

基于长短期记忆多维主题情感倾向性分析模型

滕 飞, 郑超美*, 李 文

(南昌大学 信息工程学院, 南昌 330031)

(* 通信作者电子邮箱 zhengchaomei@ncu.edu.cn)

摘 要: 针对中文微博全局性情感倾向分类的准确性不高的问题, 提出基于长短期记忆模型的多维主题模型 (MT-LSTM)。该模型是一个多层多维序列计算模型, 由多维长短期记忆 (LSTM) 细胞网络组成, 适用于处理向量、数组以及更高维度的数据。该模型首先将微博语句分为多个层次进行分析, 纵向以三维长短期记忆模型 (3D-LSTM) 处理词语及义群的情感倾向, 横向以多维长短期记忆模型 (MD-LSTM) 多次处理整条微博的情感倾向; 然后根据主题标签的高斯分布判断情感倾向; 最后将几次判断结果进行加权得到最终的分类结果。实验结果表明, 该算法平均查准率达 91%, 最高可达 96.5%; 中性微博查全率高达 50% 以上。与递归神经网络 (RNN) 模型相比, 该算法 F-测量值提升 40% 以上; 与无主题划分的方法相比, 细致的主题划分可将 F-测量值提升 11.9%。所提算法具有较好的综合性能, 能够有效提升中文微博情感倾向分析的准确性, 同时减少训练数据量, 降低匹配计算的复杂度。

关键词: 中文微博; 情感倾向分析; 长短期记忆; 多层多维模型; 主题标签

中图分类号: TP181 **文献标志码:** A

Multidimensional topic model for oriented sentiment analysis based on long short-term memory

TENG Fei, ZHENG Chaomei*, LI Wen

(College of Information and Engineering, Nanchang University, Nanchang Jiangxi 330031, China)

Abstract: Concerning the low accuracy of global Chinese microblog sentiment classification, a new model was introduced from the perspective of Multi-dimensional Topics based on Long Short-Term Memory (MT-LSTM). The proposed model was constituted by hierarchical multidimensional sequence computation, it was composed of Long Short-Term Memory (LSTM) cell network and suitable for processing vector, array and higher dimensional data. Firstly, microblog was divided into multiple levels for analysis. To upward spread, sentiment tendencies of words and phrases were analyzed by three-Dimensional Long Short-Term Memory (3D-LSTM); to rightward spread, sentiment tendencies of the whole microblog were analyzed by Multi-Dimensional Long Short-Term Memory (MD-LSTM). Secondly, sentiment tendencies were analyzed by Gaussian distribution in topic sign. Finally, the classification result was obtained by weighting above analyses. The experimental results show that the average precision of the proposed model reached 91%, up to 96.5%, and the recall of the neutral microblog reached 50%. In the comparison experiments with Recursive Neural Network (RNN) model, the F-measure of MT-LSTM was enhanced above 40%; compared with no topic division, the F-measure of MT-LSTM was enhanced by 11.9% because of meticulous topic division. The proposed model has good overall performance, it can effectively improve the accuracy of analyzing Chinese microblog sentiment tendencies and reduce the amount of training data and the complexity of matching calculation.

Key words: Chinese microblog; oriented sentiment analysis; Long Short-Term Memory (LSTM); hierarchical multidimensional model; topic sign

0 引言

随着网络新媒体的飞速发展, 大量用户已习惯于通过微博表达自己真实的想法和理念, 从而产生了庞大的数据量和很多创造性的自由、随性的表达方式。这些新鲜的方式不仅表达了微博作者的态度和想法, 还带有极高的商业、社会价值。为此, 分析这些大量且复杂的微博信息中的情感已成为当下研究热点之一。

与传统文本的情感分析不同, 微博有其独特的情感特征,

其不仅需要明白表面意思, 更需要分析字里行间的内在含义。这就需要从不同方面对微博信息的特征进行分析, 否则很难准确判断它的情感倾向, 更难以得出准确结果。其次, 微博具有篇章短小精悍、语言结构口语化、存在表情符号和创造性语言的特征, 增加了语言处理和分析的难度。

