

# HPV舆论分析报告第二版

## 第一阶段：使用深度学习分析公众舆论中HPV话题的关注热点和态度

人类习惯于通过信息交流来增进对于各种话题的认知，而当新冠肺炎疫情全球蔓延时，公众对于传染病的关注度也相应提高。其中，HPV（人乳头瘤病毒）病毒已经成为了人类面临的一种重要的健康威胁，但公众对于该病毒的认知和了解程度还比较有限。通过本研究，我们可以深入了解公众对于HPV话题的关注点和态度，建立预测模型，分析在安德森卫生模型下，公众舆论中对于HPV话题的关注热点，以及不同关注热点所对应的用户态度变化，从而建立预测模型，为未来的疫苗接种和健康管理提供参考依据。可以帮助相关机构提高公众的健康意识和素质，从而达到预防和控制HPV疾病的目的。同时，本研究也将为深度学习技术在健康领域的应用提供实践探索，对于推动健康大数据和人工智能技术的发展和具有重要的意义。

本阶段旨在利用深度学习技术，分析公众舆论中对于HPV话题的关注热点和态度，从而为未来疫苗接种和健康管理提供参考依据。整个分析过程将分为四个目标，通过渐进式研究方式对公众舆论展开分析。

### 1. 分析公众在HPV相关话题舆论中，对于安德森卫生服务模型所涉及的理论要素的关注度；

本阶段旨在了解公众在HPV话题舆论中对于安德森卫生服务模型理论要素的关注程度。HPV疫苗接种是一项重要的公共卫生服务，但是在接种率方面仍存在一些挑战，例如公众对于疫苗安全性的担忧、信息来源的不确定性等。通过了解公众对于安德森卫生服务模型所涉及的理论要素的关注度，可以更好地了解公众对于HPV疫苗接种的认知情况，进而制定更有效的宣传策略，消除公众的担忧，增加接种率，保护公众健康。此外，还能为相关机构制定相关政策和计划提供重要参考。

### 方法和步骤：

为了达成上述目标，我们拟从微博评论中爬取相关数据，并使用安德森卫生模型对用户关注热点进行建模。具体步骤分为四个部分：

#### 1. 收集数据集。

我们将采用网络爬虫技术从新浪微博平台上收集与HPV话题相关的微博评论。使用关键词，如疫苗、HPV、宫颈癌、人乳头瘤病毒、九价、四价、二价、Gardasil和Cervarix等关键词，通过网络爬虫技术和流媒体API，收集从欧诺个2018年3月19日（HPV疫苗国内批准上市日期）至当前日期的微博博文，用于研究分析。为了保证数据的质量和可靠性，我们将对数据进行清洗和筛选。具体地，我们将根据以下几个方面进行筛选：

- 评论内容必须包含HPV相关的关键词或者话题标签；

- 评论长度不能少于5个汉字；
- 评论不能含有敏感或者不适宜的内容；
- 评论不能是广告或者垃圾信息。

在收集到的评论数据中，我们将根据用户将其进行分类和组织，以便后续分析用户对于安德森卫生服务模型理论要素的关注倾向。

## 2. 预训练深度学习Bert模型。

为了对评论数据进行分析 and 挖掘，我们将使用深度学习Bert模型进行预训练。具体地，我们将利用大型中文语料库和医学语料库对Bert模型进行预训练。通过预训练，我们将获取一个具有强表征能力的模型，可以为后续分析提供更加准确和可靠的基础。具体步骤如下：

- 收集并准备数据集。预训练语料采用清华大学自然语言处理实验室推出的中文文本分类工具包中整理的新闻语料，和从互联网中收集的大量医学报告和文献，对其进行处理和清洗。
- 使用预处理工具对数据进行分词、句子分割、Tokenization等操作，生成可以用来训练Bert模型的数据。
- 使用Pytorch、Huggingface等深度学习框架，搭建Bert模型并在预处理后的数据上进行训练。训练任务采用无监督训练任务。在不增加人力劳动的情况下，充分理解句子语义信息、理解医疗实体。预训练过的词向量可以对词进行更精细化的表示，提高情感分析的性能。训练完成后，保存模型权重和参数，以备后续使用。

