一种半监督学习的金融新闻文本分类算法

张晓龙[1,2]，支龙[1,2]，高剑[3]，苗仲辰[3]，林越峰[3]，项雅丽[1,2]，熊赟[1,2]
1. 复旦大学计算机科学技术学院，上海 210438；2. 上海市数据科学重点实验室，上海 200438；
3. 上海金融期货信息技术有限公司，上海 200120

摘要
对金融文本进行分类是一项常见的用于识别金融风险的任务。传统的金融新闻文本分类方法需要大量的已知类别文本来训练分类器，然而标注金融新闻文本标签不仅需要专业的金融背景知识，而且耗时耗力。为了减少对已知类别文本的依赖，提出了一个基于半监督学习的金融文本分类算法，该算法采用有监督学习和无监督学习的一致性训练方式，以更好地利用未知类别的文本数据；针对金融领域文本引入无监督数据增强方法，即对特定任务使用特定目标的数据增强方法，以产生更有效的数据。在多个金融文本数据集上开展的实验证明，相比其他文本分类算法，提出的算法在有效性上有明显提升。

关键词
自然语言处理；文本分类；半监督学习；金融

中图分类号：TP312 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022019

A semi-supervised learning financial news classification algorithm

ZHANG Xiaolong[1,2], ZHI Long[1,2], GAO Jian[3], MIAO Zhongchen[3], LIN Yuefeng[3], XIANG Yali[1,2], XIONG Yun[1,2]
1. School of Computer Science and Technology, Fudan University, Shanghai 200438, China
2. Shanghai Key Laboratory of Data Science, Shanghai 200438, China
3. Shanghai Financial Futures Information Technology Co., Ltd., Shanghai 200120, China

Abstract
Classifying financial texts is a common task for identifying financial risks. Traditional financial news classification requires a large number of labeled texts to train the classifier. However, labeling financial news requires not only professional financial background knowledge, but also time-consuming and labor-intensive. In order to reduce the dependence on labeled text, a semi-supervised learning financial text classification algorithm- SSF (semi-supervised learning financial news classification algorithm) was proposed, which uses a consistent training method of supervised learning and unsupervised learning to improve the use of unlabeled data. And unsupervised data augmentation for financial texts was introduced, that is, use specific target data augmentation methods for specific tasks to generate more effective data. Experiments on

multiple financial news data sets were conducted to verify that the proposed SSF algorithm has a significant improvement in effectiveness compared with other text classification algorithms.

Key words
natural language processing, text classification, semi-supervised learning, finance

0 引言

文本分类是一项常见的数据任务，通过对金融领域的新闻、言论等文本数据的主题进行识别，可以有效地给金融相关部门提供技术支持。然而在针对金融领域的实际业务开发过程中，不免会遇到标注数据缺乏、类别标签不均衡等挑战。由于金融领域本身的复杂性，这些数据往往包含了大量的专业术语和特定表达方式，因此领域相关的文本标注需要由具备较高专业知识水平的人员完成，这使得金融语料的标注代价昂贵，且效率低下。

半监督学习（semi-supervised learning，SSL）[1]是利用无标签数据解决这一问题的具有代表性的一种方法，其中，基于一致性训练的半监督学习方法已经在图像领域取得了良好的效果，受到研究者的广泛关注[2-5]。与一致性训练相关的一类研究方法是在训练的过程中对输入样本[6-8]或隐藏状态[9]增加噪声，并且保持模型的预测值不会因此发生改变。例如，Laine等人[3]提出的Pseudo-ensembles方法在训练过程中应用高斯噪声和dropout噪声；Miyato等人[6]提出的虚拟对抗训练方法通过近似模型最敏感的输入空间的变化方向来定义噪声；Clark等人[8]提出的交叉视图训练方法通过掩盖部分输入数据的方法引入噪声。另一类与一致性训练相关的研究方法是在模型参数空间上实现强制一致性，如插值一致性训练[9]、MixMatch[10]和无监督数据增强（unsupervised data augmentation，UDA）[11]等方法。受到UDA方法的启发，本文将其引入金融文本分类中，以应对金融文本标记不足的挑战。但是UDA方法在对金融中文无标签文本进行数据增强时，存在增强后的中文文本质量差的问题，需要对金融中文无标签文本的数据增强方法进行研究。针对金融新闻的文本分类任务，本文提出了一个基于半监督学习的金融新闻文本分类（semi-supervised learning financial news classification，SSF）算法。本文主要贡献如下：

