# 一种基于卷积神经网络的企业舆情情感分类技术

## 技术领域

本发明是使用深度学习技术对文本数据进行企业舆情篇章级的情感方向分类，具体是基于卷积神经网络建立情感分类模型的一种方法。

## 技术背景（传统情感分析、提出卷积、结合的优点）

情感分析（Sentiment analysis）是人们对于带有情感色彩的文本类数据进行主观性的分析，形成对其的观点、情绪并评估对企业、组织、个人等实体的态度。情感分析是自然语言处理（NLP）中最活跃的研究领域之一。究其原因，随着各行各业数字化程度的不断提高，网络数据呈现爆发式增长，每天网络都会产生大量的数据，例如商品的评论信息、企业的舆情数据、微博、微信等，它在一定程度反映的人们的对某些事件的情感倾向。因此，无论是对于企业还是政府部门，可以通过情感分析或者是舆情系统来更加高效、快速地帮助政府监控群众的情感变化或者是舆论趋势，来避免发生恶性事件或者是虚假事件的发生。

传统的企业舆情数据情感分类主要是基于情感词典和基于机器学习。基于情感词典的技术是通过对人脑简单的模拟，核心是基于词典和规则，文本的情感倾向是通过情感词典来作为主要判断的方法，根据以往的经验对现有词汇做出评价的一种模型，如：通常把开心、快乐、幸福等一类的词汇作为积极词汇，把迟缓、低迷、落后等作为消极词汇。这种技术的机制也很简单，通过句子中所出现的情感词来预测句子的情感倾向。因此，情感模型的建立依赖于情感词典的构建和判断规则的质量，同时需要人工设计和先验知识。显然，这个过程是需要耗费大量的人力资源，效率低下。基于机器学习的技术建立情感分类模型的核心是模型的训练，其需要对训练的进行人为的标注，因此，标注的质量是影响模型质量的原因之一。高质量的标注也需要大量的人力资源成本。主要过程为分词的处理、特征向量化、特征选择、模型调参。其中特征的构造和选取常用的是N-gram特征、句法特征等。此方法除了需要人工标注样本在某些领域需要一定的专业知识外，特征工程也是此方法的重要环节之一，可以说，一个模型的好坏很大程度上取决于这一环节是否有效，这需要对业务场景进行深度的剖析，从而设计出相应的特征，这一过程所投入的人力和物力成本也是非常大的。

随着新媒体时代技术和网络技术的不断发展，网络舆情数据不断增长，单单按照前面所使用的传统技术进行情感分析模型的建立，显然是不能够满足实际的需要的。除了传统机器学习方法之外，近年来，随着深度学习技术的快速发展，也有许多通过深度学习技术来进行情感分类任务的实现，其与传统机器学习不同的是在与特征处理和模型结构上。首先，深度学习可以通过训练数据本身让机器自动进行特征提取，当数据量比较大时，机器自动提取特征比人工进行特征选择的效果显著。其次，主要是通过模拟人脑的智能行为，以类人脑的多层神经网络模型为主，包括卷积神经网络、循环神经网络等不同网络组织形式和训练方法。

卷积神经网络（Convolution Neural Network, CNN）在计算机视觉领域无论在学术界还是 工业界都取得了长足的进步。尤其是CNN进行图像分类取得了巨大突破，如Facebook的图像标注系统和自动驾驶的汽车系统等CNN都发挥着重要作用。与此同时，将CNN应用于自然语言处理（Natural Language Processing, NLP）领域也可以表现出比较好的效果，主要方面是由于词向量的发展，如：GloVe（Global Vectors for Word Representation）、word2vector、连续词袋模型（continuous bag of words, CBOW)、跳字模型（Skip-gram）等现有的词向量库已经能较好的表达词语的含义。基于深度学习抽象特征，不仅避免了人工提取特征的工作，也可通过词向量技术模拟词语之间的联系，具有局部特征抽象以及记忆功能在情感分类的建模中具有非常大的优势。因此，利用CNN结合词向量技术进行情感分类建模对于提升和改进使用传统机器学习方法的效率和不足具有重要的研究和实践意义。

## 发明内容（改进点、目标、执行机制、网络架构、算法、成果、使用流程、优势）

针对现有传统机器学习方法进行情感分类建模技术存在的不足，本发明的目的是基于卷积神经网络结合词向量技术实现对企业舆情分类的方法。从而解决传统机器学习方法中人工提取特征的低效问题并引入神经网络进行建模。

本发明主要通过文本预处理、神经网络模型搭建、模型训练三个主要步骤完成对企业舆情的情感分类模型的建立。利用词向量技术对输入数据进行预处理后作为网络，神经网络模型主要是卷积网络。模型最终能够实现对企业舆情数据进行二分类。

