

基于 RAE+Dropout 相结合的微博情感分析

1. 南华大学计算机科学与技术学院 衡阳 421001;
2. 国家行政学院 北京 100089

李坚¹ 肖基毅¹ 欧阳纯萍¹ 阳小华¹ 翟云²

摘要 文本情感分析是近年自然语言处理领域的研究热点之一, 其中微博情感分析受到了学术界和企业界的广泛关注。微博情感分析是指对用户针对某一事件发表的言论进行正向、负向和中性情感的判定。本文在分析了标准 RAE 模型缺点的基础上, 提出了一种基于 RAE+Dropout 的联合模型。该模型利用 Dropout 技术有效地预防过拟合问题的发生, 同时也提高了模型训练速度。RAE+Dropout 模型与 RAE+ 词性选择模型、标准 RAE 模型以及 SVM 模型的对比实验结果表明: RAE+Dropout 模型的准确率和 F1 值属于最优, 比标准 RAE 模型的准确值和 F1 值高出 0.82% 和 0.64%, 尤其是在高维词语向量中 RAE+Dropout 模型的效果更加明显。

关键词: 情感分析, Dropout, RAE, 微博情感分析

中图分类号: G35

开放科学(资源服务)标识码(OSID)



Micro-blog Emotion Analysis Based on RAE + Dropout

1. School of Computer Science and Technology, University of South China, Hengyang 421001, China;
2. Chinese Academy of Governance, Beijing 100089, China

LI Jian¹ XIAO JiYi¹ OUYANG ChunPing¹ YANG XiaoHua¹ ZHAI Yun²

基金项目: 本文受国家自然科学基金(61402220、61672178), 湖南省哲学社会科学基金(14YBA335、16YBA323), 湖南省研究生科研创新项目(CX2016B446), 浙江省自然科学基金(LY13F020024)的资助。

作者简介: 李坚(1990-), 硕士, 研究方向: 自然语言处理, 机器学习, Email: lijian2033@qq.com; 肖基毅(1962-), 硕士生导师, 研究方向: 社交网络与大数据应用技术, 数据挖掘, 数据库应用, Email: 532735539@qq.com。

Abstract Recently, text emotion analysis is one of the most popular research hotspots in the field of natural language processing. Furthermore, micro-blog emotion analysis has received extensive attention in academia and industry. Micro-blog emotion analysis aims to judge the positive, negative and neutral emotion of the speech delivered by the user for an event. In this paper, firstly, the shortcomings of the original model were illustrated according to the standard RAE model, and then an improved model, which is a joint model of RAE + Dropout was put forward. This model aimed to prevent the problem of over-fitting as well as improve the speed of model training by Dropout technology. At last, The RAE + Dropout model was compared with the RAE + POS model, the standard RAE model and the SVM model respectively. The experiment results indicated that: RAE + Dropout model has the best accuracy and F1 value, which are 0.82% and 0.64% up on the results of the standard RAE model, especially the RAE + Dropout model has an obvious advantage in the high-dimensional word vector representation.

Keywords: Emotion analysis, dropout, RAE, micro-blog emotion analysis

1 引言

随着互联网技术的发展和普及,越来越多的用户在微博、贴吧、论坛、社区等社交媒体上发表自己的观点和意见。由于微博信息具有发布及时性,内容形式随意性等^[1]特点,已成为人们最喜爱的社交媒体。根据最新统计数据显示,新浪微博的注册用户已经超过了5亿,用户每日的发博量超过1亿条。由于网络资源数据十分庞大,仅仅通过人工方式来筛选大众观点信息是一件十分困难繁琐的工作,因此,利用计算机帮助人们快速、高效地筛选和整理大众观点信息是十分迫切的。文本情感分析是大数据背景下自然语言处理的一个新兴研究领域,具有重要的学术研究价值和实际应用价值,受到了国内外学者专家的关注,其中微博情感分析在文本情感分析中占有重要的地位。