- 引入有监督学习和无监督学习的一致性训练方法，在有标签数据较少的情况下，实现金融文本的分类任务；
- 针对不同的金融领域任务，采用不同的训练信号退火（training signal annealing，TSA）收敛策略，降低模型过拟合的可能性；
- 在真实数据集上的实验结果表明，本文提出的SSF算法相比主流文本分类算法在有效性上有明显提升。

1 相关工作

1.1 预训练和微调框架

预训练和微调框架已被应用于多种自然语言处理（natural language processing，NLP）任务中[12-14]。Howard等人[15]提出在大型通用语料库上预先训练

语言模型，再对目标任务进行微调（即预训练+微调框架的方式）。这种方法相对于需要大量的标注数据的连续词袋（continuous bag-of-words，CBOW）模型[16]，即使使用少量标记数据，经过预训练的模型也能表现出较优的性能，并且基于注意力机制的预训练模型可更好地理解特征之间的相互关系。算法除了对结果的有效性有要求，对内存占用、运行速度也有一定的要求。本文在预训练模型方面采用ALBERT（a lite bert）[17]模型，ALBERT模型使用句子顺序预测（sentence order prediction）代替下一个句子预测（next sentence prediction），提升了训练效率，并且采用参数因式分解以及跨层参数共享两种技术降低资源消耗，相比于OpenAI GPT[18]和BERT[19]等规模较大的预训练模型，ALBERT模型的训练速度更快。

1.2 一致性正则

一致性正则可以被看作标签传播的一种形式，在空间表示中，相似的训练样本更有可能属于同一类别。基于这个假设，一致性正则通过某种机制可以将标签信息从样本传播到与其相邻的样本。一致性正则框架在图像领域受到了广泛关注[3,7,20-21]。现有的利用一致性进行训练的模型虽然用到了数据增强，但是它们仅仅应用了较弱的数据增强方法，如随机翻译和裁剪。与本文工作更为相关的工作有MixMatch[10]和UDA[11]，这些方法在半监督学习领域都取得了成功。然而，这些方法在处理金融领域文本等含有较多专业术语的文本时，存在数据增强后的文本质量较差等问题。本文充分利用了文本中单词的权重信息，将训练集中其他句子的非关键词替换为当前句子的非关键词，提出的SSF算法在金融领域文本的数据增强上取得了当前最佳（state-of-the-art，SOTA）的效果。除此之外，本文提出的SSF算法在提升训练速度以及减少资源消耗上也有显著效果。

2 基于半监督学习的金融新闻文本分类

将金融文本记为x，y^*是该文本的标注类别，$\hat{x}$是对无标注数据的增强样本。本节具体介绍SSF算法，SSF模型采用半监督学习的一致性训练[7,20-21]的思路，从预训练模型和数据增强两个角度对已有半监督学习模型进行优化。在预训练模型选择上，如第1.1节所述，ALEBRT预训练模型在训练过程中可以显著降低资源消耗，并缩短训练时间。在数据增强方面，由于金融领域文本存在较多专业性术语，随机替换和回译法等文本数据增强方法可能会替换掉文本中的专业术语，使增强后的样本与原样本差别较大。本文采用的数据增强方法可以选择性地替换样本中的非关键词。模型框架如**图1**所示，**图1**上半部分是有监督学习部分，下半部分是无监督学习部分。在有监督学习部分，利用有标签数据在预训练模型上进行微调；在无监督学习部分，不同于在无标注数据注入噪声的方法，通过将用于有监督学习数据增强的方法迁移至无监督学习来增强模型的鲁棒性。

