2020 MCM Problem C

Review Quantification

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.

Naive Bayes

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.

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.
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: $P(Y = c_k|X = x)$

According to Bayes formulaL

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

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:

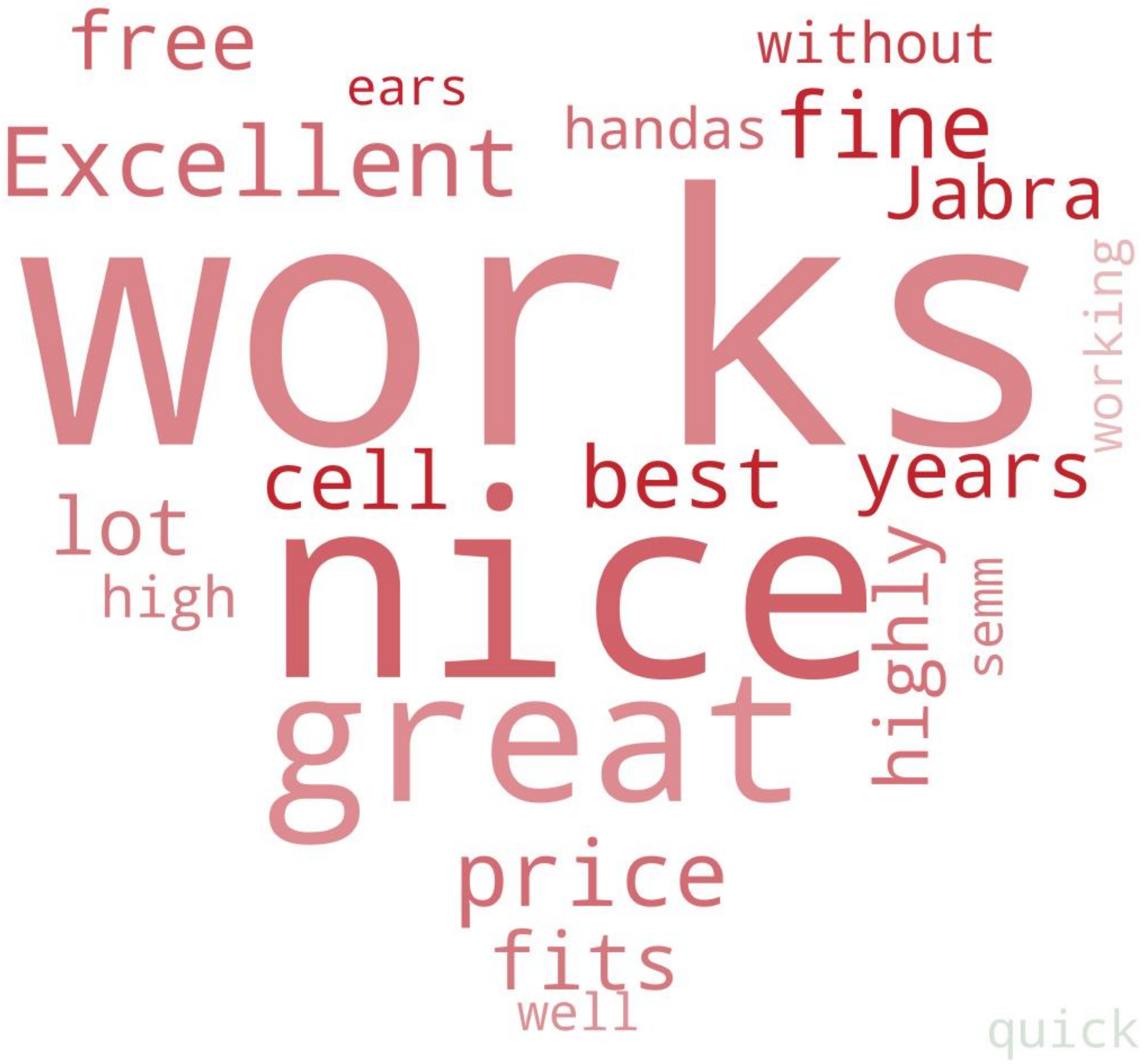
$$y = argmax_{c_k} P(Y = c_k)P(X = x|Y = c_k)$$

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.