2 相关工作

在文本情感分析领域,国内外学者已取得了一定的成果。目前针对文本情感倾向性分析主要集中于以下三种方法^[2]:基于简单统计的倾向性分类、基于机器学习的文本倾向性分类和基于相关性的文本倾向性分类。Hatzivassiloglou 和 McKeown^[3]使用监督学习方法对英文词语进行情感语义倾向性判别,其准确率达到90%左右。Peter D.Turney^[4]指出利用情感词和标记词库的互信息值来判断文本的情感倾向性。在2003年,Turney^[5]提出了利用统计信息对单词进行语义倾向判断的新方法。Kamps^[6]等利用WordNet中词语的同义结构图,通过待测词语与种子词的相似度计算得出待测词的情感倾向值。Chunhua JU^[7]则通过构建评价词集和情感词集,对情感倾向性进行细粒度分析,从而得到文本的情感倾向性判断。

近年来,随着数据挖掘和深度学习技术的发展,科研工作者将这项技术广泛用于文本情感特征提取和情感分析。在国外, Richard Socher^[8] 等人利用递归自编码器进行语句特征的提取,从而预测句子的情感特征; Wenya Wang^[9] 等人将递归神经网络与条件随机场结合的方法对句子进行情感判断; Salah Rifai^[10] 等人利用卷积自编码器对数据特征进行提取; Pascal Vincent^[11] 等人利用去噪自编码器进行训练数据特征。在国内,梁军^[12] 等人将长短期记忆扩展到基于树状结构的递归神经网络上,并根据句子前后词语间的关联性引入情感极性转移模型,用以判断文本情感极性。虽然深度神经网络模型在自然语言处理方面得到了成功的应用,但是如何提高深度神经网络模型的泛化能力和处理速度仍是一个值得深入研究的问题。

根据已有的研究成果,本文提出一种基于递归自编码器的扩展方法,利用 Dropout 方法对语料信息进行泛化处理操作,用于提高微博语料信息分类的准确率和效率。本文利用深度神经网络在图形图像方面的应用^[13-15],根据微博语言本身特有的特征,为了提高微博情感分析的泛化能力防止过拟合现象的出现,引入 Dropout 方法^[16],Dropout 方法能够有效的防止训练出来的模型发生过拟合现象,同时提高了模型训练的速度。

3 RAE模型

3.1 标准RAE模型

RAE 的主要思想:将一个句子看做由若干

个词语组成的序列,将序列中相邻的两个词语进行合并产生一个重构误差,遍历优化将重构误差最小的词语两两合并,最终将句子抽象成类似哈弗曼树结构,称之为短语树。短语树中的叶子节点为句子的词语,非叶子结点为组合词语的语义向量,最后的任务就是求得短语树最优的重构误差。

递归神经网络最早由 Goller^[17] 提出,根据句子语法的树形结构由底向上递归合并两个相邻的结点。基本公式如下所示:

$$p = f(W^{(1)}[x_1 : x_2] + b^{(1)}) \quad (\text{公式 1})$$

$$[x'_1 : x'_2] = g(W^{(2)}p + b^{(2)}) \quad (\text{公式 2})$$

$$E_{\text{rec}}([x_1 : x_2]) = \frac{1}{2} \| [x_1 : x_2] - [x'_1 : x'_2] \|^2 \quad (\text{公式 3})$$

其中 $[x_1 : x_2] \in R^{2n}$ 由 $x_1 x_2$ 联接得到, n 为词语向量维数, $W \in R^{n \times 2n}$ 为参数矩阵, $b \in R^n$ 为偏置项,函数 f 为 encoding 函数,函数 g 为 decoding 函数, E_{rec} 为重构误差函数。

为了提高微博信息特征提取的可靠性和准确性,做出以下两点规定:

