2020MCM Problem C

问题假设

- 忽略单词排列顺序对于语义分析的误差，在分析评论感情的时候不考虑单词排列顺序对整体情感带来的影响
-

文本评论量化 Review Quantification

为了从用户的评论中提取对于商品评价的信息，我们需要用自然语言处理（NLP）来提取评论中于感情相关的词，并从这些词中判断用户的情感倾向，进一步将评论内容进行量化，我们将使用这个量化后的值作为衡量用户满意度的一个衡量标准

In order to extract information about product from user reviews, we need to use Natural Language Processing (NLP) to extract emotionally relevant words in the reviews, and judge the user's emotional tendencies from these words, and further quantify the content of the reviews. We will use this quantified value as a measure of user satisfaction.

朴素贝叶斯法进行情感分析 Naive Bayes

首先我们想要得到的目标是，可以通过评论中的一个单词来预测这个评论的感情是积极的概率和消极的概率，从而来判断这个评论是积极的还是消极的。在自然语言处理领域(nlp)中，朴素贝叶斯法通常用来对文本进行分类，并且处理问题时直接而且高效。因此我们用朴素贝叶斯法来对用户评论进行情感分类。
First of all, we want to get the goal that we can predict the positive or negative probability of the emotion of a review through a word in the review, so as to determine whether the review is positive or negative. In the field of Natural Language Processing (NLP), Naive Bayes is usually used to Text-Categorization, and it is direct and efficient when dealing with problems. Therefore, we use Naive Bayes to classify user reviews.

朴素贝叶斯法对单词进行情感分类时：对输入的单词x，通过学习到的模型计算后验概率分布，将后验概率最大的类作为x的类输出。

When Naive Bayes is used to classify words into words, for the input word x, the posterior probability distribution is calculated by the learned model, and the class with the largest posterior probability is output as the class of x.

假设$x \in X$，X是评论单词组成的集合，$c_k \in Y$，$Y = \{Pos, Neg\}$, 代表着评论的积极和消极。$Y = Pos$代表这个评论是积极评论，$Y = Neg$代表这个评论是消极评论。那么我们需要得到它的后验概率：
$P(Y=c_k|X=x)

Suppose $x \in X$, X is a set of review words, $c_k \in Y$, $Y = \{Pos, Neg\}$, represents positive and negative reviews. $Y = Pos$ means this review is a positive review, $Y = Neg$ means this review is a negative review. Then we need to get its posterior probability: $P(Y = c_k|X = x)$

根据贝叶斯公式，我们有：

According to Bayes formulaL:

$$P(Y = c_k|X = x) = \frac{P(X = x|Y = c_k)P(Y = c_k)}{\sum_k P(X = x|Y = c_k)P(Y = c_k)}$$

因为
$\sum_k P(X = x|Y = c_k)P(Y = c_k) = P(X = x|Y = Pos)P(Y = Pos) + P(X = x|Y = Neg)P(Y = Neg)$是一个定值，所以我们求解的问题变为以下最优化问题：

Becaues
$\sum_k P(X = x|Y = c_k)P(Y = c_k) = P(X = x|Y = Pos)P(Y = Pos) + P(X = x|Y = Neg)P(Y = Neg)$ is a fixed value, so the problem we solve becomes the following optimization problem:

$$y = argmax_{c_k} P(Y = c_k)P(X = x|Y = c_k)$$

其中$y \in Y$是x后验概率最大的类别，我们将这个类别作为该单词的类别输出。

Among them, $y \in Y$ is the category with the largest posterior probability of x, and we output this category as the category of the word.

我们在kaggle网站上寻找带有标签的亚马逊评论词料库作为训练集训练我们的贝叶斯分类器，然后用来分析评论的情感倾向。

We searched the kaggle website for tagged Amazon review corpora as a training set to train our Bayesian classifier, and then use it to analyze the emotional tendencies of reviews.

