# **金融预训练模型 FinBERT**

# **背景及下载地址**

为了促进自然语言处理技术在金融科技领域的应用和发展，**熵简科技 AI Lab 近期开源了基于 BERT 架构的金融领域预训练语言模型 FinBERT 1.0**。据我们所知，这是国内首个在金融领域大规模语料上训练的开源中文BERT预训练模型。相对于Google发布的原生中文BERT、哈工大讯飞实验室开源的BERT-wwm 以及 RoBERTa-wwm-ext 等模型，本次开源的 FinBERT 1.0 预训练模型在多个金融领域的下游任务中获得了显著的性能提升，在不加任何额外调整的情况下，F1-score **直接提升至少 2~5.7 个百分点**。

对于深度学习时代的自然语言处理技术，我们一般认为存在两大里程碑式的工作。第一个里程碑是在2013年逐渐兴起，以 Word2Vec 为代表的的词向量技术；第二个里程碑则是在 2018 年以 BERT 为代表的深度预训练语言模型（Pre-trained Language Models）。一方面，以 BERT 为代表的深度预训练模型在包括文本分类、命名实体识别、问答等几乎所有的子领域达到了新的 state of the art；另一方面，作为通用的预训练模型，BERT 的出现也显著地减轻了NLP算法工程师在具体应用中的繁重工作，由以往的魔改网络转变为 Fine tune BERT，即可快速获得性能优秀的基线模型。因此，深度预训练模型已成为各个 AI 团队必备的基础技术。

但是，**当前开源的各类中文领域的深度预训练模型，多是面向通用领域的应用需求，在包括金融在内的多个垂直领域均没有看到相关开源模型**。熵简科技希望通过本次开源，推动 NLP技术在金融领域的应用发展，欢迎学术界和工业界各位同仁下载使用，我们也将在时机合适的时候推出性能更好的 FinBERT 2.0 & 3.0。

**项目地址**：https://github.com/valuesimplex/FinBERT

# **二、模型及预训练方式**

## **2.1网络结构**

熵简 FinBERT 在网络结构上采用与 Google 发布的原生BERT 相同的架构，包含了 FinBERT-Base 和 FinBERT-Large 两个版本，其中前者采用了 12 层 Transformer 结构，后者采用了 24 层 Transformer 结构。考虑到在实际使用中的便利性和普遍性，本次发布的模型是 FinBERT-Base 版本，本文后面部分统一以 FinBERT 代指 FinBERT-Base。