(1) 在这里我们将所有的结点进行标准化,这样做有两点好处,一是方便计算;二是防止零向量带来的误差影响。

$$p = \frac{f(W^{(1)}[x_1 : x_2] + b^{(1)})}{\sum f} \quad (\text{公式 4})$$

(2) 为了更好的衡量不同节点的权重,将节点出现的次数作为衡量权重大小的标准。

$$E_{\text{rec}}([x_1 : x_2]; \theta) = \frac{n_1}{n_1 + n_2} \|x_1 - x'_1\|^2 + \frac{n_2}{n_1 + n_2} \|x_2 - x'_2\|^2 \quad (\text{公式 5})$$

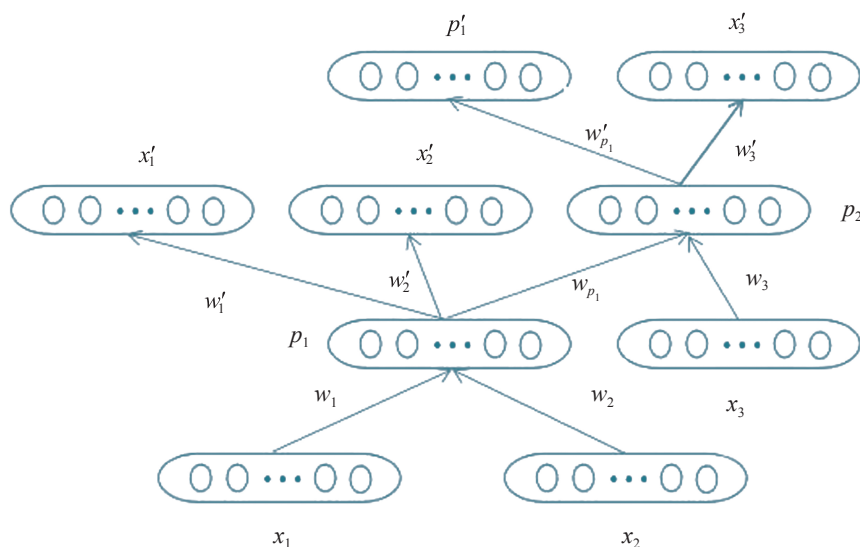


图1 短语树重构模型

3.2 基于Dropout技术的RAE模型

尽管神经网络在很多方面取得了成功应用，但是过拟合和训练周期长一直是神经网络模型的缺点。公式 6 和公式 7 为标准前向神经网络模型的公式：

$$z_i^{(l+1)} = w_i^{(l+1)} y^l + b_i^{(l+1)} \quad (\text{公式 6})$$

$$y_i^{(l+1)} = f(z_i^{(l+1)}) \quad (\text{公式 7})$$

为了提高神经网络分类器的泛化能力和处理速度，本文中引用 Dropout 方法，该方法来源于有性生殖理论^[18]。Dropout 具体指在模型训练时随机让网络某些隐层结点的权重不工作，不工作的那些节点可以暂时认为不是网络结构的一部分，但是它的权重值会被记录（暂时不更新），便于下次样本输入时重新使用这部分节点。基于 Dropout 的前向神经网络模型公式如下：

$$r_j^{(l)} \sim \text{Bernoulli}(p) \quad (\text{公式 8})$$

$$\tilde{y}^{(l)} = r^{(l)} * y^{(l)} \quad (\text{公式 9})$$

$$z_i^{(l+1)} = w_i^{(l+1)} \tilde{y}^{(l)} + b_i^{(l+1)} \quad (\text{公式 10})$$

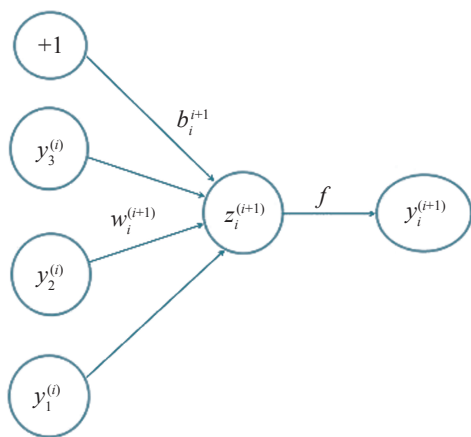
$$y_i^{(l+1)} = f(z_i^{(l+1)}) \quad (\text{公式 11})$$

