北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

情感分析综述

成员：雷志强 雷浩洁 林金坤 王江雨 程玉琪 马平川 刘宏荣　　时间：2021/10/18

目录

CONTENTS

1 概述 Introduction

2 传统方法 Traditional Methods

3 现代方法 Modern Methods

4 Demo Demo

5 总结 Summary

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

概述

Introduction

概念

利用计算机技术，对自然语言文本进行分析，挖掘其中包含的情感、态度等主观信息的研究。

背景

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- **社交媒体**
  - 微博 知乎 豆瓣
- **电子商务**
  - 淘宝 京东 亚马逊
- **领域相关平台**
  - 美团 哔哩哔哩 IMDB

越来越多的网络用户成为内容的贡献者或传播者，互联网上的主观信息急剧增长。
需要高效快速的处理甄别信息

应用与意义

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

政府与国家决策

利用基于情感分析的技术实现无感的民情调查、舆情分析，与传统的问卷式调查相比，具有更高的准确性和效率，成本也更低。

企业/金融决策

通过对各个平台的反馈建议和市场相关信息的处理，得到有效的辅助决策信息，帮助企业/金融市场做出正确的决策。

消费/产品优化

一方面可以帮助消费者形成对产品或品牌的全面评价；另一方面可以帮助企业了解产品市场评价，针对性优化。

应用导向研究

基于社交网络的选举结果预测；

研究话题

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

情感语料构建

情感语料库是研究的重要资料，用于训练模型、构建词典等工作。

情感词典构建

识别文本情感的工具。SENTIWORDNET3.0 HOWNET等

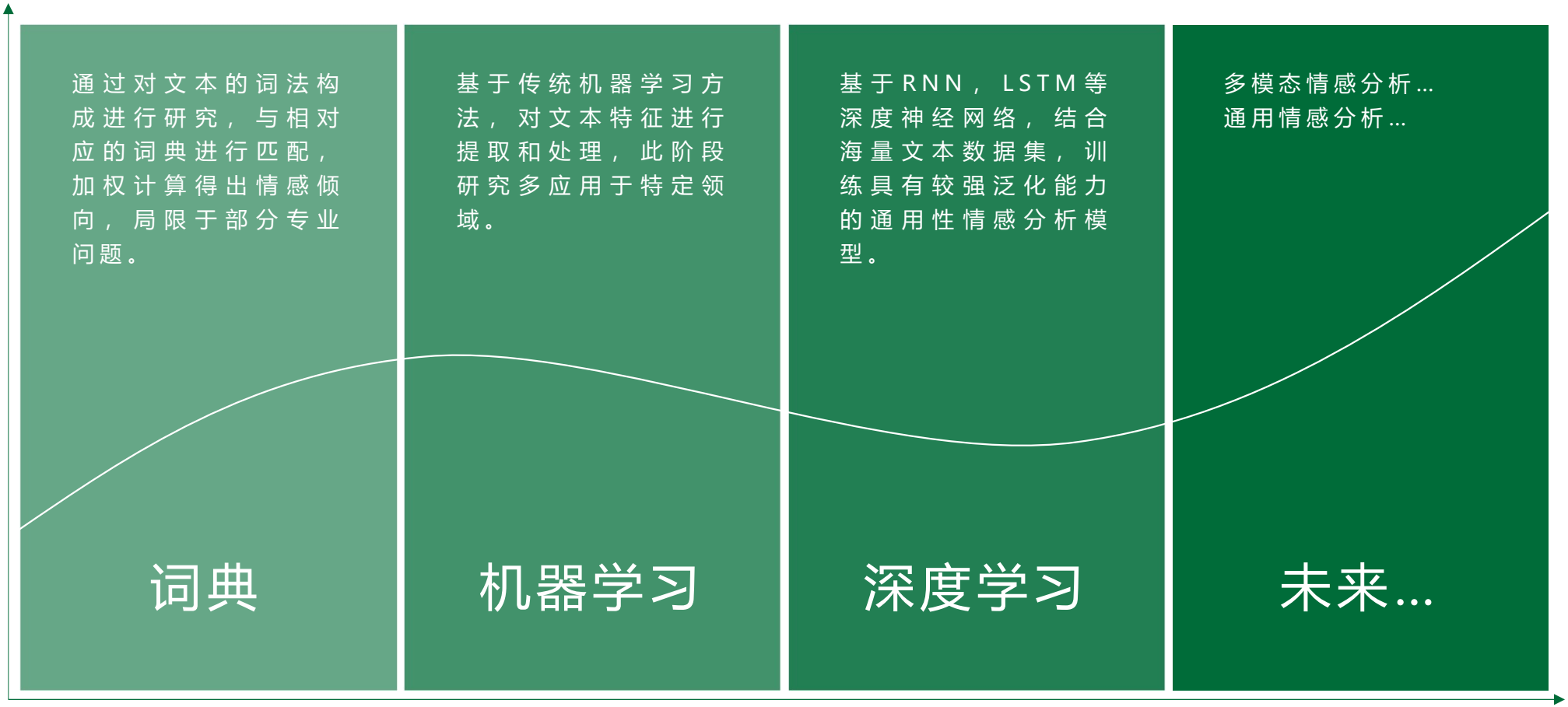
情感信息抽取

抽取文本的情感信息，包含多个细分的研究方向。最主要的研究话题。

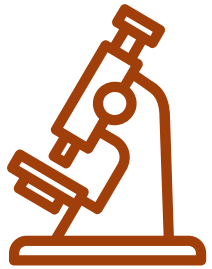
情感应用研究

研究情感分析的应用领域。

发展历程


通过对文本的词法构成进行研究，与相对应的词典进行匹配，加权计算得出情感倾向，局限于部分专业问题。
词典
基于传统机器学习方法，对文本特征进行提取和处理，此阶段研究多应用于特定领域。
机器学习
基于RNN，LSTM等深度神经网络，结合海量文本数据集，训练具有较强泛化能力的通用性情感分析模型。
深度学习
多模态情感分析...
通用情感分析...
未来...

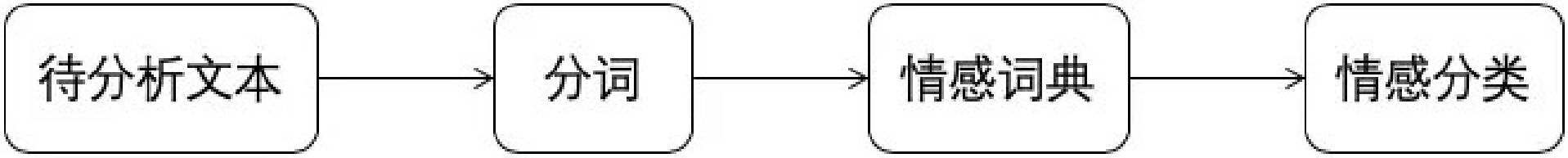
北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



传统方法

Traditional Methods

基于情感词典的情感分析方法

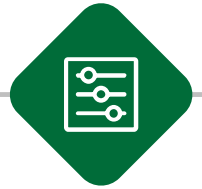


基于词典的方法主要通过制定一系列的情感词典和规则，对文本拆解、关键词提取，计算情感值，最后通过情感值来作为文本的情感倾向依据。

典型情感词典

BosonNLP

基于微博、新闻、论坛等数据来源构建的情感词典



知网(HowNet)

包含中文和英文两部分

NTSUSD

台湾大学中文情感极性词典

大连理工大学中文情感词汇本体库

基于机器学习的情感分析方法

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 将情感分析问题视作一个典型的文本分类任务
- 使用传统的机器学习方法作为分类器，利用文本数据中的语法特征解决分类问题，常见的模型包括：
  - 概率模型分类器
  - 线性分类器
  - 决策树分类器
  - 无监督分类器

问题定义：

给定一组有标签的训练样本$D = \{X_1, X_2, \ldots, X_N\}$，通过找到一个分类模型使得数据的底层特征信息与其对应的标签相关联。对于硬分类问题，一个实例只对应标签集中的一个标签；对于软分类问题，一个实例对应多个标签。

概率模型分类器

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

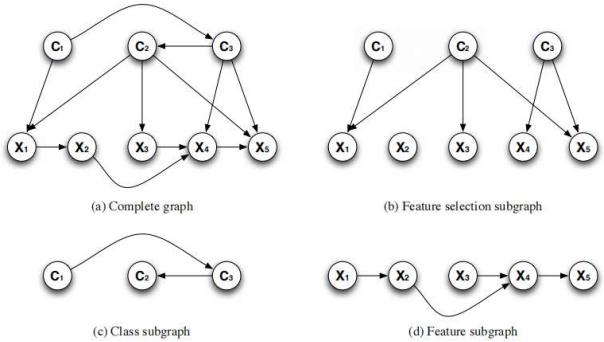
- 通常使用混合模型进行分类，并假设每个类对应混合模型中的一个生成模型，这类分类器也被称为生成型分类器（*Generative Classifiers*）

1.贝叶斯分类器[1]（*Naïve Bayes Classifier, NB*）

$$P(\text{label}|features) = \frac{P(label) * P(f1|label) * \cdots * P(label) * P(fn|label)}{P(features)}$$

2.贝叶斯网络[2]（*Bayesian Network, BN*）

假设所有特征是相互独立的，并得到一个型为有向无环图的贝叶斯网络模型，图中节点代表随机变量，边代表条件依赖。



1. Kang H, Yoo S J, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews[J]. Expert Systems with Applications, 2012, 39(5): 6000-6010.
2. Ortigosa-Hernández J, Rodríguez J D, Alzate L, et al. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers[J]. Neurocomputing, 2012, 92: 98-115.

概率模型分类器

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

3.最大熵分类器[1]（*Maximum Entropy Classifier, ME*）

- 最大熵分类器又被称为条件指数分类器（conditional exponential classifier），由一组“权重”参数化，用于组合通过“编码”从特征集和类别生成的联合特征。
- 分类器通过一组权重$X\{\text{weights}\}$进行参数化，这些权重被用于组合经过encoder编码过的联合特征，并根据以下公式给出每一个标签对应的概率：

$$P(label|fs) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))\ for\ l\ in\ labels)}$$

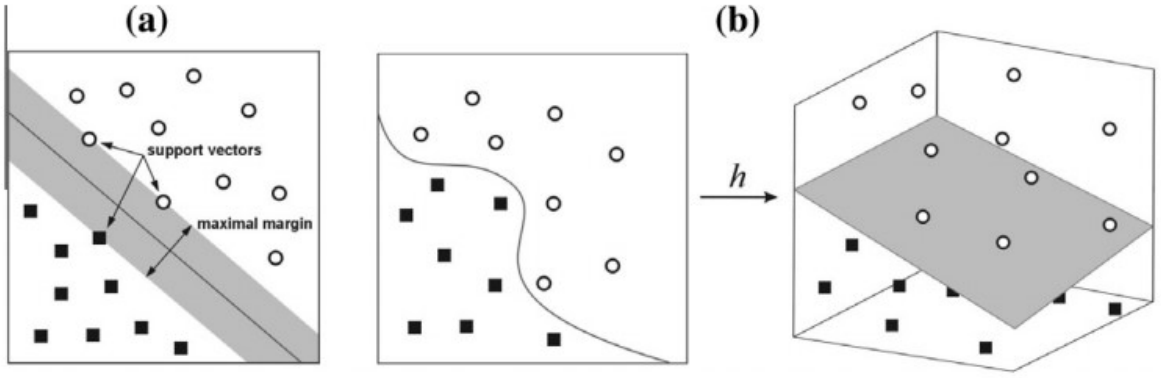
1. Kaufmann M. JMaxAlign: A maximum entropy parallel sentence alignment tool[C]//Proceedings of COLING 2012: Demonstration papers. 2012: 277-288.