我们的贝叶斯分类器经过训练后，可以根据分类准确率分别得到一个积极评价的词云和消极评价的词云： ：

After training our Bayesian classifier, we can get a positive evaluation word cloud and a negative evaluation word cloud respectively according to the classification accuracy:

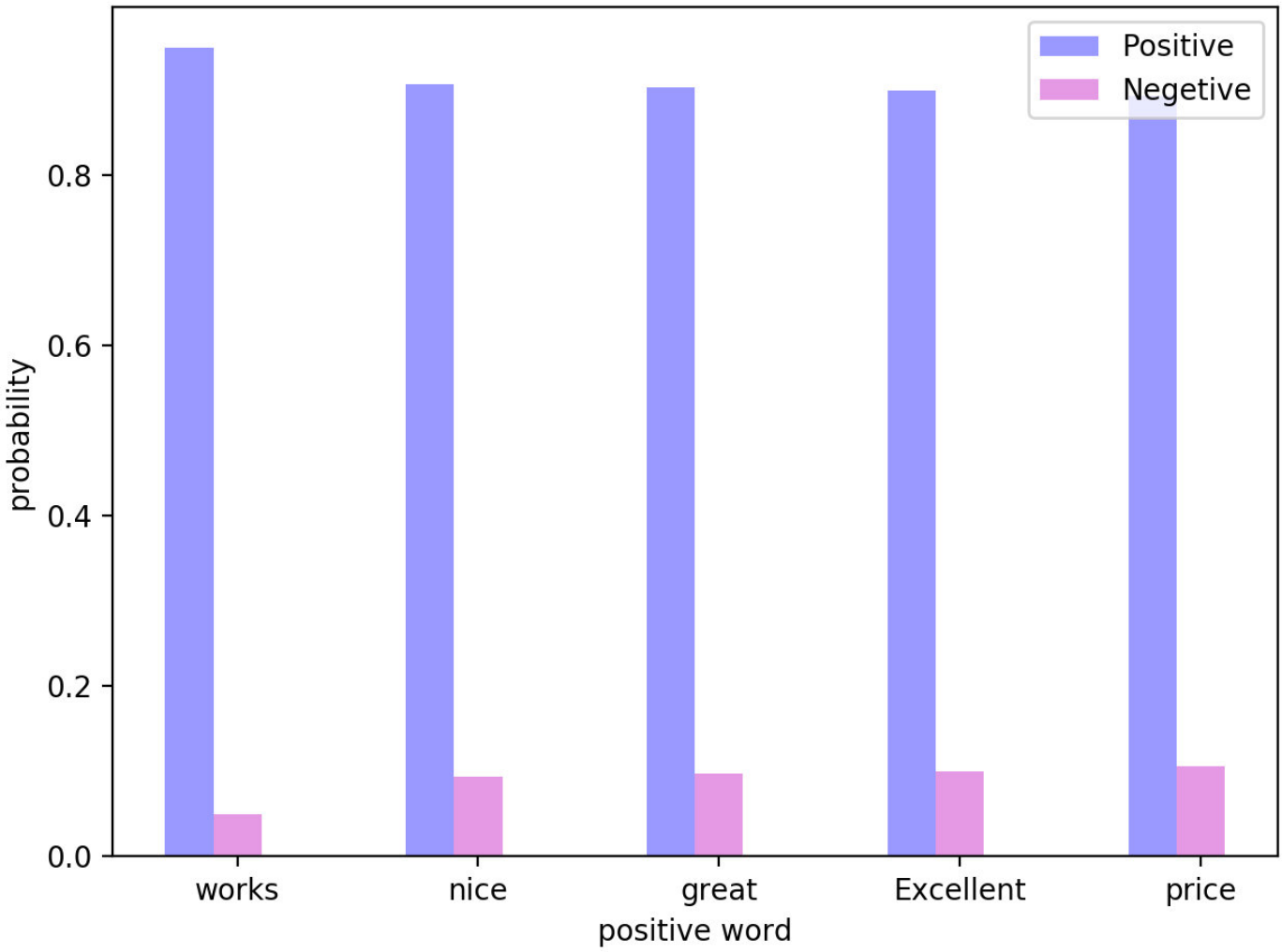
积极评价的词云：Positive word cloud:

消极评价的词云：Negative word cloud:

