机器学习方法调研

情感分析

概述

情感分析方法可以被分为以下几种：

- 机器学习方法
- 基于词典的方法`Lexicon-Based Approach`

相关介绍链接：

https://www.sciencedirect.com/topics/computer-science/lexicon-based-approach

词典计算的情感分析方法为，使用一个预训练的情感字典去给文本打分，通过聚合文本中所有词的情感分数。这里这个预训练的词典应该包括词和对应的情感分数。

> 词典中的否定词应作为独立的entry加入词典中，而且应该比非否定的词赋予更高的优先级。
>
> *这里就类似于基本加法的话，单独的like这个词可以有分数1，那么否定词don't就应该有-1.5，从而证明整个句子为否定的情感。*

情感分析中不同领域很难用同样的词典进行分析，因此对每个特定的领域需要一个新的情感分析词典。有一些工作能够从一个初始较小的词典开始，构建特定领域的领域特定情感词典。

- 混合方法。

情感分析的子类别包括：多模态情感分析`multimodal sentiment analysis`、基于方面的情感分析`aspect-based sentiment analysis`、细粒度的观点分析`fine-grained opinion analysis`以及特定语言的情感分析`language specific sentiment analysis`。

最近，深度学习的方法如`RoBERTa`和`T5`，用于训练更高性能的情感分析分类器，通过F1、recall和精度的方式进行评估。benchmark数据集包括，SST、CLUE、IMDB movie reviews。

查阅paperswithcode网站相关资料，https://paperswithcode.com/task/sentiment-analysis。大部分论文效果验证都是在更加大的数据集上，和本次作业任务规模不同，效果有一些区别。

paperswithcode网站数据

BERT

https://paperswithcode.com/paper/bert-pre-training-of-deep-bidirectional

情感分析类别数据集得分：

task	dataset	model	Metric Name	Metri Value
文本分类	DBpedia	BERT	Error	0.64
句子分类	SciCite	BERT	F1	84.4
情感分析	SST-2 Binary Classification	BERT	Accuracy	94.9

CNN类别

https://paperswithcode.com/paper/convolutional-neural-networks-for-sentence

> We report on a series of experiments with convolutional neural networks (CNN) trained on top of pre-trained word vectors for sentence-level classification tasks. We show that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results on multiple benchmarks. Learning task-specific vectors through fine-tuning offers further gains in performance. We additionally propose a simple modification to the architecture to allow for the use of both task-specific and static vectors. The CNN models discussed herein improve upon the state of the art on 4 out of 7 tasks, which include sentiment analysis and question classification.

提出了一系列采用CNN方法在为了句子级别分类的词向量上预训练，展示了一个简单的CNN组合少量的超参数调优以及静态向量就能达到多任务上的很好的效果。这个多任务包括了情感分析。

