疫情时期 twitter 数据情感分析

一、 背景介绍

在如今的新媒体时代，社交媒体平台已经成为人们表达观点、分享情感的主要渠道之一。因此，社交媒体不仅仅是一个信息分享的平台，它也是一个重要的情感传播工具。特别是在面对大规模事件时，人们倾向于在社交媒体上表达他们的情感、担忧和希望。因此，分析人们在社交媒体平台上的发表的内容，可以获得有关公众情感倾向的重要洞察。

新冠疫情是自 20 世纪以来全球范围内最严重的公共卫生危机之一。这一大流行病对人们的生活、社交和心理健康产生了深远的影响。在这个背景下，对公众情感的深入了解变得至关重要，有助于政府更好地了解人们的需求和反应，从而更好地应对危机。

因此，本研究旨在利用疫情期间 Twitter 上的文本数据，采用自然语言处理（NLP）进行情感分析，分析疫情期间大众的情感倾向。我们的目标是判断人们的情感是积极的还是消极的，通过情感分析，我们可以了解公众对疫情的情感反应。

二、 数据介绍与描述性分析

我们使用的数据是新冠期间人们在 Twitter 上发表的内容及其情感，该数据来自于 kaggle 平台（https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification）。

数据已被划分为训练数据和测试数据，其中，训练数据共 41157 条，测试数据共 3798 条。数据的属性包括：用户名称、设备名称、所在地、时间、原始内容、情感倾向共六个属性。其中，为防止隐私泄露，用户名称和设备名称用数字标注，原始内容即为用户发表的内容，是我们文本分析的重点，而情感倾向是通过人工标注的，共分为五类：非常积极、积极、中性、消极、非常消极。

我们画出在训练集中数据大于 200 条的各国家或地区如下图 1 所示，可以看到前三名的国家或地区分别为伦敦、英国和英国伦敦，说明在我们的训练集

中，来自英国的用户发表的内容最多。而紧随其后的是一系列所在地位于美国的数据。这也符合 Twitter 用户以欧美国家为主的特征。

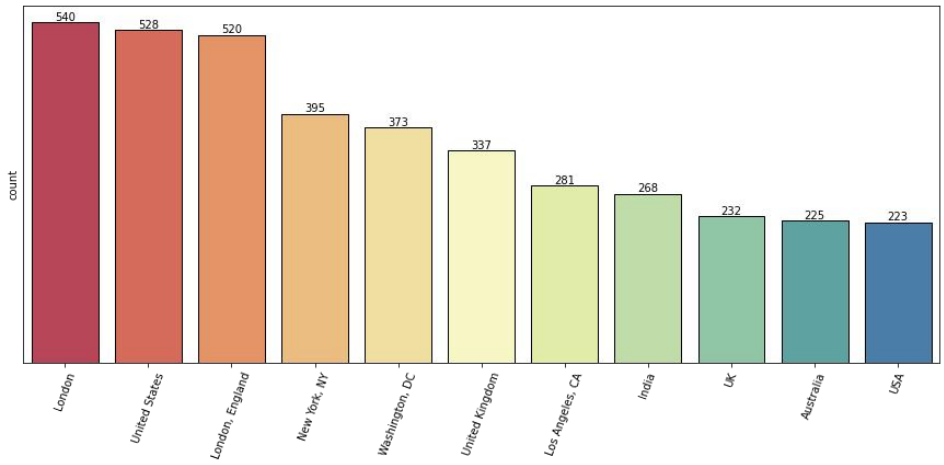

图 1 所在地

我们画出用户的情感倾向，可以发现最多的是积极情感，且正向情感总数大于负面情感，这可能说明疫情期间人们对于战胜疫情抱有更积极的期待，而不是消极态度占据社会主流思想。