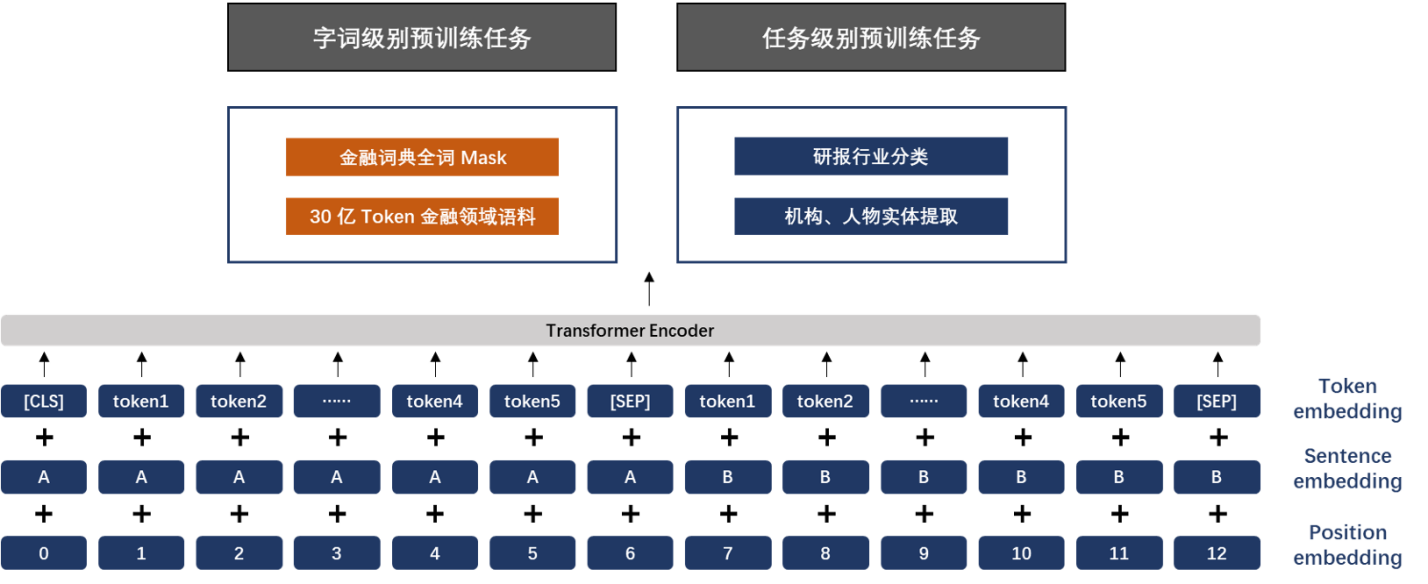
## **2.2训练语料**

FinBERT 1.0 所采用的预训练语料主要包含三大类金融领域的语料，分别如下：

* 金融财经类新闻：从公开渠道采集的最近十年的金融财经类新闻资讯，约 100 万篇；
* 研报/上市公司公告：从公开渠道收集的各类研报和公司公告，来自 500 多家境内外研究机构，涉及 9000 家上市公司，包含 150 多种不同类型的研报，共约 200 万篇；
* 金融类百科词条：从 Wiki 等渠道收集的金融类中文百科词条，约 100 万条。

对于上述三类语料，在金融业务专家的指导下，我们对于各类语料的重要部分进行筛选、预处理之后得到最终用于模型训练的语料，共包含 30亿 Tokens，这一数量超过了原生中文BERT的训练规模。

## **2.3预训练方式**



**FinBERT 预训练框架图**

如上图所示，FinBERT 采用了两大类预训练任务，分别是字词级别的预训练和任务级别的预训练。两类预训练任务的细节详述如下：

**（1）字词级别的预训练**

字词级别的预训练首先包含两类子任务，分别是 Finnacial Whole Word MASK（FWWM）、Next Sentence Prediction（NSP）。同时，在训练中，为了节省资源，我们采用了与 Google 类似的两阶段预训练方式，第一阶段预训练最大句子长度为128，第二阶段预训练最大句子长度为 512。两类任务具体形式如下：

**Finnacial Whole Word MASK（FWWM）**

Whole Word Masking (wwm)，一般翻译为全词 Mask 或整词 Mask，出是 Google 在2019年5月发布的一项升级版的BERT中，主要更改了原预训练阶段的训练样本生成策略。简单来说，原有基于WordPiece的分词方式会把一个完整的词切分成若干个子词，在生成训练样本时，这些被分开的子词会随机被mask。 在全词Mask中，如果一个完整的词的部分WordPiece子词被 Mask，则同属该词的其他部分也会被 Mask，即全词Mask。

在谷歌原生的中文 BERT 中，输入是以字为粒度进行切分，没有考虑到领域内共现单词或词组之间的关系，从而无法学习到领域内隐含的先验知识，降低了模型的学习效果。我们将全词Mask的方法应用在金融领域语料预训练中，即对组成的同一个词的汉字全部进行Mask。首先我们从金融词典、金融类学术文章中，通过自动挖掘结合人工核验的方式，构建出金融领域内的词典，约有10万词。然后抽取预语料和金融词典中共现的单词或词组进行全词 Mask预训练，从而使模型学习到领域内的先验知识，如金融学概念、金融概念之间的相关性等，从而增强模型的学习效果。