目前, 循环神经网络 (Recurrent Neural Network, RNN) 模型正应用于各种机器学习所涉及的任务中, 尤其适用于输入输出序列长度可变的环境中进行分类和生成任务; 然而在实际应用中, 由于长期目标依赖性导致训练难度极大。Socher

收稿日期: 2016-01-27; 修回日期: 2016-04-23。 基金项目: 江西省科技支撑计划项目 (20112BBE50045)。

作者简介: 滕飞 (1990—), 女, 天津人, 硕士研究生, 主要研究方向: 人工智能、数据分析; 郑超美 (1959—), 女, 江西抚州人, 教授, 主要研究方向: 人工智能、数据分析、计算机网络; 李文 (1980—2016), 女, 江西宜春人, 副教授, 博士研究生, 主要研究方向: 文本信息处理。

等^[1]使用张量形式的递归神经网络 (Recursive Neural Network, RsNN) 侧重于对整个句子的理解,但中文尤其是微博很少有完整的句子和完善的句法结构。Koutnik 等^[2]将循环神经网络的隐藏单元划分为组,采用不同频率时钟的发条循环神经网络 (Clockwork Recurrent Neural Network, CW-RNN) 模型跨时空链接信息;但不适用于正则文法表达,缺乏上下文的内在关联,使整条微博的识别性降低。近来相对有效的方法之一,是增加特殊控制单元来限制内存访问,即使用长短期记忆模型 (Long Short-Term Memory, LSTM) 来获得更持久的记忆,以及更轻松地捕获长期依赖项,减缓信息衰减的速率,增加深度计算的优势。Stollenga 等^[3]则是从线的角度出发进行扫描,代替了原先的点辐射的思想,提出金字塔型长短期记忆模型 (Pyramidal Multi-Dimensional LSTM, PMD-LSTM);但其打破了上下文的关联,且复杂度较高,影响分类效果。Li 等^[4]在 RNN 的基础上增加了自动编码模型形成了一种按等级划分的自动编码模型 HNA (Hierarchical Neural Autoencoder),是一种多维的 LSTM 模型;但其效率不高,每句话都要反复地进行编码和解码的工作。

针对以上问题,笔者根据中文微博的特性,提出了基于 LSTM 的多维主题模型 (Multidimensional Topic LSTM, MT-LSTM),以提高微博情感倾向预测的准确率。它不依赖于句子的标签和形式,通过分层的方式增强词与词之间的联系,以及义群与义群、句与句之间的联系。最后,通过主题分类判断情感倾向,再将每一层结果进行加权求和得到最终的情感倾向。由此,增强了句子的特征,解决了因时间迁移导致数据模糊而无法计算的问题,降低了因长期记忆影响导致遗忘速率过快而对结果产生的不利影响,增强了分类的准确性,且更适用于口语化的中文微博。

隐藏序列和记忆序列的计算与传统 RNN 不同,通过 Python 予以实现^[5]。本文通过输入序列得到标准 RNN 计算出的隐藏序列和记忆序列。由于目标类会与逻辑序列产生联系,所以这种表示不会产生逻辑衰退。实验表明,通过这种组合方式进行情感分析得到的结果准确率更高。

1 相关工作

1.1 循环神经网络

RNN 模型^[6]支持可变长度的输入,即句子长度与类型可以不同,在语言模型^[1]和机器翻译^[4]方面都展现了很强的优越性。所有的训练数据构成 RNN 的词汇表,每个词均随机初始化。作为模型的参数,采用非线性的预测模型为其赋权值。当前时间节点的隐藏层状态与当前时间 t 输入 x_t 和前一时间隐藏层状态 h_{t-1} 有关:

$$h_t = \varphi(Wx_t + Uh_{t-1}) \quad (1)$$

其中: W 是输入过程的权重矩阵; U 是状态转移的循环权重矩阵; φ 为逻辑 S 形函数或双曲正切函数,由此构成的函数,即式 (1),得到输出序列来预测下一时刻输入 x_{t+1} 的分布。