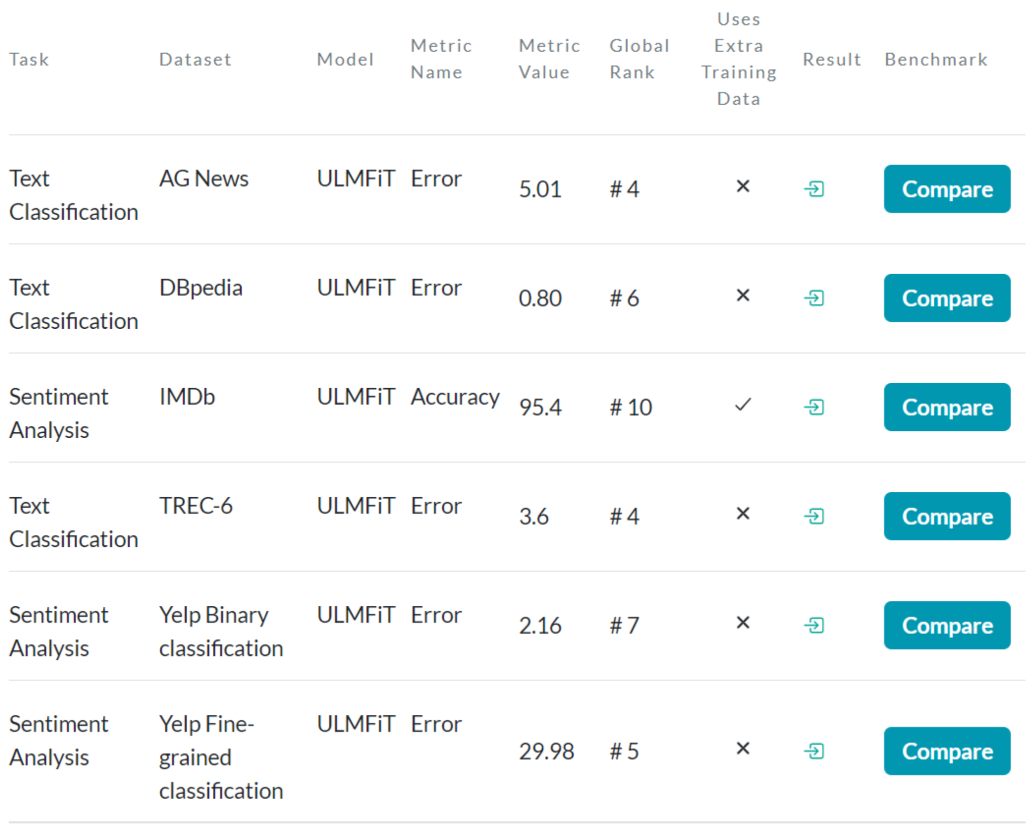
task	dataset	model	Metric Name	Metri Value
情感分析	SST-2 Binary Classification	CNN-MC	Accuracy	88.1

用于文本分类的通用语言模型微调 Universal Language Model Fine-tuning for Text classification

https://paperswithcode.com/paper/universal-language-model-fine-tuning-for-text

> We propose Universal Language Model Fine-tuning (ULMFiT), an effective transfer learning method that can be applied to any task in NLP, and introduce techniques that are key for fine-tuning a language model. Our method significantly outperforms the state-of-the-art on six text classification tasks, reducing the error by 18-24% on the majority of datasets. Furthermore, with only 100 labeled examples, it matches the performance of training from scratch on 100x more data. We open-source our pretrained models and code.

提出了通用语言微调模型，这个方法可以有效的应用到NLP中的任何领域。在多个文本分类任务有较好的性能。



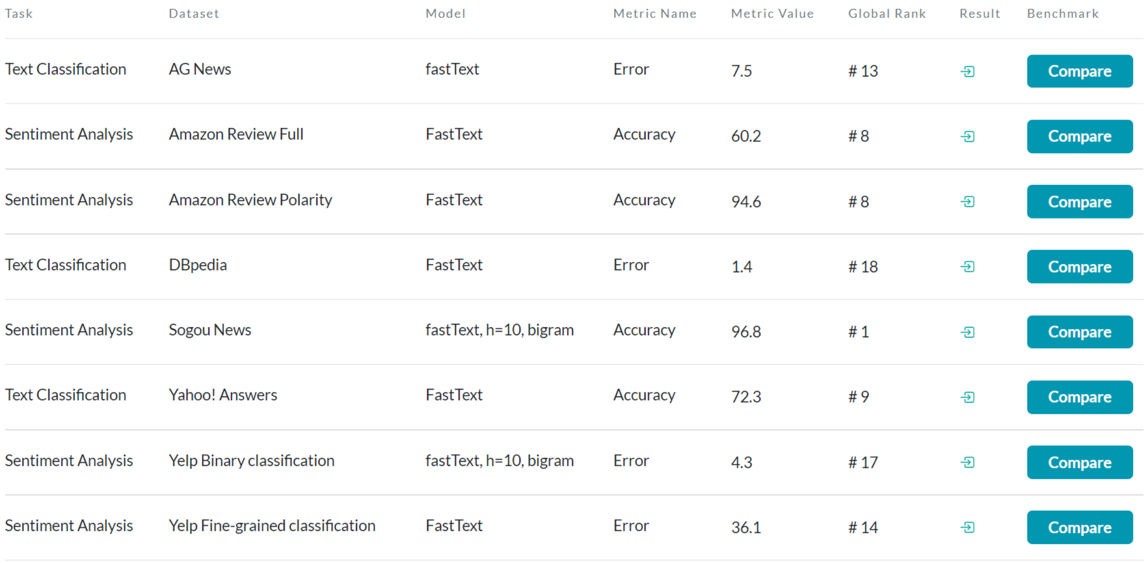
Task	Dataset	Model	Metric Name	Metric Value	Global Rank	Uses Extra Training Data	Result	Benchmark
Text Classification	AG News	ULMFiT	Error	5.01	# 4	×	⇥	Compare
Text Classification	DBpedia	ULMFiT	Error	0.80	# 6	×	⇥	Compare
Sentiment Analysis	IMDb	ULMFiT	Accuracy	95.4	# 10	✓	⇥	Compare
Text Classification	TREC-6	ULMFiT	Error	3.6	# 4	×	⇥	Compare
Sentiment Analysis	Yelp Binary classification	ULMFiT	Error	2.16	# 7	×	⇥	Compare
Sentiment Analysis	Yelp Fine-grained classification	ULMFiT	Error	29.98	# 5	×	⇥	Compare

