|  |  |  |  |
| --- | --- | --- | --- |
| **发明名称** | **基于迁移学习的细粒度情感倾向分析** | | |
| **所属项目** | 互联网广告 | | |
| **所在部门** | 精准广告平台中心 | **提交日期** |  |
| **提交人（撰写人）** | 陈飞宇 | **提交人联系电话** | 18712718250 |
| **发明人** | 陈飞宇 | **提交人电子邮件** | chinachenfeiyu@outlook.com |

**专利技术方案交底书模板**

**技术交底书填写注意事项：**

**1. 整体要求**

专利交底书是技术方案的实现说明，主要写明发明目的，实现方案，带来的技术效果，进一步可以通过流程图等方式说明技术的架构和实现方法步骤，不能仅有原理或功能介绍。

**2. 具体要求**

专利的技术方案写的尽可能详细，以便于后续评估和申报，并提高授权可能。按照发明所处的开发状态以及与业界现有技术的区别大小来给予不同程度的揭示，对于不涉及具体的代码或者公式，只需写明技术实现流程即可，具体如下：

（1）已上线的技术需将技术实现细节进行介绍，对技术诀窍可以采取一定的手段隐藏；

（2）目前正在研发中的业界领先技术，重点介绍技术实现方法，或者提供技术实现的开发文档即可；

（3）仅有创意的业界尚无的技术，可以只介绍技术实现思路，以及可应用的场景。

**3. 对现有技术的初步检索要求：**

由于专利授权的前提是申请时不存在跟发明内容相同或者近似的国内外现有技术，因此，需要提交人撰写技术方案时，初步检索跟本发明内容相关的国内外现有技术（包括：已发表的论文、期刊，已经公开的专利、已经公开使用的介绍等），以及业界类似产品，分析跟现有技术的区别。

如果该发明评估通过，后续会由外部律师进行补充的现有技术检索，进一步评估专利授权的可能性，如发现相同或者类似的现有技术，我们也无法找出区别方案的，则终止申请。

专利检索资源请见附件1。

**以下是交底书内容部分：**

**发明名称：** **基于迁移学习的细粒度情感倾分析**

**一、 发明目的**

**1.1解决的技术问题以及应用该发明的产品或项目：**

**目前在汽车领域，用户针对汽车评价信息数据非常庞大，同时也在不断增长。如果不对这些评价信息加以处理，各种各样的评论语句在一起就会显得非常庞杂。而对其进行有效地分类，不仅能够帮助消费者参考车辆各方面信息，避免盲目消费，还能帮助制造商根据用户反馈改进车辆性能，对于经销商而言，还能做到趋利避害，因此，这项工作非常重要。**

**1.2 现有技术的实现方案：**

**目前，汽车领域的情感分析主要实现方案是在发布评论的时候，首先强制用户去针对汽车的各个方面进行评价，从而达到分类的目的。然后，在上诉分类结果的前提下，通过机器学习的方式进一步分析用户具体评论的情感倾向，如是积极态度还是消极态度。当前对的主要方式包括基于规则（无监督），基于机器学习（有监督）的方式还有基于深度学习的，当前汽车领域的企业，如汽车之家、爱卡、易车网等，都是通过这几类方法。**

**1.3 现有技术的缺点：**

**虽然上诉方案得到了不错的效果，也得到了广泛使用，但是仍然存在一些缺点。首先，强制用户分类的方式本身就阻碍了一大批“嫌麻烦”的用户，因而损失了大量数据。其次，假设是用户自身分类错误，自然得到的结果也是错的。而且，分类粒度较大，不够细致，基本都是用户分成什么样就是什么样。**

**而且，基于机器学习的方式实现起来比较耗时耗力，需要大量的标注数据和复杂特征提取过程，对于深度学习的方式，虽然不需要特征提取，一般只需要词向量即可，但还是需要大量的标注数据训练模型，而标注数据的获取本身就是一个极大的难题。**

**二、发明内容**

**2.1 本发明提供的完整技术实现方案**

**本专利提出一种基于迁移学习的情感倾向分析方案，直接对评论语句进行细粒度的情感倾向分析，从而得到用户对车辆各方面信息的评价倾向。在模型训练过程中，由于利用了开放领域的先验知识，因此不需要太多的领域标签数据，使得不仅效果得到显著提升，而且训练时间也大大缩短。**

**细粒度情感倾向分析一般分为两个过程：评价对象抽取和情感倾向分析，针对整个过程，本专利的具体内容包括以下四点：**

* **共享信息的表示：**

**包含三个部分，其一是开放领域命名实体识别共享信息的表示，其二是开放领域情感倾向分析共享信息的表示,其三是汽车领域新闻信息的表示。**

**基于迁移学习的评价对象的抽取**

**采用命名实体识别和词性标注的方法，结合开放域命名实体识别共享信息来识别评价对象，利用到了第一步中汽车领域新闻信息的知识。**

* **基于迁移学习的情感倾向分析**

**将开放领域情感倾向分析语料与汽车领域评论语料结合，建立基于深度学习的短文本分类器模型，实现情感分析，同样利用到了第一步中汽车领域新闻信息的知识。**

* **长评论语句的划分**

**专利训练模型的语料中每个评论语句只有一个评价对象，模型的训练只需得出整句的情感倾向，但是实际的评论语句中一般包含较多的评价对象，要想得到各个评价对象的评价情感词，一个较为容易的方式是将长句划分为子句，子句中仅包含一个评价对象，这样就变成对子句的粗粒度情感分析，更为方便。**