1.2 长短期记忆模型

LSTM 模型^[7]构建了专门的记忆存储单元,便于发现和建立输入值之间的长期依赖关系,适合上下文相关的语言学习。它可以解决 RNN 的一个重要问题:隐藏层在新的时间状态下将不断叠加输入序列而导致前面的信息模糊,因而无法继续向后传播。

LSTM 模型包括:输入序列 $X = \{x_1, x_2, \dots, x_n\}$,每个步长与其对应的输入,输入门 i_t ,遗忘门 f_t 和输出门 o_t ,记忆单元

c_t 控制着长期记忆单元的记忆与遗忘。第 j 个 LSTM 的单位时间 t 的记忆单元 c_t^j 为经过输入门 i_t^j 和遗忘门 f_t^j 调整的新内容 \tilde{c}_t^j 和早期的记忆内容 c_{t-1}^j 之和。依据 Zaremba 的版本^[8],给出公式如下:

$$c_t^j = f_t^j c_{t-1}^j + i_t^j \tilde{c}_t^j \quad (2)$$

其中:

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1}) \quad (3)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1}) \quad (4)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1}) \quad (5)$$

其中:输入门 i_t 和遗忘门 f_t 控制新内容的输入和旧内容的遗忘; σ 为对应元素相乘的逻辑 S 形函数或双曲正切函数。一旦记忆单元更新,隐藏层会根据当前输出门得到的结果计算当前隐藏层 h_t^j :

$$o_t = \sigma(W_o x_t + U_o h_{t-1}) \quad (6)$$

$$h_t^j = o_t^j \tanh(c_t^j) \quad (7)$$

上述控制门和记忆细胞允许 LSTM 单元自适应地忘记、记忆和展示记忆内容。遗忘门的开闭可以同时发生在不同的 LSTM 单元。基于 RNN 的多重 LSTM 单元可以同时捕捉在网络中快速和缓慢移动的数据。

2 构建模型

2.1 模型架构

与英文相比,中文的语法不够严谨,而微博语言的随意性更强,使得依据细致的语法分析进行句子的倾向性分析比较困难。为此,考虑放弃复杂的语法分析,而对句子的内部构造进行整合。目前的研究多是将整条微博当成一个句子进行处理,或仅处理微博中的一句话。为此,可以将整条微博视为一个整体,探讨其内在的逻辑和最终的情感倾向;再加上对微博主题倾向的逻辑划分,形成细粒度的微博情感模型。以一条微博为例,其情感分析的框架结构如图 1 所示。

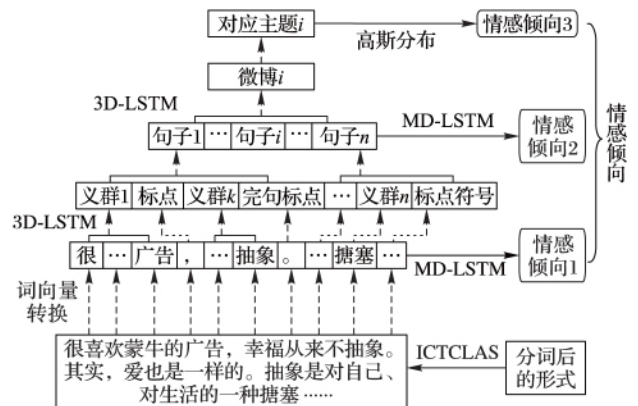


图1 主题情感倾向分析模型示意图

其中最底层的句子为预处理后的结果。由于计算时间会随着维度的增加呈指数级上升,为避免形成维数灾难,本文模型纵向传播采用三维长短期记忆模型 (Three-Dimensional Long Short-Term Memory, 3D-LSTM), 横向传播采用多维长短期记忆模型 (Multi-Dimensional Long Short-Term Memory, MD-LSTM)。

图1给出了处理某一条微博的详细过程,其详细内容如下:

- 1) 对语料库进行预处理,去掉无关部分。依照 ICTCLAS 分词系统将句子进行词语划分,并保留标点以及各种符号和符号集(多个符号组成的表情符号)。
- 2) 通过谷歌的 word2vec 工具进行词语向量化表示,并将

向量化的词语调整格式进行输入。

3) 随时间推移, 每条微博的处理方式:

a) 向上传播: 使用 3D-LSTM 模型, 在不同句子粒度上进行分析;

b) 向右传播: 使用 MD-LSTM 模型, 在不同句子层次上进行分析。

4) 当分析完一整条微博后, 根据原本的微博主题或人工分类的主题进行主题匹配, 并根据高斯分布确定情感倾向。

5) 对输出的所有情感倾向进行加权运算, 得到最终的情感倾向。

2.2 三维长短期记忆模型

LSTM 虽然通过增加部分长期记忆元素可以解决 RNN 中重要的序列依赖问题, 但在解决实际问题时, 无论短期还是长期的记忆和遗忘都应该得到相同的重视, 解决该问题的有效思路之一是缩短句子长度。为此, 考虑将长句子拆分成短句, 同时还可以减少反复记忆和遗忘的时间, 提高处理速度。基于上述思路, 考虑将微博语言数据扩展为 3 维进行处理, 更加有效地利用图形处理器 (Graphics Processing Unit, GPU) 的处理功能。

MT-LSTM 中涉及的隐藏层和记忆单元可抽象表示为图 2 和图 3。图 2 中不包括遗忘门的输入输出, 仅为一层中的一次输入及其输出; 图 3 为立体结构的整体模型示意图。以图 3 所使用的框架模型的第一层为例, 将得到的词向量按照句子的标点将其划分成多个分句, 以每个分句的长度作为向量空间的划分依据。第一层不考虑标点符号, 只以分句为单位进行输出, 则每一个分句都可以根据这种标准构成一个二维向量矩阵; 再加上 LSTM 中的时间坐标, 构成 3 维的长短期记忆模型。

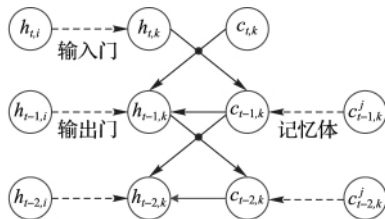


图 2 3D-LSTM 组成模型

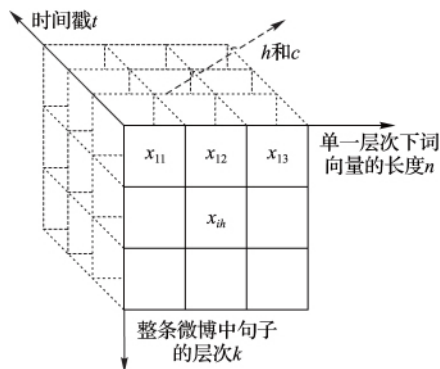


图 3 3D-LSTM 立体模型

与之类似, 当进入到以义群为单位的第二层时, 加入标点符号的成分, 以句子较长时间的停顿符号 (句号、分号或省略号等) 作为向量空间划分的依据, 即: 将原有句子中的逗号、顿号或加号等, 连同上一层的训练结果作为本层的输入矩阵。句子层则可以看作是普通的二维 LSTM 模型, 可以通过公式表达为:

$$c_{t,k}^j = f_{t,k}^j c_{t-1,k}^j + i_{t,k}^j \tilde{c}_{t,k}^j + f_{t,k}^j c_{t,k-1}^j + i_{t,k}^j \tilde{c}_{t,k-1}^j \quad (8)$$