## 3. 使用训练好的Bert模型来提取微博评论中的实体。

具体步骤如下：

- 对每个微博评论进行预处理，包括分句、分词、Tokenization等。
  - 对处理后的数据进行编码，并使用Bert模型进行特征提取。
  - 将提取到的特征输入到实体识别模型中，使用CRF等算法进行实体识别和提取。
  - 最终得到的实体包括人物、疾病、药品等与HPV话题相关的实体。
4. 根据安德森卫生模型对实体进行分类。将使用预先定义的规则和分类方法，将提取到的实体根据安德森卫生模型的理论要素进行分类。具体步骤如下：
- 定义一个基于安德森卫生模型的分类规则，将实体按照其所属的理论要素进行分类。
  - 对每个实体进行特征提取，并根据分类规则进行分类。
  - 统计每个用户关注的理论要素及其关注程度，为后续分析用户态度提供依据。

## 2. 分析公众在HPV话题舆论中持有的情感态度。（中性态度，负面态度，正面态度）；

情感是人类在沟通交流中所表达的、包含情绪和态度的非语言化信息。情感通常是无意识的，但可以通过表述的语法、用词等多种方式挖掘出来。通过情感态度分析，了解公众对于HPV话题的潜意识态度倾向，包括中性态度、负面态度和正面态度。HPV疫苗是预防人类乳头瘤病毒感染和相关癌症的有效措施之一，但由于种种原因，如缺乏知识和误解，一些人可能不接受疫苗接种。因此，了解公众在HPV话题中对于疫苗的态度倾向，可以帮助医学工作者和公共卫生官员更好地了解公众对于疫苗的需求和关注点，有助于制定更有针对性的疫苗推广策略，提高疫苗接种率，进一步预防HPV相关疾病的发生。

## 方法和步骤：

使用预训练好的Bert模型，可以将用户在微博上的评论转化为计算机可理解的向量表示。然后可以使用这些向量作为输入来进行情感分析。

具体实现上，可以使用开源的中文情感分析库（如THULAC、SnowNLP等）或者基于深度学习模型的情感分析库（如TextCNN、LSTM、Bert等）。其中，使用Bert模型进行情感分析的方法较为高效和准确。

对于用户微博评论的情感分析，一般采用以下步骤：

1. 数据预处理：将采集到的微博评论进行清洗、过滤和分词处理，得到符合BERT输入格式的数据。
2. 使用预训练好的BERT模型：将预处理后的数据输入到BERT模型中，获取每个微博评论对应的BERT向量表示。
3. 评论情感分类：将BERT向量输入到情感分类模型中，对每个微博评论进行正面情感和负面情感分类。可以采用多种分类算法，如逻辑回归、支持向量机等。
4. 统计用户态度：将情感分析的结果分为中性态度、负面态度和正面态度。对每个用户的微博评论进行情感分类后，可以统计每个用户在HPV话题舆论中表达的不同情感态度的频率，分析出公众对于HPV话题的整体情感态度倾向。

## 3. 分析所关注不同安德森卫生模型的理论要素的人群，在HPV话题舆论中所持有情感态度之间的联系。（如安德森理论模型中要素相似的人群，所持有的情感态度）

分析安德森卫生模型和情感态度之间的联系，可以为制定更有针对性的宣传策略提供参考。在HPV接种宣传中，不同的人群可能对于安德森卫生模型中的理论要素有不同的关注点。因此，通过对不同人群的聚类分析，可以了解到每个人群的特点和需要重点关注的问题。同时，还可以根据不同人群所持有的情感态度，制定更加个性化的宣传语言和方式，从而更好地促进公众的HPV接种意识和接种行为。