上面公式 8 中 Bernoulli 函数，是为了以概率 p 随机生成一个 0、1 的向量。

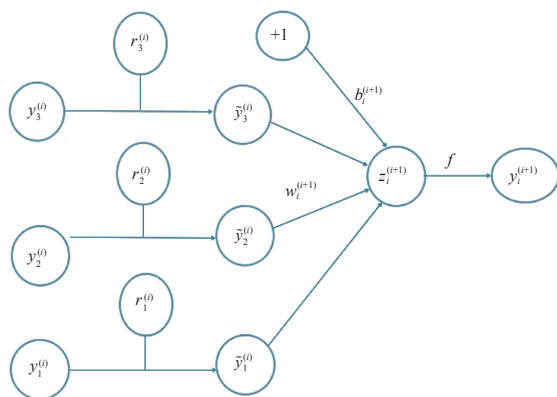
为了更加形象比较标准前向神经网络模型与基于 Dropout 的前向神经网络模型的差异，本文对比了两种前向神经网络的模型图，如图 2 所示。

Dropout 技术主要用于防止实验结果过拟合，其主要原理是忽略掉模型中某些神经节点，相当于在模型中注入了一些噪声，这样做的主要目的是防止数据训练过程中模型发生过拟合现象，同时提高模型数据训练的速度。Dropout 主要使用在当训练样本较小，试验样本较大的数据集中，这一特性正好符合微博信息的特点。基本模型如图 3 所示。

通过观察 Standard Neural Net 和 Dropout Neural Net 的结构图，可以很清楚的看出使用 Dropout Neural Net 不仅可以提高程序的运行速度，而且还可以提高模型的泛化能力，从而提高结果的准确率和速度。

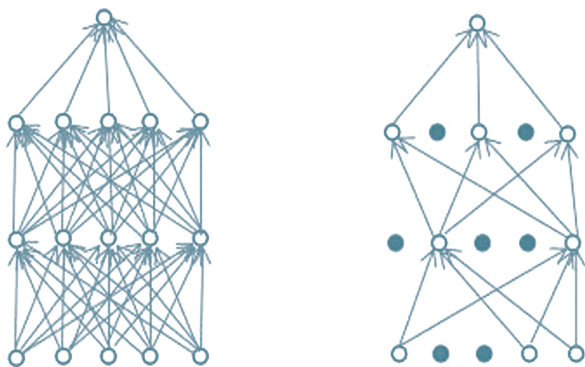


(a) 标准神经网络模型



(b) Dropout神经网络模型

图2 公式对应的神经网络模型



(a) Standard Neural Net

(b) After applying dropout

图3 标准神经网络模型和Dropout模型

4 实验设置及结果分析

4.1 数据预处理

本文使用的是NLPCC-ICCPOL 2016比赛中所提供的语料数据集,其中正面情感信息1165条,负面情感信息1209条和中性情感信息626条,总共有3000条数据。这3000条数据全是标记好的数据,总共包含五个主题,分别为: IphoneSE、春节放鞭炮、俄罗斯在叙利亚的反恐行动、开放二胎和深圳禁摩限电,每个主题为600条信息,包含正面信息、负面信息和中性信息。

在训练与实验阶段,为了更好获得实验结果的正确性,本文采用的评价指标为准确率(Accuracy)和F1值,同时采用交叉验证的方式进行训练和实验。

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (\text{公式 } 12)$$

$$F1 = \frac{2TP}{2TP + FN + FP} \quad (\text{公式 } 13)$$

其中, TP 为被检索到相关的样本数量, FP 为被检索到无关的样本数量, FN 为未被检索到相关的样本数量, TN 为未被检索到无关的样本数量;

前期我们对数据进行了以下预处理:

(1) 首先对语料集中的文本进行预处理,去除掉文本中的网络符号。

(2) 然后对预处理后的实验数据进行分词。为了验证词语词性对结果的影响,本文中将对分词后的句子进行词性筛选^[19-21],只保留形容词、副词、情感词,用词向量表示文章并保持语序不变。