线性分类器

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

1. SVM

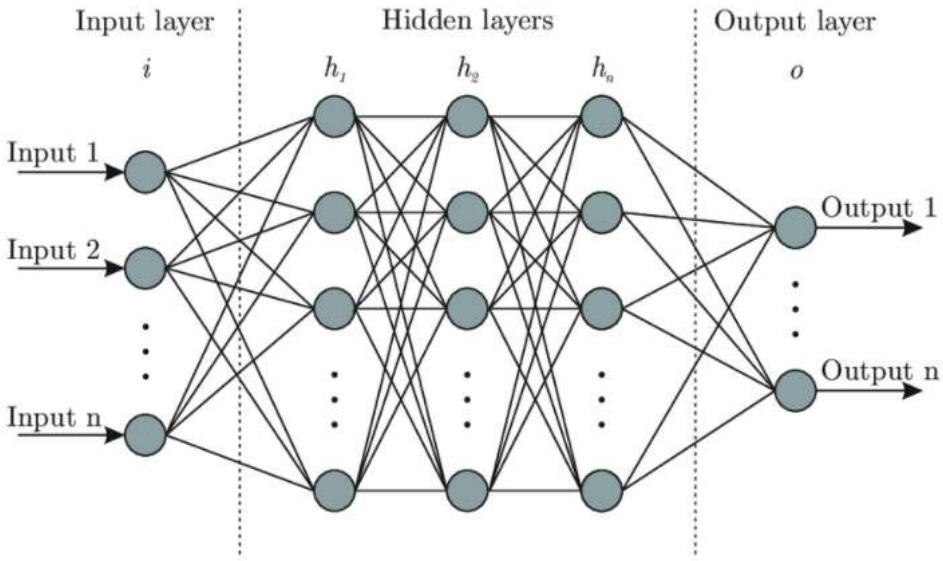
- 文本特征的稀疏性决定了很少有特征是不相关的，并通常可以组织成线性可分的类别。
- 具体来说，将文本数据实例非线性地映射到线性可分的特征空间中，使用SVM找到线性分类的超平面以达到在原始特征空间中构造一个非线性的决策曲面。



线性分类器

2. ANN

- 从输入数据的线性组合中推导出特征信息，并将输出建模为这些特征的非线性函数。
- 输入特征通常为文档中的单词及其出现频次。



决策树分类器

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

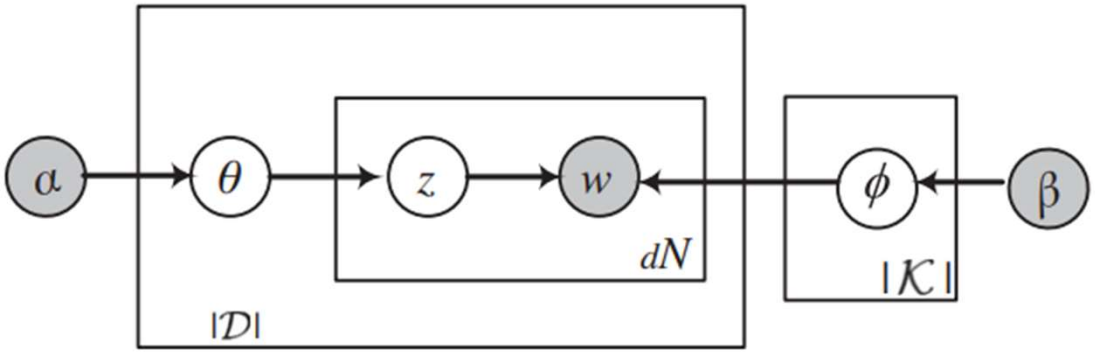
- 决策树分类器提供了训练数据空间的层次分解，其中使用属性值上的一个条件来划分数据，决策树的叶子节点代表用于分类的一个具体类别。
- 决策树中执行分割的方式通常有如下几种：

1. 单属性分割：在树中的特定节点根据特征单词或短语（通常具有强烈的情感倾向）的存在与否作为分割条件[1]。
2. 基于相似度的多属性分割：使用文档或者频繁出现的单词簇，通过衡量文档与这些单词簇之间的相似性来执行分割过程。
3. 基于判别器的多属性分割：例如使用Fisher判别器来进行分割过程[2]。

1. Lewis D D, Ringuette M. A comparison of two learning algorithms for text categorization[C]//Third annual symposium on document analysis and information retrieval. 1994, 33: 81-93.
2. Chakrabarti S, Roy S, Soundalgekar M V. Fast and accurate text classification via multiple linear discriminant projections[J]. The VLDB journal, 2003, 12(2): 170-185.

无监督分类器

通过无监督的机器学习方法解决情感分类问题，从而避免人工标注大量文本数据。



图：使用LDA模型进行情感分析

Xianghua F, Guo L, Yanyan G, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon[J]. Knowledge-Based Systems, 2013, 37: 186-195.

情感分析平台

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

NLPIR-ICTCLAS汉语分词系统

Python Jieba

thunlp/THULAC-Java

An Efficient Lexical Analyzer for Chinese

語言

msr_test（560KB）

Algorithm	Time	Precision	Recall	F-Measure
LTP-3.2.0	3.21s	0.867	0.896	0.881
ICTCLAS(2015版)	0.55s	0.869	0.914	0.891
jieba(C++版)	0.26s	0.814	0.809	0.811
THULAC_lite	0.62s	0.877	0.899	0.888

pku_test（510KB）

Algorithm	Time	Precision	Recall	F-Measure
LTP-3.2.0	3.83s	0.960	0.947	0.953
ICTCLAS(2015版)	0.53s	0.939	0.944	0.941
jieba(C++版)	0.23s	0.850	0.784	0.816
THULAC_lite	0.51s	0.944	0.908	0.926

图：ICTCLAS，LTP，jieba，THULAC性能对比

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

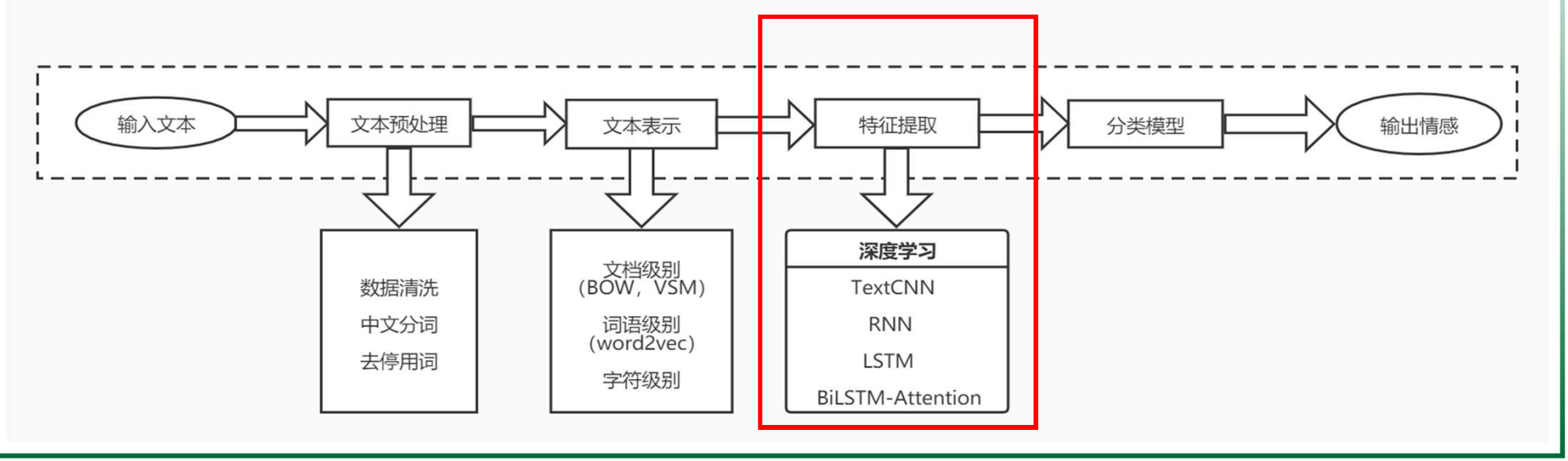
现代方法

Modern Methods

理论基础

基于深度学习的情感分析方法

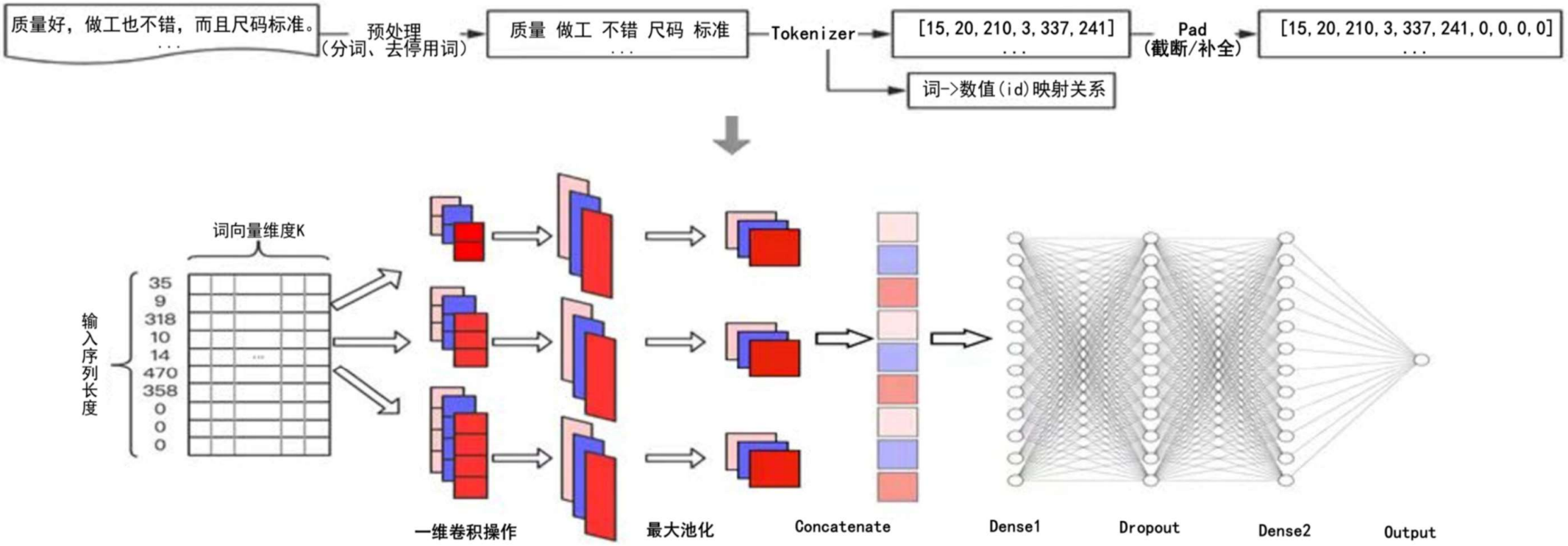
基于深度学习的分析方法，经典模型有：**TextCNN**(Text Convolutional Neural Networks)、**RNN**（Recurrent Neural Network）、**LSTM**（Long Short-Term Memory）、**BiLSTM+Attention**等，近几年基于**Transformer**的预训练-微调训练模型火热。



TextCNN

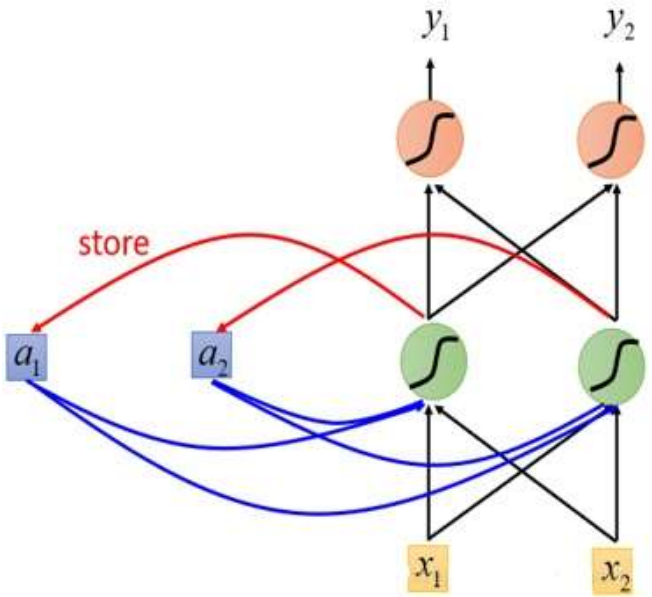
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

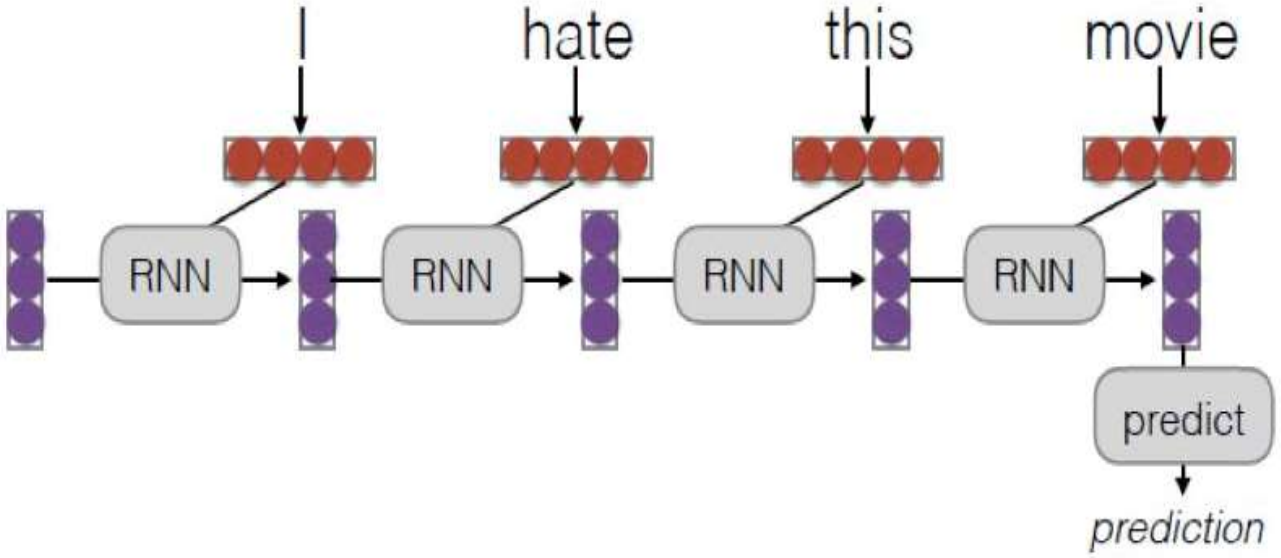
- TextCNN开创的将源于计算机视觉的卷积神经网络CNN 应用于NLP文本分类中。请模型结构比较简单，由输入表征->卷积层->最大池化->全连接层->输出softmax组成。