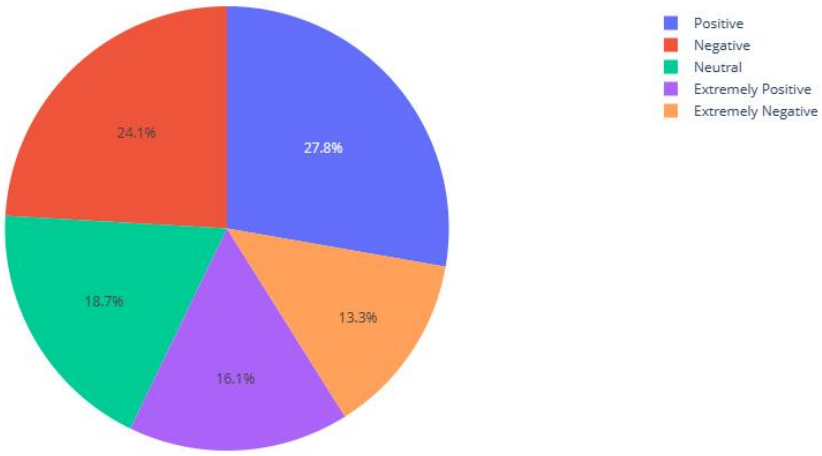

图 2 情感倾向

我们接着使用数据量大于 200 条的六个地区的数据，画出了所在地与情感

倾向的旭日图。图中三个美国地区与三个英国地区中，最多的情感倾向都是积极，表明在主要城市与地区中积极情感是占主流的。此外，三个美国地区的非常积极数量都多于非常消极，而三个英国地区中有两个都是非常消极的更多，说明相比之下美国民众可能比英国对于抗疫更有信心。

图 3 所在地与情感倾向

三、 数据处理

对于文本数据，处理成建模需要的数据是非常重要的环节。

对训练集数据分析发现，数据中在原始内容和情感倾向两个属性下不存在缺失数据，且原始内容中也没有相同的数据，因此无需删除数据。

我们接着定义了一个类 data_clean，用来实现我们对原始数据的处理。

通过新类 data_clean 我们将原始数据转换成数据处理的对象，在初始化函数中只保留了数据中原始内容和情感倾向这两个属性，并将原始内容转化为字符串格式，如果格式转换遇到错误会返回异常。

然后我们在类中定义了标签转化的函数，将情感倾向由文本转化为数字，我们只保留了积极、消极和中立，而不再区分程度，即非常积极和积极我们都共同处理为 1。

我们语料清洗的工作也是通过定义的函数实现的。具体来说，由于 Twitter

的内容中@其他用户以及分享网址非常常见，而这些又不能用于情感识别，因此我们首先通过正则表达式将这些内容去除。此外，我们还删除了非 ASCII 字符，删除了标点符号，将多余的空格仅保留一个空格，并将所有的大写字母转换成小写，实现了基本的数据清洗。

由于中英文的差异，英文的分词往往非常简单，我们上述的数据清洗步骤事实上也完成了英文的分词。

我们接下来删除数据中的停用词。停用词就是没有什么实际含义的词，删除它们对于理解句意没有影响，例如句子中大量存在的虚词、代词和冠词等等，在英文中表现为‘a’，‘the’，‘i’，‘very’等单词，为了简化模型我们可以直接删除。

英文文本处理另一个与中文有差异的处理在于，我们需要将英文单词进行词干提取。因为在英文中同一个词存在大量的不同形态，包括单数、复数、被动语态、进行时语态等等，这些形态的差异对于建模没有作用，在文本处理时我们仅需要当做同一个单词处理，因此我们将这些不同形态的词全都还原为同一个词干。

到此我们得到了比较纯净的文本数据，数据均为小写的英文词干，虽然还没有将文本转换为模型需要的数字数据，但是数据的基本处理到此都已实现，因为更进一步的特征工程需要根据不同模型的需求来进行，我们在下面建模分析时再对数据深入处理。

四、 建模分析

4.1 逻辑回归

我们首先使用逻辑回归来进行分类。

我们基于词袋模型来处理文本数据，词袋模型不考虑单词在句子中的顺序，直接将所有文本中出现的词放入一个词袋，通过统计单词出现的频率来提取特征。我们使用的 TF-IDF 方法则是基于词袋模型的经典方法，通过词频和包含该单词的文本数量，来将文本数据转化为数值特征。

我们首先只保留文档频率大于等于 5 的单词，使用训练集进行拟合，最后得到的特征数为 7681。使用拟合后的模型，我们将测试集的文本也转换为数值

矩阵。

由于我们的问题是一个三分类问题，因此我们使用多类别逻辑回归模型来进行拟合。

使用拟合好的模型，我们对预测集数据进行预测，得到准确率为 78.57%，表现较好。而在训练集上逻辑回归的准确率为 85.35%，没有明显的过拟合，具有良好的泛用性。