**Next Sentence Prediction（NSP）**

为了训练一个理解句子间关系的模型，引入一个下一句预测任务。具体方式可参考BERT原始文献，Google的论文结果表明，这个简单的任务对问答和自然语言推理任务十分有益，我们在预训练过程中也发现去掉NSP任务之后对模型效果略有降低，因此我们保留了NSP的预训练任务，学习率采用Google 官方推荐的2e-5，warmup-steps为 10000 steps。

**（2）任务级别的预训练**

为了让模型更好地学习到语义层的金融领域知识，更全面地学习到金融领域词句的特征分布，我们同时引入了两类有监督学习任务，分别是研报行业分类和财经新闻的金融实体识别任务，具体如下：

**研报行业分类**

对于公司点评、行业点评类的研报，天然具有很好的行业属性，因此我们利用这类研报自动生成了大量带有行业标签的语料。并据此构建了行业分类的文档级有监督任务，各行业类别语料在 5k~20k 之间，共计约40万条文档级语料。

**财经新闻的金融实体识别**

与研报行业分类任务类似，我们利用已有的企业工商信息库以及公开可查的上市公司董监高信息，基于金融财经新闻构建了命名实体识别类的任务语料，共包含有 50 万条的有监督语料。

整体而言，为使 FinBERT 1.0 模型可以更充分学习到金融领域内的语义知识，我们在原生 BERT 模型预训练基础上做了如下改进：

1、训练时间更长，训练过程更充分。为了取得更好的模型学习效果，我们延长模型第二阶段预训练时间至与第一阶段的tokens总量一致；

2、融合金融领域内知识。引入词组和语义级别任务，并提取领域内的专有名词或词组，采用全词 Mask的掩盖方式以及两类有监督任务进行预训练；

3、为了更充分的利用预训练语料，采用类似Roberta模型的动态掩盖mask机制，将dupe-factor参数设置为10；

## **2.4预训练加速**

当前，对于所提供的一整套软硬件深度学习炼丹系统，英伟达提供了丰富的技术支持和框架优化，其中很重要的一点就是如何在训练中进行加速。在 FinBERT 的训练中，我们主要采用了 Tensorflow XLA 和 Automatic Mixed Precision 这两类技术进行预训练加速。

### **Tensorflow XLA 进行训练加速**

XLA 全称为加速线性运算，如果在 Tensorflow 中开启了 XLA，那么编译器会对 Tensorflow 计算图在执行阶段进行优化，通过生成特定的 GPU 内核序列来节省计算过程对于硬件资源的消耗。**一般而言，XLA 可以提供 40% 的加速**。

### **Automatic Mixed Precision**

一般深度学习模型训练过程采用单精度（Float 32）和双精度（Double）数据类型，导致预训练模型对于机器显存具有很高的要求。为了进一步减少显存开销、加快FinBERT预训练和推理速度， 我们实验采用当前最新的Tesla V100GPU进行混合精度训练。混合精度训练是指FP32和FP16混合的训练方式，使用混合精度训练可以加速训练过程同时减少显存开销，兼顾FP32的稳定性和FP16的速度。在保证模型准确率不下降的情况下，降低模型的显存占用约一半，提高模型的训练速度约 3 倍。

# **三、下游任务实验结果**

为了对比基线效果，我们从熵简科技实际业务中抽象出了四类典型的金融领域典型数据集，包括句子级和篇章级任务。在此基础上，我们将 FinBERT 与 Google 原生中文 BERT、哈工大讯飞实验室开源的 BERT-wwm 和 RoBERTa-wwm-ext 这三类在中文领域应用广泛的模型进行了下游任务的对比测试。在实验中，为了保持测试的公平性，我们没有进一步优化最佳学习率，对于四个模型均直接使用了 BERT-wwm 的最佳学习率：2e-5。

