# Text Classification System Based on Fields and Methods

**He Ying**

**Sept 27th, 2019**

- Target: to build a well-trained text classification system based on different fields and methods

- Methods for **short text** classifications:

  (1) LR

  (2) Naive Bayes

  (3) SVM

  (4) XGBoost

  (Using tf-idf values of words or dense word embeddings)

- Target: to build a well-trained text classification system based on different fields and methods

- Methods for **long text** classifications:

  (1) Word representation: word2vec, GloVe, ELMo, GPT, BERT

  (2) Model architecture:

  · Text-CNN (Convolutional Neural Networks for Sentence Classification, 2014)

  · RCNN (Recurrent Convolutional Neural Networks for Text Classification, 2015)

  · HAN (Hierarchical Attention Networks for Document Classification, 2016)

  · FastText (Bag of Tricks for Efficient Text Classification, 2017)

  · Transformer (Attention Is All You Need, 2017)

- Data:

(1) Chinese news data from toutiao (more than 3 millions with 16 fields)

(https://github.com/fate233/toutiao-text-classfication-dataset)

(2) Chinese news data from google (more than 2.5m in the json format)

(https://github.com/brightmart/nlp_chinese_corpus)

(3) Financial dataset

(https://github.com/smoothnlp/FinancialDatasets)

(4) Sentiment dataset

(https://github.com/z17176/Chinese_conversation_sentiment)

(5) Zhidao QA pairs crawled from Baidu Zhidao

(https://github.com/liuhuanyong/MiningZhiDaoQACorpus)

- **子任务**

  (1) 新闻文本分类

  (2) 语义相似性分类

  (3) 问答类文本分类

  (4) 情感分析

- **数据集**

(1) 今日头条新闻数据集（短文本：38w）

(2) 清华开源新闻数据集（长文本，74w，14类）

(3) 百度知道问答（长文本，580w问题数，平均每个问题有1.7个回答）

(4) 美版知乎问题匹配度数据集（长文本，40w，已翻译为中文）

(5) 情感分析

A. 新浪微博数据（长文本，2分类/4分类，10w/36w）

B. 大众点评餐馆评论数据（长文本，5分类，440w）

C. Amazon商品评论数据（长文本，5分类，720w）

- **词嵌入获取**

基于维基百科开源的中文语料（1.8G，90w文章）训练文本表征提取器

注：对于中文语料，需要一些特殊的处理，包括分词、文本提取信息补全、标点符号、繁简转换等

(1) word2vec (dim = 200)

(2) GloVe (dim = 200)

(3) BERT中文预训练模型（dim = 768）

(5) GPT中文预训练模型（dim = 512）

- **系统设计的性能目标**

  (1) 系统文本分类离线acc > 0.8，在线acc > 0.7 / 0.6

  (2) 系统返回结果的响应时间不超过5秒