我们对 TF-IDF 方法的参数进行调整，将文档频率小于 30 的单词都删除，此时剔除了大量的特征，保留的特征数降低为 2164 个，但模型在训练集上的准确率为 78.17%，与原模型的准确率相比几乎没有变化。而当我们将文档频率小于 5 或者文档频率大于 100 的单词删除时，保留的特征数为 6539，表明文档频率较大的单词占比较少，但是模型的准确率降低到了 52.96%，说明这些频率较大的单词在预测中具有重要的作用，而频率较低的单词虽然占据了非常多的特征数，但带来的信息是不足的。

4.2 循环神经网络

我们接下来使用循环神经网络（RNN）来进行拟合。RNN 具有循环结构，使得信息可以在不同时间之间传递，适合处理文本数据这种具有上下文信息传递结构的数据。

我们将训练数据进一步划分为训练集和验证集，通过验证集评估模型拟合效果。我们重新对文本数据进行特征提取，使用完整的训练数据构建词汇表，将文本数据转化为单词在词汇表中的索引，再将这些索引序列全部转为统一长度的序列作为模型的输入。此外，我们将情感倾向数据转化为独热编码。

我们设置嵌入层来将词汇嵌入到一个固定长度的向量空间，嵌入层的维度为 300。而在循环神经网络层，我们设置了 100 个隐藏单元用以处理输入序列。在输出层，应对三分类问题，我们设置了 3 个神经元，并使用 Softmax 作为激活函数，来进行多类别分类。

接着在编译模型时，我们设置交叉熵为损失函数，梯度下降方法选择 Adam，使用准确率来在训练过程中检测模型性能。

使用训练集拟合数据，迭代次数为 5 次，批量大小设置为 100，并通过验证集进行评估，我们得到模型的准确率为 74.17%，而在训练集上的准确率达到

了 99.52%，存在严重的过拟合现象。

因此，我们在模型中加入 L2 正则项和 Dropout 层来减轻过拟合问题。L2 正则项的权重设置为 0.3，Dropout 层随机关闭神经元的概率设为 0.2。此外，我们使用验证集上的准确率来监测迭代中的性能，并加入早停策略来防止过拟合，我们设置当验证集准确率连续两次迭代中降低后就停止迭代。由于正则项和 Dropout 层，降低了模型收敛的速度，且加入了早停策略可以监测性能，因此我们增加迭代次数为 10 次，批量大小增加到 200。

此时模型在验证集上的准确率为 86.61%，在训练集上的准确率则达到了 78.04%，过拟合问题有所减轻，提高了准确率。因此我们将此时的模型作为拟合的最终模型，用于测试集数据，得到最终的测试准确率为 76.33%。

4.3 长短期记忆网络

我们再使用长短期记忆网络（LSTM）模型来进行拟合，LSTM 模型在 RNN 的基础上，能够捕捉长期依赖关系，在自然语言处理中更具优势。

我们模型的初始设置保持和 RNN 一致，只将中间的 RNN 层改为同样具有 100 个神经元的 LSTM 层。我们得到 LSTM 模型在验证集上准确率为 79.49%，而在测试集上准确率为 95.88%，准确率比初试的 RNN 模型高，但同样存在明显的过拟合问题。

我们使用和 RNN 模型类似的方式调参来减轻过拟合问题。我们将 Dropout 层随机关闭神经元的概率设为 0.1，批量大小设置为 300。

得到的模型在验证集和训练集上的准确率都降低了，缓解了过拟合问题，但降低了准确率。因此我们将原来的模型作为拟合的最终模型，最终的准确率 79.49%。

五、结论

我们建模分析的结果表明，使用用户发表内容的文本数据判断用户的情感倾向是可行的，各模型具有较高的准确率，其中长短期记忆网络（LSTM）具有最高的准确率，高于机器学习方法的逻辑回归，但是我们的数据量对于循环神经网络（RNN）和长短期记忆网络（LSTM）可能有些偏少，因此循环神经网络（RNN）的准确率略低于没有使用文本上下文信息的逻辑回归，长短期记

忆网络（LSTM）的准确率在更大的样本量下也有提升的空间。而传统统计方法下的逻辑回归表现出了优秀的准确率和可迁移性，在保持准确率的同时没有出现过拟合的问题。

附录

GitHub 链接：luxiaoxue23/nlp_programing (github.com)