所有实验结果均为五次实验测试结果的平均值，括号内为五次测试结果的最大值，评价指标为 F1-score。

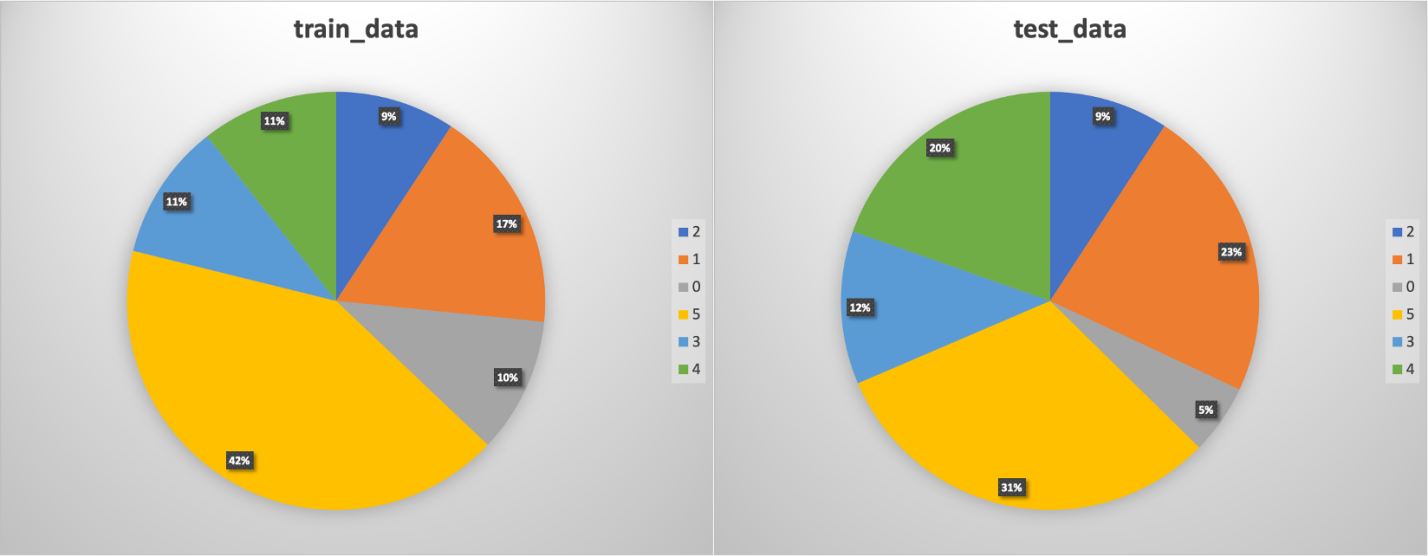
## **3.1实验一：金融短讯类型分类**

### **（1）实验任务**

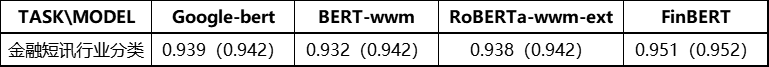
此任务来自于熵简科技信息流相关的产品，其核心任务是对金融类短文本按照文本内容进行类型分类，打上标签，从而方便用户更及时、更精准地触达感兴趣的内容。我们对原任务进行了简化，从原始的 15个类别中抽离出难度最大的 6个类别进行实验。

