



基于 LSTM 深度学习模型的中国电信官方微博用户情绪分析

蔡鑫, 娄京生

(中国电信股份有限公司上海研究院, 上海 200122)

摘 要: 中国电信官方微博(简称电信官微)作为一种互联网媒体,是中国电信对外发布信息并获取用户反馈的重要渠道。用户对电信官微消息内容的评论,反映了用户对中国电信品牌、产品和服务的不同态度。爬取了电信官微的消息内容和评论数据,对经过清洗的数据采用 Word2vec 进行文本信息表示,并使用深度学习平台基于 LSTM 深度神经网络模型进行用户交互文本的正负面情感分类,实现电信官微访问用户情绪分析。

关键词: 深度学习; LSTM; 情绪分析; 微博

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017340

Sentiment analysis of telecom official micro-blog users based on LSTM deep learning model

CAI Xin, LOU Jingsheng

Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

Abstract: As an internet media, China Telecom official micro-blog is an important channel for the company to publish information and get feedback from users. Users' comments on telecom official micro-blog messages reflect different attitudes towards telecom brand, products and services. The message content and comment data of the micro-blog was crawled, and the Word2vec was used to express the text information after data cleaning, and the deep learning platform was chosen to carry out the positive and negative emotional classification of the user interaction text based on the LSTM deep neural network model, and sentiment analysis of telecom official micro-blog users was realized.

Key words: deep learning, LSTM, sentiment analysis, micro-blog

1 引言

官方微博是企业或机构通过身份认证的具有法律效应的微博平台账号。对于用户众多的电信运营商而言,官方微博是进行产品/套餐展示、营销、客服以及客户沟通的平台,成本低、辐射广、传播快是这一互联网媒体渠道的重要特征。利用

官方微博及时发现用户的诉求,快速处理好用户关注的问题,就能够赢得用户称赞,促成舆论转向,树立品牌形象。

微博具有互动的特性,粉丝或其他微博用户都可以对中国电信官方微博(简称电信官微)发布的消息进行评论,而这些反馈内容反映了用户对电信品牌、产品和服务的主观态度。如果对用

户交互数据进行深入的分析,可以准确获取对中国电信而言非常有价值的用户意见和建议,对于改善企业的产品和服务大有裨益。

传统方式下,需要设立官方微博运营专员,对用户互动内容进行人工分析和处理,耗费大量人力成本。本文尝试采用基于 LSTM 深度学习模型的自然语言处理技术,对电信官微的访问用户交互信息进行用户情绪分析,辅助开展微博运营工作。

2 基于深度学习的自然语言处理关键技术

对于文本形式的微博评论数据,计算机处理过程中需要用到自然语言处理(natural language processing, NLP)的相关技术。自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法^[1]。简而言之, NLP 是让计算机能够理解人类语言的一种技术。

2.1 Word2vec 文本信息表示技术

自然语言领域最细粒度的是词语,词语组成句子,句子再组成段落、篇章、文档。所以要想实现让计算机处理自然语言,首先要解决的是词语的二进制信息表示问题。

词向量的意思就是通过一个数字组成的向量来表示一个词, Word2vec 是其中一种向量表示技术, Google 于 2013 年将其开源。 Word2vec 使用分布式表示(distributed representation)的词向量表示方式。其基本思想是通过训练将每个词映射成 K 维实数向量(K 为模型中的超参数),通过词与词之间的距离来判断它们之间的语义相似度。词向量来源于上下文的词预测模型, Mikolov 等人^[2]提出 CBOW (continuous bag-of-word) 和 Skip-gram (continuous skip-gram) 两个模型,分别使用三层神经网络对语言建模预测上下文单词出现的概率。 CBOW 是依据词所在的上下文中的词作为输入预测词本身;而 Skip-gram 则是依据词

本身作为输入预测上下文中的词。

Word2vec 模型在实现上下文预测的同时,获得了一项副产品,即单词在向量空间上的表示,这是 Word2vec 真正的最重要的贡献。 Word2vec 算法使得具有相似语义的词可以表示为相互靠近的向量, Mikolov 等人^[2]给出了一个著名等式“king-man+woman=queen”,形象地说明了词向量处理文本的魔力。