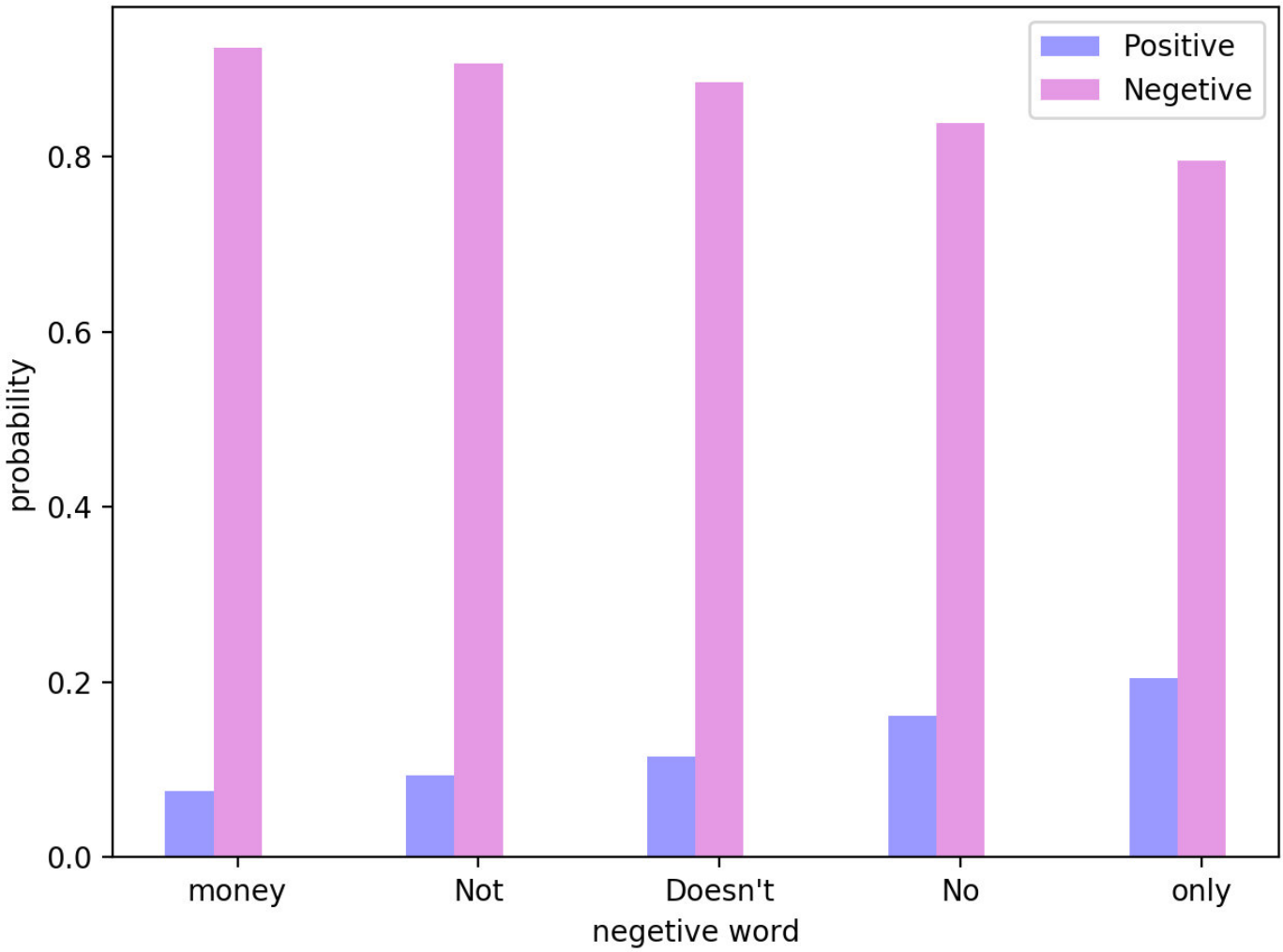
其中单词的体积越大，说明这个单词越具有分类效果，根据以上这些分类效果好的单词，我们可以对其和对应的后验概率进行可视化：

The larger the volume of a word, the more classification effect this word has. According to the above words with good classification effect, we can visualize their corresponding posterior probability.

积极单词：Positive word:



消极单词：Negative word:



基于这些具有显著分类效果的单词，我们可以对评论的感情基调进行量化。

Based on these words with significant classification effects, we can quantify the emotional tone of the review.

对评论感情进行量化 Review Emotion Quantification

我们在对评论进行量化的时候，对于评论中的每一个单词，我们都会计算在它存在于评论中的情况下，该评论是积极的和评论是消极的概率。然后我们根据以下公式来赋予评论一个“用户满意度”分数。

When we quantify a review, for each word in the review, we calculate the probabilities that the review is positive or negative if it exists in the review. We then give reviews a "user satisfaction" score based on the following formula.