本发明对文本预处理的执行流程描述如下：首先将原始输入的舆情文本通过分词后转换成词序列，其次将词序列转换成带有唯一词编号的元素序列，最后将词的编号序列中的每个元素展开为词向量的形式。具体就是加载新闻语料后构建所需的词典，去掉文本中的标点符号及特殊字符后经过结巴分词对语料进行分词后进形成了词典。下一步对词典构造映射表：word到tokens和tokens到word。最后进行文本转换，有了映射表的基础上，我们就可以对原始文本进行转换，即将文本转换为机器可识别的编码。除此之外，为了保证句子有相同的长度，需要对句子长度进行处理。我们在描述性统计阶段可以发现，语料中句子的平均长度为2465个字符，因此我们就设置2465作为文本的的标准长度，取tokens平均值并加上两个tokens的标准差，假设tokens长度的分布为正态分布，则max\_tokens这个值可以涵盖95%左右的样本，我们对长度不足的进行padding补全，超长的进行截断。用于语义相似度的深度学习模型的预训练的词向量知乎问答语料库训练的300维度——sgns.zhihu.bigram作为模型的word embeddings。每一个字符对应一个300维的向量，我们将基于pre-trained word embeddings构造一个vocab\_size \* embedding\_size大小的矩阵。本文使用的矩阵大小为15538\*300。以上过程就完成了数据的预处理部分，主要完成了将原始新闻舆情文本转换为tokens并构建了word embeddings。总结来说，数据预处理就是为了克服文字长短不均和将词与词之间的联系纳入模型中的困难，人们使用了一种技术——词嵌入。词嵌入最经典的就是Word2Vec，它通过对具有数十亿词的新闻文章进行训练而得到词向量。其主要思想依然是把词表示成向量的形式，而不是One Hot 编码。简单说来，就是给每个词赋一个向量，向量代表空间里的点，含义接近的词，其向量也接近，这样对于词的操作就可以转化为对于向量的操作了，在深度学习中，这被叫作张量（tensor）。用张量表示词的好处在于：第一，可以克服文字长短不均的问题，因为如果每个词已经有对应的词向量，那么对于长度为N 的文本，只要选取对应的N 个词所代表的向量并按文本中词的先后顺序排在一起，就是输入张量了，其中每个词向量的维度都是一样的。第二，词本身无法形成特征，但是张量就是抽象的量化，它是通过多层神经网络的层层抽象计算出来的。第三，文本是由词组成的，文本的特征可以由词的张量组合。文本的张量蕴含了多个词之间的组合含义，这可以被认为是文本的特征工程，进而为机器学习文本分类提供基础。

其次，本发明深入研究了卷积神经网络模块的执行机制，CNN捕捉的局部特征，在图像处理中表现出不错的效果，这是因其卷积和池化操作可以捕捉到图像的局部特征。同样，我们将CNN用于文本处理上，也可捕捉到文本的局部信息。基于CNN的企业舆情情感分类的网络模型结构如图1所示，输入文本长度为2465个字符，长短不同的文本可采用Padding 技术取最长的文本长度作为文本的输入长度。而不足长度的都用空格填满，即把空格当成一个特殊字符处理。空格本身一般也会被赋予词向量，这可以通过机器学习训练出来。输入句子序列后，经过embedding，获得每个字符的词向量（假设词向量维度为300），则我们就得到一个text\_len \* embedding\_size的矩阵，如上图中embedding layer所示。我们可将它看做是width=embedding\_size，height=text\_len，channel=1的一张图片，可以用filter去做卷积操作。实现CNN在文本上使用。本网络结构中使用3中大小不同的过滤器其filter-size分别是2、3、4，每种filter-size有100个，如上图所示，红色的filter为height=2，width=embedding\_size，channel=1；绿色height=3，蓝色height=4。filter在width上要保持和embedding\_size一致，width代表的是词向量的大小，对于一个词来说，其本身的词向量分割是没有意义的，卷积操作的目的是在height方向滑动来捕捉词与词之间的局部关系。经过卷积操作后，得到了如上图中所示convolutional layers的输出，多个列向量，完成对文本中局部信息的捕捉；再经过max-pooling操作来提取每个列向量中的最重要的信息。最终经过全连接层fully-connected layer后再经过Sigmoid函数得到输出结果。

最后是模型的训练阶段。将文本数据经过词向量转化后分为训练集和测试集，对构建文本卷积神经网络模型进行训练和测试，最终训练出稳定的企业舆情情感分类模型；卷积神经网络模型的第一层为嵌入层（Embedding Layer）。第二层为卷积层（convolutional Layers），第三层为最大池化层（max-pooling Layer），第四层为全连接层（fully-connected Layer），最后一层用Sigmoid函数，训练的预测结果是一个[0, 1]区间的连续的实数，而程序默认情况下会将0.5设为阈值，也就是将大于0.5的结果判断为正，将小于0.5的结果判断为负。训练过程对训练集中每一个样本进行处理，对输出结果与标签分类结果求损失用于训练，测试集用于得到输出概率结果并根据训练实际情况人为设置阀值与标签分类结果进行准确率计算，不参与训练；对于每一个样本数据（即每一条舆情的数据）采用步骤4）~11）的方法进行训练，得到该样本数据（即该文本）的情感倾向及概率值；