2.2 RNN/LSTM 深度神经网络技术

深度学习的概念来源于人工神经网络模型。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征,以发现数据的分布式特征表示。传统的神经网络一般由输入层、隐藏层、输出层组成,其中隐藏层的数目按需确定。深度神经网络是目前最常见的深度学习模型,而且包括多种不同类型的网络,例如: CNN(convolutional neural network, 卷积神经网络)、 DBN(deep belief network, 深度置信网络)等。

在 NLP 领域, RNN(recurrent neural network, 循环神经网络)因其对上下文的兼顾而显得更为适用。在自然语言处理中,上下文关系非常重要,句子里相邻词并不独立,不同的排列组合会有不同的意义,例如“微信比短信方便”“短信比微信方便”,只是变换一下次序,意义就会完全相反。 RNN 中当前时刻的输出,除与上一层的输出有关之外,还考虑了上一时刻隐藏层的输出的影响。 RNN 是非常强大的动态系统,但由于梯度消失/梯度爆炸的问题,传统 RNN 在实际中很难处理长期依赖。

为了解决这个问题,采用了特殊隐式单元的 LSTM(long short-term memory, 长短期记忆)网络被首先提出,其自然行为便是长期的保存输入。 LSTM 包含一种称作记忆细胞的特殊单元类似累加器和门控神经元:它在下一个时间步长将拥有一个权值并连接到自身,复制自身状态的真实值和累积的外部信号,但这种自联接是由另一个单



元学习并决定何时清除记忆内容的乘法门控制的^[3]。LSTM 网络随后被证明比传统的 RNN 更加有效，特别是在语音识别以及机器翻译应用中表现良好。简而言之，LSTM 也是一种 RNN，并且有效解决了传统 RNN 未解决的长期依赖问题，能够存储远距离上下文对当前时刻输出的影响。

3 基于 LSTM 的用户情绪分析实现过程

3.1 官方微博内容爬取

首先需要获取语料数据，电信官微数据属于公开数据集，可以通过官微网址进行获取。实际中一般采用网络爬虫软件执行内容爬取任务。

网络爬虫是获取网站公开数据的一种有效工具。网络采集器属于网络爬虫的一种，利用采集工具批量采集网页中的内容，其中包括论坛、贴吧、媒体网站等，从中摘取某些数据形成统一的本地数据库。网络采集器不需要技术成本，可被用作初级的数据采集工具，常见的有 Gooseeker、八爪鱼等。此外，针对特定网站的特定数据采集，例如本文中的官方微博用户评论数据，可基于 Scrapy 等开源框架自行开发爬虫程序，这样可以满足更加自主和灵活的采集需求。

爬取下来的微博用户交互数据，还需要进行必要的清洗和转换。主要包括：去重，即去掉同一消息内容下同一用户的重复评论；去噪，一些低质量的内容，如文本过短只包含一个汉字的评论可视为噪声数据并简单剔除；拼接，将同一消息内容下同一用户的多条评论按时序进行拼接，以形成更为完整的上下文语义。经过上述预处理过程的文本数据，便可以进入后续的 NLP 环节。

3.2 文本分词与词向量表示

中文以字为基本书写单位，而词是一个或多个字的组合。与西方语言不通，词和词之间没有明显的区分标记。因此中文自然语言处理，首先要进行中文分词。通俗地说，中文分词就是要由机器在中文文本中词与词之间加上标记。具体地，

输入“中国电信致力于做领先的综合智能信息服务运营商”；输出“中国电信/致力于/做/领先的/综合/智能/信息/服务/运营商”。

分词技术被归为以下 3 个主要类别：基于词典的分词技术；基于统计的分词技术；基于规则的分词技术。3 类分词技术有各自的优缺点，具体工作中，可采用较为成熟的结巴（Jieba）分词或 CRF++ 分词。这两个分词算法都不能简单地归属于某一大类的分词技术中，而是采用了交叉的技术策略。算法支持用户自定义词库，可以将具有行业特征的专有词添加到自定义词库，以便分词引擎能够识别和处理，如“运营商”“天翼”等。