假设y_i代表评论的分数，P_{ij}代表着第i个评论中第j个单词的后验概率，C是一个常数，那么量化公式为：

Suppose y_i represents the score of the review, P_{ij} represents the posterior probability of the j^{th} th word in the i^{th} comment, and C is a constant, then the quantization formula is:

$$y_i = log(\sum_{j}^{n} P_{ij} + C) - log(C)$$

其中

Among,

$$P_{ij} = \begin{cases} P(Y = Pos|X = x_j) & P(Y = Pos|X = x_j) \geq 0.5 \\ -P(Y = Neg|X = x_j) & P(Y = Pos|X = x_j) < 0.5 \end{cases}$$

$$C = min_i \sum_{j} P_{ij}$$

我们提出这个量化公式主要基于以下考虑和假设：

We propose this quantitative formula based on the following considerations and assumptions:

- 虽然评论中并不是每一个词都对评论情感判断有着很显著的帮助，但与此同时根据这些词计算出来的后验概率一般会接近0.5，比根据显著效果的词计算得到的后验概率（一般接近0.9或0.1）要小，在整体上这些无关紧要的词不会对整体量化的值影响很大。
- 对于情感为积极的词，我们施加一个“奖励”；对于情感为消极的词，我们施加一个“惩罚”。这便是我们为什么当判断为积极的后验概率大于0.5的时候整体分数加上这个概率，当这个概率小于0.5的时候减去对应的判断为消极的后验概率。
- 为了让这个分数在整体上更加平缓，我们用对数函数来减少极端值对整体带来的影响，也可以减少评论整体上的方差，因为对文本情感进行精确量化是很困难的。为了实现这个操作，我们在对评论中单词的后验概率进行求和之后加上一个常数C，使得所有的求和都大于0，之后再减去这个常数的对数，为了让这个分数在$x = 0$上下波动，如果分数>0，代表着该评论倾向于积极，如果分数<0，代表着该评论倾向于消极。

- Although not every word in the review has a significant effect on the sentiment judgment of the review, at the same time the posterior probability calculated based on these words will generally be closer to 0.5, smaller than the posterior probability calculated based on the significant effect words (generally close to 0.9 or 0.1), so these insignificant words will not affect the overall quantified value.

- For words with positive emotions, we impose a “reward”; for words with negative emotions, we impose a “punishment”. This is why we add the probability to the overall score when the positive a posteriori probability is greater than 0.5, and subtract the corresponding posterior probability when the probability is less than 0.5.

- In order to make this score more smooth overall, we use a logarithmic function to reduce the impact of extreme values on the whole, and also reduce the overall variance of the review, because it is difficult to accurately quantify the sentiment of the text. In order to achieve this operation, we add a constant C after summing the posterior probabilities of the words in the review, so that all the summations are greater than 0, and then subtract the logarithm of this constant. If the score > 0, it means that the comment tends to be positive, and if the score < 0, it means that the comment tends to be negative.

信息熵 Information Entropy

信息熵是表示随机变量不确定性的度量，设X是一个取有限个值的离散随机变量，其概率分布为：

Information entropy is a measure of the uncertainty of a random variable. Let X be a discrete random variable with a finite number of values. Its probability distribution is:

$$P(X = x_i) = p_i, i = 1, 2, \ldots, n$$

则随机变量X的信息熵定义为：

The information entropy of the random variable X is defined as:

$$H(X) = -\sum_{i=1}^{n} P_i log P_i$$

信息熵的值越大，随机变量的不确定性越大。而一个随机变量所包含的信息量与其不确定性有着直接的关系。比如说我们要搞清楚意见非常不明确的事，或者我们一无所知的事情，就需要了解大量的信息。所以从这个角度来看，可以认为信息量等于不确定性的多少。因为我们可以用信息熵来衡量一个随机变量所包含的信息。

The larger the value of the information entropy, the greater the uncertainty of the random variable. The amount of information contained in a random variable is directly related to its uncertainty. For example, we need to know a lot of information if we want to figure out something that is very ambiguous or we know nothing. So from this perspective, the amount of information can be considered equal to the amount of uncertainty. Because we can use information entropy to measure the information contained in a random variable.