RNN

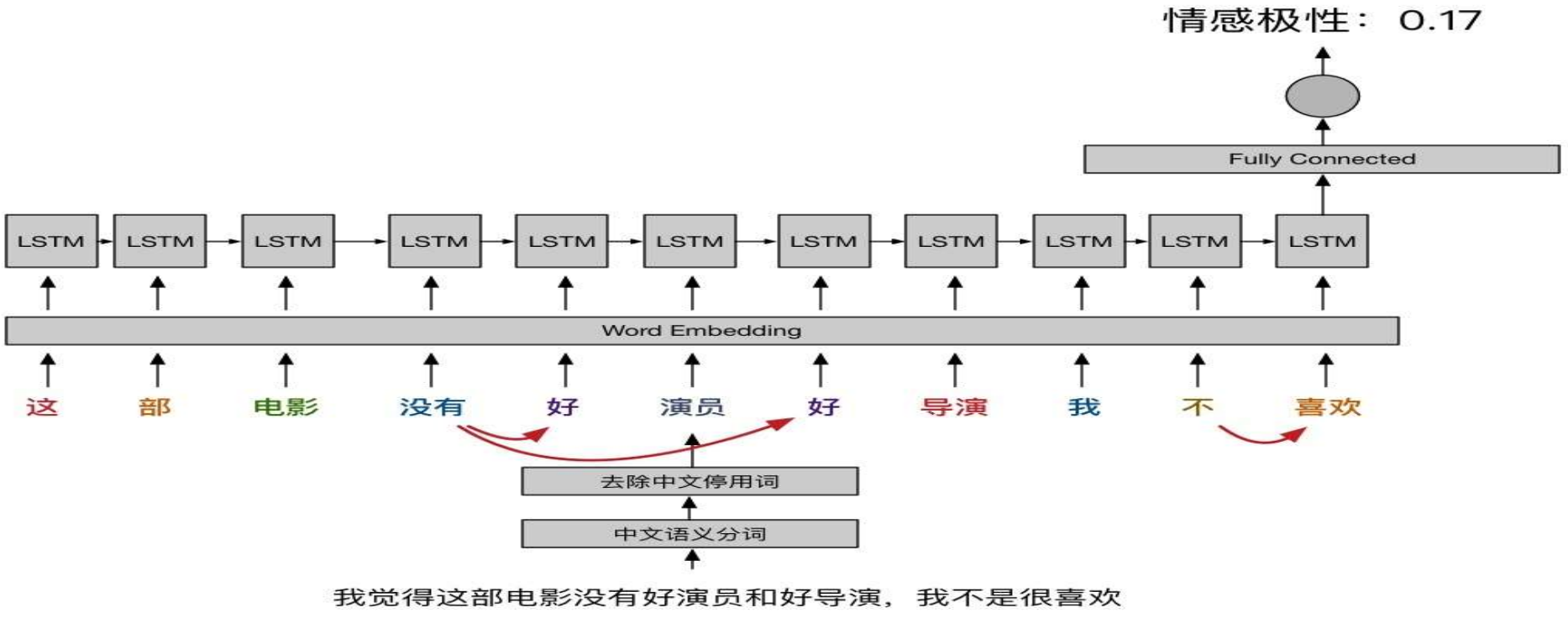
- RNN是处理文本任务的经典模型。一个典型的 RNN 结构由输入层、隐藏层和输出层组成。RNN实现了类似于人脑的这一机制，对所处理过的信息留存有一定的记忆。





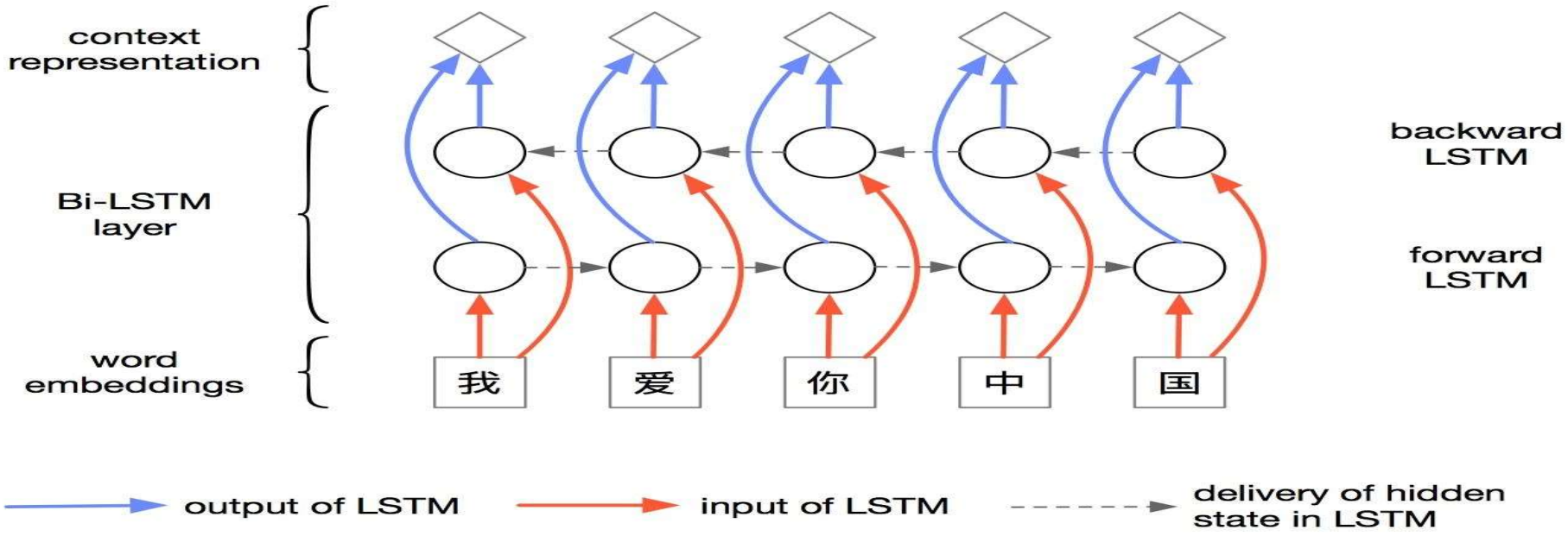
LSTM

- LSTM主体结构沿用RNN的样子，只是把RNN cell改成了LSTM cell，引入了记忆单元和三个门：输入门、遗忘门、输出门。实现了保留较长序列中的重要信息，忽略不重要信息。



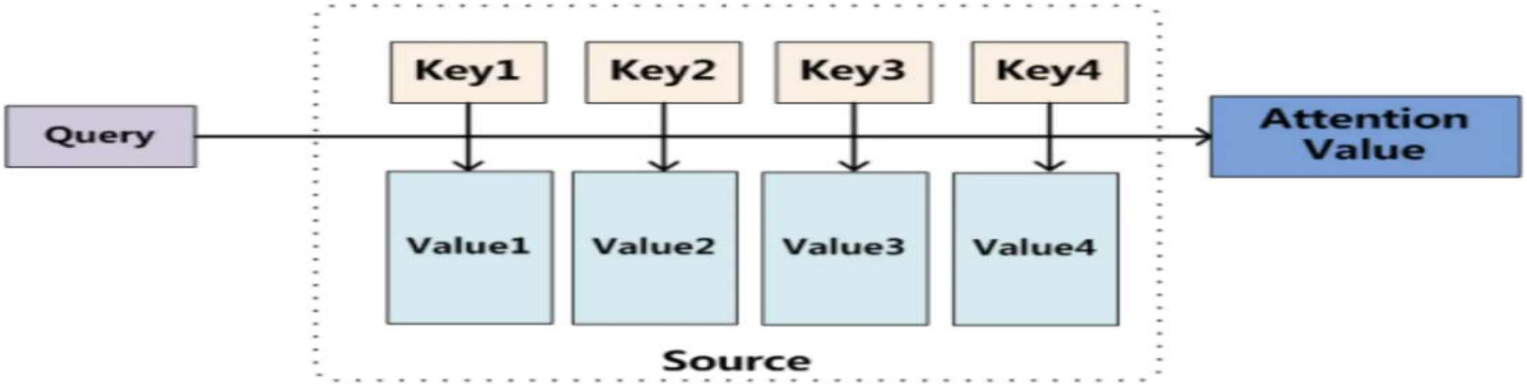
Bi-LSTM

- BiLSTM由前向的LSTM与后向的LSTM结合而成,输入序列分别以正序和逆序输入至2个LSTM神经网络进行特征提取，将2个输出向量（即提取后的向量）进行拼接后形成的词向量作为该词的最终特征表达。



Attention

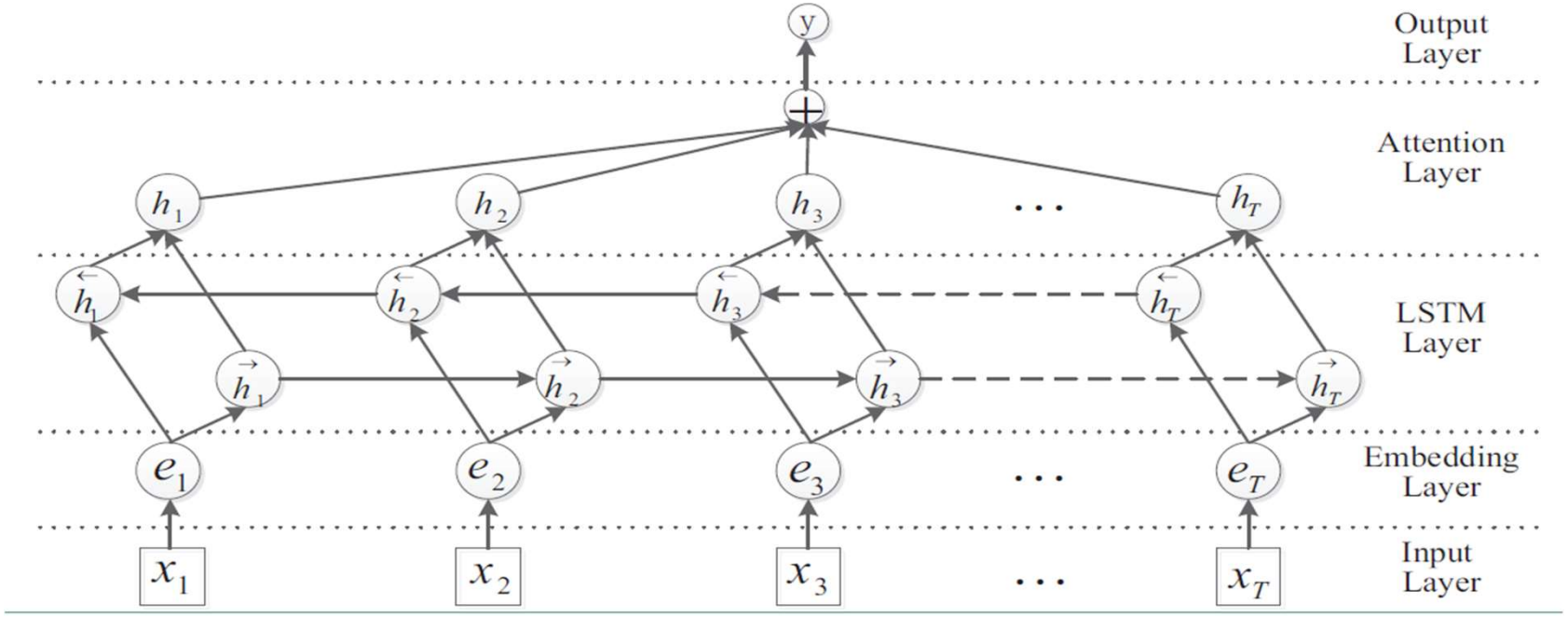
- Attention机制可以看作，给定一个Query，计算Source中每个单词对Target中每个单词的重要程度（权重），并对权重大的词语给予更多的关注，对权重小的词语进行远离。



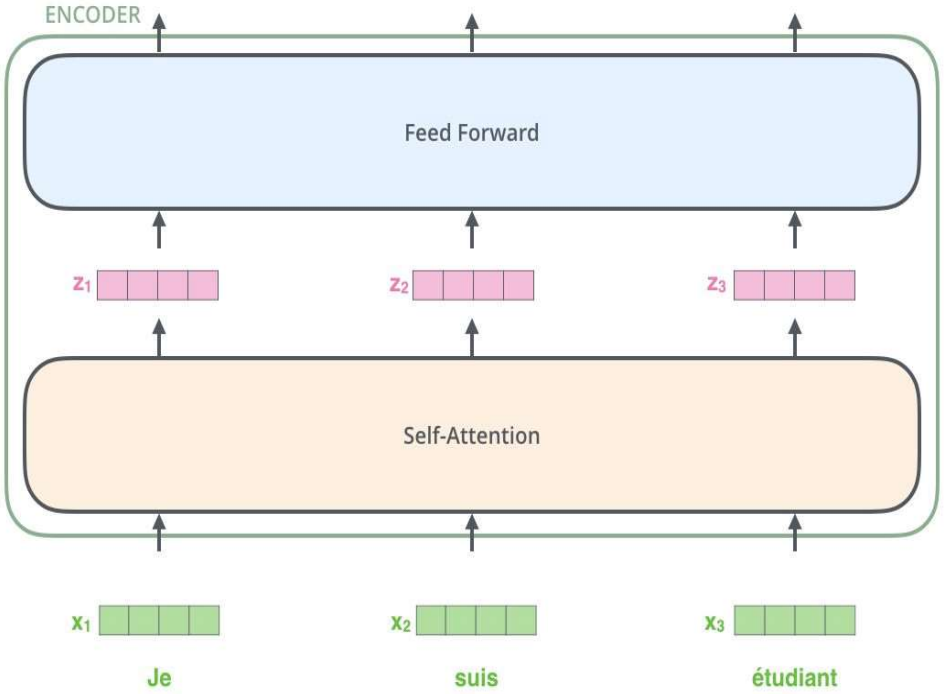
Bi-LSTM+Attention

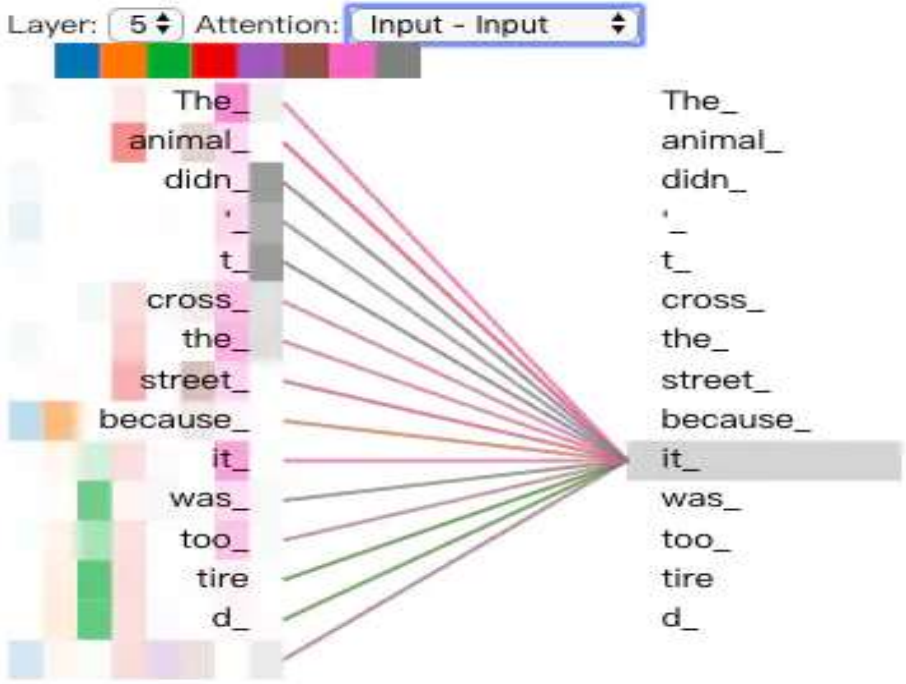
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- Bi-LSTM+Attention，通过双层LSTM进行特征提取，输入到Attention中，先计算每个时序的权重，并将所有时序的向量进行加权和作为特征向量，然后进行softmax分类，得到情感倾向。实现了**参数少、速度快、 效果好。**