(3) 将训练数据中所有的词语构成词库,并标记序号,根据词库将词语向量化表示并作为输入数据。

表1 数据集样例表

主题	正面	负面	中性
IphoneSE	已用两周，满意，尤其电池续航！	这是翻新 5s 清理库存吧。	iPhone SE 的上市，新款系列也应运而生了。
春节放鞭炮	今天是大年初一，传统意义上的春节，放鞭炮接财神吃饺子一样不能少。	# 春节不放鞭，沈阳见蓝天 # 为了我们的呼吸更加清新 要少放鞭炮。	# 春节健康养生 # 欢度春节放鞭炮，惊吓小儿会哭闹。可以用手指由轻及重，按四神聪、百会、风池穴，能让孩子安静下来。
俄罗斯在叙利亚的反恐行动	极端势力必须清除。	新武器的试验场啊。	重大消息叙利亚和平有望。
开放二胎	19 号顺产二胎男宝，8.42 斤，母子平安。凑个好字。	别开放二胎了好不好。急死我啦！挤死我得了！车站快塌了我感觉。	真受不了啦！今天大学群成立，才知道有好多人都是二胎了。让我情何以堪！
深圳禁摩限电	深圳管的挺好	为了好管理，就一禁了之，这样的管理水平，懒政真是懒到了一定程度了。	分享视频：深圳禁摩限电首日拘留三百余人！！ http://t.cn/RqLQaP2 秒拍视频

4.2 参数的选择

Dropout 技术能够在模型训练阶段防止过拟合现象发生，同时提高模型训练的速度。本文中在训练阶段和实验阶段的 p 值取 $0.5^{[15,22,23]}$ ，然后对改进后的 RAE 模型参数进行选优，其中包括训练迭代次数、词语向量表示维数和交叉验证文件数。

本文采取以下参数选优方案：首先将 3000

条数据集根据主题等比例随机选择 2800 条数据作为训练集，200 条作为实验集；然后对训练集进行训练，期间将训练集作为训练模型的实验集，选出结果较好的模型参数值；最后根据选出的模型参数进行组合，来作为 200 条实验集的实验模型参数。