* **评价对象和情感分析的关联**

**在划分长短句的时候需要注意每个评价对象词和情感词的关系，因为长句形式一般比较复杂，不能把所有情感词都计算到某一个对象上，而要分开，通过依存分析找到真实的情感词加以计算。**

* 1. **附图说明**

**基于迁移学习的细粒度情感分析过程包含如下七个过程：**

1. **语料收集：包括开放领域语料（非汽车相关），汽车领域语料，汽车领域新闻语料以及情感词典，其中开放领域语料包含情感和命名体识别语料，汽车领域语料较少，包含汽车领域命名体识别和情感分析标注，汽车新闻语料用来引入汽车行业背景信息。**
2. **训练词向量模型：采用word2vec工具，在开放域语料（去除标签）和汽车领域语料的混合语料上训练词向量模型。**
3. **获取开放领域实体标注语料和汽车领域新闻语料的共享信息：采用BiLSTM构建AutoEncoder模型网络结构，用上述两种语料分别训练两个模型LM1和LM2，并将AutoEncoder的模型参数分别保存下来，其基本结构如下:**



**本图是以BiLSTM为基本结构的AutoEncoder模型图,通过将X在BiLSTM中编码，在解码，最终希望得到的X’和X尽可能相同。**

**即: h(x) = BiLSTM(W,X)**

**X’=BiLSTM(W’,h(x))**

**X’≈X**

**符号说明：**

**X:输入句子对应的词向量序列**

**xi:X句子中每个词的词向量**

**xi’：经过解码后编码向量**

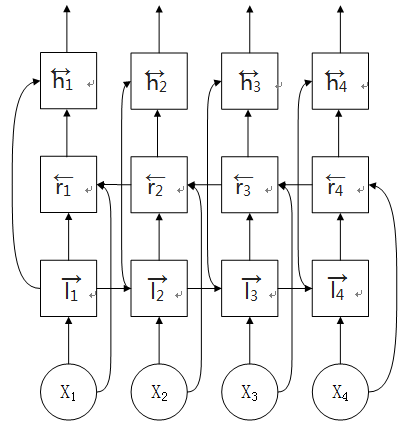
**X’：经过解码后的句子向量**

**将汽车领域实体标注语料S输入LM1中的BiLSTM模型，取BiLSTM正向最后一个向量和反向最后一个向量，并组合为：**

**LM1(s)= <,> 记为**

**LM2(s)= <,> 记为**

**BiLSTM结构如下:**



**符号说明：**

**S:汽车领域实体识别句子对应的词向量序列**

**xi：组成S的每个词的词向量**

**：表示包含单词i和其上文信息的向量**

**: 表示包含单词i和其下文信息的向量**

**:表示和的组合**

**评价对象抽取： 将汽车领域实体标注语料S输入BiSLTM-CRF模型中开始训练，在BISLTM输出向量的时候，取=max(,…),max是取每个向量各个维度最大的数值，组合<,,>，输入CRF，并在训练过程中优化参数,冻结和。**

1. **情感倾向分析：在3)中将开放领域实体识别语料改为开放领域情感分析语料,将BiLSTM-CRF改为BiLSTM-SoftMax，重复3)到5)的步骤。**
2. **评价对象与情感分析的关联：基于开源的依存关系和角色标注模型，将抽取的评价对象和情感倾向分析结果关联起来。**

**训练流程图如下：**

**训练流程（实体识别）**



**符号说明:**

**Si：上诉三种语料原句子的词向量序列**

**Lm1和Lm2：分别使用Open Source Ner语料和汽车领域新闻语料训练得到AutoEncoder得到的模型**

**：分别使用Lm1和Lm2,输入汽车领域NER语料得到的向量  
 :训练Bilstm最终每个向量各个维度最大的值得向量**

**情感分析的流程只需将上诉步骤中Open Source NER语料改为Open Source 情感分析语料，然后将虚线框中CRF改为SoftMax即可。**

**测试流程**



**符号说明:**

**Model1：评价对象抽取训练得到的模型**

**Model2：情感倾向分析训练得到的模型**

**三、业界相关产品及现有技术检索：**

**3.1 与该技术相关的竞争对手或相关产品：**

**（请列出竞争对手的名称、相关产品的名称，对于相关产品也可以采用截图等方式）**

**3.2 相关现有技术，重点是已公开专利及科技论文**

**（如专利/论文/标准，可以用附件、链接、文献名称及出处等方式提供）**

* **请列出使用的中英文关键词（或组合）：**
* **跟本发明相关的现有专利或者论文名称及链接：**

* **跟上述专利或者论文内容的核心区别点：**