$$\tilde{c}_{t,k} = \tanh(W_{c,k} x_{t,k} + U_{c,k} h_{t-1,k}) \quad (9)$$

$$i_{t,k} = \sigma(W_{i,k} x_{t,k} + U_{i,k} h_{t-1,k}) \quad (10)$$

$$f_{t,k} = \sigma(W_{f,k} x_{t,k} + U_{f,k} h_{t-1,k}) \quad (11)$$

$$o_{t,k} = \sigma(W_{o,k} x_{t,k} + U_{o,k} h_{t-1,k}) \quad (12)$$

$$h_{t,k}^j = o_{t,k}^j \tanh(c_{t,k}^j) \quad (13)$$

其中 k 表示整条微博被划分后句子的层次, 即微博中存在的义群数或句子数。

2.3 多维长短期记忆模型

多维长短期记忆模型 (MD-LSTM) 不需要对整篇微博进行细致的划分。它是一个相对独立的模型, 可以将总体模型中的某一层作为输入, 直接得到该层对应的情感倾向。

MD-LSTM 与 3D-LSTM 的区别是, MD-LSTM 将一整条微博视为一个整体, 更侧重对全局的考虑, 故这里的多维度仅针对隐藏层进行设置。根据前面的经验公式, 可以知道记忆单元的维度也会随着隐藏层维度的增加而增长, 也就意味着记忆周期更长。为抵消短周期记忆的缺失, 定义每个义群或句子中的标准输入为:

$$x_{j,norm} = f_1(x_j) = \frac{\frac{1}{n} \sum_{k=1}^n |x_k|}{|x_j|} \cdot x_j \quad (14)$$

同样, 以图 3 中第一次使用 MD-LSTM 模型的层为例。其中: x_j 为某一义群中的第 j 个词向量 n 为该义群中全部向量的长度。中间第二次使用 MD-LSTM 模型时, x_j 为某一句子中得第 j 个义群。最后一次由于不存在群体概念, x_j 为其本身, 此时 MD-LSTM 模型即为标准 LSTM 模型。

由式 (14) 可见, 输入队列中每个词或义群与其周围词或义群相互之间的关系更加紧密, 可以减轻 MD-LSTM 短期记忆的负担。

同时, 为简化计算, 将隐藏层表示为以下形式:

$$h = \begin{bmatrix} I^* f_1(x) \\ h \end{bmatrix} \quad (15)$$

其中 I 为转移矩阵。

3 实例分析

3.1 实验数据

实验数据来自新浪网提供的应用程序编程接口 (Application Programming Interface, API) 根据主题获取微博, 共随机获取到 20 个不同主题的微博, 其中正向主题 10 个 (如“刚出生的双胞胎手牵手”), 负向主题 10 个 (如“每天一剂负能量”)。去除转发和无文字内容的微博, 每类主题约有 3000 条微博进入预处理。同时由 10 人进行人工分类, 每条微博的情感倾向均由 3 人评价打分的结果确定。最终得到的情感分类结果如表 1 所示, 该结果作为实验中 MT-LSTM 与比较模型训练及检验的标准。

表 1 数据集的情感分配

情感倾向	数量	比例/%
正向	30391	49.01
负向	28428	45.84
中性	3195	5.15
总计	62014	100.00

3.2 预处理

由于微博更倾向于口语表达, 存在较多噪声, 因而需要预处理, 其主要工作是对微博信息进行清洗, 为此参考文献 [9] 并根据最新微博版本进行调整, 去掉微博中不存在情感的噪

声数据,包括:话题、标题、回复、统一资源定位器(Uniform Resource Locator, URL)、来源等。

此外,还需将表情符号改为文字,以便后续处理。其中,表情符号为微博官方表情符号库,转为文字形式时使用符号库中表情对应文字;不存在于官方表情库中的表情,如“(: 3 | ∠) _”则以原格式保留,作为标点符号处理。

3.3 使用词向量表示词语

使用词向量可以使模型变得更加客观,目标词向量不依赖 RNN 的权重。Turney 等^[10]使用词向量作为特征进行有监督的训练和测试,但词袋(bag-of-words)模型^[11]已经不能准确地捕获词语的含义。为获取情感倾向性分析的基本依据,国内往往将整条微博进行拆分,仅保留已知的情感词作为整条微博情感倾向判断的依据。实际上,中文表达十分丰富,很多名词或网络用语也存在主观情感,如用“小凯”或“凯凯”作为对凯迪拉克轿车的称呼,表现了使用者喜爱的情绪,也是积极情感倾向的一种。