- 能够同时处理句子中的所有词，并且任意两个词之间操作的距离都是1；通过自注意力机制，句子中的某个词对于本身的所有词做一次Attention，每个词都考虑了上下文的关系。这样很好的解决了情感分析问题中的效率和依赖问题。





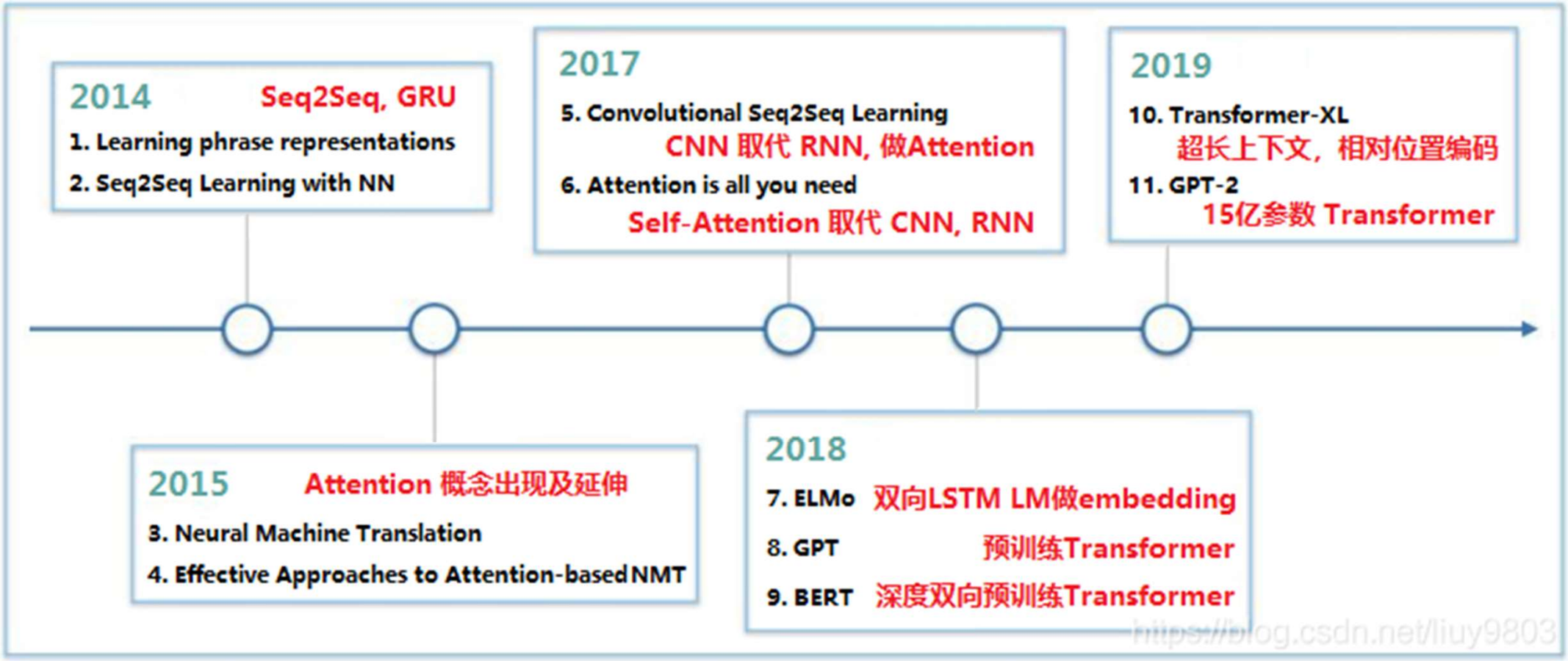
北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

现代方法

Modern Methods

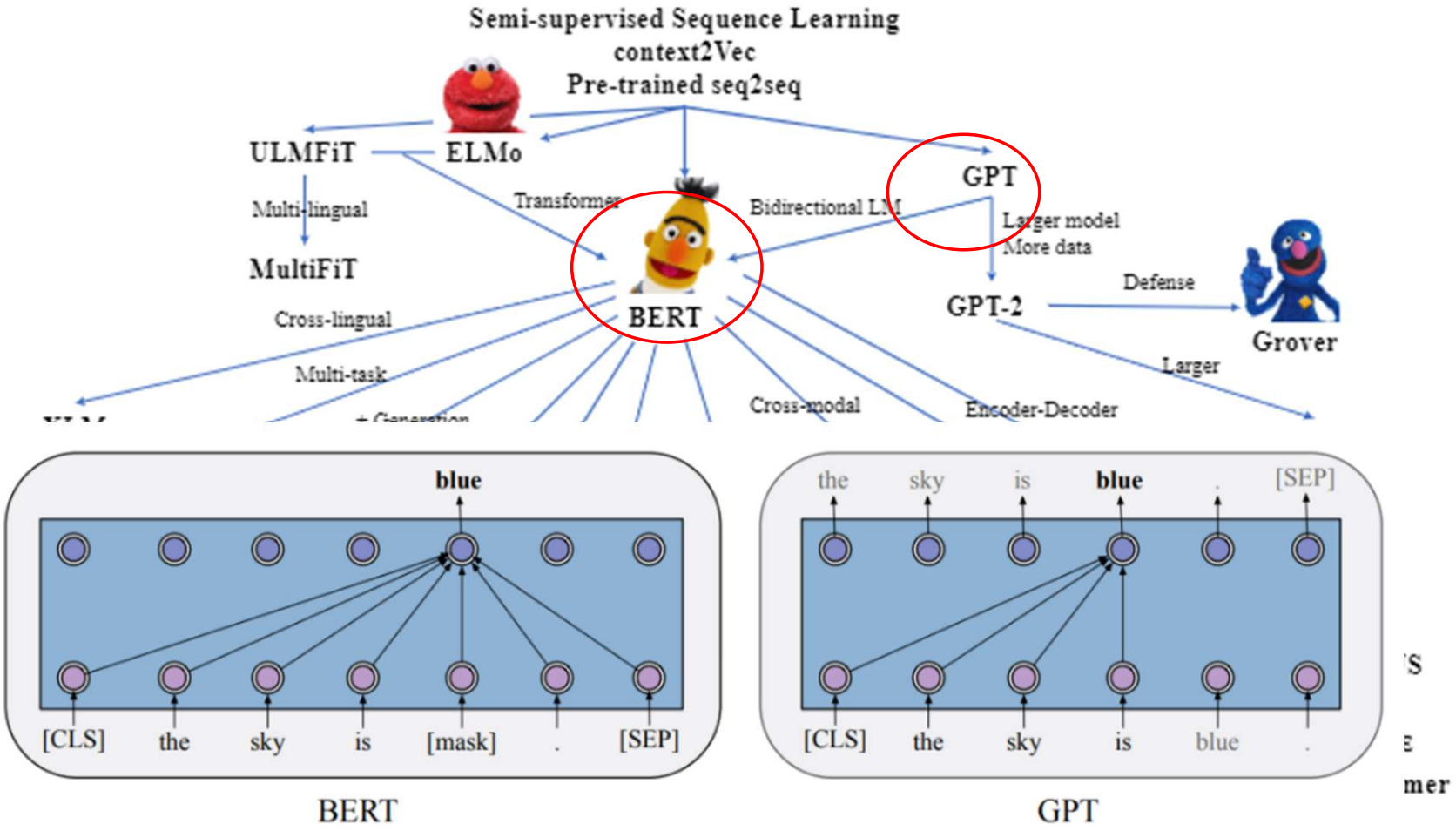
预训练与SOTA

发展历程



https://blog.csdn.net/liuy9803

预训练模型


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
GPT
Larger model
More data
GPT-2
Defense
Grover
Larger
BERT
Cross-lingual
Multi-task
Cross-modal
Encoder-Decoder
blue
[CLS] the sky is [mask] . [SEP]
BERT
the sky is blue . [SEP]
[CLS] the sky is blue .
GPT

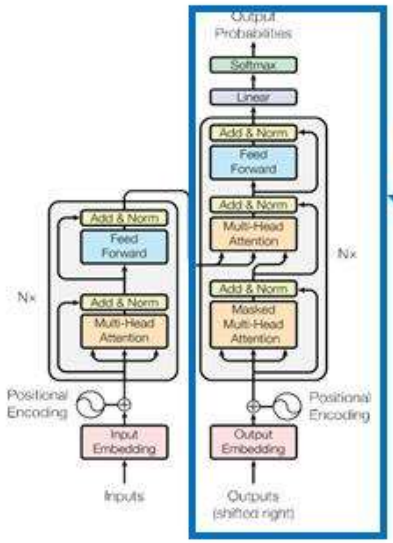
Bert的改进

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 优点：能够获取上下文相关的双向特征表示
- 缺点：
  - 预训练过程和生成过程不一致，生成任务表现不佳
  - 没有考虑预测[MASK]之间的相关性
  - 输入噪声[MASK]，造成预训练-微调两阶段的差异
  - 无法文档级别的NLP任务
- 生成任务
  - MASS UNILM
- 引入知识
  - ERNIE 1.0 ERNIE（THU）
- 多任务学习
  - MTDNN ERNIE 2.0
- MASK策略
  - Bert-WWM ERNIE SpanBert
- 精细调参
  - RoBerta
- XLNet

GPT-2

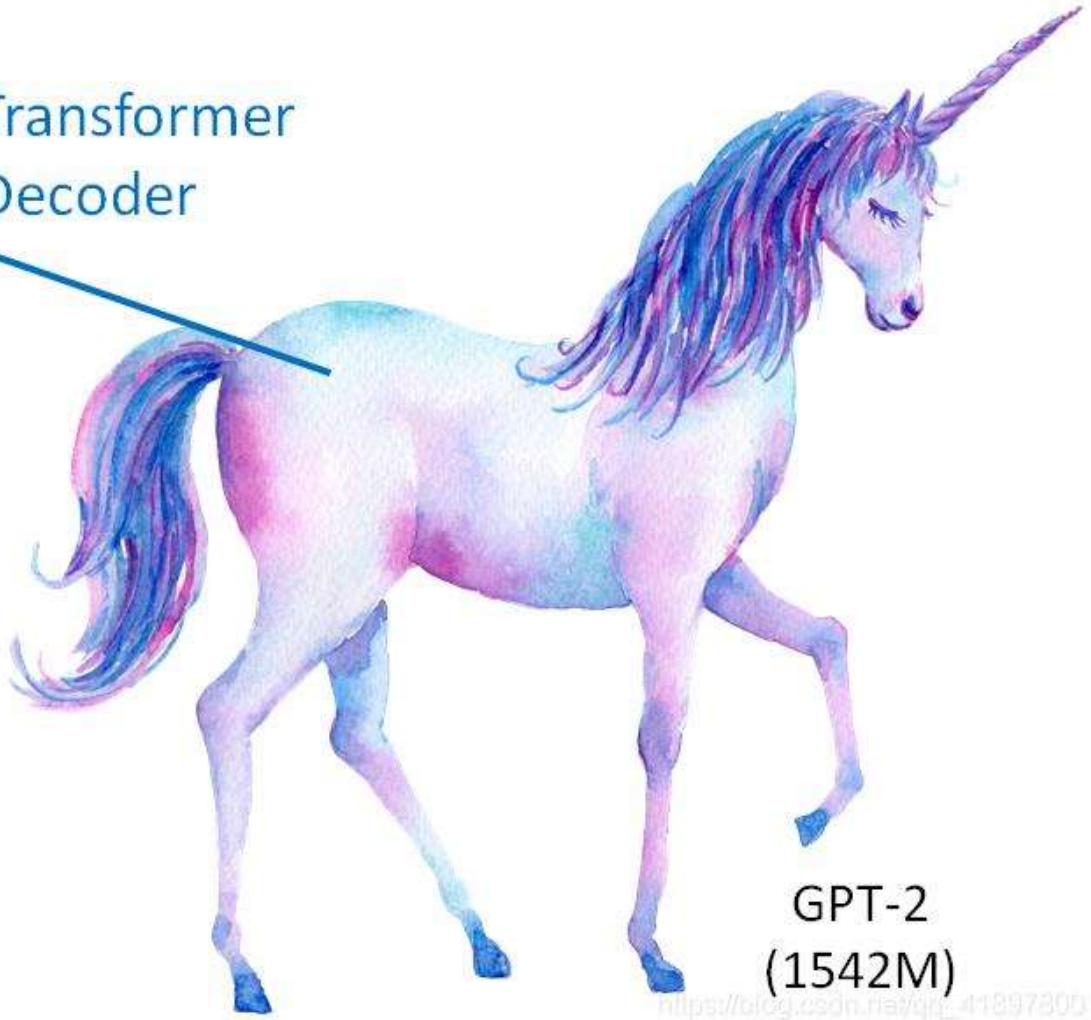
参数比较



Transformer Decoder

BERT (340M)

