

题目：

消极和积极情绪分类

成员：

全林献 2006500034

郭尔霜 2006500033

周文杰 2006500005

Kaggle 链接：

[Comment Crush | Kaggle](#)

科学意义：

社交媒体上的评论已经成为人们表达情感的主要方式之一，并且评论也可以成为公共舆论的重要来源。因此，情感分析已经成为了一个重要的研究领域，可以帮助人们了解某个话题在大众心目中的印象和态度。

为了提高情感分析的效率和准确度，本研究选择了 Kaggle 上的评论情感分析竞赛数据集，该数据集包括大量的人类生成的评论，可用于建立情感分析模型的训练和测试。

科学问题：

本研究旨在建立有效的情感分析模型，以自动化地判断社交媒体上的评论表达了积极或消极的情感。从而帮助人们更好地了解网络上的舆情，并为企业、政府等机构提供决策依据。

数据集：

该竞赛提供了一个数据集，包含了来自不同来源的评论，共计 2112 条。每条评论都标注了情感极性，即正面、负面。数据集分为训练集和测试集，其中训练集包含 1962 条评论，测试集包含 150 条评论。数据集的格式为 CSV。

内容：

Id - comment id.

Comment - The comment to be analysed.


Positive - The comment is positive or not.

Train.csv

Detail Compact Column

Id	Comments	# Positive
 <p>1 1926</p>	<p>1913 unique values</p>	 <p>0 1</p>
1	So there is no way for me to plug it in here in the US unless I go by a converter.	0
2	Good case, Excellent value.	1
3	Great for the jawbone.	1
4	Tied to charger for conversations lasting more than 45 minutes.MAJOR PROBLEMS!!	0
5	The mic is great.	1
6	I have to jiggle the plug to get it to line up right to get decent volume.	0
7	If you have several dozen or several hundred contacts,	0

Test.csv

Detail	Compact	Column
🔗 Id	⌵	⌵
 1150		150 unique values
1		299 confirmed kills awesome fight scenes 🍷
2		The fact is, this film is a wonderful, heartwarming tale about two people chasing their dreams.
3		The best part about "Nurse Betty" is it's unpredictability.
4		I felt asleep the first time I watched it, so I can recommend it for insomniacs.
5		There aren't death scenes like in previous movies and the f/x are terrible.
6		The story is lame, not interesting and

打算用的方法：

1. 文本预处理：对原始评论文本进行去除噪声、分词、词形还原等文本预处理操作，以消除干扰，提高质量。
2. 文本向量表示：将经过预处理后的文本转化为向量的形式，使得机器可以更好地理解和处理文本。
3. 特征选择：根据文本的统计特征选取最具代表性的特征。
4. 用基于 transformer 的模型进行实现。

预期目标：

预期目标是构建一个能够自动分类评论标签的模型，精度越高越好。在此基础上，要完成以下任务：

1. 建立适合该数据集的分类模型。
2. 进行预测，获得测试集的标签分类结果。
3. 对模型进行优化调整，提高分类精度。