* 1. 卷积网络中的嵌入层将训练集中的企业舆情数据进行变换，得到高维特征词向量表示；
  2. 生成模块将步骤4）得到的高维特征表示存入记忆单元中，每一家企业对应高维特征存入一个时间槽中；
  3. 当在训练集中有一个新的企业特征输入时，先将其变换成与记忆单元中特征大小相同的高维特征，再将记忆单位中由步骤5）得到的每一个时间槽的高维特征与之进行内积运算，对内接结果通过softmax函数进行计算，得到两者相似性；
  4. 输出模块将相似性较高的记忆插槽的高维特征与新的企业的特征直接相加，将得到的新特征直接输出；
  5. 将步骤7）得到的输出特征，经过输出门输入到计算单元并且在重复9次后进入下一层继续计算，其中输出门系数是由输出特征的线性变换的sigmoid非线性变换来决定的；
  6. 将步骤8）的输出作为下一层的输入，下一层即与步骤5）到步骤8）所描述的网络模型一致，即步骤5）到步骤8）所描述网络为一层，整个网络模型为步骤5）到步骤8）所描述的网络的多层叠加，重复步骤5）到步骤8）的过程；
  7. 最后一层的（即步骤5）~步骤8）重复9次后的第10层）输出特征输入计算单元，多个计算单元（可以由卷积神经网络全连接网络构成）分别进行计算，每个计算单元的结果（每个结果为1\*5的one-hot向量）送入Softmax函数进行分类概率计算，并进行投票，从而得到最终的结果；
  8. 在训练阶段，将预测结果与标签进行比较，根据损失函数求得损失，并对模型进行训练；在测试阶段，选择概率最大值作为当前样本数据（即当前企业）的最终风险预警评估级别； 将待分析的企业舆情数据作为模型的输入，对企业舆情进行情感分类，输出为企业舆情情感正负向分类，并将结果反馈给用户。

进一步的，输入文本的使用词嵌入技术转化为词向量表示，本发明使用word2vec构建词向量，向量化文本数据，其原理如下：

本发明执行过程如下,如图4所示：

（1）数据预处理，词向量编码。主要步骤有加载词向量。词向量可以从网上下载或者自己训练。网上下载的词向量获取简单，但往往缺失特定场景的词语。本发明选择将两种方法结合起来，也就是加载下载好的词向量，然后利用补充语料进行增量训练；建立词语到词向量的映射，也就是找到文本中每个词语的词向量；对文本进行词向量编码。

（2）构建卷积神经网络模型。确定模型的基本结构，包括嵌入层、卷积层、池化层和全连接层等不同层及其相互之间的基本结构和参数构成，根据模型训练的结果来不断调整训练迭代的次数、学习率、过滤器的数量和大小等网络超参的设定。

（5）对构建的网络模型进行训练和测试，形成稳定的企业舆情情感分类模型。对企业、的舆情数据进行随机划分，分为训练集和测试集，以训练集数据为输入，对卷积神经网络模型进行迭代训练，然后以测试集对模型进行测试，并根据测试结果进行参数调整，最终得到趋于稳定企业舆情情感分类模型。

（6）对新的舆情数据进行输入，并通过构建的卷积神经网络模型进行企业舆情情感分类，将分类和分析结果通过Web浏览器进行展示。

本发明提出的一种 基于卷积神经网络企业舆情情感分类技术的方法，与现有传统技术相比，具有以下优势：

1. 高效性。卷积神经网络具有自动提取特征的功能，避免了传统机器学习方法人工进行特征选择。
2. 泛化能力强。利用词嵌入技术可对词之间的相关性进行表示
3. 低门槛。用户无须关心数据的预处理、模型的构建与训练和参数等底层的细节。可直接通过系统的web界面直观的查看对企业舆情情感分类的结果和分析结果，方便用户使用此模型进行企业的舆情的监测。

## 附图说明

图1为本发明方法的卷积神经网络模型图；

图2为本发明词向量机制图；

图3为本发明的整体流程图；

图4为 本发明执行过程示意图。

## 具体实施方式

下面结合附图和具体实施案例，进一步阐明本发明，应理解这些实施案例仅用于说明本发明而不用于限制本发明的范围，在阅读了本发明之后，本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。







如图4所示，第一，本发明需要获取企业舆情数据，具体可通过网络爬虫进行获取也可直接使用企业已有的舆情文本数据。

第二，需要对输入的舆情数据文本进行预处理。具体包括对文本分词和为文本词汇建立索引

第三，构建卷积网络。

第四，网络训练。

第五，返回结果并为新的舆情数据进行分类。

以上实施仅用以说明本发明的技术方案而非对其进行限制，本领域的普通技术人员可以对本发明的技术方案进行修改或者等同替换，而不脱离本发明的精神和范围，本发明的保护范围应以权利要求书所述为准。