**Introduction**

**在当今的数字时代，智能手机已成为现代生活的一个重要方面，改变着通信、娱乐和生产力。随着移动技术的进步，智能手机的采用率大幅提高，改变了个人与数字世界互动和接触的方式。Carbonell、Oberst 和 Beranuy（2013 年）强调，手机已成为个人身份和日常生活中不可或缺的一部分，人们与他们的设备形成了强烈的情感纽带。这使得人们越来越离不开手机，这也说明了人们对于手机产品的强烈需求，手机产业的市场具有良好的发展空间。Storbacka、Strandvik 和 Grönroos（1994 年）提出了一个服务质量模型，强调了服务质量与客户满意度之间的重要联系。他们进一步强调，客户满意度在培养客户忠诚度方面起着举足轻重的作用，而客户忠诚度反过来又会提高公司的盈利能力。鉴于这种关系，客户反馈是决定产品成败的关键因素，因为它直接影响服务改进和业务绩效。通过客户评论了解消费者情绪，这能够帮助手机公司识别关键问题。根据这些问题公司能够了解改进方向，从而为客户提供更优质的决策和产品。这种数据驱动的决策方式不仅可以提升品牌竞争力，同时也能促进市场份额的增长和产品销量的提升（Kim, Lee & Ahn, 2006）。当用户体验得到提升和用户需求得到满足时，最终公司能够达到提高市场竞争力和利润的目的。**

**本研究利用了从 Kaggle 获得的数据集。该数据集由亚马逊客户对十大品牌智能手机的评论组成： 华硕、苹果、谷歌、华为、摩托罗拉、诺基亚、OnePlus、三星、索尼和小米（Kaggle，2019）。通过分析这些评论，我们可以探索消费者对智能手机行业的整体看法。然后通过消费者的想法去提供一些商业见解和商业分析，帮助企业和行业发展改进**

**本研究拟结合情感分析与主题建模方法，讨论消费者对手机行业的关注和潜在的需求：情感分析将量化用户对产品的态度倾向，主题建模则通过文本聚类识别高频讨论议题。两者的协同应用有助于揭示行业发展的核心问题。从而为厂商优化产品策略和精准响应用户需求提供支持。**

**数据收集**

**本研究所使用的数据集为 Amazon Cell Phones Reviews。该数据集来自 Kaggle，包含来自亚马逊平台的智能手机用户评论和多个国际知名手机品牌。下面的几个原因促使了我选择这个数据集：**

**1. 数据的丰富性**

**本研究聚焦于智能手机用户的反馈及其对产品质量、功能和体验的评价。因此所选数据集需具备充足的数据量和品牌多样性，which能够反映市场整体趋势。该数据集包含超过60,000条用户评论，覆盖多个知名品牌，数据量充足且具有代表性。这使得研究者能够从多样化数据中提取全面的消费者的消费行为信息。从而帮助企业更准确地分析用户评价并且从中得到用户的诉求。从而根据实际的市场诉求改进企业策略和方法。此外，数据来源于亚马逊——全球最大的电子商务平台之一（Moriset, 2018），其评论系统是消费者表达购买体验和产品意见的重要渠道，也是在手机消费行业中占据重要地位的平台。这确保了数据来源的广泛性和可靠性。这些评论为研究消费者行为模式和行业竞争格局提供了有力支持。**

**2. 数据集符合研究目标要求**

**本研究旨在分析消费者对智能手机产品的情感倾向及其关注主题。所选数据集包含详细的文本评论和用户评分，能够有效支持自然语言处理技术进行情感分析和主题建模。具体而言，用户评分为情感分类提供了明确的标注依据，而丰富的评论文本则为深入挖掘消费者关注点提供了充足的语言素材。这些数据特征不仅符合研究目标的要求，也为社交媒体分析提供了可靠的基础。通过该数据集，我们可以系统性地探索消费者情感表达与产品特征之间的关联，从而得出具有实践意义的结论。**

**3. 公开数据集的便利性**

**相比于使用网页爬取获取社交媒体数据，该数据集来自 Kaggle，属于公开可用的二次数。这有效的避免了数据采集过程中数据访问限制的问题。除此之外，公开数据数据的完整性化和合理性也有所保证，这大量减少了在收集数据中的工作量，使研究能够专注于数据分析。**