笔者使用中国科学院计算技术研究所开发的 ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) 分词系统^[12]对已经预处理的文档进行分词;使用谷歌的 word2vec 工具^[13]对完成分词的文档进行词向量转换工作;使用词向量表示词语。由此,摆脱了传统方法的束缚,更适用于微博这种灵活的语言形式,可以更全面地反映句子中存在的情感倾向。

3.4 微博主题分类

在微博使用过程中,用户可以根据提示添加已有主题或自己添加主题。实验发现,一般情况下,同一微博话题下的情感倾向呈高斯分布。大多数带有主题的微博,其情感一般都趋近于相关主题,当主题情感为正向时,极少出现负向情感,反之亦然。

此外,虽然存在情感倾向的微博数量比例较高,而仅仅表达中性或无明确意义的微博相对存在数量较少,但在大数据分析中也不能忽视。笔者采集多类不同主题的语料进行分析,发现详细的主题划分有助于微博情感倾向的判断。

为验证主题分类的有效性,使用不同方法将主题分为不同数量的类别,如表 2 所示。

表 2 主题分类方式

分类数量	分类方式
1	—
2	主题情感倾向(正、负)
5	主题类别(经济、娱乐等)
20	主题名称

3.5 训练模型

对于划分层次后离散的文本数据,如果该条微博输入向量 x_i 共有 k 个层次,则第 k 层在时间 t 进行推送,其输入值为 1,其余均为 0;每个 x_i 对应输出一个预测值 y_i 。因此 $\Pr(x_{t+1} | y_t)$ 为多项分布,可以采用 softmax 函数进行参数化,计算每一层的概率,然后以期望值作为预测标准。

$$\Pr(x_{t+1} | y_t)_k = \frac{\exp(x_{t+1}(h_{i,k}))}{\sum_{i \in [1, \mu]} \exp(x_i(h_{i,k}))} \quad (16)$$

$$\Pr(x_{t+1} | y_t) = \frac{1}{K} \sum_{k=1}^K \Pr(x_{t+1} | y_t)_k \quad (17)$$

其中: $x_i(h_{i,k})$ 为输入序列第 k 层的激活函数; $x_{t+1}(h_{i,k})$ 为与上一隐藏层中第 k 层相关的当前输入,输入某一目标层中第 k

层的输入。对于一个固定的 k 来说,每组权值都是独立的,所以当去掉下标 k ,一个输入与其对应隐藏层的更新与其他更新的过程一样。

同样,在计算序列损失函数时,仍以负对数的形式来训练网络。

$$L(x) = - \sum_{i=1}^T \text{lb} \Pr(x_{t+1} | y_t) \quad (18)$$

由此可以方便地计算反向传播^[14],并使用梯度下降训练网络。

根据文献[4]设置训练中用到的参数,具体细节如下:

- 1) 统一初始化 3D-LSTM 和 MD-LSTM 中的参数,参数值设置区间为 $[-0.08, 0.08]$;
- 2) 随机梯度下降使用固定学习速率 0.1,训练了接近 7 个周期;
- 3) 最低批处理文件数为 20;
- 4) 漏码率为 0.2;
- 5) 当梯度规模超过临界值 5,进行梯度裁剪;
- 6) 模型框架中每层的权重为 $[0.3, 0.4, 0.3]$ 。