图 4，5，6 是利用训练集合作为实验集合并得出的实验结果：

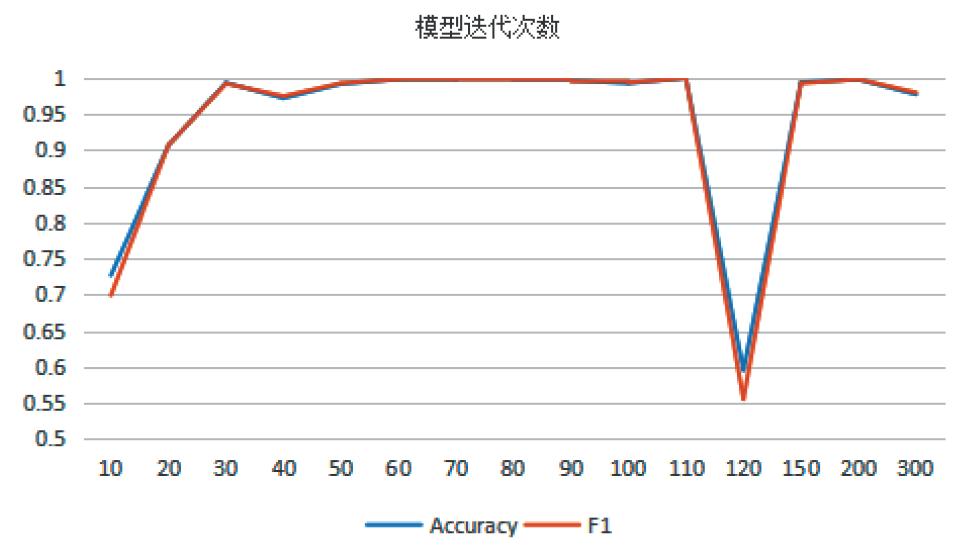


图4 模型训练迭代次数

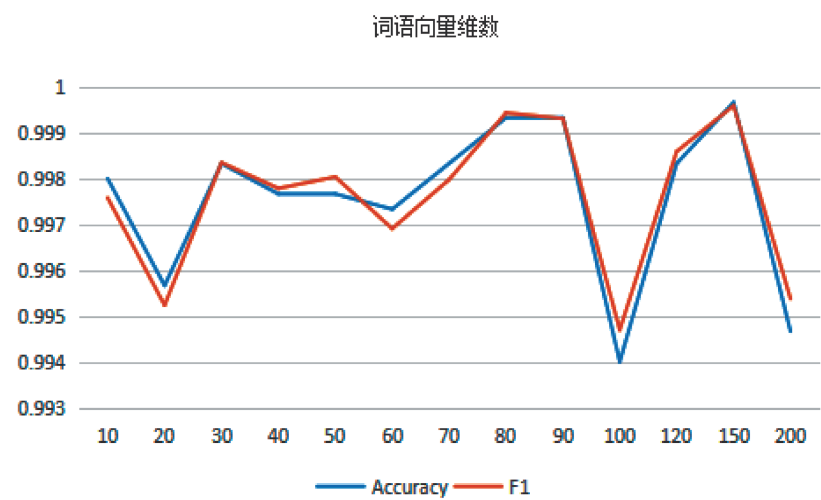


图5 模型词语向量维数

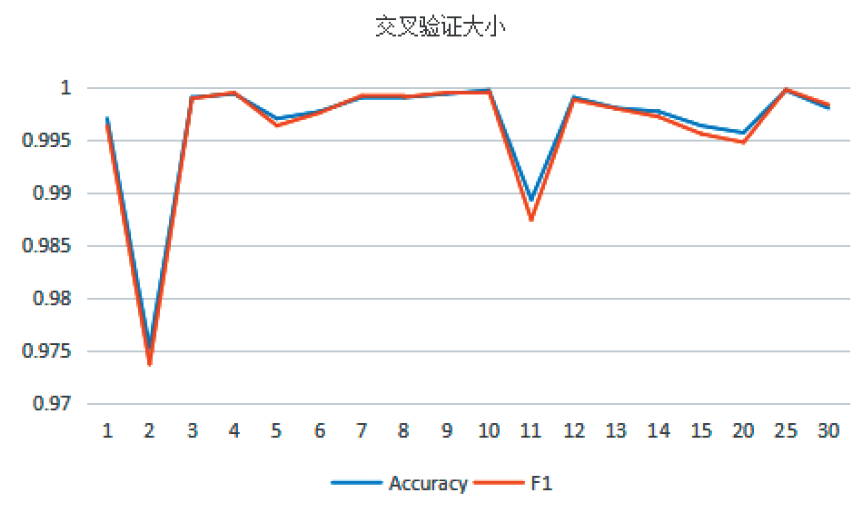


图6 模型交叉验证

观察上图可以得出：模型训练的迭代次数在 70、110，词语向量表示维数在 80、150 和交叉验证文件数在 10、25 处具有较高的 A 值和 F1 值。

4.3 不同RAE模型的比较

将模型训练阶段训练好的参数进行组合，共 8 组情况，如表 2 所示。为了能够说明实验结果的正确性，在对照试验过程中，我们

保持一个变量变化，其他变量一致，进行实验对比。

表2 模型参数组合表			
编号	NumFolds	MaxIterations	EmbeddinSize
1	10	70	80
2	10	70	150
3	25	70	80
4	10	110	80
5	25	110	150
6	10	110	150
7	25	70	150
8	25	110	80

实验结果如图 7 所示：

通过 RAE+ 词性选择模型与 RAE 模型比较可以得出：在 8 种组合 RAE 模型中 A 值始终比 RAE+Dropout 模型的 A 值大，平均大出 1.568%；同时 RAE 模型中的 F1 值只有在 2 情况下小于 RAE+ 词性选择模型的 F1 值，其他情况均大于它。