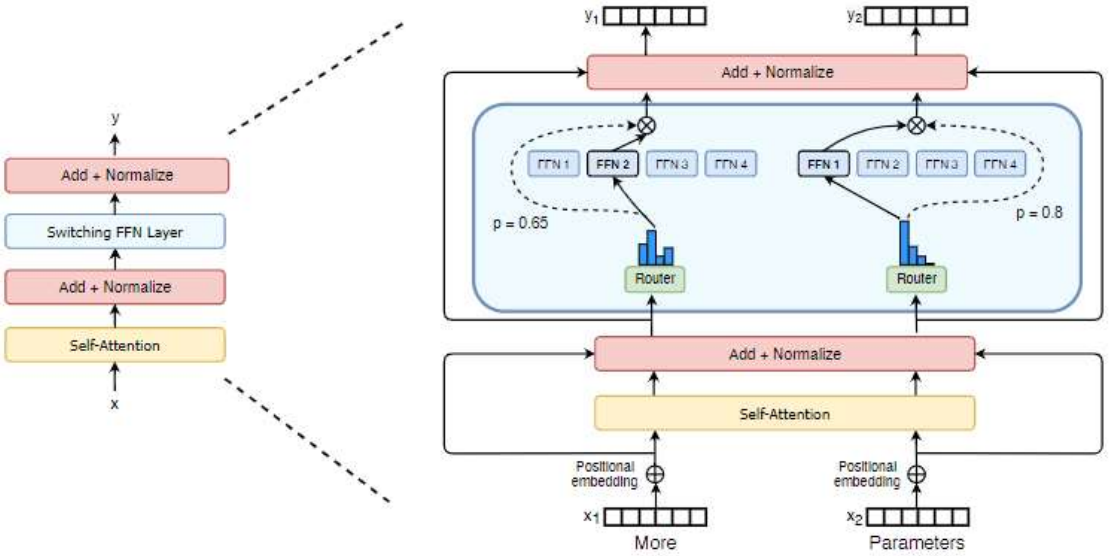
ELMO (94M)



GPT-2 (1542M)

后起之秀

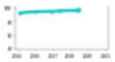
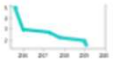
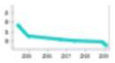
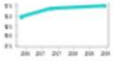
T5



T5 110亿参数　　GPT3 1750亿参数　　Switch Transformer 1.6万亿参数

SOTA

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Trend	Dataset	Best Model	Paper Title
	SST-2 Binary classification	SMART-RoBERTa Large	SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization
	IMDb	NB-weighted-BON + dv-cosine	Sentiment Classification Using Document Embeddings Trained with Cosine Similarity
	SST-5 Fine-grained classification	RoBERTa-large+Self-Explaining	Self-Explaining Structures Improve NLP Models
	Yelp Binary classification	XLNet	XLNet: Generalized Autoregressive Pretraining for Language Understanding
	Yelp Fine-grained classification	XLNet	XLNet: Generalized Autoregressive Pretraining for Language Understanding
	MR	EFL	Entailment as Few-Shot Learner
	Amazon Review Polarity	BERT large	Unsupervised Data Augmentation for Consistency Training
	Amazon Review Full	BERT large	Unsupervised Data Augmentation for Consistency Training

SST-2数据集

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Rank	Model	Accuracy↑	Paper	Code	Result	Year
1	**SMART-RoBERTa Large**	97.5	SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization			2019
2	**T5-3B**	97.4	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer			2019
3	**MUPPET Roberta Large**	97.4	Muppet: Massive Multi-task Representations with Pre-Finetuning			2021
4	**ALBERT**	97.1	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations			2019
5	**T5-11B**	97.1	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer			2019
6	**StructBERTRoBERTa ensemble**	97.1	StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding			2019
7	**XLNet** (single model)	97	XLNet: Generalized Autoregressive Pretraining for Language Understanding			2019

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

现代方法

Modern Methods

前沿进展

属性级(Aspect)情感分析

属性级情感分析旨在捕捉用户生成的评论中对产品、电影、公司等评价对象的不同方面/属性所表达的情感的情感极性。

输入文本：这家店环境不错，位置也好，就是服务员态度一般

抽取

分析结果：

- 环境　不错
- 位置　好
- 服务员态度　一般

分类

属性级(Aspect)情感分析

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

1. A Unified Generative Framework for Aspect-Based Sentiment Analysis. Yan et.al. ACL2021

本文认为Aspect-based Sentiment Analysis(ABSA)任务包含七个子任务，目前现有工作大多只是完成上述部分任务，缺少一个**统一的框架**。本文把这7个子任务的目标转化成了包含指针索引和情感类别索引的序列，从而统一了上述 7个任务的输出。本文使用预训练模型**BART**，用端到端的方式解决ABSA的所有7个子任务。

2. Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis. Li et.al. ACL2021

本文提出一个双重的**图卷积网络**DualGCN，可以同时考虑到语法结构的互补性以及语义的相关性。使用SynGCN模块捕获更丰富的句法信息，使用包含自注意力机制的SemGCN捕获语义关联。

3. Aspect-Category-Opinion-Sentiment Quadruple Extraction with Implicit Aspects and Opinions. Li et.al. ACL2021

本文提出一个**新的任务**，抽取（属性词、属性类别、观点、情感极性）**四元组**，同时也关注到那些隐含的属性和观点。

属性级(Aspect)情感分析

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

4. A Joint Training Dual-MRC Framework for Aspect Based Sentiment Analysis. Mao et.al. AAAI2021

本文将ABSA转换为两个**机器阅读理解问题MRC**，通过对两个BERT-MRC模型进行参数共享联合训练，解决了所有的子任务。。

5. Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis. Phan and Ogunbona. ACL2020

本文将属性词抽取和属性情感分类任务分开来研究，结合**BERT**和句法信息抽取出属性词，再根据属性词实现情感分类。

6. Aspect Sentiment Classification with Document-level Sentiment Preference Modeling. ACL2020

本文提出了**文档级**的属性情感分类方法，他们认为无论是微博等社交文本还是购物平台的评价文本，句子都不是单独出现，而是几句含义较为集中、情感较为一致的句子共同出现。由此，提出了一种**联合图注意力网络CoGAN**方法，分别在对象内和跨对象两个层级收集情感信息，并将这两种情感信息在图注意力网络上优化，在联合分析后得出针对对象的情感倾向

对话情绪识别

机器与人的情感交流

对话背景　上下文　说话人

对话情绪识别

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

1. ***Directed Acyclic Graph*** *Network for Conversational Emotion Recognition. Shen et.al. ACL2021*

2. *Topic-Driven and Knowledge-Aware* ***Transformer*** *for Dialogue Emotion Detection. Zhu et.al. ACL2021*

3. *DialogueCRN:* ***Contextual Reasoning Networks*** *for Emotion Recognition in Conversations. Hu et.al. ACL2021*

4. *DialogXL: All-in-One* ***XLNet*** *for Multi-Party Conversation Emotion Recognition. Shen et.al. AAAI2021*

5. ***Quantum-inspired*** *Neural Network for Conversational Emotion Recognition. Li et.al. AAAI2021*

6. *Relation-aware* ***Graph Attention Networks*** *with Relational Position Encodings for Emotion Recognition in Conversations. Ishiwatari et.al. EMNLP2020*

7. ***DialogueGCN****: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. Ghosal et.al. EMNLP2019*

8. ***DialogueRNN****: An Attentive RNN for Emotion Detection in Conversationsn. Majumder et.al. AAAI 2019*

对话情绪识别

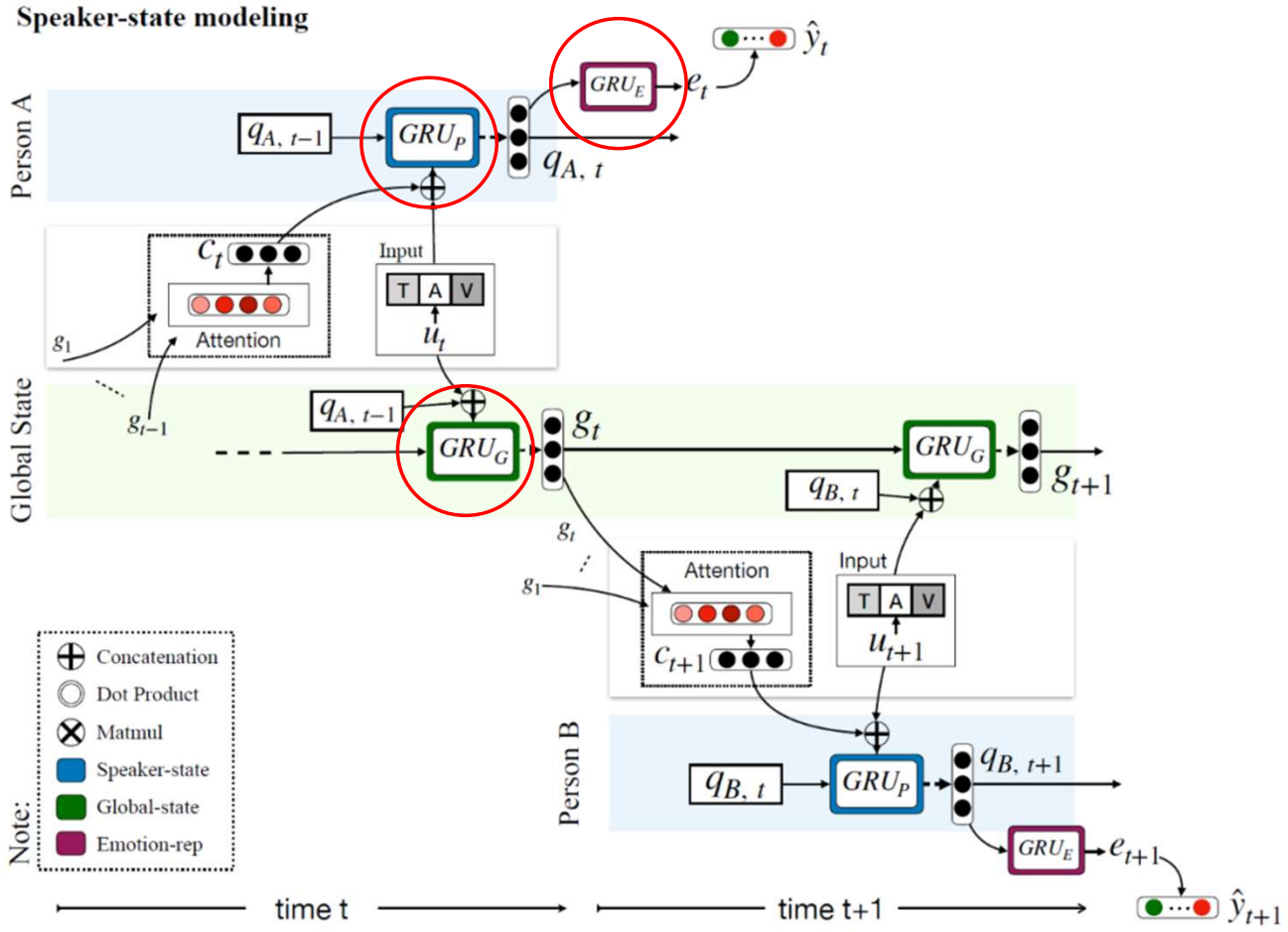
DialogueRNN

- 全局状态
- 说话人状态
- 情绪状态

Person A

Global State

Person B



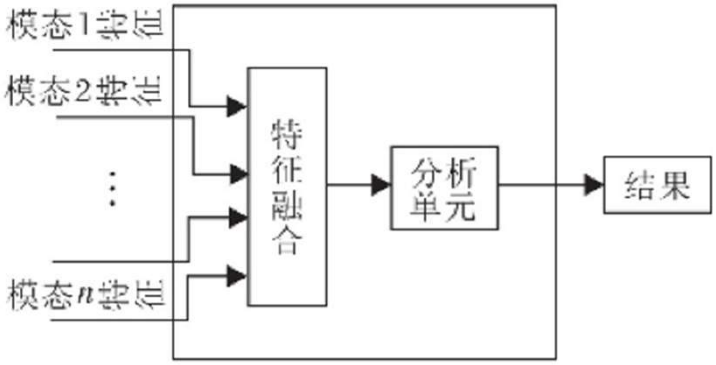
DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. Majumder et.al. AAAI 2019