## 方法和步骤：

1. 数据准备：根据第一阶段的方法，将用户根据对安德森卫生模型理论要素的关注倾向进行聚类，得到用户对安德森模型理论要素的关注倾向。根据第二阶段的方法，对用户评论进行情感分析，得到用户在HPV话题中的情感倾向。
2. 安德森实体要素建模：根据用户对于安德森模型理论要素的关注倾向进行聚类。根据安德森卫生模型的理论要素，我们可以将用户的关注倾向分成几类，比如“健康信念”，“社会支持”，“自我效能”，“知识认知”等等。针对每个理论要素，可以统计每个用户在微博评论中对相关实体的提及频次，然后将每个用户对不同理论要素的提及频次向量化。
3. 安德森用户聚类分组：对于向量化后的数据，可以使用传统的聚类算法，如K-means聚类、层次聚类，或者使用基于深度学习的聚类算法，如变分自编码器聚类。将具有相似向量的用户归为同一类别。聚类算法的目标是最小化同一簇内的距离，同时最大化不同簇之间的距离。

离。经过聚类后，我们可以得到一些聚类簇，每个簇代表着一个用户群体，该群体对不同安德森卫生模型理论要素的关注倾向相似。

4. 不同聚类簇内的情感态度分布的差异分析：如果不同簇之间的情感态度分布差异较大，则说明在某些安德森卫生模型的理论要素上有所关注的人群，在情感态度上存在差异。这一信息可以帮助相关机构更精准地制定有针对性的宣传策略，提高HPV疫苗的接种率。

## 4. 分析不同安德森卫生模型的理论要素对于公众对HPV态度的影响程度。（即哪些因素的提升更能让人群获得积极态度）

HPV话题中的态度倾向与安德森卫生模型理论要素的关系，有助于我们更准确地识别和理解公众的态度，从而采取更加精准有效的干预措施，以提高HPV接种率。通过机器学习模型的可解释性分析，可以深入了解哪些因素对于公众在HPV话题中的态度具有更强的影响力，为政策制定者提供有力的指导。例如，如果模型分析出知识水平的提升对于促进公众对HPV的积极态度具有最大的影响，那么政策制定者可以考虑加强公众的HPV知识宣传和教育。

### 方法和步骤：

1. 建立随机森林决策树模型：随机森林是一种机器学习模型，可以用于分类和回归分析。在此研究中，我们需要建立一个分类模型，以分析安德森卫生模型的要素对于态度倾向的影响程度。我们需要先将数据集划分为训练集和测试集，然后使用训练集来训练随机森林模型，以预测测试集中每个用户的态度倾向。
2. 可解释分析：在这个步骤中，我们需要使用可解释分析方法来理解随机森林模型中每个变量的重要性。这可以帮助我们确定哪些安德森卫生模型的要素对于态度倾向具有重要影响。常用的可解释分析方法包括基尼重要性、平均减少不纯度和排列重要性等。
3. 分析卫生模型要素对于态度倾向的影响程度：在这个步骤中，我们将分析每个安德森卫生模型的要素对于态度倾向的影响程度。例如，我们可以分析“感知障碍”、“行为信念”、“态度”等因素对于态度倾向的影响程度。这可以帮助我们确定哪些要素能够对态度倾向产生积极影响。
4. 总结能对态度产生重大积极影响的卫生模型要素：在这个步骤中，我们将总结哪些卫生模型要素能够对态度产生重大积极影响。这些要素可能包括促进个人信仰和态度变化的行为信念、提高认知水平的教育和宣传、降低接种障碍的政策和服务等。

## 第二阶段：使用深度学习模型分析民众接种转化率分析。

本阶段目的是使用深度学习模型分析民众接种转化率的影响因素。通过问卷调查和深度学习模型相结合的方法，综合分析了民众接种HPV疫苗的态度和行为之间的关系，同时还探讨了公众舆论中的关注热点和态度，旨在揭示公众对HPV疫苗接种的态度和行为之间的关系，为提高疫苗接种率提供科学依据，提高公众对疫苗接种的认知和接种率。

