

DOI: 10.7652/xjtuxb201609013

基于短文本情感分析的敏感信息识别

李扬, 潘泉, 杨涛

(西北工业大学自动化学院, 710072, 西安)

摘要: 针对现有的敏感信息识别是基于敏感关键词匹配方式判断的, 准确度不是很高, 且具有较高的误检率等问题, 提出了敏感关键词与情感极性协同分析的敏感信息识别方法。在真实数据集上, 利用监督学习的方式对微博的情感极性进行了度量, 得到具体的情感极度, 并将文本分为正情感极性和负情感极性两类。通过定义色情、暴力、违禁、邪教、反动等 5 大类 2 639 个敏感关键词和在数据集里面所呈现的 Zipf 分布特性, 发现含有负情感极性的微博内容具有较高的敏感性, 由此深入研究了敏感关键词对情感极性的动力因素, 构建了含有情感极性因素的敏感度模型, 提出了敏感信息的判别方法, 敏感信息检测的准确率由传统方法的 31.25% 提高到了 58.75%, 召回率则由 95% 提升到 96%, 综合指标 F 值从 47.0% 提升到了 72.3%。

关键词: 社交网络; 情感分析; 敏感信息

中图分类号: TP271 **文献标志码:** A **文章编号:** 0253-987X(2016)09-0080-05

Sensitive Information Recognition Based on Short Text Sentiment Analysis

LI Yang, PAN Quan, YANG Tao

(School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: The existing sensitive information recognition is based on the sensitive keyword matching method, so the accuracy is low and the miss rate is high. We presented a collaborative method by using the sensitive keywords and sentiment polarities to identify the sensitive information. In the real dataset, we used the supervised way to measure the sentiment polarities of the blogs, and divided the blogs into two categories, namely the blogs are with positive or negative sentiment polarities. Five kinds of 2 639 sensitive keywords, including pornography, violence, illegality, cult and reactionary, were defined, and it was found that according to the Zipf distribution of these words in the dataset, the contents of blogs with negative sentiment polarities exhibited high sensitivities. Then we studied the contribution of the sensitive keywords to the sentiment polarity, and constructed the model of sensitivity degree that contains the sentiment polarity factor. Based on this, we proposed a new way to identify the sensitive information, which makes the accuracy and miss rate improved from 31.25% to 58.75% and from 95% to 96%, respectively, and the F -measure was improved from 47.0% to 72.3%.

Keywords: social networks; sentiment analysis; sensitive information

近几年来, 微博的出现和发展给信息发布、传播方式带来了巨大变革, 以往大量的信息接受者一跃成为信息的发布者。数以亿计的微博用户每天发布

大量微博, 或记录生活, 或表达看法, 或仅仅宣泄一种情绪。网络环境复杂, 如何有效识别色情、暴力等敏感信息的文本对构建和谐社会具有重要意义。当

前微博敏感信息的识别及过滤多以敏感关键词进行,这种方式将所有含有敏感关键词的内容统统滤除,具有一定的局限性。在社交网络平台上,用户的信息内容蕴含一定的情感,有些是传达积极高兴的心情,而有些则是表达消极郁闷的情绪。目前,可以通过分类的方式对微博进行情感分析。

情感分析已经成为一个重要的研究领域,对中文文本的情感分析也日渐成熟^[1-2]。情感分析主要是研究用户对某一商品、事件、话题、服务或者组织所流露出的情感、评价、态度及情绪等^[3]。就文本的情感分析而言,一般涉及3个问题:极性分类(polarity classification),分为积极的和消极的两个极性;主客观分类(subjective/objective classification);情感强度分类(rating inference)等^[3-4]。其中对文本的极性分类最为重要,通过分析用户广播的某种状态信息或者评论信息,来判断文本内容是正面的赞赏还是负面的批判^[5-6],是舆情分析的重要研究内容。情感分析的方法主要分为有监督的学习训练模型、无监督模型及半监督方式,文献[7-12]利用深度学习的思想进行情感学习训练,很大程度上去除了特征向量的选取,文献[7]利用 tree-bank 的方式对句子的结构进行描述,并利用 RNTN 的方式对句子的情感进行了分析,文献[12]直接利用音形文字的字符作为输入,从根本上去除了分析过程中的分词、词干化等步骤,降低了信息的丢失率,提高了情感分析的准确率。