下面针对模型的各个部分展开叙述。

2.1 有监督学习

如**图1**上半部分所示，对于有标签金融文本x，模型将其送入预训练模型ALBERT得到文本的嵌入表示，再经过全连接层得到文本的预测标签。这部分的损失函数是标准有监督训练中预测标签和真实标签的交叉熵，记为：

$$L_L(\theta) = E_{x\sim p_L(x)}[-\log p_\theta(f^*(x)|x)] \quad (1)$$

其中，P_L为有标签数据的分布，$f^*(x)$是预测函数。

2.2 一致性训练

如图1下半部分所示，对于无标注数据x，一方面，模型通过预训练模型ALBERT得到无标签文本的嵌入表示，计算其分布$p_\theta(y|x)$；另一方面，模型通过对无标签样本进行数据增强，得到$\hat{x}$。$\hat{x}$经过预训练模型得到嵌入表示，再计算该增强版本的分布$p_\theta(y|\hat{x})$。模型最小化两个分布之间的差异，使两者尽可能相似，从而优化模型的参数。模型保持增强样本的预测值与无标签样本的预测值一致，这使模型对噪声不敏感，因此算法相对于输入（或隐藏）空间的变化更平滑，更具鲁棒性。其损失函数为两个分布之间的交叉熵损失，形如：

$$L_U(\theta) = E_{\hat{x}\sim p_U(x)} E_{\hat{x}\sim q(\hat{x}|x)}[\mathrm{CE}(p_{\tilde{\theta}}(y|x) \| p_\theta(y|\hat{x}))] \quad (2)$$

其中，CE表示交叉熵损失函数，P_U表示无标记数据的样本分布，$q(\hat{x}|x)$是一个数据增强函数，$\tilde{\theta}$是当前训练参数θ的复制，反向传播时不会更新$\tilde{\theta}$。本文针对金融文本分类任务，考虑到文本中金融领域的关键词对预测标签的影响较大，采用随机替换和删除可能会损失文本中的关键信息，因此采用了TF-IDF（term frequency-inverse document frequency）进行同义词替换，兼顾词频与新鲜度，替换一些常见词，同时保留能提供更多信息的关键词。