特定质量描述与评级水平的关系 Relationship Between Specific Quality Descriptions and Rating Levels

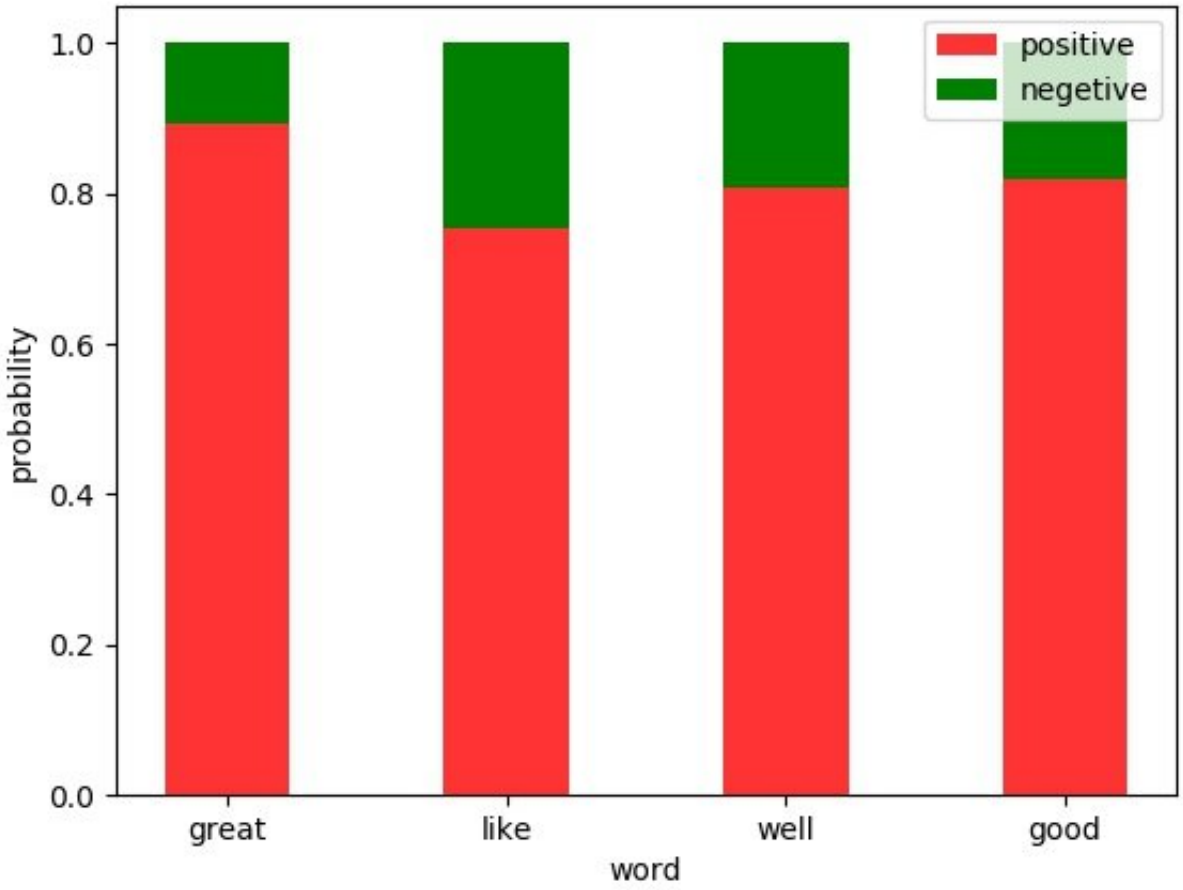
我们对于一些带有感情色彩的词（比如great, like, well, good, awesome），统计了当评论中出现这些单词的时候评级为优（4-5星）和劣（1-3星）的概率，我们发现，对于一些带有强烈感情色彩的词，评级水平往往与这些单词有着紧密的联系。