模型训练过程中使用单独 GPU (Tesla K40m, 1 Kepler GK110B),处理速度约为每秒 600 ~ 1 200 条微博。

3.6 结果分析

为保证分析的客观性,选取目前公认较先进的四种模型与 MT-LSTM 进行比对,分析比对的主要性能评估指标为查准率和查全率。查准率定义为正确判别为该类的测试样本占判别为该类测试样本的比例,而查全率定义为正确判别为该类的测试样本占该类总测试样本的比例^[15]。然而,这两个指标往往相互矛盾,为此一般采用 F-测量值作为综合评估标准,其定义如下:

$$F = 2PR / (P + R) \quad (19)$$

其中: P 为查准率, R 为查全率。

由表 3 数据可以看出,MT-LSTM 可以较准确地查出微博的情感倾向,同时可准确全面识别 50% 以上的中性微博。

表 3 MT-LSTM 在 20 组主题 10% 训练数据时的分类性能

情感	查准率	查全率	F-测量值
正向	0.943	0.427	0.595
负向	0.965	0.412	0.577
中性	0.822	0.501	0.623
平均	0.910	0.446	0.599

观察表 4 可以发现:通过第 3 层判断,即增加主题分类,可以有效提高微博情感倾向的准确率;适当增加第 2 层的权重,可以提升模型整体的查全率,对提升总体的 F-测量值起到至关重要的作用。

表 4 每个层次情感倾向判断的正确率对比

层次	查准率	查全率	F-测量值
第 1 层	0.728	0.351	0.474
第 2 层	0.892	0.426	0.577
第 3 层	0.951	0.417	0.580
总体	0.916	0.433	0.588

实验中,分别采用 10% 和 1% 的训练数据占比(从实验数据中随机取样的训练数据比例)进行训练,并采用 10 折交叉验证技术,得到的 F-测量值的结果如图 4 所示。由图 4 可见,与四种先进模型相比,当训练数据占比为 10% 时,通过 MT-LSTM 进行情感分析得到的 F-测量值与表现最好的 HNA 不相上下;

当占比减少到 1% 时, MT-LSTM 的 F-测量则比其他模型至少提高了 40.2%。值得注意的是, 占比越小, 意味着所需要的训练数据越少, 还可以有效降低计算复杂度。当训练数据减少时, 其他模型的 F-测量值都相对较低且结果大致相同, 应该是因为它们仅仅纵向使用模型, 而未考虑到层次间的联系。

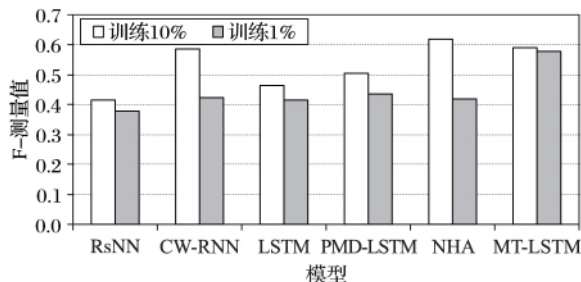


图 4 训练数据占比不同时的 F-测量值

由图 5 可知, 主题的细致划分有助于提高分类的准确性。当主题数量达到 20 时, 与无主题分类 (即主题数量为 1 时) 相比, F-测量值提高了 11.9%。

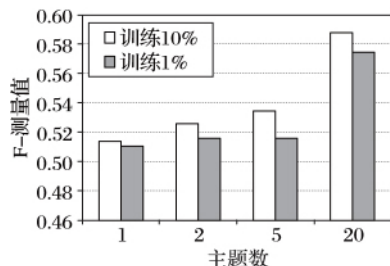


图 5 不同数量主题下 MT-LSTM 的 F-测量值

上述分析表明, MT-LSTM 可以较准确地划定情感倾向, 同时对中性微博有较强的分辨能力; 而且当训练集数据较少时, 结果依然令人满意; 同时主题数量对 F-测量值有较大影响。因此, 与目前的几种先进模型相比, 在对中文微博的情感倾向性进行分析时, MT-LSTM 具有更好的综合性能。