2.3 TF-IDF文本数据增强

数据增强方法能够生成多样且有效的样本，文本数据增强方法可以被设计为保留关键词，并用其他非关键性

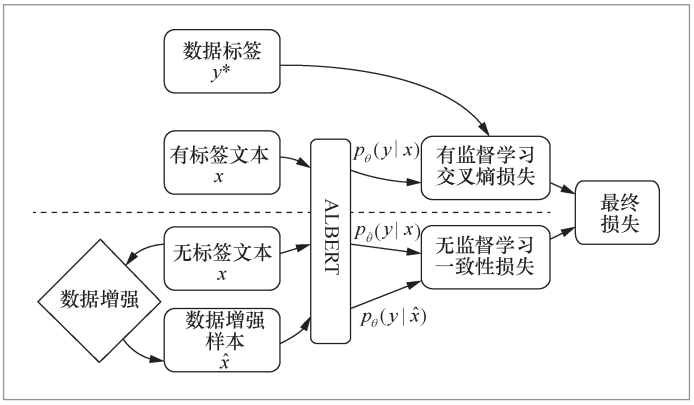

图1 SSF算法框架

单词替换句子中的非关键性单词。本文将TF-IDF信息应用到数据增强中。具体而言，IDF(w)是单词w在整个语料库中的IDF分数。TF(w)是单词w在每个句子中TF分数。每个单词的TF-IDF分数计算如下：TF-IDF(w)= TF(w)×IDF(w)。假定在一个句子x中，最大的TF-IDF分数为$C=\max_i \text{TF-IDF}(x_i)$。为了使句子中被替换的单词与单词的TF-IDF分数负相关，将单词替换的概率设置为$(\min(p/C-\text{TF-IDF}(x_i))/Z,1)$，其中，$p$是超参数，用于控制数据增强的程度，$Z=\sum_i(C-\text{TF-IDF}(x_i))/|Z|$是平均分数，从整个词汇表中抽取另一个单词来替换原文中的单词。直观地讲，采样的单词不应当是别的词汇表中的关键词，以防止更改句子的标签。为了衡量一个单词是否是关键词，计算整个语料库中每个单词的分数，即计算分数$S(w)=\text{freq}(w)\text{IDF}(w)$，freg($w$)是单词$w$在整个语料库中出现的频率。采样单词$w$的概率设置为$\left(\max_{w'} S(w') - S(w)\right)/Z'$，其中$Z' = \sum_w \max_{w'} S(w') - S(w)$是归一项。数据增强方法实例如图2所示。

2.4 半监督学习

SSF将有监督学习与无监督学习结合

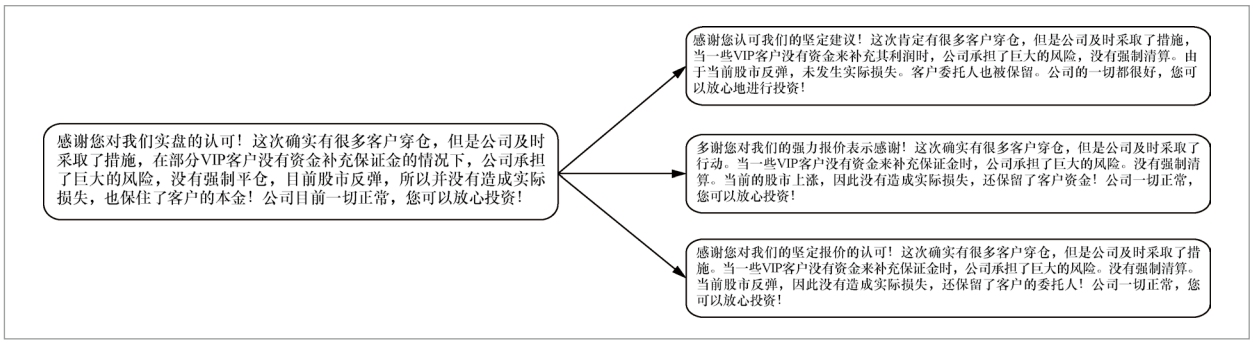

图 2 TF–IDF 数据增强示例

起来，其最终的损失函数为：

$$\min_{\theta} L(\theta) = L_L(\theta) + \lambda L_U(\theta) \qquad (3)$$

其中，权重因子λ用于控制无监督损失和有监督损失的重要程度，一般情况下设置为1。同时无标签样本的批次大小大于有标签样本的批次大小。

将有监督学习与无监督学习结合后，SSF模型既利用了有限的有标签数据，又利用无标签数据丰富了模型的表达能力。在有监督训练、无监督训练与增强样本的训练过程中，三者的ALBERT模型一致，且参数共享，因此，有监督训练过程与无监督训练过程相辅相成。SSF框架通过引入TF-IDF数据增强方式，无标签样本中的一致性损失项得到更严格的保证，并将模型的共享参数传递到有监督训练部分，使整个模型更具有鲁棒性。从另一个角度来看，将一致性损失降至最低会逐渐将标签信息从标记的样本传播到未标记的样本，某种程度上这是在为某些无标记数据打标签，提高了未标记数据的利用率。

2.5 针对样本不均衡的模型设定

本节旨在说明SSF框架在处理文本半监督问题时遇到的问题以及解决方法。

（1）置信度阈值

在无监督训练过程中，要排除掉那些模型预测不确定的样本。例如，在小批次训练过程中，过滤预测值小于输出阈值的样本，从而使余留样本的预测标签更加接近真实值。

（2）熵正则化[10]