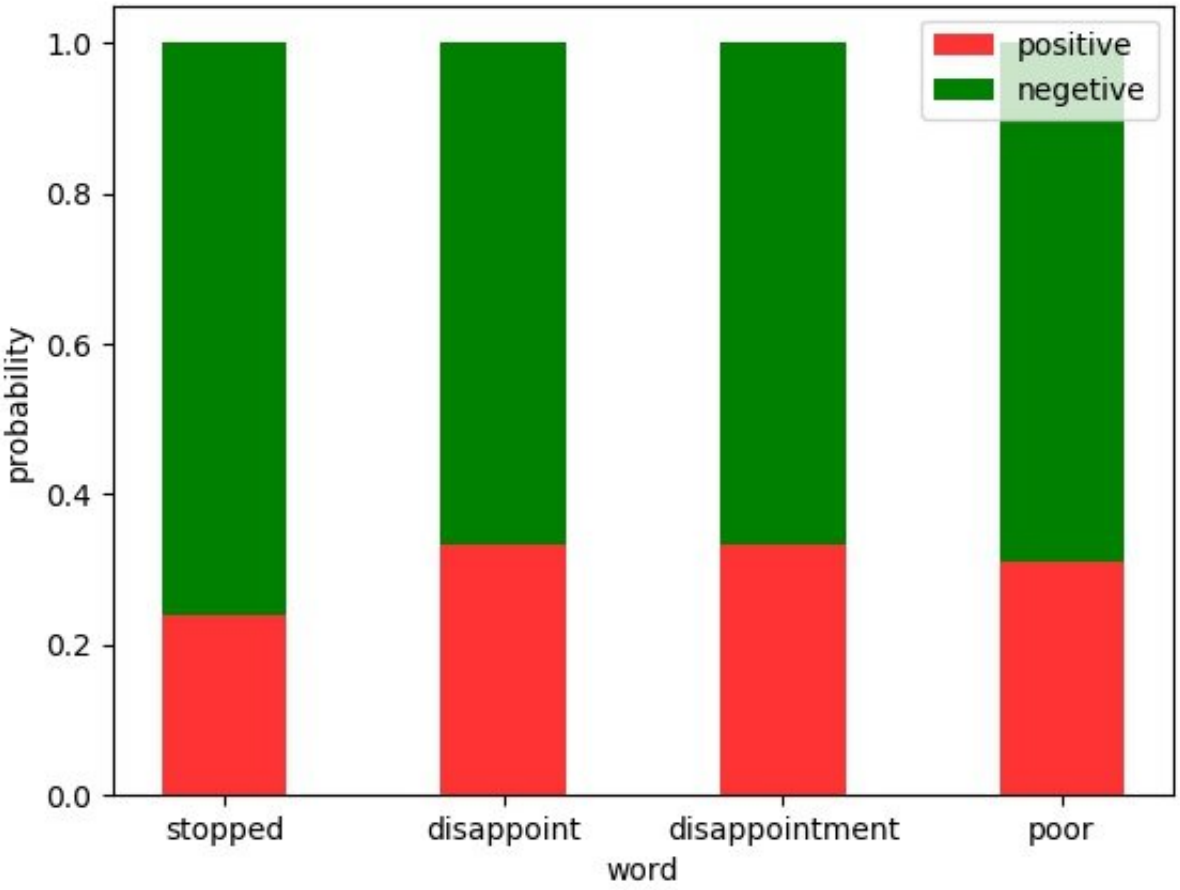
For some words with emotional colors (such as great, like, well, good, awesome), we counted the star ratings of excellent (4-5 stars) and poor (1-3 stars) when these words appeared in the reviews. Probability, we found that for some words with strong emotions, the rating rating is often closely related to these words.

我们对于几个带有积极感情色彩的词和消极感情色彩的词，统计了评论中出现这些词其评分质量为高和低的概率：

For several words with positive emotions and negative emotions, we counted the probability that these

words appear in the reviews with high and low quality star rating:





从图中我们可以看到，这些单词对于评分的区分度比较大，如果在评论中出现这些单词，那么相对应的评分往往会与其积极性或消极性挂钩。但并不是所有的词语都具有类似的区分度，比如一些与情感不相关的单词（price），或者一些中性词语（much），这些单词对于评分的区分度并不高，因此我们认为这些单词与评级水平并没有直接的关系。

Not all words have similar discrimination, such as some words that are not related to emotion (price), or some neutral words (much). These words are not very distinguishable for scoring, so we think there is little relationships between these words and star rating. In addition, we can see from the figure that there are some words that are more distinguishable for star ratings. If these words appear in the reviews, the corresponding ratings are often linked to their positivity or negativeness.