然而,只分析微博文本的情感是不够的,如何表示用户发表信息的敏感性及煽动性是舆情分析关注的重点^[5,13-14]。例如,2011年爆发的埃及革命就是不法分子利用 Twitter、Facebook 等社交网络媒体大肆造谣,煽动民意,传播恶意信息并组织犯罪活动,在社交媒体的帮助下,骚乱被极度放大并快速演变^[15]。信息敏感性对事件有很强的推动性,因此对社交媒体信息的敏感性及情感极性态势的掌握有重要的意义。

本文主要针对传统敏感信息检测误检率和漏检率高的问题,提出了基于情感分析的敏感度模型,以增强敏感信息的检测。首先,将敏感信息分类,根据与敏感相关的先验信息,诸如“作弊器”、“统一教”等词,将其分为色情(518个)、违禁品(391个)、不文明用语(122个)、邪教(209个)及政治(1399个)等5大类敏感关键词,并构建敏感关键词种子库;然后,介绍了敏感关键词在实际数据中具有分布特性,对微博短文本进行了情感极性分析,介绍了敏感关

键词对情感极性的动力因素,并在此基础上构建了 DS(degree of sensitive)模型,分析验证了 DS 模型的有效性;最后,利用 DS 模型对微博的敏感程度做了评测及效果展示。

1 微博短文本关键词分布特性

因为含有敏感关键词并不意味着真正含有敏感信息,所以只利用敏感关键词对敏感信息进行过滤欠缺一定的合理性。随机提取 2014 年 3 月 1 日到 4 月 13 日的 150 386 条微博进行关键字抽取,并将其分为含有敏感关键词和不含关键词两类,给出了微博文本含有敏感关键词个数的分布,如图 1 所示。

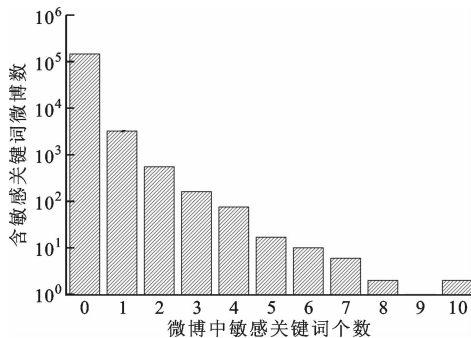


图1 微博含有敏感关键词个数的分布

由图 1 可得,不含敏感关键词的微博占 97.29%,只有 2.71%的微博文本含有敏感关键词,而这 2.71%的微博中,约 79.80%只含有一个敏感关键词,20.20%含有两个以上的敏感关键词,呈现出 Zipf 分布模式,对这 79.80%的微博进行分析,发现此类微博约 60%并没有包含敏感信息,只具有敏感关键词。

文献[3]提出了对文本情感极性分类比较有效的方法,一般通过最优化的方式解决,即

$$\min_w = \|Cw - y\|^2 \quad (1)$$

式中: C 为训练的样本数据; w 为学习到的特征系数; y 为样本数据的情感标签。

2 微博短文本情感分析

情感分析是一个分类问题,通过抽取文本情感特征,并构造相应的特征向量,然后通过有监督或无监督的方法进行训练及参数学习,得到一个表现优良的分类器,进而实现情感的基本分类^[3,16]。一般可将情感简单分为正极性及负极性情感两类,而 Dong 等通过构建词语级的情感语料库,并利用快速贝叶斯方法构建了情感分类器,将微博情感细分为愤怒、厌恶、高兴和低落等^[2]。

本文研究的是如何在情感基础上对微博的敏感性进行定义、分析,使用有监督的方法,例如支持向量机、朴素贝叶斯等对文本的情感进行分类,文献[17]验证了支持向量机在分类问题上具有很好的效果。对含有不同敏感关键词的微博进行情感分类,训练情感模型的数据是从提取的微博当中选取具有明显情感倾向符号的微博,共33 211条,其中含有正情感符号的24 373条,含有负情感符号的8 838条,根据含有表情符号进行标准情感分类,在此基础上计算每条微博的情感得分,计算得到的得分区间为 $(-3,3)$,如图2所示。