熵正则化已经被证明在半监督学习上具有很好的效果，SSF模型也采取熵正则化来进行训练。如前文所述，无监督损失项$L_U(\theta) = E_{\hat{x}\sim p_U(x)} E_{\hat{x}\sim q(\hat{x}|x)}[\mathrm{CE}(p_{\tilde{\theta}}(y|x) \| p_{\theta}(y|\hat{x}))]$中$p_{\tilde{\theta}}$的计算如下：

$$p_{\tilde{\theta}}(y|x) = \frac{\exp(z_y / \tau)}{\sum_{y'} \exp(z_{y'} / \tau)} \qquad (4)$$

其中，τ是超参数，Z_y是对样本x预测的Logit值。

（3）TSA

在半监督学习中，无标签数据量远远大于有标签数据量往往导致模型在少量的有标签样本下过拟合，但在无标签样本中却尚未产生收敛。TSA方法可以解决这个问题，即当有标签数据过少时，对预测值设定阈值，高于阈值的预测值不会参与反向传播，从而确保模型不会因为标签数据过少而产生过拟合。针对金融领域任务以及数据集的不同，采用不同的TSA策略，具体将在第3.4.2节中展开说明。

3 实验分析

本节通过实验验证SSF模型的有效

性，分析讨论实验中的场景数据，以及相关的参数设置。

3.1 数据集

实验使用了3份来源于某金融机构的金融领域文本数据集。按照主题可分为违规类别数据集、期货期权数据集和机构相关数据集，各类别数据的数量见**表1~表3**。将数据按照8:1:1的比例随机划分为训练集、验证集以及测试集。这些数据集均存在不同程度的类别不均衡，且针对某些业务场景的有标签样本数目稀少。

- 违规类别数据集：来源于某金融机构从社交媒体平台爬取的数据集，任务是预测一条文本是否违规以及违规类别，违规类别分别为恶意抹黑监管机构、非法荐股、诱导开户、煽动维权诈骗。
- 期货期权数据集：数据集来源于某新闻机构，任务类型为分类任务，预测任务是判断一条文本是否属于某一主题。
- 机构相关数据集：数据集来源于某金融机构，任务类型为分类任务，预测任务是判断一条文本的主体是哪个私募机构，其中，文本中可能包含多个私募机构。

3.2 对比算法

为了测试本文提出的方法的有效性，将其与几种主流的文本分类模型进行了比较，具体如下。

- GloVe[22]：GloVe模型将基于奇异值分解（singular value decomposition，SVD）的潜在语义分析（latent semantic analysis，LSA）算法和word2vec算法结合到一起，既使用了语料库的全局统计特征，也使用了局部的上下文特征，得到文本词向量后经过逻辑回归得到分类结果。
- ELMo[23]：ELMo事先用语言模型在一个大的语料库上学习好词的表示，接着用下游任务中的无标签数据来微调预训练好的ELMo。相比GloVE，ELMo在多义词的表示方面取得了改善，得到文本词向量后经过逻辑回归得到分类结果。
- FastText[24]：FastText模型架构与word2vec中的CBOW很相似，不同之处是FastText预测的是标签，而CBOW预测的是中间词，即两者模型架构相似，但是模型的任务不同。
- VAMPIRE[25]：VAMPIRE模型是一种基于预训练半监督的文本分类轻量型模

表1 违规类别数据集

违规类别	合规	违规类别1	违规类别2	违规类别3	违规类别4	无标签数据
数量/条	307	68	52	35	60	57 000

表2 期货期权数据集

主题类别	豆粕期权	贵州茅台	国泰君安	股指期权	汇率波动	区块链	中国石油	贷款基础利率（LPR）	无标签数据
数量/条	16	273	102	200	79	1 325	46	42	29 969