高效文本分类技巧包 Bag o Tricks for Efficient Text Classification

https://paperswithcode.com/paper/bag-of-tricks-for-efficient-text

> This paper explores a simple and efficient baseline for text classification. Our experiments show that our fast text classifier fastText is often on par with deep learning classifiers in terms of accuracy, and many orders of magnitude faster for training and evaluation. We can train fastText on more than one billion words in less than ten minutes using a standard multicore~CPU, and classify half a million sentences among~312K classes in less than a minute.

探索了一个简单高效的文本分类方法，这个文本分类器`fastText`在准确度方面与深度学习分类器相当，而且在训练和评估方面会快很多个数量级。可以在10分钟内使用一个标准多核CPU训练超过100万个词，1分钟内将50万个句子分成30多万个类。



Task	Dataset	Model	Metric Name	Metric Value	Global Rank	Result	Benchmark
Text Classification	AG News	fastText	Error	7.5	# 13	⇥	Compare
Sentiment Analysis	Amazon Review Full	FastText	Accuracy	60.2	# 8	⇥	Compare
Sentiment Analysis	Amazon Review Polarity	FastText	Accuracy	94.6	# 8	⇥	Compare
Text Classification	DBpedia	FastText	Error	1.4	# 18	⇥	Compare
Sentiment Analysis	Sogou News	fastText, h=10, bigram	Accuracy	96.8	# 1	⇥	Compare
Text Classification	Yahoo! Answers	FastText	Accuracy	72.3	# 9	⇥	Compare
Sentiment Analysis	Yelp Binary classification	fastText, h=10, bigram	Error	4.3	# 17	⇥	Compare
Sentiment Analysis	Yelp Fine-grained classification	FastText	Error	36.1	# 14	⇥	Compare

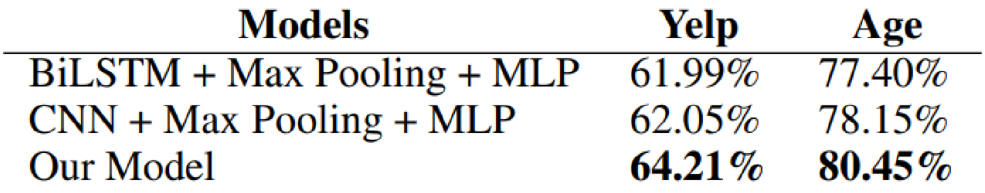
结构化的自注意力句子Embedding A Structured Self-attentive Sentence Embedding

https://paperswithcode.com/paper/a-structured-self-attentive-sentence

> This paper proposes a new model for extracting an interpretable sentence embedding by introducing self-attention. Instead of using a vector, we use a 2-D matrix to represent the embedding, with each row of the matrix attending on a different part of the sentence. We also propose a self-attention mechanism and a special regularization term for the model. As a side effect, the embedding comes with an easy way of visualizing what specific parts of the sentence are encoded into the embedding. We evaluate our model on 3 different tasks: author profiling, sentiment classification, and textual entailment. Results show that our model yields a significant performance gain compared to other sentence embedding methods in all of the 3 tasks.