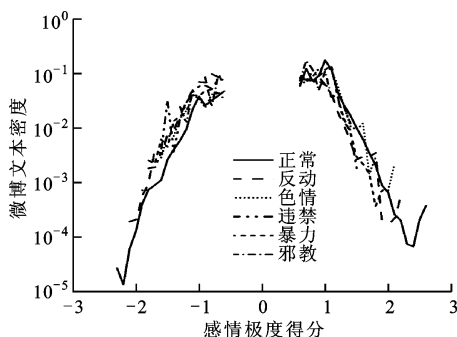


图2 情感得分分布

在同等比例情况下,含有敏感关键词较不含敏感关键词微博的情感极度得分的值要低,说明敏感关键词对文本的情感极度有内在动力因素,使含有敏感关键词微博的情感极度得分偏低。

3 DS建模

3.1 DS模型分析

假设微博的敏感性与其含有的敏感关键词个数有正相关性,即微博含有敏感关键词个数越多,表达的敏感程度(以下简称敏感度) S 就越大。

为了定量表达敏感关键词与情感极度之间的关系,定义全局敏感关键词频

$$F_s = \sum_{i=1}^n \frac{s_i}{w_i} / n \quad (2)$$

式中: n 为微博总条数; s_i 为一条微博中敏感关键词的个数; w_i 为该条微博中总词数。

微博的情感极度、正(负)平均情感极度为

$$D = LP \quad (3)$$

$$\bar{D} = \frac{\sum_{i=1}^m |D_i|}{m} \quad (4)$$

式中: P 为该标签得分; m 为正(负)情感极性微博数; L 为模型预测出的情感标签, $L \in \{-1, 1\}$,当 $L = 1$ 时,微博的情感极度属于正情感极度 D_P ,而当 L

$= -1$ 时,则属于负情感极度 D_N 。

相关关系是一种非确定性的关系,相关系数是研究变量之间线性相关程度的量。因此,可根据相关系数来定性分析出两者之间的关系,计算公式为

$$\rho = \frac{\text{cov}(X, Y)}{(D(X))^{1/2} (D(Y))^{1/2}} \quad (5)$$

式中: $D(X)$ 为变量 X 的方差; $\text{cov}(X, Y)$ 为变量 X 、 Y 的协方差。分析3月份微博的敏感关键词频与 $\bar{D}$ 之间的相关性,结果如图3所示。

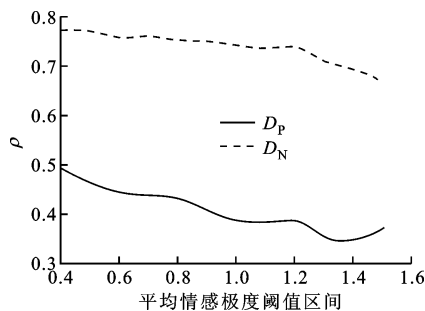


图3 $\bar{D}$ 与敏感关键词频的相关性

由图3可知, D_P 与敏感关键词频之间一直处于低度相关的状态,而 D_N 与敏感关键词频之间的相关度一直处于0.75左右,其相关度明显高于 D_P 与敏感关键词频之间的相关度,定量证明了微博敏感度 S 与情感极度 D 之间的关系,即负情感极性的微博出现敏感关键词的可能性大于正情感极性的微博,定义敏感度模型

$$S = P t f_s^{-5} \quad (6)$$

$$f_s = \frac{s}{L+2} \quad (7)$$

式中: t 为敏感关键词底数,本文取 $t=2$; s 为敏感关键词的个数; f_s 为敏感关键词频,假定文本中最多出现的敏感关键词不超过5个。当 $L=-1$ 时, $f_s=s$;当 $L=1$ 时, $f_s=s/3$ 。通过式(7),降低了 D_P 对 S 的贡献,而加强了 D_N 对 S 的贡献,使得 S 与 D_N 之间的相关性加强,即保证了敏感关键词对负情感极性微博的动力更强。

3.2 DS模型的验证

为了验证模型的有效性,利用式(6)对3月的12万多条微博进行敏感度 S 的计算,将所得结果按 S 区间进行分类,并计算相应区间所有微博的 $\bar{D}$,结果如图4所示。