表3 机构相关数据集

机构名称	淡水泉	敦和资管	高毅资产	汉和资本	景林资产	凯丰投资	林园投资	明汯投资	千合资本	石锋资产	无标签数据
数量/条	20	20	20	20	20	20	20	20	20	20	4 000

型，旨在解决由大量数据和高昂计算力导致的资源不足问题。

- BERT[19]：BERT代表Transformers的双向编码器。它被设计为通过对左右的上下文的联合来预训练未标记文本，从而得到深层的双向表示。这里使用BERT-base-Chinese预训练模型，并在下游任务上进行微调得到分类结果。
- UDA[11]：UDA采用一致性训练框架，在文本分类任务上，采用BERT预训练模型，在数据增强方面，基于WMT'14英法翻译模型，通过回译法对无标签数据产生噪声进行数据增强。

3.3 实验结果

实验中将有标签数据集按照8:1:1划分为训练集、验证集和测试集。测试集实验结果见**表4**。

从**表4**可以发现，SSF模型在3个数据集上的精度和召回率均超过了先前的对比模型。与GloVe、ELMo和FastText文本分类算法相比，采用一致性训练框架的VAMPIRE、UDA和SSF算法取得了较优的表现。与VAMPIRE和BERT算法相比，SSF模型在精度和召回率上都取得了更好的结果，这表明引入无监督数据增强方法可以带来更好的性能。与UDA模型相比，SSF模型在精度和召回率上也取得了更好的表现。可以得出结论，相对于UDA中对无标签数据通过回译法进行数据增强，SSF通过TF-IDF数据增强方法可以针对性地在中文金融新闻文本分类上获得更好的表现。

通过改变有标签文本的数量，将有标签数据的数量降为原来的50%，对比SSF算法与其他文本分类算法的性能，实验结果见**表5**。

在这部分实验中，笔者针对有标签数据的数量进行了调整。见**表5**，给定相同的无标签数据，将有标签数据的数量减少50%，实验结果表明，本文所提文本分类算法在F1值上都有下降。值得一提的是，

表 4 SSF 模型及其基准模型实验结果

模型	违规类别数据集			期货期权数据集			机构相关数据集		
	精度	召回率	F1值	精度	召回率	F1值	精度	召回率	F1值
GloVe	0.709	0.684	0.696	0.785	0.789	0.787	0.710	0.704	0.707
ELMo	0.705	0.712	0.708	0.769	0.781	0.775	0.724	0.718	0.721
FastText	0.743	0.729	0.736	0.756	0.749	0.752	0.789	0.801	0.795
VAMPIRE	0.749	0.735	0.742	0.778	0.762	0.770	0.838	0.853	0.845
BERT	0.789	0.784	0.786	0.878	0.873	0.875	0.919	0.928	0.923
UDA	0.811	0.802	0.806	0.920	0.915	0.917	0.938	0.951	0.944
SSF	0.819	0.807	0.813	0.939	0.932	0.935	0.973	0.976	0.974

表 5 SSF 模型及其基准模型实验结果

模型	违规类别数据集		期货期权数据集		机构相关数据集	
	F1值	F1值下降率	F1值	F1值下降率	F1值	F1值下降率
GloVe	0.385	44.68%	0.521	33.80%	0.525	25.74%
ELMo	0.407	42.51%	0.536	30.84%	0.548	23.99%
FastText	0.561	23.78%	0.589	21.68%	0.631	20.63%
VAMPIRE	0.406	45.28%	0.531	31.04%	0.550	34.91%
BERT	0.583	25.83%	0.633	27.66%	0.683	26.00%
UDA	0.684	15.14%	0.782	14.72%	0.821	13.03%
SSF	0.772	5.04%	0.883	5.56%	0.887	8.93%