经过分词处理的文本语料，便可送入 Word2vec 进行训练，以获得词向量表示，这一过程也成为词嵌入（word embedding）。在互联网上有一些基于海量语料（例如中文维基百科等）的预训练词向量，这些语料内的词基本可以覆盖本文所需的电信官微使用词汇。但是一方面由于词库过于庞大，势必要用更多维数去表示，词向量就会更加冗长和稀疏；另一方面，某些词在电信行业具有特指的含义，区别于普通文档，例如“终端”一词，在电信行业中与“手机”的含义就比一般意义上更为接近。因此，并不适合将通用的预训练词向量直接拿来使用，而应基于场景问题域本身的语料重新进行词向量训练。

除了 Google 自己的 Word2vec 工具，开源社区中还存在有对词向量感兴趣的第三方开发者所编写的不同版本，其中 Python gensim 主题模型包中的 Word2vec 对算法的封装比较完善。训练 Word2vec 模型主要需要设置词向量的维数，即上文所述的模型超参数 K 值。 K 的取值与语料内包含的不同词汇个数相关，可依据经验或由封装函数进行优化设置。

本文对高维数的词向量做了主成分分析（principal component analysis, PCA），选取部分词

以及与它们距离最接近的 Top10 的词,在权重最高的两个主成分构成的二维平面中进行了可视化展示,效果如图 1 所示。图 1 中结果印证了词向量距离以及词之间语义相似度二者的一致性关系。

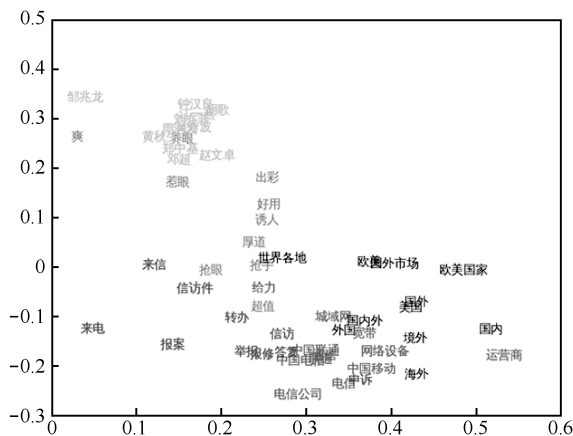


图1 Word2vec 词向量距离图示

3.3 深度学习模型构建

本场景的分析目标,是根据用户评论文本数据判别微博用户对电信的情绪偏好(包含:“正面”“负面”和“无关”),这在 NLP 基本问题归类中属于文本分类问题。即根据给定文本的内容,将其判定为事先确定的若干个文本类别中的某一类或者某几类的过程。

从机器学习的技术层面看,分类问题是比较典型的监督学习(supervised learning)。百度百科将监督学习定义为利用一组已知类别的样本调整分类器的参数,使其达到所要求性能的过程,也称为监督训练或有教师学习。监督学习是从标记的训练数据来推断一个功能的机器学习任务^[4]。对于监督学习而言,有高质量标注的学习样本是模型训练成功的重要前提。为此,本文从全部语料中随机地抽取 20% 的样本,分 3 组独立进行人工标注,再通过投票机制确定每个样本的分类标注,从而保证学习样本的标注质量。

本文采用 Word2vec 即 K 维词向量表示一个词,而句子可以看成多个词所组成的序列。句子

的长短即包含的词语个数可能不同,如果将词向量序列简单拼接作为句子的信息表示,那么句子的维数就会长短不一,无法作为后续深度神经网络的直接输入。考虑到微博评论的一般表述方式都以短句为主,因此设定一个固定长度 L (本文取值 34) 作为文本的限定长度。当评论文本长度(包含词数) $> L$ 时,对 L 之后的文本进行截断;当文本长度 $< L$ 时,通过类似缺失值处理的方式补齐,本文采用空字符填充以做到不失原文语义。经过上述处理,文本就可以表示成 L 个 K 维向量序列形式,输入深度学习网络。