**4. 数据集具有良好结构**

**该数据集包含评论文本、评分、产品品牌及评论时间等结构化变量，为品牌对比分析和消费者情感动态研究提供了基础。据研究报道结构化数据集有助于信息提取和商业分析，提高数据处理效率并增强分析结果的可靠性（Chen, Chiang & Storey, 2012）。一份具有良好结构的数据集能够帮助研究者进行深入的消费者思维分析和消费趋势分析。**

**数据清理**

**在数据清理的部分中，我遵从以下几个步骤完成我的数据预处理工作：**

**1. 数据加载、去除冗余、合并文本和缺失数据**

**我删除了不必要的字段，例如 asin、name、date、verified（是否为验证用户）以及 helpfulVotes（点赞数），其目的是简化数据集。除此之外，我移除了 body（评论正文）为空的行，以解决后续分析中的数据缺失问题。部分评论仅包含标题（title）或正文（body），它们不是一个完整的评论文本。我将 title 和 body 进行合并，添加一个新列full\_review作为完整的字段。并先用 fillna('') 处理 title 可能的缺失值。结果如下图所示：**

**Figure 2：去除多余的列和缺失值的结果**

**图形用户界面, 文本, 应用程序, 电子邮件

AI 生成的内容可能不正确。**

**2. 构建一个数据清理函数讲文本标准化**

**为了确保文本的可用性，我们设计了 clean\_text 函数进行预处理，包括：转换为小写，统一格式；移除 URL、@用户标签、#话题标签，去除干扰信息；去除标点符号和特殊字符，仅保留字母和数字；规范空格，确保文本格式整洁。过程如下图所示：**

**Figure 3：clean\_text 函数细节**

**文本

AI 生成的内容可能不正确。**

**最终我得到了一个只包含字母和数字的文本新列cleaned\_text：**

**Figure 4：clean\_text 函数运用结果**

**图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。**

**3.** **停用词去除与文本标记化**

**停用词是高频出现但几乎不携带任何信息的词，,比如“is”、“the”、“and”。** **根据 Ghag 和 Shah（2015）的研究表明，其去除有助于减少无意义的特征，提高机器训练和分类的效率。因此，我结合NLTK库的标准停用词和自定义停用词对文本进行清理。文本标记化的过程我采用的是自己构建函数，加入列表推导式的方法去标记文本的同时检测是否属于停用词。如下图所示：**

**Figure 5：停用词的创建和标记化函数的构建细节**

**文本, 散点图

AI 生成的内容可能不正确。**

**最终，tokenized\_text列是文本标记化的结果：**

**Figure 6：标记化结果**

**图形用户界面, 文本

AI 生成的内容可能不正确。**

**4. 词形还原**

**词形还原是 NLP 处理中用于将单词还原为其基本形态的重要步骤。与词干提取不同，词形还原依赖词典和语言规则，使其在语义保留方面更具优势以及提供更高的准确性（Pramana et al., 2022）。在本研究中，我使用 WordNetLemmatizer 进行词形还原，具体实现如下：**

**Figure 7：词形还原实现过程**

**图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。**

**最终得到了一个新列*lemmatized\_review：***

**Figure 8：词形还原结果**

**图片包含 图形用户界面

AI 生成的内容可能不正确。**

**首先，我要确定主题建模中最优的主题数量。我从数据集中随机抽取 2000 条样本，并进行分词和字典构建。然后，定义一个函数 *compute\_coherence\_values*，在不同的主题数量范围内训练 LDA 模型。然后计算每个模型的 Coherence Score，以评估主题质量。这个函数遍历多个主题数量，记录 Coherence Score，并绘制随主题数量变化的折线图。最后，选择 Coherence Score 最高的主题数量作为最优主题数，并打印结果。**

**从下面的结果图来看，在本数据集中最优的主题数量是4：**

这段代码使用 LDA 主题建模分析文本数据，并识别出 4 个主要主题。**Topic 1** 关注 **手机问题**，比如 *phone, issue, problem*，反映用户对设备故障的反馈。**Topic 2** 主要涉及 **用户体验**，比如 *great, good, love*，反映正面评价。**Topic 3** 讨论 **手机功能**，比如 *sim, unlocked, brand*，涉及解锁状态和品牌。**Topic 4** 关注 **技术与硬件**，比如 *android, fingerprint, video*，涉及安卓系统、指纹识别和视频质量。