结合前一个阶段的研究，本阶段旨在结合用户个体更详尽的主客观情况，对HPV最终接种率的影响展开分析。整个阶段分为两个目标。



# 1.根据个体HPV情感态度和个体安德森模型要素，叠加个体主客观因素，分析最终对转化率造成显著影响。

本目标将分析个人评论中非意识流露的主观因素、个人有意识流露的主观因素和客观因素对最终接种率的影响程度，以及这些要素之间的不一致性。这有助于深入挖掘能够起到促进作用的因子，并帮助政府和相关机构更好地理解公众对于HPV疫苗的态度和接种行为。同时通过分析个人主客观因素和情感态度等要素的影响，可以更全面地了解公众对于HPV话题的态度和行为。这对于政府和相关机构制定更加精准和有效的宣传和推广策略具有重要意义。

## 方法和步骤：

本目标的完成，将需要发放问卷获取评论中无法获取到的用户信息，并对信息展开分析。

1. 确定问卷内容和问题：首先确定需要收集的个体主观和客观因素，客观因素如学历水平、经济水平等；主观因素如接触HPV时间、对政府医疗的信任度等等。然后制作问卷，并确保所有问题清晰、简洁明了，并能准确地反映被调查者的观点。
2. 确定被调查者：根据前期研究的结果和数据，确定需要收集的被调查者，例如评论过的用户，或者在特定社交媒体平台上发表过相关内容的用户。确保被调查者是符合研究目的和方向的，以确保数据的有效性和可靠性。
3. 发放调查问卷：将问卷链接或二维码通过社交媒体平台、电子邮件等途径发送给被调查者。可以采用抽样方式，确保样本具有代表性。同时，应该提供足够的信息来解释问卷的目的和保证被调查者的隐私和匿名性。
4. 收集问卷数据和评论数据：收集所有回答并进行统计分析。对收集到的数据进行清洗和预处理，包括缺失值填充、异常值处理、特征标准化等操作，保证后续建模的数据质量和准确性。
5. 建立用户画像：使用机器学习算法建立用户画像模型：
  1. 对于客观因素，如年龄、性别、学历、收入水平等，可以将其作为基础信息直接加入用户画像中。
  2. 对于主观因素，如对HPV的认知程度、对疫苗的信任程度等，可以将其转化为分数进行量化，作为用户画像中的主观因素部分。
  3. 对于情感态度和个体安德森模型关注要素，可以通过文本分析等技术将评论转化为特征向量，加入用户画像中。
6. 使用机器学习算法进行用户画像建模：生成概率模型是一种基于统计的概率模型，它可以通过分析数据集中的数据分布，来建立一个可以生成新数据的概率分布模型。在用户画像建模中，可以使用生成概率模型来建立用户的概率分布模型，从而预测用户的行为，例如是否会接种HPV疫苗等。生成概率模型包括多种模型，如高斯混合模型、朴素贝叶斯模型、隐马尔可夫模型等。以生成概率模型为例，可以将以上各因素作为因子，训练模型以预测用户最终接种行为的概率，并得到每个因子对最终接种行为的影响程度，从而得到更为准确的用户画像。
7. 模型评估和解释：对建立的模型进行评估和解释，包括模型性能评估、模型参数解释和特征重要性分析等。根据机器学习的可解释分析，分析用户画像中对最终接种影响最大的因子。

## 2. 分析评论中的个人所关注的要素，和最终造成接种行为提高的要素的联系和差异。

分析评论中的个人所关注的要素和最终造成接种行为提高的要素之间的联系和差异有助于提高公众对于疫苗接种的信任度。通过了解公众的关注点和所重视的因素，可以制定更加符合公众需求和期望的疫苗推广策略，提高公众对于疫苗接种的信心和意愿。

### 方法和步骤：

1. 对于生成概率模型中使用的特征，使用现有的特征重要性分析方法（如决策树算法、随机森林算法等）对特征的重要性进行评估，得到各个特征对最终接种行为的贡献度。
2. 对于个体所关注的要素，将其转化为特征，然后再使用同样的方法进行特征重要性分析，得到各个关注要素对最终接种行为的贡献度。
3. 对比上述两个分析结果，判断个体所关注最大的要素是否与对接种率影响最大的要素一致。

通过这种方法，可以分析个体所关注最大的要素与对接种率影响最大的要素之间的关系，并确定是否存在不一致性，为后续制定促进接种的策略提供参考。