多模态情感分析

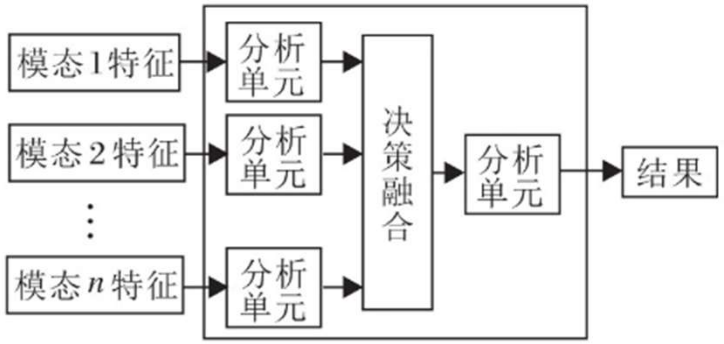
总统先生是对的

中联部发言人发推特反讽：总统先生是对的。有些人确实需要注射消毒剂，或至少用它漱口。这样他们在交谈时就不会传播病毒、谎言和仇恨。

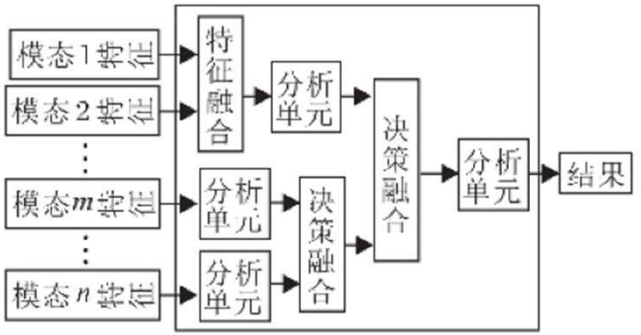
多模态情感分析



特征级融合



决策级融合



混合融合

多模态情感分析

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

1. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. Yu et.al. AAAI2021

本文认为有效的多模态表示应该同时考虑不同模态的一致性和差异性，但统一的多模态标签限制了差异性，因此提出**自监督的标签生成模块来获得单模态标签**，并通过多模态任务和单模态任务联合训练来学习模态一致性和差异性。

2. Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing. Zhang et.al. AAAI2021

本文研究**多标签场景**下的多模态情绪识别问题，不仅考虑了标签之间的依赖，还进一步考虑了特征与标签和模态与标签之间的依赖，提出了异构层次化消息传递网络HHMPN。

3. Humor Knowledge Enriched Transformer for Understanding Multimodal Humor. Hasan et.al. AAAI2021

本文研究**多模态幽默识别**问题，提出了幽默知识增强的Transformer模型，该模型整合了上下文和外部知识来捕获多模态幽默表达的要点

多模态情感分析

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

4. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. Wu et.al. ACL2021

以往的研究将多模态特征视为同等重要，然后隐式地对不同模态之间的交互进行建模。本文认为多模态情感分析中，文本模态占据主要地位，其他模态提供共享语义和私有语义两种信息。因此，提出了一种基于跨模态预测的**以文本为中心的共享私有框架。**

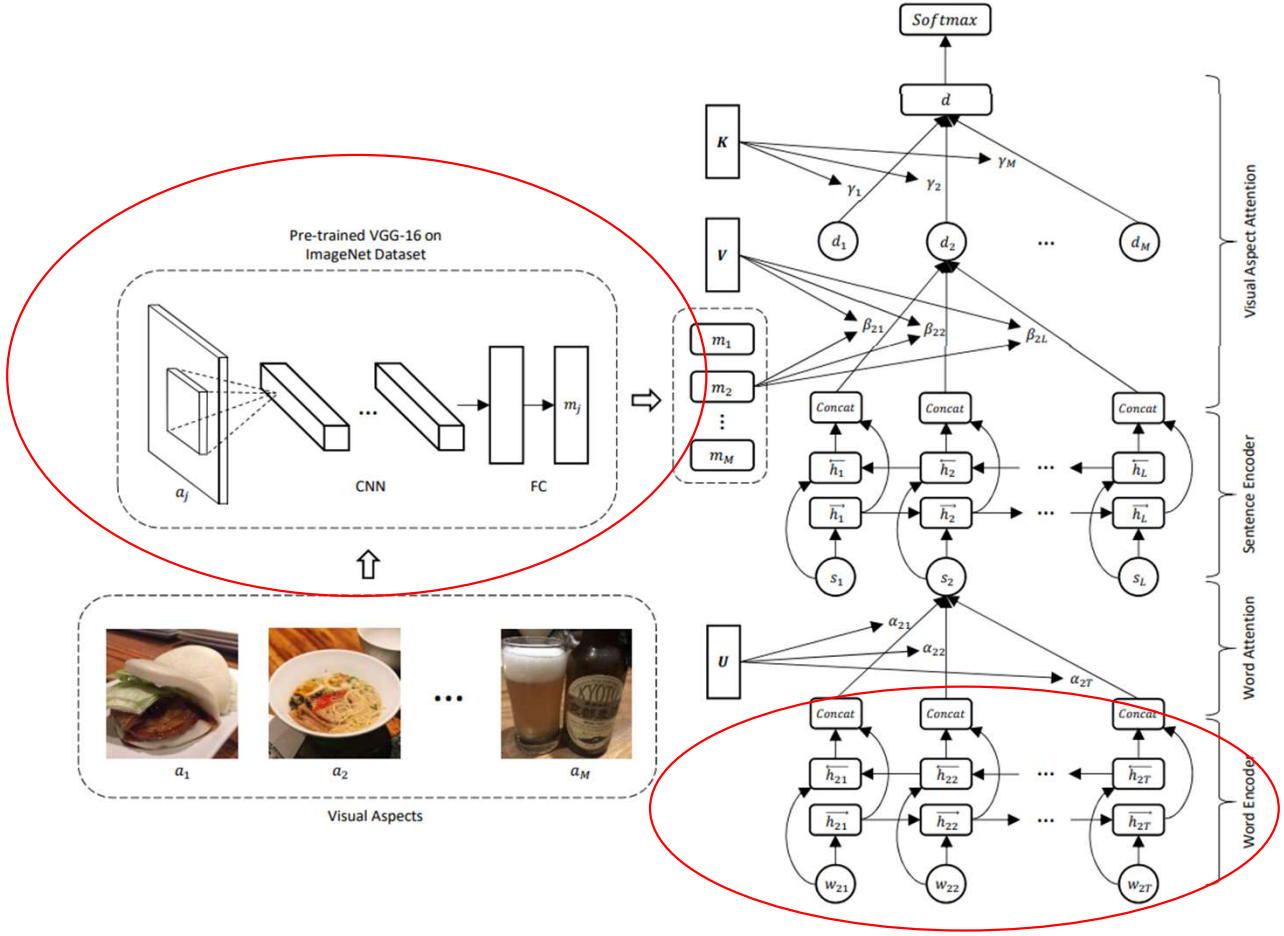
5. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotations of Modality. Yu et.al. ACL2020

本文给出一个**新的中文多模态情感分析数据集CH-SIMS**，并提出了一个基于晚期融合的多任务学习框架作为数据集的基线。

6. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. Cai et.al. ACL2019

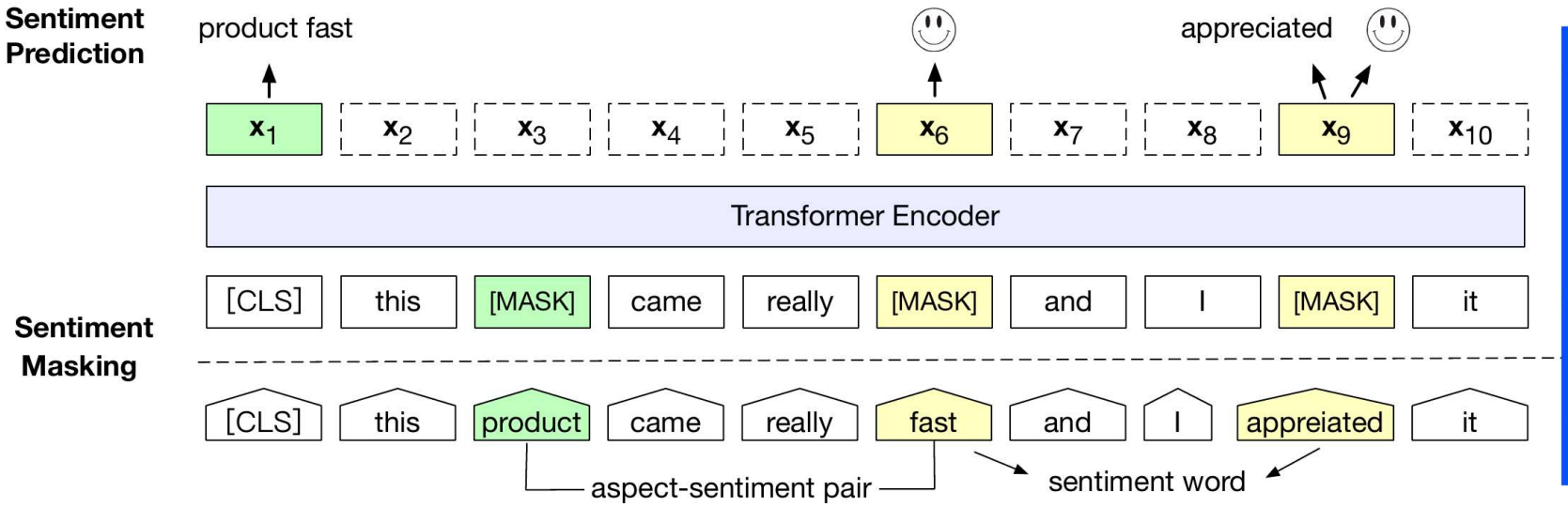
本文研究面向**图文的反讽识别**问题，提出了多层融合模型HFM，本文认为图像属性能够将图像和文本的内容联系起来，具有桥梁的作用，因此在文本和图像双模态的基础上，增加了图像的属性模态。

多模态情感分析



VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis. Truong and Lauw. AAAI 2019

情感知识增强的预训练



创新方法

基于无监督的情感知识挖掘
情感知识屏蔽
情感预训练目标

模型特点

无监督挖掘情感知识，简单高效
情感知识屏蔽&情感预训练：使得模型对情感知识有更好的感知

SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. Tian et al. ACL2020

情感知识增强的预训练

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

14个中英公开数据集刷新业界SOTA，在PaddleNLP中开源

任务	数据集合	语言	指标	原SOTA	SKEP
句子级情感分类	SST-2	英文	ACC	97.50	97.60
	Amazon-2	英文	ACC	97.37	97.61
	ChnSentiCorp	中文	ACC	95.80	96.50
	NLPCC2014-SC	中文	ACC	78.72	83.53
评价对象级的情感分类	Sem-L	英文	ACC	81.35	81.62
	Sem-R	英文	ACC	87.89	88.36
	AI-challenge	中文	F1	72.87	72.90
	SE-ABSA16_PHNS	中文	ACC	79.58	82.91
	SE-ABSA16_CAME	中文	ACC	87.11	90.06
观点抽取	MPQA-H	英文	b-F1/p-F1	83.67/77.12	86.32/81.11
	MPQA-T	英文	b-F1/p-F1	81.59/73.16	83.67/77.53
	COTE_BD	中文	F1	82.17	84.50
	COTE_MFW	中文	F1	86.18	87.90
	COTE_DP	中文	F1	84.33	86.30

主要实验结果

飞桨 PaddleNLP

pip v2.0.0rc22 | python 3.6 | 3.7 | 3.8 | downloads/month 7k | python 3.6+ | os linux, win, mac | license Apache-2.0

简介

PaddleNLP 2.0拥有覆盖多场景的模型库、简洁易用的全流程API与动静统一的高性能分布式训练能力，旨在为飞桨开发者提升文本领域建模效率，并提供基于PaddlePaddle 2.0的NLP领域最佳实践。

特性

- 覆盖多场景的模型库
  - PaddleNLP集成了RNN与Transformer等多种主流模型结构，涵盖从词向量、词法分析、命名实体识别、语义表示等NLP基础技术，到文本分类、文本匹配、文本生成、文本图学习、信息抽取等NLP核心技术。同时针对机器翻译、通用对话、阅读理解等系统应用提供相应核心组件与预训练模型。更多详细介绍请查看PaddleNLP应用示例。
- 简洁易用的全流程API

SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. Tian et al. ACL2020

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

Demo

Demo

BERT-WWM-Chinese

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

WWM（Whole Word Masking）

[Original Sentence] 使用语言模型来预测下一个词的probability。 **[Original Sentence with CWS]** 使用 语言 **模型** 来 **预测** 下 一个 词 的 **probability** 。
[Original BERT Input] 使 用 语 言 **[MASK]** 型 来 **[MASK]** 测 下 一 个 词 的 **pro [MASK] ##lity** 。 **[Whold Word Masking Input]** 使 用 语 言 **[MASK] [MASK]** 来 **[MASK] [MASK]** 下 一 个 词 的 **[MASK] [MASK] [MASK]** 。

Figure 1: Examples of whole word masking in BERT.