引入自注意力方法，提出了一种提取可解释的句子embedding的方法。使用二维的矩阵而不是向量去表示embedding，每一行矩阵代表句子的一个不同的部分。同样在模型中提出了一个注意力机制以及特别的正则化项。这个嵌入也提供了一种简单的方法，可视化句子的哪些特定部分被编码到了embedding之中。

结果显示这个模型在情感分类方面比其他的句子embedding效果更好，在情感分类数据集Yelp上的准确率为64.21%。



Models	Yelp	Age
BiLSTM + Max Pooling + MLP	61.99%	77.40%
CNN + Max Pooling + MLP	62.05%	78.15%
Our Model	**64.21%**	**80.45%**

RoBERTa: A Robustly Optimized BERT Pretraining Approach

https://paperswithcode.com/paper/roberta-a-robustly-optimized-bert-pretraining

> We present a replication study of BERT pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD. These results highlight the importance of previously overlooked design choices, and raise questions about the source of recently reported improvements. We release our models and code.

展示了一个BERT预训练模型，仔细的衡量了很多个关键超参数的影响以及训练数据的大小。我们发现BERT是很明显训练不足的，可以匹配或超过所有后提出模型的性能。训练结果强调了以前忽视的设计选择的重要性，并且对最近报告性能改进的来源提出了质疑。

情感分类方面得分：

task	dataset	model	Metric Name	Metri Value
情感分析	SST-2 Binary Classification	RoBERTa	Accuracy	96.7

机器学习方法

参考IEEE论文：

Twitter Sentiment Analysis Using Machine Learning Algorithms: A Case Study

https://libcon.bupt.edu.cn/https/77726476706e69737468656265737421f9f244993f20645f6c0dc7a59d50267b1ab4a9/document/9213011

（通过校内VPN访问，不太确定上面那个网站能不能用）

非监督学习：加入无标签的数据集以及一些内置的库如TextBlob以及VADER，用预测一句话的情感。

监督学习：涉及到有标签的数据集如Naive Bayes、SVM等。

贝叶斯

基于可能性的方法去计算一个类别。可能性方法计算数据集中所有词的概率，而后将句子分类到特定的类别中，计算基于贝叶斯规则。

训练方法为：

- 数据集划分。
- 基于training数据集中的词建立词汇表。
- 将tweet content内容与词汇表匹配。
- 创建一个特征向量。
- 根据特征向量训练分类器。
- 用测试集测试模型效果。

SVM

> SVM is a supervised machine learning algorithm that has been used for both classification and regression problems. SVM classifies by determining a hyperplane to classify the data distributed in the ndimensional space. The classification is done based on the mathematical functions called kernels and these kernels are used to determine a hyperplane. Two different classes exist on the opposite sides of the hyperplane and thus this plane could be

> considered a decision boundary which could help in simple classification of the data points available.

监督的机器学习算法，在N维空间里设置一个高维平面在多维空间里区分数据。这个分类是基于数学的kernel，用这些kernel去定义一个高维平面，这个平面可以认为是一个决策边界。

训练方法为：

- 收集训练和测试数据集。
- 数据向量化。
- 创建一个训练的SVM。
- 将模型应用到测试集之上。

随机森林分类器

> This is a classifier that is composed of several Decision Trees as a single Decision Tree suffers from several disadvantages like noise which can affect the overall results and performance. However, Random Forest classifier has several advantages over the Decision Tree like it is robust when compared to a Decision Tree and the reason being, a Random Forest Classifier uses the concept of Bootstrapping i.e. each tree is trained on a different training data as we divide the training data into subsets equivalent to the number of trees. This ensures that each tree has a different and its own training data.

由多个决策树组成的分类器。将数据划分成和决策树数量相同的自己，因此每棵决策树都有不同的、专属的训练数据。

LSTM

LSTM设计了输入门、遗忘门和输出门，这些门是结果为0-1之间的sigmoid函数。