4 结语

本文在传统 LSTM 模型基础上提出了一个多层多维主题情感分析模型。与原序列模型相比, MT-LSTM 模型对每条微博进行逐层分析, 在增加词与词相关性的基础上, 增加了义群与句子和句子与句子的逻辑结构; 其次, 在保留了句子的一致性和完整性的同时, 增加了对主题的考虑, 可以更真实地反映用户对热点事件的态度; 第三, 可以自动学习中文口语表述, 在多个层次上对整条中文微博的情感倾向进行判断, 提高了中文微博情感分类的准确性。值得指出的是, 此模型还可以应用到更广泛的领域, 如翻译和文字识别等。

虽然 MT-LSTM 模型可以根据上下文较准确地推断微博的情感倾向, 但网络词语和较少出现的古代文体对准确率造成一定影响。在今后的工作中, 希望构建一个不需要分词的神经网络模型, 处理上下文关联较弱的文本内容。

参考文献

- [1] SOCHER R, PERELYGIN A, WU J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]// EMNLP 2013: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2013: 1631–1642.
- [2] KOUTNIK J, GREFF K, GOMEZ F, et al. A clockwork RNN [C]// ICML 2014: Proceedings of the 31st International Conference on Machine Learning. [S.l.]: International Machine Learning

Society, 2014: 1863–1871.

- [3] STOLLENGA M F, BYEON W, LIWICKI M, et al. Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation [C]// NIPS 2015: Proceedings of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015.
- [4] LI J, LUONG M T, JURAFSKY D. A hierarchical neural autoencoder for paragraphs and documents [EB/OL]. [2015-11-09]. <http://arxiv.org/pdf/1506.01057v2.pdf>.
- [5] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C]// NIPS 2014: Proceedings of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 3104–3112.
- [6] GRAVES A. Generating sequences with recurrent neural networks [EB/OL]. [2015-08-24]. <http://arxiv.org/pdf/1308.0850v5.pdf>.
- [7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735–1780.
- [8] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization [EB/OL]. [2015-08-24]. <http://arxiv.org/pdf/1409.2329v5.pdf>.
- [9] 袁丁, 周延泉, 鲁鹏, 等. 多方法融合的微博情感分析 [C]// 第六届中文倾向性分析评测报告. 昆明: 中国中文信息学会信息检索专业委员会, 2014: 35–39. (YUAN D, ZHOU Y Q, LU P, et al. Sentiment analysis of microblog combining multi-methods [C]// Proceedings of the sixth Chinese Orientation Analysis Evaluation Report. Kunming: China Computer Federation and Chinese Information Processing Society of China, 2014: 35–39.)
- [10] TURNEY P D, PANTEL P. From frequency to meaning: Vector space models of semantics [J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141–188.
- [11] PANG B, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques [C]// EMNLP 02: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2002: 79–86.
- [12] 张华平. NLP 汉语分词系统 [CP/OL]. [2014-12-11]. <http://ictclas.nlpir.org/>. (ZHANG H P. Chinese lexical analysis system [CP/OL]. [2014-12-11]. <http://ictclas.nlpir.org/>.)
- [13] Google. word2vec [CP/OL]. [2015-03-25]. <http://word2vec.googlecode.com/svn/trunk/>.
- [14] WILLIAMS R J, ZIPSER D. Gradient-based learning algorithms for recurrent networks and their computational complexity [M]// Back-propagation: Theory, Architectures and Applications. Hillsdale, NJ: L. Erlbaum Associates Inc., 1995: 433–486.
- [15] 张启蕊, 董守斌, 张凌. 文本分类的性能评估指标 [J]. 广西师范大学学报 (自然科学版), 2007, 25(2): 119–122. (ZHANG Q R, DONG S B, ZHANG L. Performance evaluation in text classification [J]. Journal of Guangxi Normal University (Natural Science Edition), 2007, 25(2): 119–122.)

Background

This work is partially supported by Science and Technology Plan Project of Jiangxi Province (20112BBE50045).

TENG Fei, born in 1990, M. S. candidate. Her research interests include artificial intelligence, data analysis.

ZHENG Chaomei, born in 1959, professor. Her research interests include artificial intelligence, data analysis, computer network.

LI Wen, born in 1980, Ph. D. candidate, associate professor. Her research interests include text information processing.