一般股评的标题也属于预测性观点句，例如“下探格局仍会延续”，“本周大盘有望再度挑战 3000 点”，标题中非常明确的包含了对未来的预测，像这样的标题就是符合标题模式的. 而像“大盘或将在 2310 一线继续震荡”这样的标题中出现了“整理”、“震荡”、“调整”等词，这类标题通通归为不明确类别，是不符合标题模式的.

而在股评文章的正文中，根据预测词词表和时间词词表中的未来时间词在待分类股评 r_i 中提取预测句组成预测语句集 F .

3.1.1 构建模式类

这里定义了六大模式类，模式类的符号解释对照表如表 1 所示:

表 1　模式类的符号对照表

模式类		符号	词类解释	词语示例
倾向词类	看多倾向词类	PW	表示股票后市良好的词	上行、高走、反弹、井喷
	看平倾向词类	NW	表示股票横盘震荡的词	震荡
	看空倾向词类	GW	表示股票后市惨淡的词	下挫、下跌、低迷、走低、减仓
否定词类	前置否定词	NP	位置在否定中心倾向词前面的否定词	不
	后置否定词	NX	位置在否定中心倾向词后面的否定词	不
转折词类	中心转折词	AC	后面带有转折中心句	但是
	非中心转折词	AP	后面所带的句子不是该转折复句阐述重点	虽然
时间词类	过去时间词	TP	表示过去时间的词	昨天、昨日、上周
	现在时间词	TN	表示现在、当前时间的词	现在、目前、今天
	未来时间词	TF	表示将来时间的词	明天、下周、将来
预测词类	预测动词	PR	表示预测的动词	预测、预计、有望
	人称代词		人称代词	我们、笔者
	其他短语		股评文章中表示预测观点的短语	认为、觉得
辅助词类	极性消除词	DE	带有否定意义的动词	排除、改变、告一段落、封杀、扼杀
	极性反面词	RE	反转倾向词的语义	扭转、转折、逆转、反转
	极性放大词	LA	放大了倾向词的极性程度	加强、增加、放大、升温
	极性缩小词	DU	缩小了倾向词的极性程度	下降、缩小、减弱、放缓
	极性正面化	PP	使倾向词的极性正面化	有助于、有利于、利于、重视、完善
	极性负面化	NN	使倾向词的极性负面化	压力、风险、挫伤、打击

1) 否定词类: 否定词前置和后置的情况，用符号表示如下:

否定词前置:

$$\{P,\cdots,NP,w_1,w_2,\cdots,w_n,PL,\cdots,P\}\quad (n\le\lambda_{\max})$$

否定词后置:

$$\{P,\cdots,PL,w_1,w_2,\cdots,w_n,NX,\cdots,P\}\quad (n\le\lambda_{\max})$$

其中 P 指标点符号，n 为否定词和倾向词之词长，$\lambda_{\max}$ 指否定中心的最大词长.

2) 转折词类: 根据中心句和非中心句中是否含有倾向次可分为三种情况，用符号表示如下:

$$\{P,\cdots,AP,\cdots,PL,P,AC,\cdots,PL,\cdots,P\}$$

$$\{P,\cdots,AP,\cdots,P,AC,\cdots,PL,\cdots,P\}$$

$$\{P,\cdots,AP,\cdots,PL,P,AC,\cdots,P\}$$

其中 AC 和 AP 不一定同时出现.

3) 时间词类: 时间词词表的构建方法如图 2 所示.

本文中，分别收集了否定词类 47 个、转折词类 17

个、时间词类 33 个, 其中一部分词来源于 HowNet, 另一部分为人工整理.

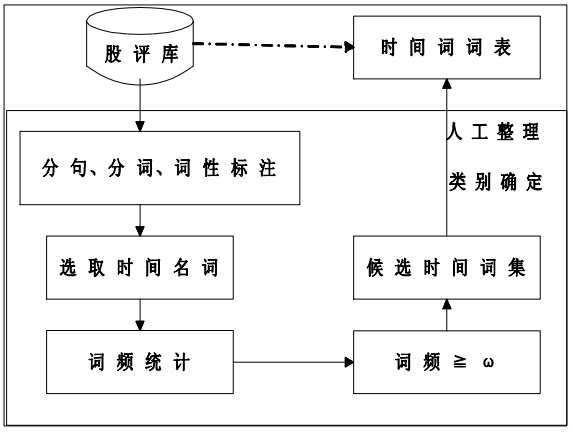

图 2 时间词词表的构建方法-

3.1.2 模式获取

股评文章可以看成是由各种模式类组合而成, 而最终的结果是整篇股评的倾向性——看多、看平或看空. 用符号 $token_j(1\le j\le N)$ 表示各种模式类, 不同的符号排列代表不同的模式. 如以下关系:

$$\{token_1 \wedge token_2 \wedge \cdots \wedge token_j \wedge \cdots \wedge token_n\} \Rightarrow \{PS|NU|NG\}$$