网络的具体结构如下:文本经过词嵌入进行向量化之后,进入标准 LSTM 层进行上下文学习,之后 LSTM 的输出结果经过全连接(dense)层将维度降至目标变量类别个数 3,再迭加 softmax 层就可以得到输入文本在各个类别的概率,从而实现最终的分类目的。整个深度学习模型构建流程如图 2 所示。

4 Tensorflow 深度学习平台建模实证

4.1 验证平台环境搭建

伴随着人工智能的热潮,很多互联网公司和研究机构都积极参与研发深度学习框架,并将其开源以扩大影响力,因此目前市面上深度学习框架数目众多。其中,TensorFlow 由 Google 于 2015 年推出,基于架构灵活的计算图进行设计,依托 Google 强大的研发实力和业界影响力,Tensorflow 在社区活跃度、文档丰富性、生态完备性方面都体现了优势。

基于此,在本文实际验证中采用 TensorFlow 作为深度学习平台。同时为了简化编程开销,更专注于深度学习模型本身,还使用了 Keras 作为高层框架。Keras 是一个极简化、高度模块化的神经网络库,可充分利用 TensorFlow 通用计算能力。另外,为了更好地实现 Python 环境包和多版本管理,引入了 Anaconda 作为管理工具。

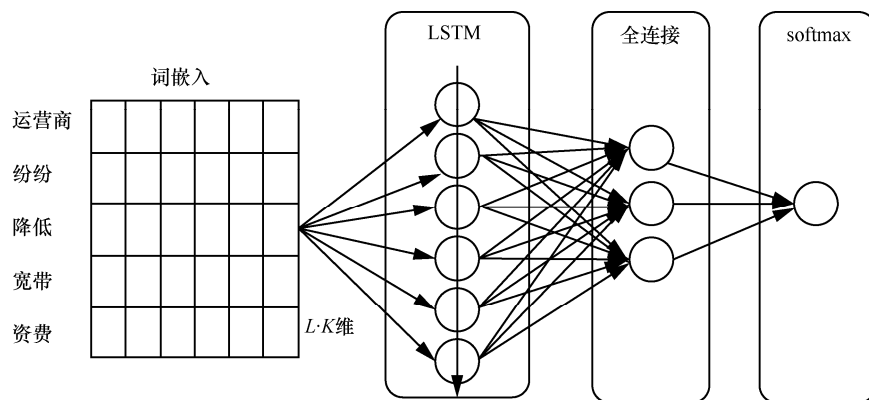


图2 深度学习模型建模示意

验证平台环境主要安装过程如下。

步骤1 Anaconda 安装配置

下载 Python 版本为 2.7 对应的 Anaconda:

https://repo.continuum.io/archive/Anaconda2-5.0.0.1-Linux-x86_64.sh

0.0.1-Linux-x86_64.sh

安装 Anaconda:

```
$ bash Anaconda2-5.0.0.1-Linux-x86_64.sh
```

步骤2 Tensorflow 安装配置

执行下面命令，准备环境:

```
$ conda create -n tensorflow python=2.7
```

激活上述名字为『tensorflow』的环境:

```
$ source activate tensorflow
```

安装最新版本 Tensorflow:

```
(tensorflow)$ pip install tensorflow
```

步骤3 Keras 安装配置

```
(tensorflow)$ pip install keras
```

4.2 深度学习网络训练

本文使用 Python 脚本完成深度学习网络构建。在数据预处理阶段，对于经过分类标注的训练文本，使用结巴分词进行中文分词。再通过字符截断和填充，形成长度限定为 L 的等长分词序列。之后，Python 调用 Keras 的 API 构建本文第 3.3 节所述的深度学习模型。

深度神经网络建模主要语句如下:

```
model = Sequential()
```

```
model.add(Embedding(len(dict)+1, 256, input_
```

```
length=34))
```

```
model.add(LSTM(128, dropout_W=0.2, dropout_U=0.2))
```

```
model.add(Dense(3,activation='relu'))
```

```
model.add(Activation('softmax'))
```

```
mod-
```

```
el.compile(loss="categorical_crossentropy",optimizer="rmsprop",metrics=['accuracy'])
```

```
model.fit(x, y, batch_size = 128, nb_epoch=20)
```