SSF算法在更少的标注数据上的表现大幅优于其对比算法。

通过上述在3个标签数量少的数据集上的实验可以得出，在金融领域中文文本分类任务中，本文提出的SSF框架在有监督数据样本缺乏的场景下有更好的表现。

3.4 消融实验分析

本节从数据增强方面和模型阈值设置两个方面开展实验。

3.4.1 数据增强维度的影响分析

不采用数据增强机制时的SSF变种模型为SSF-w/o-aug，实验结果见**表6**。

表6的结果显示，数据增强机制在3个数据集上都为模型的性能带来了提升。其中，在违规类别数据集上，数据增强为模型带来了1.74%的精度增值和2.28%的召回率增值；在期货期权数据集上，数据增强机制为模型带来了2.40%的精度增值和1.30%的召回率增值；在机构相关数据集上，数据增强机制给模型带来了2.21%的精度增值和2.09%的召回率增值。因为数据增强机制可以帮助模型保留文本中的关键信息，所以它在含有专业词汇较多的金融文本领域分类效果更好。

3.4.2 模型阈值设置维度的影响分析

考虑不同TSA策略对实验结果的影响，实验结果见**表7**。

表7的结果显示，在违规类别数据集上，有标签数据和无标签数据的比例为1:109（**表1**），在无监督训练时较快的收敛策略得到了较高的准确率；而在期货期权数据集上，有标签数据和无标签数据的比例约为1:14（**表2**），对数增长的TSA策略取得了最佳的效果；在机构相关数据集上，有标签和无标签的比例约为1:20（**表3**），采用线性增长的TSA策略取得了最佳的效果。这表明在有标签数据和无标签数据比例不同时，采用不同的TSA策略可以有效地避免模型过拟合。

4 结束语

本文围绕金融领域的业务需求，针对中文金融领域数据集提出了SSF半监督学习框架，通过使用针对性的数据增强方法对样本中的无标签数据进行了数据增强，在真实数据集上的实验表明，本文提出的SSF方法适用于金融领域下标签样本少的

表 6 去除数据增强时的实验结果

变量	违规类别数据集			期货期权数据集			机构相关数据集		
	精度	召回率	F1值	精度	召回率	F1值	精度	召回率	F1值
SSF-w/o-aug	0.805	0.789	0.797	0.917	0.920	0.918	0.952	0.956	0.954
SSF	0.819	0.807	0.813	0.939	0.932	0.935	0.973	0.976	0.974

表 7 采用不同 TSA 策略的实验结果

变量	违规类别数据集			期货期权数据集			机构相关数据集		
	精度	召回率	F1值	精度	召回率	F1值	精度	召回率	F1值
对数增长	0.762	0.741	0.751	0.939	0.932	0.935	0.954	0.961	0.958
线性增长	0.812	0.779	0.795	0.922	0.931	0.926	0.973	0.976	0.974
指数增长	0.819	0.807	0.813	0.906	0.914	0.910	0.947	0.938	0.942

文本分类任务，并且性能优于先前的工作。由于硬件以及ALBERT预训练模型本身的限制，笔者在实验中采用的最大序列长度为256，但是相关数据集的长度一般为1 000左右，需要指出，即使在如此有限的文本输入上，SSF模型的表现能力也是较为理想的。但是，更好地处理长文本信息使得模型感知到尽可能多的内容，将有助于模型的效果提升，因此，长文本数据上的模型优化是进一步的研究工作。

参考文献:

[1] CHAPELLE O, SCHOLKOPF B, ZIEN E. Semi-supervised learning (Chapelle, O. et al. Eds.; 2006) [book reviews][J]. IEEE Transactions on Neural Networks, 2009, 20(3): 542.

[2] RASMUS A, VALPOLA H, HONKALA M, et al. Semi-supervised learning with ladder network[J]. arXiv preprint, 2015, arXiv:1507.02672.

[3] LAINE S, AILA T M. Temporal ensembling for semi-supervised learning[J]. arXiv preprint, 2016, arXiv:1610.02242.

[4] TARVAINEN A, VALPOLA H. Weight-averaged, consistency targets improve semi-supervised deep learning results[Z]. 2017.