### **（2）数据集**

该任务的数据集共包含 3000 条样本，其中训练集数据约 1100 条，测试集数据约 1900条，各类别分布情况如下：



### **实验结果**



## **3.2实验二：金融短讯行业分类**

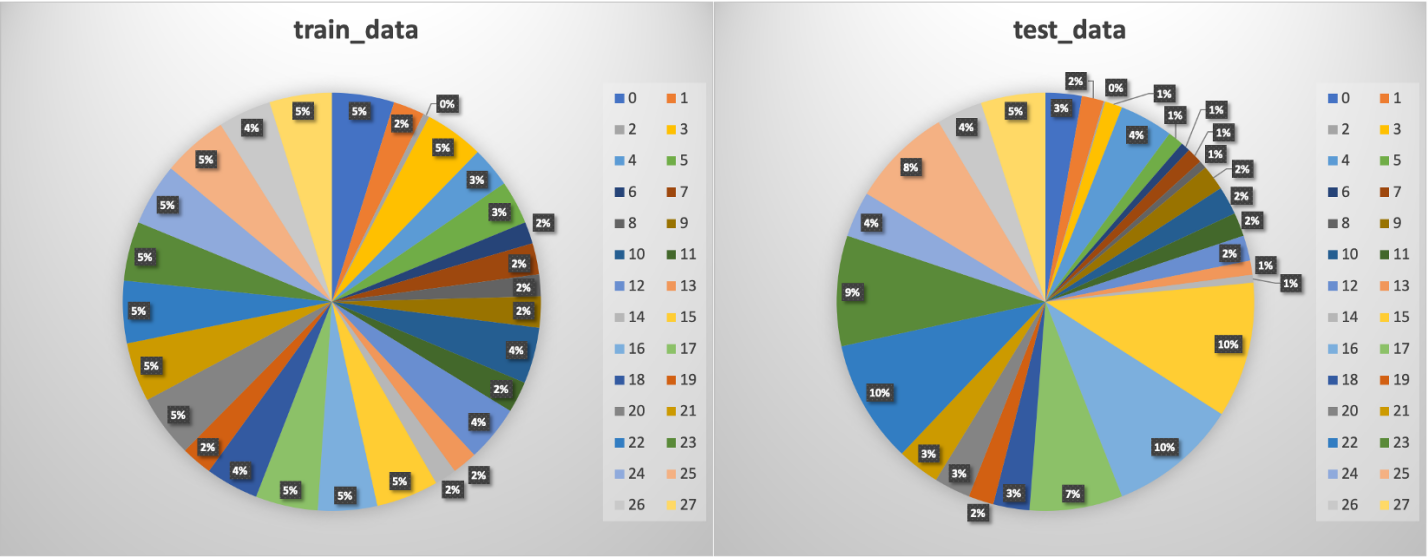
### **（1）实验任务**

此任务核心任务是对金融类短文本按照文本内容进行行业分类，以中信一级行业分类作为分类基准，包括餐饮旅游、商贸零售、纺织服装、农林牧渔、建筑、石油石化、通信、计算机等 28 个行业类别，可以用在金融舆情监控、研报/公告智能搜索等多个下游应用中。