通过标准 RAE 模型和 RAE+Dropout 模型的实验结果进行对比可以得出：对比实验在参数 2、5、6 和 7 情况下，RAE+Dropout 模型的 Accuracy 和 F1 值比标准 RAE 模型更优，平均值高出 0.82% 和 0.64%，可以看出 Dropout 技术达到了降维的目的，提高了模型的准确率和 F1 值。

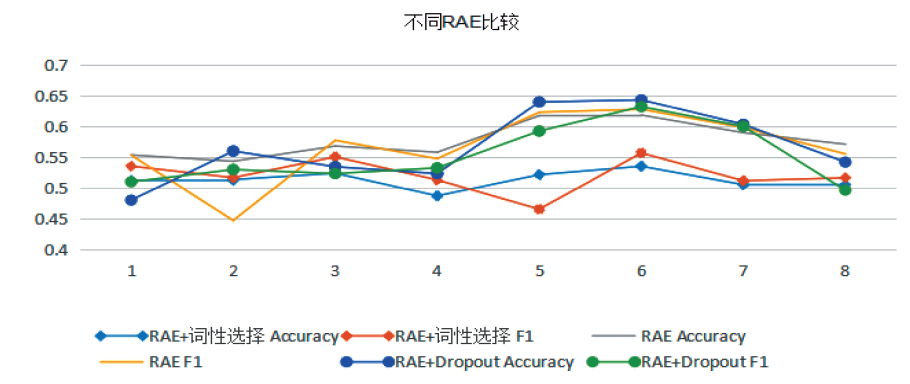


图7 不同RAE实验结果

4.4 RAE和SVM算法比较

SVM 作为有监督学习的经典算法，在本文

中为了比较半监督学习和监督学习的性能，做以下对比试验，实验结果如图 8 所示：

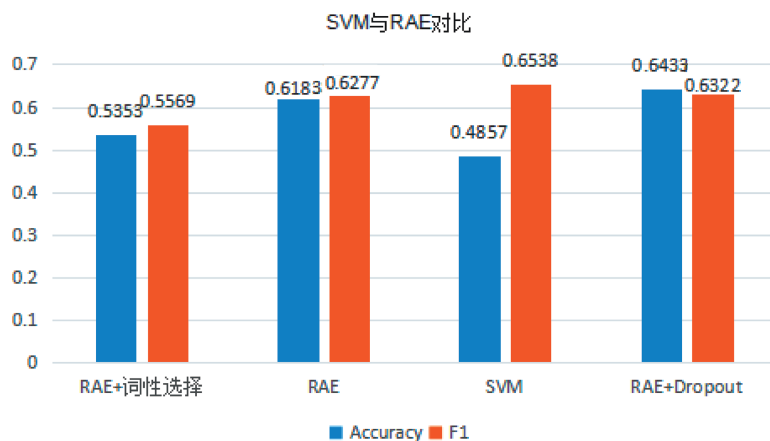


图8 SVM与RAE模型比较

通过图 8 可以得出：监督学习 SVM 的 Accuracy 低于半监督学习的 RAE+ 词性选择、RAE、RAE+Dropout 的 Accuracy，但是 F1 值

却比这三种半监督学习模型的 F1 值高，可以看出监督学习 SVM 算法在实际的应用中具有一定的优势，单纯的比较 Accuracy 值却不是很理

想, 应该与 SVM 的特征选择有一定的关系。

RAE+Dropout 模型中的 Accuracy 值高出 SVM 模型的 Accuracy 值 15.67%, 针对高维词语向量模型优势更加明显, 其中 4 组高维词语向量的实验结果显示出了 Dropout 方法对于预防模型训练过程中出现过拟合现象具有良好的能力。