[5] BACHMAN P, ALSHARIF O, PRECUP D. Learning with pseudo-ensembles[J]. arXiv preprint, 2014, arXiv: 1412.4864.

[6] MIYATO T, MAEDA S I, KOYAMA M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1979-1993.

[7] SAJJADI M, JAVANMARDI M, TASDIZEN T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning[J]. arXiv preprint, 2016, arXiv:1606.04586.

[8] CLARK K, LUONG M T, MANNING C D, et al. Semi-supervised sequence modeling with cross-view training[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018.

[9] VERMA V, KAWAGUCHI K, LAMB A, et al. Interpolation consistency training for semi-supervised learning[J]. arXiv preprint, 2019, arXiv:1903.03825.

[10] BERTHELOT D, CARLINI N, GOODFELLOW I J, et al. MixMatch: a holistic approach to semi-supervised learning[J]. arXiv preprint, 2019, arXiv:1905.02249.

[11] XIE Q Z, DAI Z H, HOVY E, et al. Unsupervised data augmentation for consistency training[J]. arXiv preprint, 2019, arXiv:1904.12848.

[12] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[Z]. 2019.

[13] CHEN J A, CHEN J S, YU Z. Incorporating structured commonsense knowledge in story completion[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6244-6251.

[14] AKBIK A, BERGMANN T, VOLLGRAF R. Pooled contextualized embeddings for named entity recognition[C]// Proceedings of the 2019 Conference of the North. Stroudsburg: Association for Computational Linguistics, 2019.

[15] HOWARD J, RUDER S. Universal language model fine-tuning for text classification[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018.

[16] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, 2013, arXiv:1301.3781.

[17] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised

learning of language representations[J]. arXiv preprint, 2019, arXiv:1909.11942.
[18] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[Z]. 2018.
[19] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv:1810.04805.
[20] BEYER L, ZHAI X H, OLIVER A, et al. S4L: self-supervised semi-supervised learning[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 1476-1485.
[21] OLIVER A, ODENA A, RAFFEL C, et al. Realistic evaluation of deep semi-supervised learning algorithms[J]. arXiv preprint, 2018, arXiv:1804.09170.
[22] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014.
[23] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv preprint, 2018, arXiv:1802.05365.
[24] JOULIN A, GRAVE E, BOJANOWSKI P, et al. FastText.zip: compressing text classification models[J]. arXiv preprint, 2016, arXiv:1612.03651.
[25] GURURANGAN S, DANG T, CARD D, et al. Variational pretraining for semi-supervised text classification[J]. arXiv preprint, 2019, arXiv:1906.02242.

作者简介

张晓龙(1998-),男,复旦大学计算机科学技术学院硕士生,主要研究方向为自然语言处理、机器学习。

支龙(1996-),男,复旦大学计算机科学技术学院硕士生,主要研究方向为自然语言处理、机器学习。

高剑(1978-),男,上海金融期货信息技术有限公司总工程师,主要从事多项前沿科技在金融期货行业的技术研究与创新实践应用工作。

苗仲辰（1988– ），男，博士，就职于上海金融期货信息技术有限公司，主要研究方向为AI算法、数据挖掘、科技监管场景分析等。

林越峰（1990– ），男，博士，就职于上海金融期货信息技术有限公司，主要研究方向为自然语言处理、时序预测等。

项雅丽（1995– ），女，复旦大学计算机科学技术学院硕士生，主要研究方向为数据挖掘、网络表示学习。

熊赟（1980– ），女，博士，复旦大学计算机科学技术学院教授、博士生导师，主要研究方向为数据科学、数据挖掘和大数据处理。

收稿日期： 2021-03-17
通信作者： 高剑，gaojian@cffex.com.cn
基金项目： 国家自然科学基金资助项目（No.U1636207，No.U1936213）
Foundation Items: The National Natural Science Foundation of China (No.U1636207, No.U1936213)