2017年8月8日周报

1. **参考文献**

《基于长短记忆多维主题情感倾向性分析模型》

作者：滕飞、郑超美、李文、南昌大学信息工程学院

关键词：中文微博、情感倾向分析、长短记忆（LSTM）、多层多维模型、主题标签

1. **收获**
2. **微博特征**

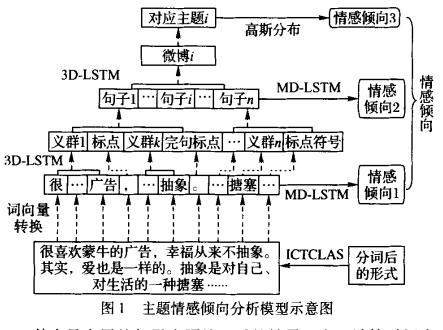
具有独特情感特征，既需明白表面意思，又需分析内在含义，篇幅短小精悍，语言结构口语化，存在表情符号和创造性语言特征，因此需从各个方面对微博信息进行分析

1. **RNN的缺陷**
2. 长期目标依赖性导致训练难度大；（2）RNN侧重于对整个句子的理解，微

博很少有完整的句子或完善的语法结构；（3）CW—RNN（CW-RNN用一组几何级数频率的时钟去驱动神经网路，如频率为f, 2f, 4f, 8f,）不适用于正则法表达，缺乏上下文内在关联，使微博识别性降低；

1. **使用LSTM原因**
2. 轻松捕获是长期依赖项，减缓信息衰缓速率，增加深度计算优势；（2）多维主题情感分析，提高微博情感倾向分辨率；（3）增强分类准确性，不依赖句子标签和形式，采用分层方式增强词与词，义群与义群，句与句的联系，再将主题进行分类；（4）解决因时间迁移导致数据模糊而无法计算问题；

4.**微博情感分析过程**



1. 对语料库进行预处理，对微博信息进行清洗，去掉微博中不存在的情感噪声数据（话题，标题，回复，统一资源定位器，来源），表情符号转为文字，采用ICTCLAS分词系统将句子进行词语划分，保留标点和各种符号，表情符号
2. 神经网络输入都为向量，采用worf2vec工具将词语进行词向量化和分层
3. 随着时间的推移，每条微博的处理方式：
4. 向上传播采用3D-LSTM模型进行不同句子粒度（数据仓库的数据单位中保存数据的细化或综合程度的级别，是不是指词向量或者内存的细化程度？）分析得到情感倾向1；

* 3D-LSTM模型：使用3D-LSTM模型，将长句拆分成短句，减少反复记忆和遗忘的时间，提高处理速度

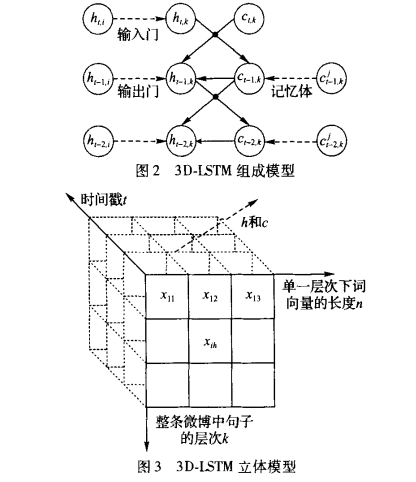
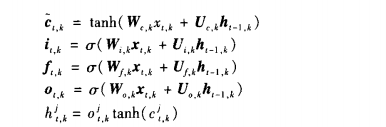
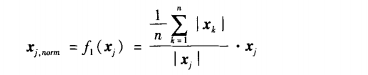


图2，图3包含隐藏层和记忆单元。图2每一层仅含一层中的一次输入及输出；图3将词向量按句子标点将其分为很多个分句X11,X12,X13……，分句的长度作为向量空间的划分依据，图中每一小格代表图2的hi或ci ，第一层的情感倾向1中只要遇见符号就进行分句，且不包含标点符号，第二层情感倾向2中包含符号，以句号或分号等较长停顿号作为划分依据 ，相关公式，其中k表示义群数或句子数



1. 向右传播使用MD-LSTM模型进行不同句子层次上的分析得到情感倾向2。

* MD-LSTM模型：多维长短记忆模型，将总体模型某一层作为输入，直接得到该层对应的情感倾向，相关公式

第一层情感倾向1的Xj表示某一义群中第j个词向量，n为该义群中全部向量的长度，第二层情感倾向2的Xj表示第j个义群，

1. 分析完一条整微博后，将原本的微博主题进行主题匹配，根据高斯分布（即正态分布）确定情感倾向3
2. 对所有输出的情感倾向1，2，3……进行加权运算，得到最终情感倾向