注: x_i 存储训练样本分词字典下标序列, y_i 存储训练样本分类标签, dict 为语料库词典。

训练过程可以输出经过每次迭代 (epoch) 之后的损失函数值 (loss) 以及模型准确率 (accuracy, 简称 acc), 通过观察 loss 和 acc 的不断变化, 可以判断模型是否收敛。本文的训练中经过 20 次迭代, 在训练集上 loss 下降到 0.062 2, acc 达到 0.977 1; 在验证集上也分别达到 0.288 4、0.921 9, 是较为理想的结果。loss 和 acc 的变化趋势如图 3 所示。

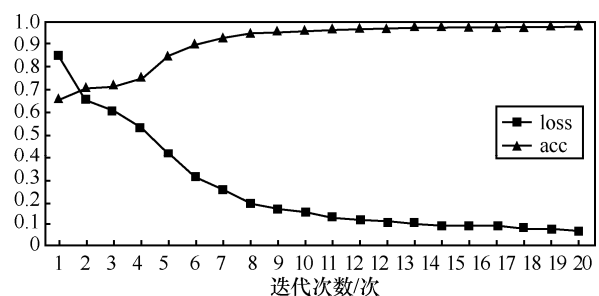


图3 深度学习网络训练迭代过程 loss 与 acc 变化趋势

由实际验证结果可知,结合词向量词嵌入与 LSTM 深度学习模型的文本分类方法,对于电信官微用户情绪分析效果十分有效。在有高质量训练样本标注的前提下,这一方法可以很容易地延伸到其他 NLP 文本分类应用场景中,如客服投诉文本分类、客服服务质量评价等。

5 结束语

传统的文本处理方法(例如 TF-IDF)基于词袋(bag-of-word)特征,基本原理还是词频统计。主要有两个缺陷,一是丢失了词的上下文顺序,二是忽略了词的语义差异。对于类似微博评论的短文本,语义信息本就不够丰富,信息丢失就会对结果产生更为明显的影响。

本文采用基于深度学习的 NLP 技术进行文本分类,工作基础是词向量(Word2vec)或称词嵌入技术,将有上下文语义信息的词向量的序列作为文本特征表示,从源头上基本消除了语义信息丢失的问题。词向量不仅可以用作文本分类,还可以被作为很多 NLP 相关工作的数据输入,如文本聚类、文本相似度分析、seq2seq 文本生成等。

另外笔者注意到,参考文献[5]中提到一种使用固定长度向量表征段落文本的方法,该方法是在 Word2vec 基础上的进一步延伸,也被称为 Sentence2vec,可以规避本文提到的文本截断和补齐过程,值得进一步研究和尝试。

近年来,基于深度学习的 NLP 技术进展迅速,在语音识别、自动问答和写作、机器翻译等场景

都有了突破性的成果。随着算法的进一步研究和改进,笔者相信基于深度学习技术一定能催生出更多相关的人工智能创新产品。

参考文献:

- [1] BIRD S, KLEIN E, LOPER E. Python 自然语言处理[M]. 陈涛, 张旭, 崔杨, 等译. 北京: 人民邮电出版社, 2014.
BIRD S, KLEIN E, LOPER E. Natural language processing with Python[M]. Translated by CHEN T, ZHANG X, CUI Y, et al. Beijing: Posts & Telecom Press, 2014.
- [2] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]// 1st International Conference on Learning Representations (ICLR 2013), Sep 3, 2013, Scottsdale, Arizona, USA. [S.l.:s.n.], 2013.
- [3] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(5): 436-444.
- [4] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
LI H. Statistical learning method[M]. Beijing: Tsinghua University Press, 2012.
- [5] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]//The 31st International Conference on International Conference on Machine Learning, June 21-26, 2014, Beijing, China. New York: ACM Press, 2014.

[作者简介]



蔡鑫(1975-),男,中国电信股份有限公司上海研究院高级工程师,主要研究方向为数据分析挖掘、人工智能、数据规划和信息管理。

娄京生(1975-),男,中国电信股份有限公司上海研究院助理工程师,主要研究方向为数据分析挖掘、人工智能。