本文首先构建模式类, 然后通过如图 3 所示的流程提取模式:

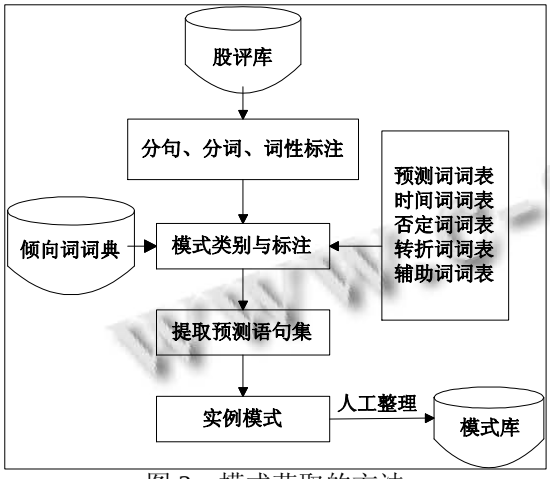

图 3 模式获取的方法

3.2 倾向性分析

3.2.1 位置权值、预测权值的计算

根据上面整理好的词表中的未来时间词在股评文章中提取预测句, 这些句子组成预测语句集, 用 F 表示, 我们用 n 表示 F 中预测句子总数, 用 f_i 表示 F 中第 i 个预测句, 则有关系: $f_i \in F(1<i<n)$. f_i 的权值由句子在全文中的位置(开头、中间、结尾)即位置权值 J_i 和 f_i 对未来大盘预测程度即预测权值 D_i 共同决定.

分别设位于开头、中间、结尾的位置权值为 ω_1、ω_2、ω_2, 设 N_d 是股评文章的句子总数, 并设 k 代表句子的序号, 则通过下面的公式(1)计算位置权值:

$$J_i=\begin{cases}\omega_1 & \left(\dfrac{k}{N_d}<\dfrac{N_d+6}{5\times N_d}\right)\\ \omega_2 & \left(\dfrac{N_d+6}{5\times N_d}\le\dfrac{k}{N_d}<\dfrac{4\times N_d-6}{5\times N_d}\right)\\ \omega_3 & \left(\dfrac{4\times N_d-6}{5\times N_d}\le\dfrac{k}{N_d}<1\right)\end{cases}\quad (N_d>1) \tag{1}$$

设 n 为预测词数目, m 为未来时间词数目, b_1、b_2、b_3 分别是三种情况下的预测权值. 则通过下面的公式(2)计算预测权值:

$$D_i=\begin{cases}b_1 & (n>0,m=0)\\ b_2 & (n=0,m>0)\\ b_3 & (n>0,m>0)\end{cases} \tag{2}$$

3.2.2 股评篇章级倾向性分析

本文对股评倾向性分类问题描述如下:

给定一个股评文本集 R, 观点分类器将每一个文档 $r_i\in R$ 分成三个类别: PS、NU、NG, 分别是指股评 r_i 预测短期未来走势是看多、看平、看空.

如果符合标题模式, 则直接通过标题的所属倾向性类别来获得整篇股评的极性.

如果标题不符合标题模式, 则提取待分类股评 r_i 预测语句集 F. 通过模式获得每一个预测句 f_i 在看多、看平、看空这三个类别上的极性值即 $Sim(f_i,PS)$, $Sim(f_i,NU)$, $Sim(f_i,NG)$. 分别设 J_i 和 D_i 代表预测句的位置权值和预测权值, 它们可由 3.2.1 节中的公式(1)和公式(2)计算出来. 设 μ_1 和 μ_2 两个自定义的阈值分别用来衡量看多和看空之间的极性值差距、看多和看空之间的极性值差距. 则股评 d 的极性 $Orientation(d)$ 计算公式如下:

$$\begin{cases}Sim(d,PS)=\sum_{i=1}^{n}J_i\times D_i\times Sim(f_i,PS)\\ Sim(d,NU)=\sum_{i=1}^{n}J_i\times D_i\times Sim(f_i,NU)\\ Sim(d,NG)=\sum_{i=1}^{n}J_i\times D_i\times Sim(f_i,NG)\end{cases} \tag{3}$$

$$Orientation(d)=\begin{cases} Sim(d,PS) & \begin{pmatrix} Sim(d,PS)-Sim(d,NG)>\mu_1, \\ Sim(d,PS)-Sim(d,NU)>\mu_2 \end{pmatrix} \\ Sim(d,NU) & \begin{pmatrix} Sim(d,NU)-Sim(d,PS)>\mu_2, \\ Sim(d,NU)-Sim(d,NG)>\mu_2 \end{pmatrix} \\ Sim(d,PS) & \begin{pmatrix} Sim(d,NG)-Sim(d,PS)>\mu_1, \\ Sim(d,NG)-Sim(d,NU)>\mu_2 \end{pmatrix} \\ Orientation(d)=0 \end{cases} \quad (4)$$

4　实验分析

4.1　实验结果

我们应用本文设计的股评观点挖掘系统对 2011 年 1 月份到 2011 年 9 月份从和讯网股票频道上采集下来的 3000 篇股评作为测试文本，图 2 为该系统股评采集模块采集下来的股评截图. 其中看多、看平、看空三个类别的文章各 1000 篇，实验结果如表 2 所示.

表 2　实验结果

	看多	看平	看空	查准率
看多	630	17	25	93.75%
看平	35	503	8	92.12%
看空	23	51	611	89.20%
查全率	63.0%	50.3%	61.1%	

4.2　实验对比

在测试样本一样的情况下，与参考文献[7]中基于篇章结构的 SVM 分类算法(算法 1)以及参考文献[8]中的 SO-PMI 算法(算法 2)相比，本文中的基于模式的算法(算法 3)在准确率上比算法 1 有一定的改善，在准确率和查全率上比算法 2 都有很大的提高. 由于本系统设计的模式不够全面所以在查全率上比基于篇章结构的SVM分类算法上稍逊一筹，只要对股评文章的模式进行进一步的总结，设计出更多、更合理的模式，那么这种算法的查全率将会有很大的提高.

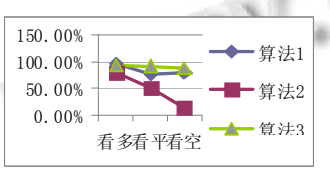

图 4　三种算法的查准率对比图

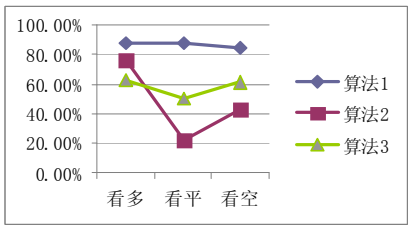

图 5　三种算法的查全率对比图

5　结语

本文构建了一种针对互联网股评的观点挖掘系统，其中包括了五个模块：股评资讯定制，股评资讯采集，预处理，股评倾向性分析，系统服务. 其中，采集模块利用了基于 DOM 的 Web 信息抽取的方法；预处理模块则对抽取下来的股评文章进行分词等预处理工作；股评倾向性分析模块则是本系统中最核心的部分，该模块混合了分别对股评的标题和正文中的预测性观点句子进行倾向性分析的方法，通过基于模式的方法获得股评文章最终的倾向性. 实验表明，基于该倾向性分析的方法的观点挖掘系统，查准率达到了 91.7%.

参考文献

1 姚天昉,等.一个用于汉语汽车评论的意见挖掘系统.中文信息处理前沿进展—中国中文信息学会二十五周年学术会议论文集.北京:清华大学出版社,2006,260–281.

2 Dave K, Lawrence S, Pennock DM. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. Proc. of the 12th International World Wide Web Conference(WWW2003). Budapest, Hungary, 2003.

3 Yi J, Niblack W. Sentiment Mining in WebFountain. Proc. of the ICDE205, the 21st International Conference on Data Engineering. IEEE Computer Society. Tokyo, Japan, 2005.

4 Wilson T, Hoffmann P, Somasundaran S, Patwardhan et al. Opinion Finder: A system for subjectivity analysis. Demonstration Description in Conference on Empirical Methods in Natural Language Processing. Vancouver, 2005.