哈工大讯飞联合实验室

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, Guoping Hu. Pre-Training with Whole Word Masking for Chinese BERT

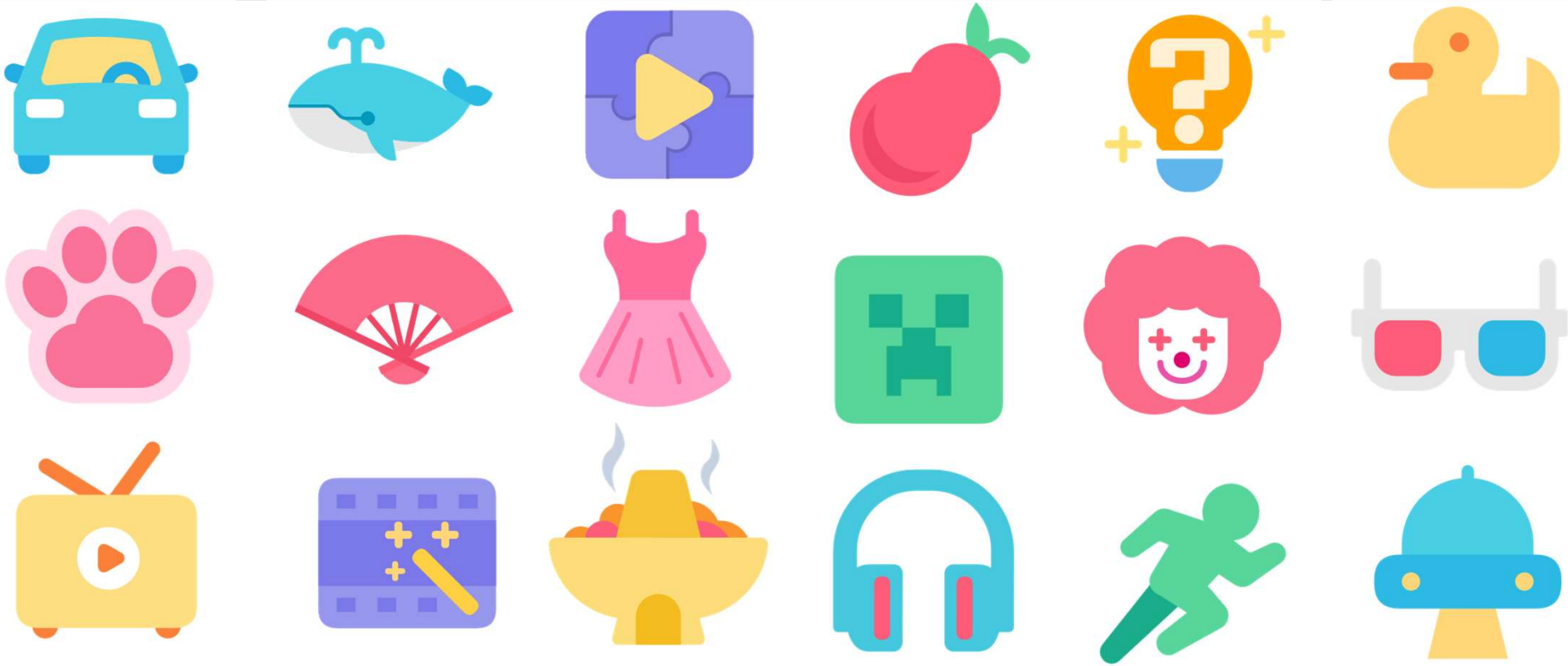
BERT-WWM-Chinese

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

模型	语料	Training Tokens #	Training Steps	Batch Size
BERT-WWM-Chinese	中文维基	0.4B	$100K^{MAX128}$ +$100K^{MAX512}$	2,560 / 384

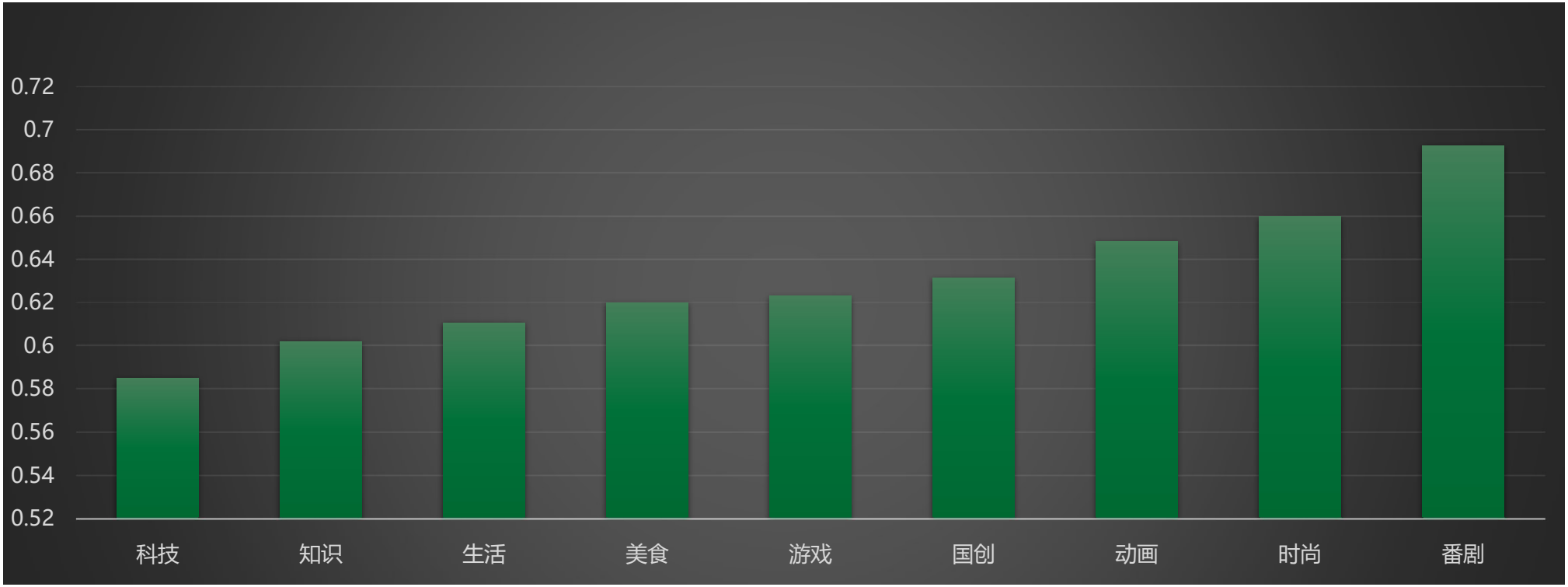
任务	开发集	测试集	评测指标	数据集
简体中文阅读理解	66.3 (65.0) / 85.6 (84.7)	70.5 (69.1) / 87.4 (86.7)	EM / F1	CMRC 2018
繁体中文阅读理解	84.3 (83.4) / 90.5 (90.2)	82.8 (81.8) / 89.7 (89.0)		DRCD
自然语言推断	79.0 (78.4)	78.2 (78.0)	Accuracy	XNLI
情感分析	95.1 (94.5)	95.4 (95.0)		ChnSentiCorp
篇章级文本分类	98.0 (97.6)	97.8 (97.6)		LCQMC
篇章级文本分类	98.0 (97.6)	97.8 (97.6)		THUCNews