5 总结

本文提出了一种基于 RAE+Dropout 相结合的微博情感分析模型, 该模型能够有效的预防过拟合现象的出现, 同时缩短了模型训练的时间。实验结果表明, 基于 RAE+Dropout 相结合的微博情感分析模型的准确率高于标准 RAE 模型, 同时其准确率也高于监督学习 SVM 算法。由于基于 RAE+Dropout 相结合的微博情感分析模型参数众多, 本文中只选取了几个重要的参数指标进行了局部选优, 下一步工作将会把更多的模型参数考虑到实验中, 同时对 SVM 模型的特征选择方式进一步优化。

参考文献

- [1] 吴晓瑾. 微博语言的特征研究 [D]. 广州: 华南师范大学, 2012.
- [2] 厉小军, 戴霖, 施寒潇, 等. 文本倾向性分析综述 [J]. 浙江大学学报 (工学版), 2011(7): 1167-1174.
- [3] Hatzivassiloglou V, McKeown K. Predicting the Semantic Orientation of Adjectives[A]. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL[C]. 1997: 174-181.
- [4] Turney P D. Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews[C]// Proceedings of Annual Meeting of the Association for Computational Linguistics. 2002: 417-424.
- [5] Peter T, Michael L. Measuring Praise and Criticism : Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems, 2003, 21(4): 315-346.
- [6] Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure Semantic Orientations of Adjectives[C]// Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal. 2004: 1115-1118.
- [7] Ju C, Ma Y, Fu X. A sentiment orientation analysis method based on emotional vocabulary[J]. Journal of Computational Information Systems, 2015, 11(5): 1745-1755.
- [8] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, Uk, A Meeting of Sigdat, A Special Interest Group of the ACL. DBLP, 2011: 151-161.
- [9] Wang W, Pan S J, Dahlmeier D, et al. Recursive Neural Conditional Random Fields for Aspect-based Sentiment Analysis[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 616-626.
- [10] Rifai S, Vincent P, Muller X, et al. Contractive Auto-Encoders: Explicit Invariance During Feature Extraction[C]// Proceedings of the 28th International

Conference on Machine Learning (ICML-11). 2011: 833-840.

[11] Vincent P, Larochelle H, Bengio Y, et al. Extracting and Composing Robust Features with Denoising Autoencoders[C]// International Conference. 2008: 1096-1103.

[12] 梁军, 柴玉梅, 原慧斌, 等. 基于深度学习的微博情感分析 [J]. 中文信息学报, 2014, 28(5): 155-161.

[13] Turaga S C, Murray J F, Jain V, et al. Convolutional Networks can Learn to Generate Affinity Graphs for Image Segmentation[J]. Neural Computation, 2010, 22(2): 511-538.

[14] Netzer Y, Wang T, Coates A, et al. Reading Digits in Natural Images with Unsupervised Feature Learning[C]// NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2011. 2011: 1-9.

[15] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012: 1097-1105.

[16] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors[J]. Computer Science,

2012, 3(4): 212-223.

[17] Goller C, Kuchler A. Learning Task-dependent Distributed Representations by Backpropagation Through structure[C]// IEEE International Conference on Neural Networks. IEEE Xplore, 1996: 347-352.

[18] Livnat A, Papadimitriou C, Pippenger N, et al. Sex, Mixability, and Modularity[J]. Proceedings of the National Academy of Sciences, 2010, 107(4): 1452-7.

[19] 万常选, 江腾蛟, 钟敏娟, 等. 基于词性标注和依存句法的 Web 金融信息情感计算 [J]. 计算机研究与发展, 2013, 50(12): 2554-2569.

[20] 张建华. 基于 LDA 和词性句法规则的用户评论情感分析研究 [D]. 南宁: 广西大学, 2014.

[21] 邱云飞, 陈艺方, 王伟, 等. 基于词性特征与句法分析的商品评价对象提取 [J]. 计算机工程, 2016, 42(7): 173-180.

[22] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[23] Tobergte D R, Curtis S. Improving Neural Networks with Dropout[J]. Journal of Chemical Information and Modeling, 2013, 53(9): 1689-1699.