由图4可知, $\bar{D}$ 随着 S 呈现先增后减的趋势。当一条微博里没有出现敏感关键词时, S 的最大值为0.031 25,所以规定 $S < 0.031\ 25$ 的微博属于正常微博,此区间的 S 即为 D ; $S > 0.031\ 25$ 时,随着 S

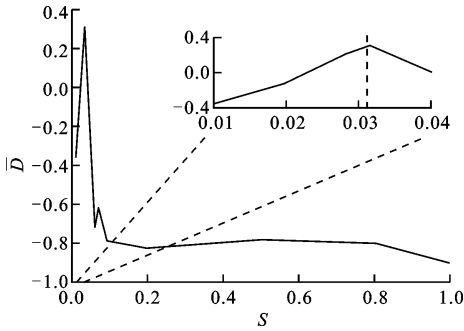


图 4 $\bar{D}$ 和 S 之间的关系曲线

的增加 $\bar{D}$ 呈现递减趋势,微博内容越敏感,则情感值越低。

对于一个事件集合 $\{A,B\}$,其中 A 的发生频次为 F_0 , B 的发生频次为 F_1 ,则事件 A 的支持度函数为

$$\text{supp}(A) = \frac{F_0}{F_0 + F_1} \tag{8}$$

条件概率定义为

$$\text{conf}\{A \rightarrow B\} = \text{supp}(\{A,B\})/\text{supp}(\{A\}) \tag{9}$$

由式(8)可计算出 3 月份每天微博 $S>0.05$ 、 $D_p>0.5$ 时的支持度,结果如图 5 所示。

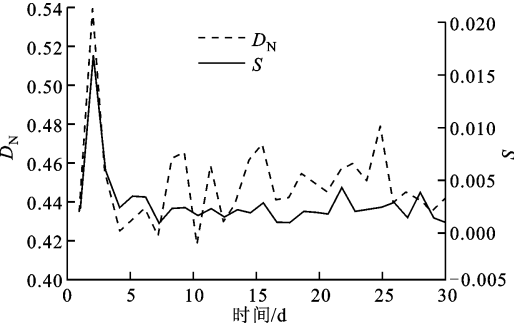


图 5 $D_N>0.5$ 、 $S>0.05$ 时的支持度曲线

由图 5 可看出,两条曲线的走势几乎相同,说明负情感极性的微博数量越多,越有可能出现敏感性的微博。在一定 S 阈值范围内情感极性的条件概率可得到两条曲线之间的条件概率,结果如图 6 所示。

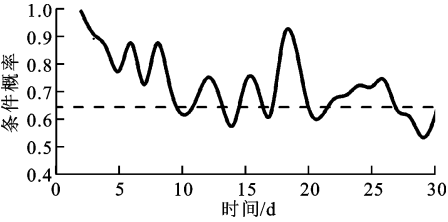


图 6 $S>0.05$ 、 $D_N>0.5$ 时的概率分布

由图 6 可知, $S>0.05$ 、 $D_N>0.5$ 时的条件概率一般维持在 0.65 左右,敏感性微博和负情感极性微博同时出现的概率较大。因此,在海量微博文本中

定位敏感型微博时,可先利用情感极性做初步筛选,从而可缩小搜寻范围。

3.3 实例分析

为了能够真实展示网上微博内容的情感及 S 的分布情况,利用敏感关键词库对,对 2014 年 3 月 1 日至 3 月 31 日之间的微博进行了 S 与 D 的计算,关键词匹配结果是:准确率为 31.25%,召回率为 95%,综合指标 F 值为 47.0%;DS 模型结果:准确率为 58.75%,召回率为 96%, F 值为 72.3%。可以发现,DS 模型在敏感信息的查找方面较敏感关键词匹配的方式有较好的效果,在损失较小召回率的情况下,能够有效提高准确率,大大提高了敏感信息的检测效率,结果如图 7 所示。