5 Liu B, et al. Opinion observer: analyzing and comparing opinions on the web. Proc. of WWW'OS, the14th International Conference on WWW. Chiba, Japan, 2005. 342– 351.

6 Jin W, Ho H, Srihari R. OpinionMiner: a novel machine learning system for web opinion mining and extraction. Proc. of KDD. 2009. 1195–1204.

7 胡航丽,莫倩.利用篇章结构改进股评观点分类的研究.小型微型计算机系统,2009,30(5):899–902.

8 刘晓,莫倩,张政.网络评论观点分类研究.北京工商大学学报(自然科学版),2008,25(3):61–65.

(下转第 51 页)

表 1 毕业设计教务管理系统比较

开发单位	全过程监控	双向选题	安全性	量化评分	评语生成	学科适应性	出题选题统计分析
英国诺丁汉大学	否	是	弱	无	无	无	无
浙江大学	否	是	强	无	无	无	无
山东大学	否	是	较强	无	无	无	无
哈尔滨工程大学	否	是	弱	无	无	无	无
西南交通大学	是	是	较强	无	无	无	无
湖南科技大学	是	是	强	有	有	有	有

5 结语

基于 B/S 模式的毕业设计教务管理信息系统, 实现了毕业设计课题的双向选择, 对毕业设计各环节实施监控管理, 具有规范性、学科适应性和易用性的特点, 融入了 TQM 思想和工作流技术, 具有量化评分和自动生成评语的创新特色, 为师生提供了一个良好的毕业设计交互平台, 使得毕业设计的管理工作从以前繁重的手工操作中解脱出来, 提高了毕业设计过程管理的效率, 保证了学生设计论文质量, 为高校毕业设计信息化管理提供了先进的解决方案.

参考文献

1 郭秀娟,王春光.基于 B/S 模式的毕业设计管理系统开发与实现.计算机技术与发展,2010,20(3):239–242.

2 丁光惠,等.基于 B/S 的毕业设计管理系统开发.湖北汽车工业学院学报,2006,20(4):71–73.

3 李静梅,刘文佳.基于 J2EE 的毕业设计管理系统的设计与实现.应用科技,2010,37(1):45–49.

4 李章兵,刘建勋,赵肆江,龚波.基于 Web 的毕业设计教务管理系统的安全设计.信息安全与保密通信,2007,(5):90–92.

5 赵肆江,李章兵,刘建勋.基于量化规则的毕业设计成绩量化评定及评语生成.湘潭师范学院学报(自然科学版),2009,31(1):176–178.

6 李章兵,刘建勋,赵肆江,龚波.基于 TQM 思想和工作流的毕业设计过程质量监控管理.教师,2008,(5):37–38.

7 赵肆江,李章兵.计算机相关学科毕业设计课题的分类研究.福建电脑,2008,(8):33,49.

8 李章兵,郑明才,刘定.Windows DNA 和工作流技术支持的电子政务系统实现研究.计算机系统应用,2005,14(3):9–11.

(上接第 42 页)

9 莫倩,张渝杰,胡航丽,等.一种混合的股评观点倾向性分析方法.计算机工程与应用,2011,47(19):222–225.

10 Prabowo R, Thelwal M. Sentiment Analysis: A Combined Approach. Journal of Informetrics, April 2009,3(2):143–157.

11 胡航丽,莫倩.基于 Web 的股评观点倾向性分析研究[硕士学位论文].北京:北京工商大学,2010.

12 Liu X, Mo Q, Zhang Z. Research on opinion classification of internet reviews. Journal of Beijing Technology and Bussiness University (Natural Science Edition), 2008, 26(3): 61–65.

13 Tumey PD. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Association for Computational Linguistics 40th Anniversary Meeting(ACL), Philadelphia, 2002. 417–424.

14 Schapire RE, Singer Y. BoosTexter: A boosting-based system for text categorization. Machine Learning, 2000, 39(2/3):135–168.

15 Jiang L, Yu M, Zhou M, Liu X, Zhao T. Target-dependent twitter sentiment classification. ACL. 2011.

16 Liu B. Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data (2nd ed.). Berlin Heidelberg: Springer, 2008.

17 Akkaya Cem, Conrad A, Wiebe J, Mihalcea R. Amazon Mechanical Turk for subjectivity word sense disambigua-码tion. Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Los Angeles, CA. 2010. 195–203.