### **（2）数据集**

该任务的数据集共包含 1200 条样本，其中训练集数据约 400 条，测试集数据约 800条。训练集中的各类别数目在 5~15 条之间，属于典型的小样本任务。

各类别分布情况如下：



### **（3）实验结果**

IMG_260

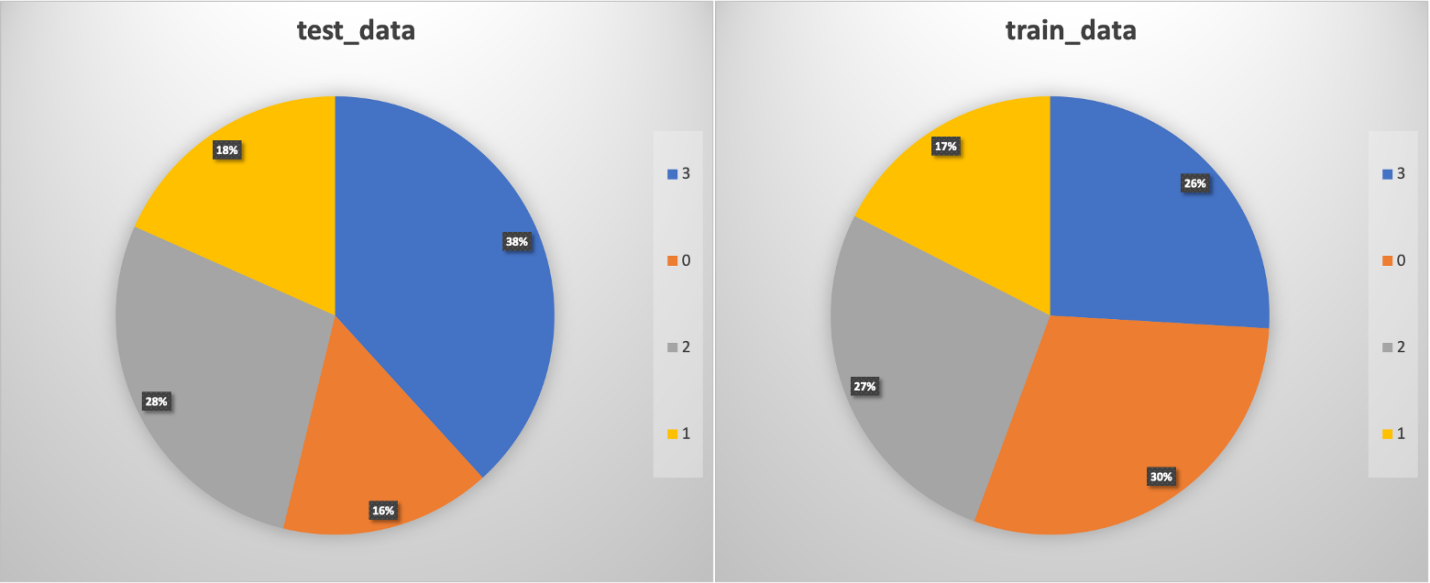
## **3.3实验三：金融情绪分类**

### **（1）实验任务**

此任务来自于熵简科技金融质控类相关产品，其核心任务是针对金融事件或标的的评述性文本按照文本内容进行金融情感分类，并用在后续的市场情绪观察和个股相关性分析中。该任务共有 4个类别，对应不同的情绪极性和强度。

### **（2）数据集**

该任务的数据集共包含 2000 条样本，其中训练集数据约 1300 条，测试集数据约 700条，各类别分布情况如下：



### **（3）实验结果**

IMG_262

## **3.4实验四：金融领域的命名实体识别**

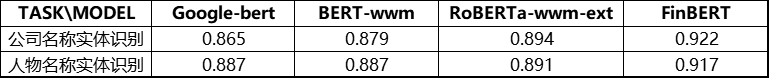
### **（1）实验任务**

此任务来自于熵简科技知识图谱相关的产品，其核心任务是对金融类文本中出现的实体（公司或人名）进行实体识别和提取，主要用在知识图谱的实体提取和实体链接环节。

### **（2）数据集**

数据集共包含 24000 条样本，其中训练集数据共3000条，测试集数据共21000条。

### **（3）结果展示**



## **3.5总结**

在本次基线测试中，我们以金融场景中所遇到的四类实际业务问题和数据入手进行对比实验，包括金融类短讯类型分类任务、金融文本行业分类、金融情绪分析任务以及金融类实体识别任务。对比 FinBERT 和 Google 原生中文BERT、 BERT-wwm、RoBERTa-wwm-ext 这三种通用领域的预训练模型可知，FinBERT 效果提升显著，在 F1-score 上平均可以提升 2~5.7 个百分点。

# **四、结语**

本文详细介绍了 FinBERT 的开源背景、训练细节和四类对比实验结果，欢迎其他从相关领域的团队提供更多、更丰富的对比实验和应用案例，让我们共同推进自然语言处理技术在金融领域的应用和发展。接下来，熵简 AI 团队会从预料规模、训练时间、预训练方式上进行更多的创新和探索，以期发展出更懂金融领域的预训练模型，并在合适时机发布 FinBERT 2.0、FinBERT 3.0，敬请期待。

任何问题，欢迎与团队负责人联系，邮箱：[liyu@entropyreduce.com](mailto:liyu@entropyreduce.com)

# **参考文献**

* [1]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018). <https://doi.org/arXiv:1811.03600v2> arXiv:1810.04805
* [2]Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics
* [3]Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv:1904.05342.
* [4]Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In Proceedings ofEMNLP.
* [5]Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:1906.08101, 2019.
* [6]Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pre-training approach. arXiv preprint arXiv:1907.11692, 2019.
* [7]Micikevicius, Paulius, et al. “Mixed precision training.” arXiv preprint arXiv:1710.03740 (2017).
* [8]<https://github.com/ymcui/Chinese-BERT-wwm/>
* [9][https://github.com/huggingface/transformers](https://github.com/huggingface/transformers.git)