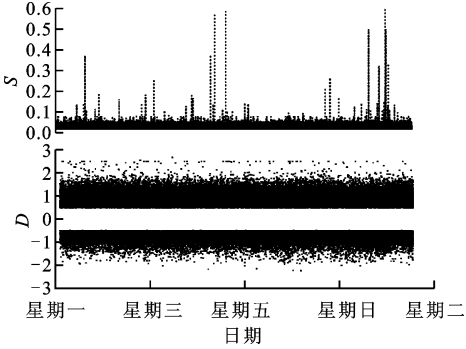


图 7 S 和 D 的时间序列图

由图 7 可知,微博平台上的正情感极性的微博要多于负情感极性的(0 轴上方为积极情感部分),由此可以推断该月微博平台上的正能量成分较多,大部分微博的 S 值处在 0.03 以下。分析这些微博可知,有些虽然也可能含有一个敏感关键词,但是所述内容很正常,而 S 值在 0.03 以上的微博的内容或多或少都含一定的敏感信息,例如“这个作弊神器真专业”,其 S 值为 0.035,具有一定的敏感性,诸如“开山队辛苦的用炸药开山”的 S 值为 0.023,虽然含有“炸药”这一敏感关键词,但整句话的内容却是正常的。因此,可通过经验设定敏感阈值,从而能有效观察到一个月内敏感性微博的分布。

4 结 论

本文定量分析了微博敏感性与其情感极度之间的关系,得出负情感极性的微博更易含敏感信息的结论,并且分析了敏感关键词的个数在微博中的 Zipf 分布特性,完成了 DS 模型的构建,并对模型进行了验证。通过实例分析,该模型使得敏感信息的识别与检测更加智能有效,利用负情感极性微博易含敏感信息的特点,能在海量数据中快速进行敏感

信息筛选。与传统敏感关键词匹配的检测方法相比,本文方法在检测的准确率、召回率及 F 值方面均有所提升。

参考文献:

- [1] WU K, ZHANG B, ZHENG J, et al. Sentiment classification for topical Chinese microblog based on sentences' relations [C]// The IEEE International Conference on Cyber, Physical and Social Computing. Piscataway, NJ, USA: IEEE, 2013: 2221-2225.
- [2] ZHAO J, DONG L, WU J, et al. Moodlens: an emoticon-based sentiment analysis system for Chinese Tweets [C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2012: 1528-1531.
- [3] LIU B. Sentiment analysis and opinion mining [J]. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1-167.
- [4] WANG G, SUN J, MA J, et al. Sentiment classification: the contribution of ensemble learning [J]. Decision Support Systems, 2014, 57(1): 77-93.
- [5] 张鲁民, 贾焰, 周斌. 一种基于情感符号的在线突发事件检测方法 [J]. 计算机学报, 2013, 36(8): 1659-1667.
ZHANG Lumin, JIA Yan, ZHOU Bin. Online bursty events detection based on emoticons [J]. Chinese Journal of Computers, 2013, 36(8): 1659-1667.
- [6] CAO J, ZENG K, WANG H, et al. Web-based traffic sentiment analysis: methods and applications [J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(2): 844-853.
- [7] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013: 1631-1642.
- [8] SANTOS C, GATTIT M. Deep convolutional neural networks for sentiment analysis of short texts [C]// Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland: Coling, 2014: 69-72.
- [9] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification [C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Menlo Park, California, USA: AAAI, 2015: 2267-2273.
- [10] ZHOU C, SUN C, LIU Z, et al. A C-LSTM neural network for text classification [EB/OL]. [2016-02-20]. <http://arxiv.org/abs/1511.08630>.
- [11] WANG P, XU B, XU J, et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification [J]. Neurocomputing, 2015, 174: 806-814.
- [12] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification [EB/OL]. [2016-02-20]. <http://arxiv.org/abs/1509.01626>.
- [13] WANG X, ZHU F, JIANG J, et al. Real time event detection in Twitter [J]. Lecture Notes in Computer Science, 2013, 7923: 502-513.
- [14] WENG J S, YAO Y X, LEONARDI E, et al. Event detection in Twitter [C]// Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Menlo Park, California, USA: AAAI, 2011: 401-408.
- [15] ZHOU Donghao, HAN Wenbao. Diffrank: a novel algorithm for information diffusion detection in social networks [J]. Chinese Journal of Computer, 2014, 37(4): 884-892.
- [16] BOLLEN J, PEPE A, MAO H. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena [C]// Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011). Menlo Park, California, USA: AAAI, 2011: 450-453.
- [17] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

(编辑 赵炜)