哔哩哔哩



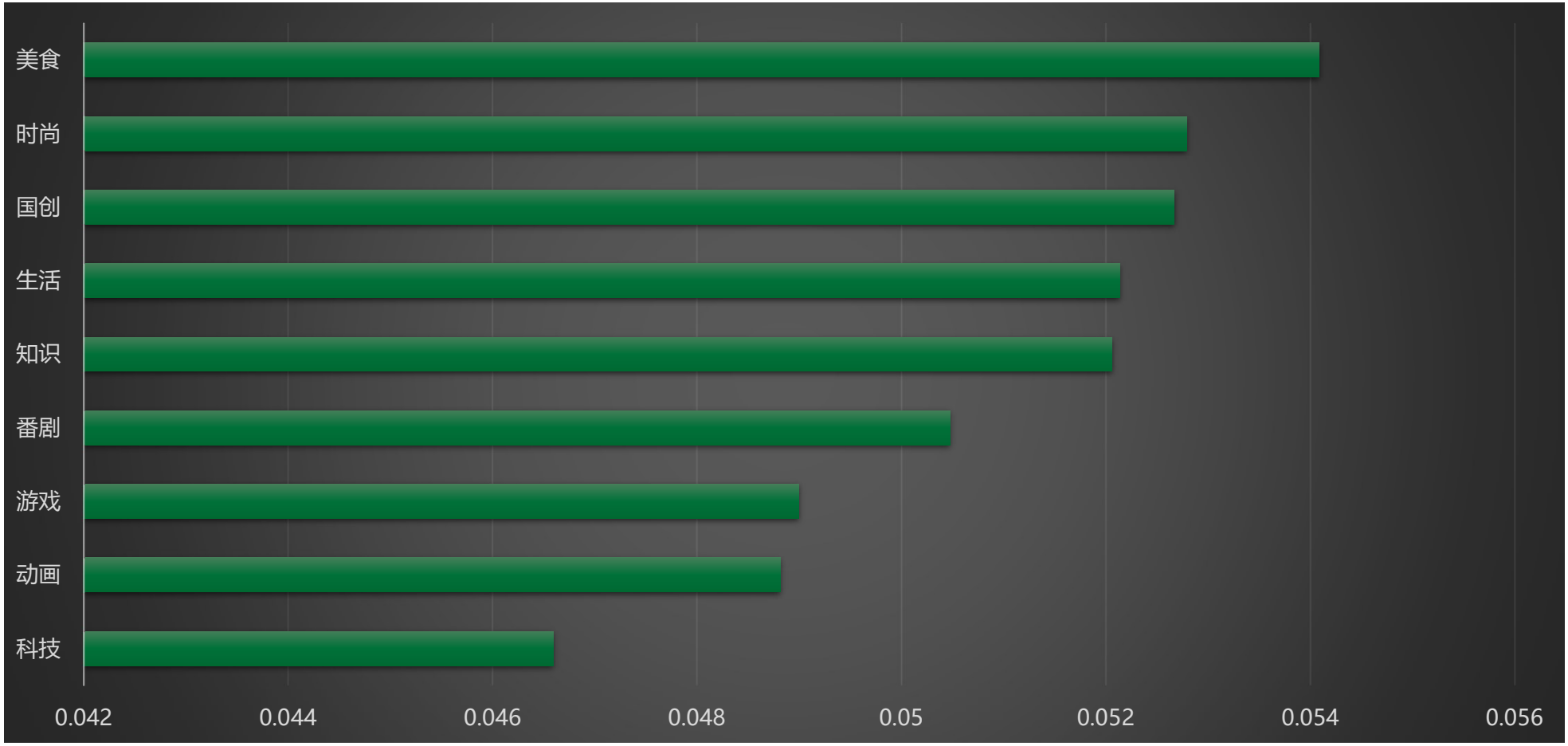
以视频内容为依据划分为二十余个大分区，近百个小分区

分区情感分析

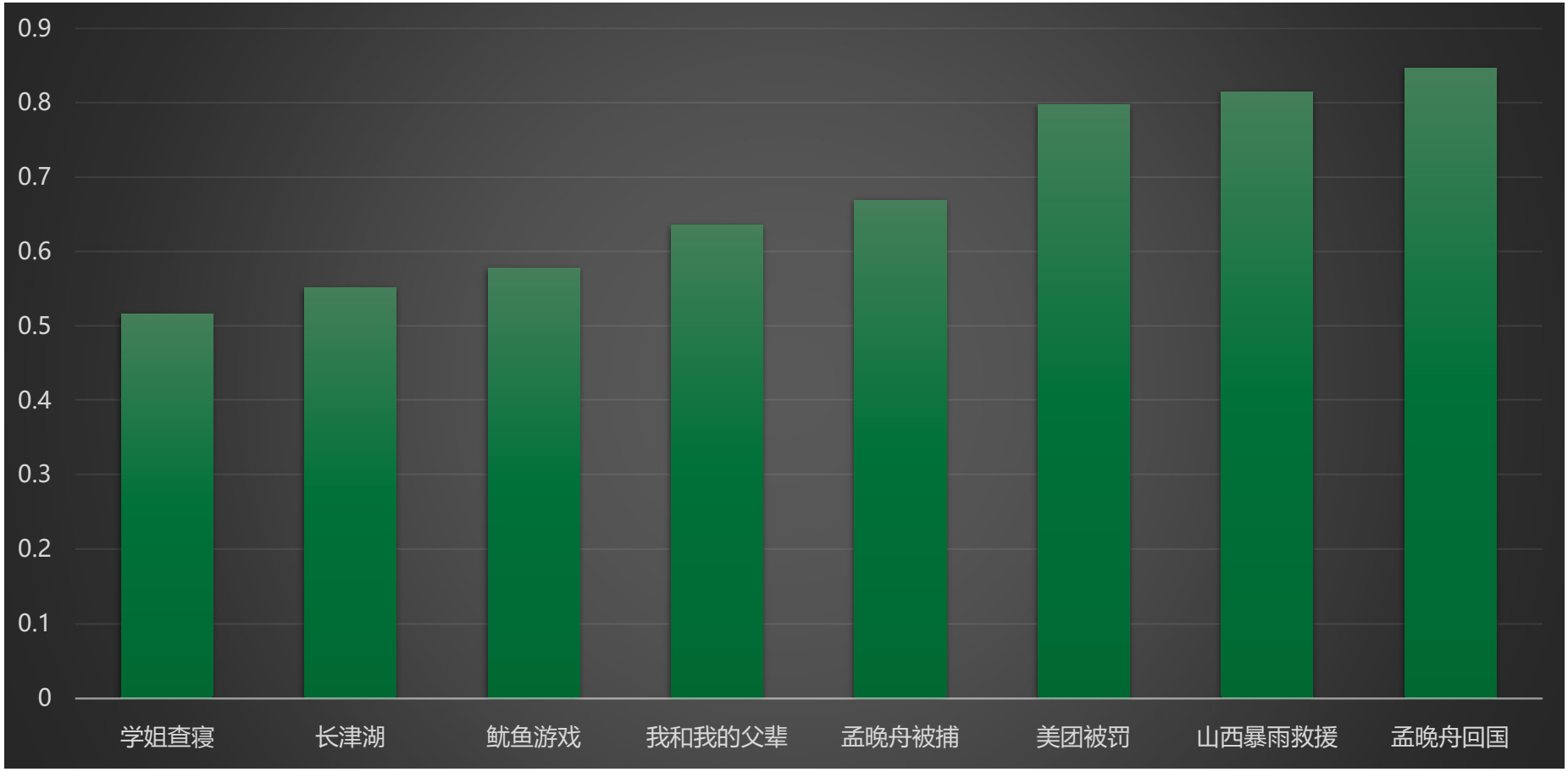


整体偏正向，科技区偏向客观中立，番剧区主观正面

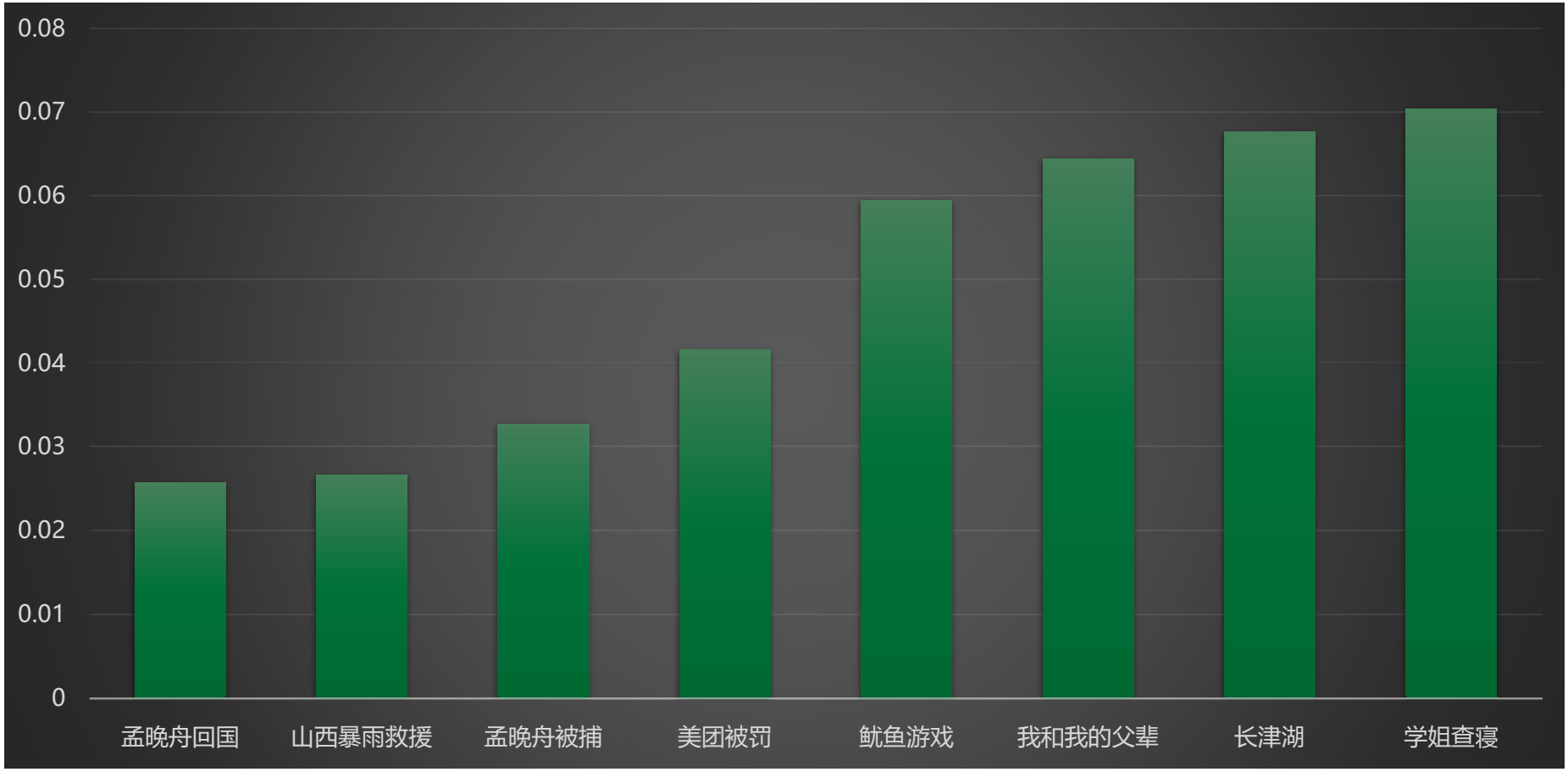
分区争议性分析



热点事件态度

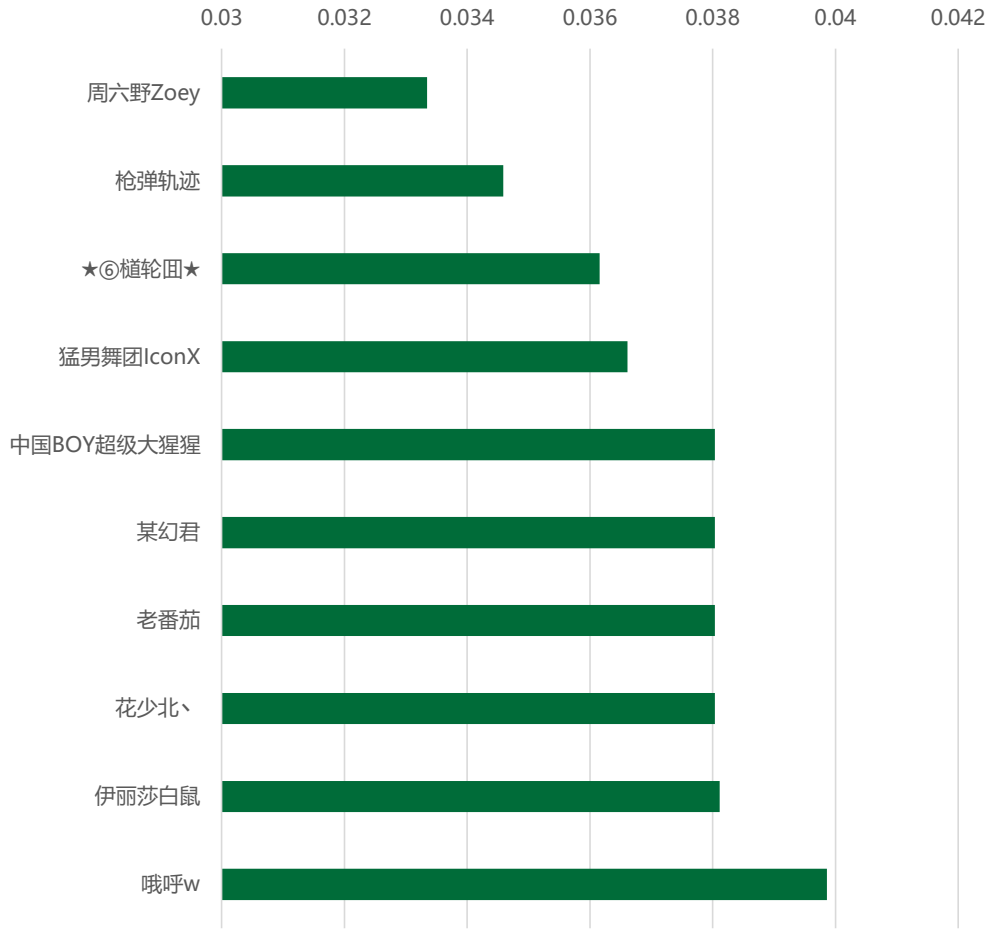

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
学姐查寝
长津湖
鱿鱼游戏
我和我的父辈
孟晚舟被捕
美团被罚
山西暴雨救援
孟晚舟回国

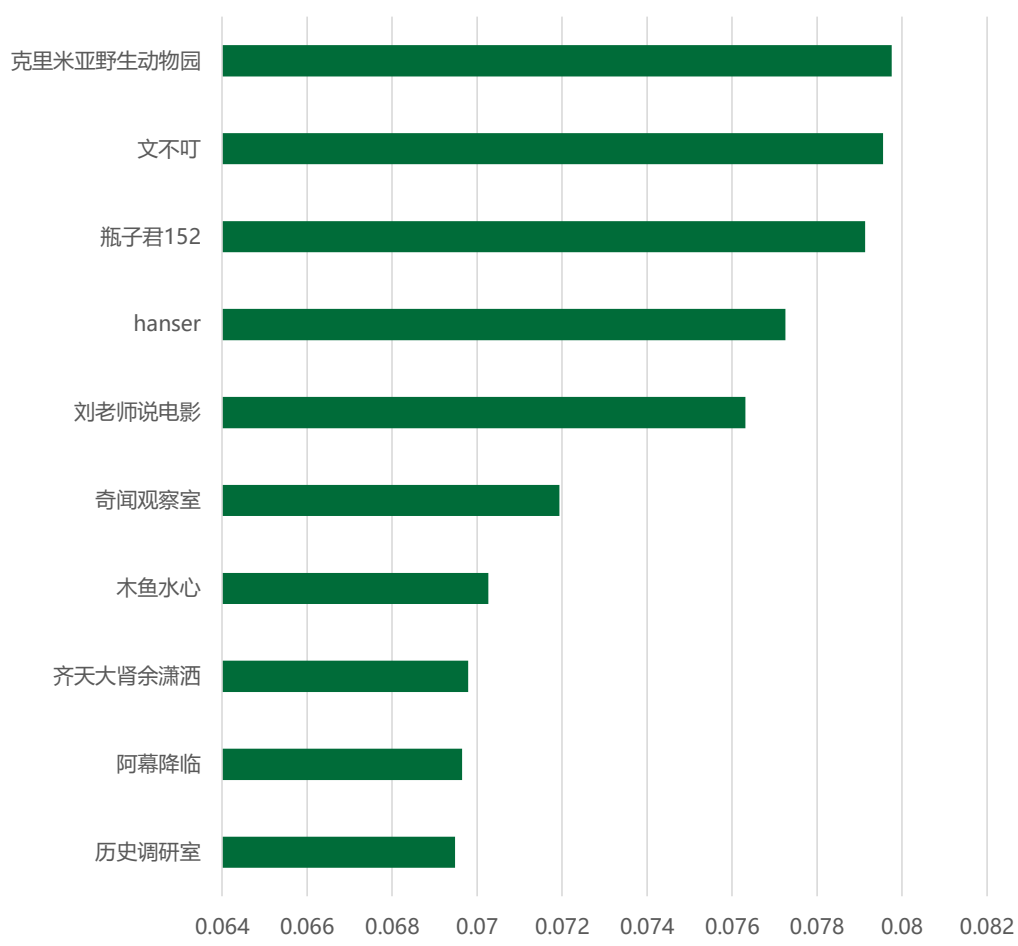
热点事件争议


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
孟晚舟回国
山西暴雨救援
孟晚舟被捕
美团被罚
鱿鱼游戏
我和我的父辈
长津湖
学姐查寝

2020百大up弹幕情感分析

2020百大up弹幕争议性





北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

总结与展望

Summary And Outlook

总结

1 **概述**

情感分析的基本信息。

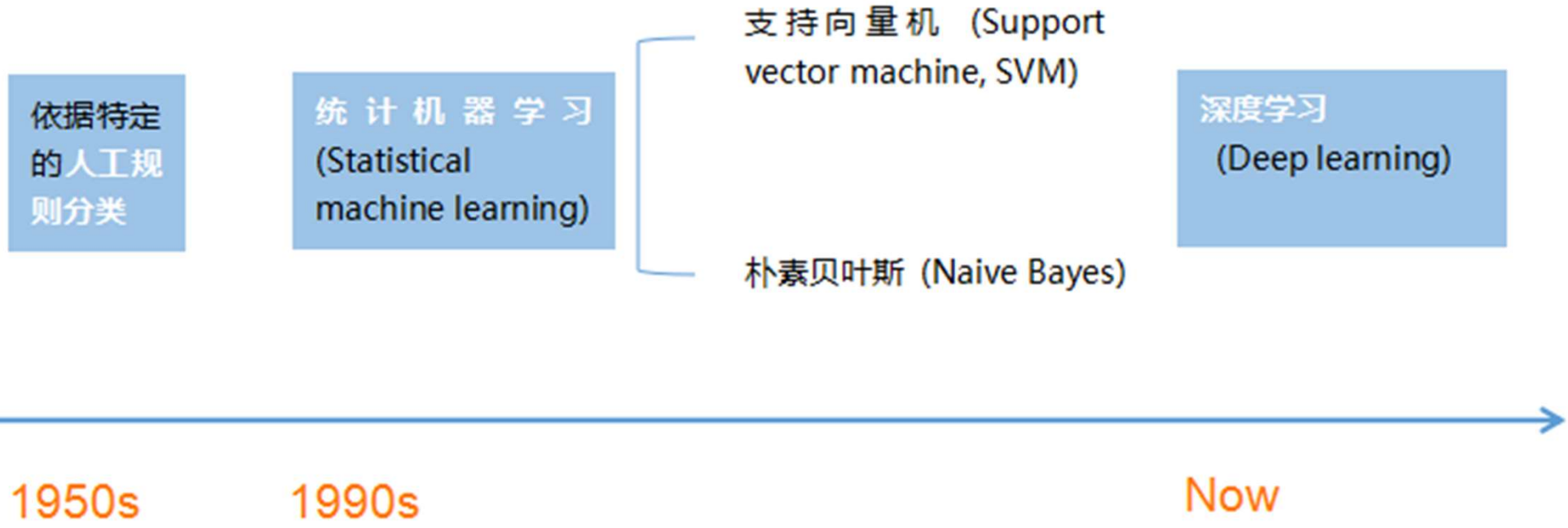
2 **方法**

情感分析研究方法发展史。

3 **应用**

情感分析的应用举例。

发展历程


依据特定的人工规则分类
统计机器学习 (Statistical machine learning)
支持向量机 (Support vector machine, SVM)
朴素贝叶斯 (Naive Bayes)
深度学习 (Deep learning)
1950s
1990s
Now

情感分析的更多可能

图片

视频

语音

广义上的情感分析还包括图片、视频、语音等多媒体信息进行分析判断，需要结合计算机视觉、语音识别等技术进行。

情感分析的未来

- 通用化
- 智能化
- 多元化
- 交互性

德以明理　　学以精工