推特情感分析

|  |
| --- |
| 实验环境：  python 3.9  pycharm  anaconda |
| **一、引言**  美国总统大选是全球瞩目的重大政治事件，不仅影响美国国内政策，也对全球政治、经济、金融市场等方面产生深远影响。随着社交媒体的普及，特别是推特（Twitter）的广泛使用，越来越多的政治人物和选民通过推特表达观点和意见。推特上的文本数据为我们提供了一个丰富的资源，可以用来分析公众情绪和意见。同时，联邦基金利率（Federal Funds Rate）彰显了美联储对于经济形式的判断，及执政党的经济政策。因此，我对联邦基金利率与美国总统大选的民意趋势进行综合分析，探求两者之间是否存在相关性。  利用自然语言处理（NLP）技术进行情感分析，可以从大量的推特数据中提取公众对候选人的态度和情感倾向。这种方法不仅可以实时反映选民的情绪变化，还可以作为预测选举趋势的一个有力工具。特别是在选举期间，公众情绪的变化可能预示着选举结果的潜在趋势。因此，通过情感分析来预测美国总统大选的趋势具有重要的现实意义。  本报告主要涵盖以下几个方面的内容：从推特API获取关于美国总统大选的推特数据，对数据进行清洗和预处理，包括去除噪音、提取相关文本等。使用自然语言处理技术对推特文本进行情感分析，将情感分析结果与美国联邦基金有效利率数据进行整合，使用双坐标轴绘制情感得分和利率的时间序列图。再计算情感得分与国债收益率之间的皮尔逊相关系数，进行假设检验，评估两者之间的相关性。  **二、Web数据特点及其挖掘价值**  1. 海量性  数据量大：推特是一个全球性的社交平台，每天都会产生数以亿计的推文。尤其是在美国总统大选这样的重要事件期间，相关推文的数量更是急剧增加。我们爬取了大量关于候选人、政党、竞选活动等相关的推文，为情感分析提供了丰富的数据来源。  高频更新：推特用户随时随地发布推文，数据实时更新。这种高频更新使得我们能够获取最新的公众情绪变化，为选举预测提供及时的信息支持。  2. 动态性  实时性强：推特数据具有高度的动态性，能够实时反映公众情绪的变化。尤其是在选举期间，候选人的言论、事件的发生都可能引发公众情绪的剧烈波动。  时间敏感性：由于推文是按时间顺序发布的，我们可以通过时间序列分析，观察公众情绪在不同时间点的变化，捕捉到重要事件对情绪的即时影响。  3. 异构性  多样化的数据格式：推特数据不仅包括文本信息，还可能包含图片、视频、链接等多种形式。虽然本研究主要关注文本内容，但这些多样化的数据形式为未来的多模态情感分析提供了可能。  语言多样性：尽管我们的分析主要集中在英文推文上，但推特平台上也存在大量的多语言数据。这些数据的多样性增加了处理和分析的复杂性。  4. 噪声和不一致性  噪声数据多：推特上的数据质量参差不齐，包含大量的噪声，例如拼写错误、语法错误、非正式语言、表情符号等。需要进行严格的数据清洗和预处理，以保证分析结果的准确性。  情感表达的多样性：用户表达情感的方式多种多样，有些情感表达是隐含的或讽刺性的，给情感分析带来了挑战。  **三、Web数据收集和预处理**  1. 数据采集  本次采集任务的内容主要来自调用推特的官方开发者平台提供的API——X API v2、    在python开发环境下，选择使用twarc作为库和工具，首先用twarc2 configure进行确权，按照提示输入自己的token和key，成功后生成twarc.log如下    通过X API收集到JSON数据，通过pandas将json数据转化为xlsx方便查看    2. 数据清洗和预处理  数据清洗和预处理中，先使用pandas库的dropna处理缺失值，去除 content 列的特殊字符和标点符号，同时因为评论内容中可能会存在网址的情况，通过正则表达式将所有网址内容去除，防止影响到后续的文本分析。    接下来因为不确定评论的对象，因此使用了基于spaCy框架的预训练模型包en\_core\_web\_sm-3.0.0来进行句子主语的提取。选取了如下的关键词，其中第一组和第三组代表共和党为主语，第二组和第四组为民主党为主语，若主语为下面的关键词，则统计其情感得分。      到此数据采集和预处理完成  **四、主要采用的Web数据挖掘技术**  首先对得到的评论进行词云分析，得到下面的结果    通过nltk中已预训练好的模型SentimentIntensityAnalyzer，可以对一个句子进行语义分析，得到如下结果：    再结合上面对关键词进行筛选后，对compound进行均值化得到结果如下    out\_sentiment越小，支持共和党和Trump的倾向就越大，越大则支持民主党和Biden的倾向就越大，越趋近于0则越中立  **五、案例分析**  再对于民众的支持率变化作时间序列的分析，通过可视化和假设检验的方法来验证民意变化与联邦基金有效利率之间是否存在关联  首先通过滑动平均来对数据进行平滑处理，选取滑动窗口为120天。    再用matplotlib.pyplot进行图像的绘制，得到结果如下：    由此可以直观的看出，随着加息的进行，美国民众对拜登政府的民意在下降。  同时，为了较为严谨地探究其相关性，故通过皮尔森相关系数并进行假设检验，得到结果两者具有显著的负相关性。    **六、总结**  本项目的特点主要包括将推特的数据进行了主语提取，由于进行爬取的关键词为US Presidential Election 2024，并不能确定每一条语句的评价对象是哪一方。因此要确保得到的情感倾向是正确的，就需要进行主语提取。同时本项目还将文本情感数据与其他维度的数据进行了关联分析，得到了存在负相关的结论。  从逻辑和政治诉求上来讲，共和党主张自由竞争、小政府，因此可能会采用较为温和的货币政策，而目前的执政党民主党主张通过较为猛烈的加息来抑制通胀，因此带来了支持率的下降。虽然说美联储和联邦政府之间理应相互独立，但两者仍存在密切的联系，例如FOMC中包括美联储主席的七人由总统任命，因此大部分情况下还是与联邦政府保持一致。从逻辑上来说负相关是有迹可循的。但大选是复杂、多因素的，因此不能只从这一个角度来理解。而且美联储的加息其实也受很多其他因素的影响，可能不仅仅是满足政府的诉求，也有外部因素和其他经济情况产生的影响。  对于数据挖掘来说，推特是主流网站中最难直接爬取的之一。它的代码是动态生成且结构不规则的，用传统的scrapy之类的使用xpath很难爬取，后续尝试了官方的api的第二个工具才成功。同时这次的数据挖掘，还有更多可以拓展的点，比如其实爬取的信息中不止文本，还有影音的链接，未来通过图像识别，卷积神经网络等可以对其进行进一步的分析，增加可以挖掘的内容。  **参考文献**  [1] Economic Policy Uncertainty and the 2020 U.S. Presidential Election (Scott R. Baker .et)  [2] How Do Expectations About the Presidential Election Affect Interest Rates (David J. Handley Francis X. Diebold .et) |