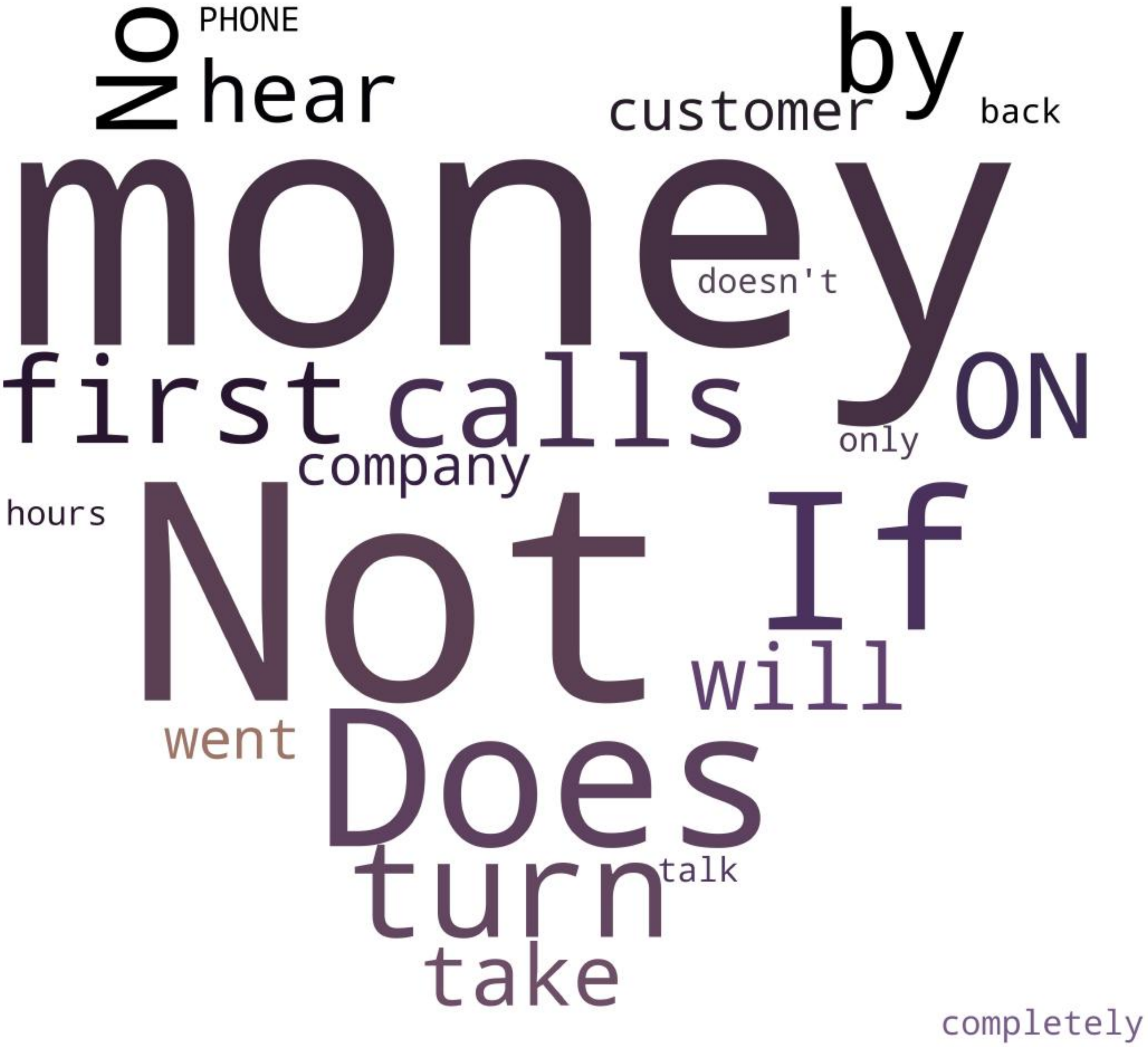
We searched the kaggle website for tagged Amazon review corpora as a training set to train our Bayesian classifier, and then use it to analyze the emotional tendencies of reviews.

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:

Positive word cloud:

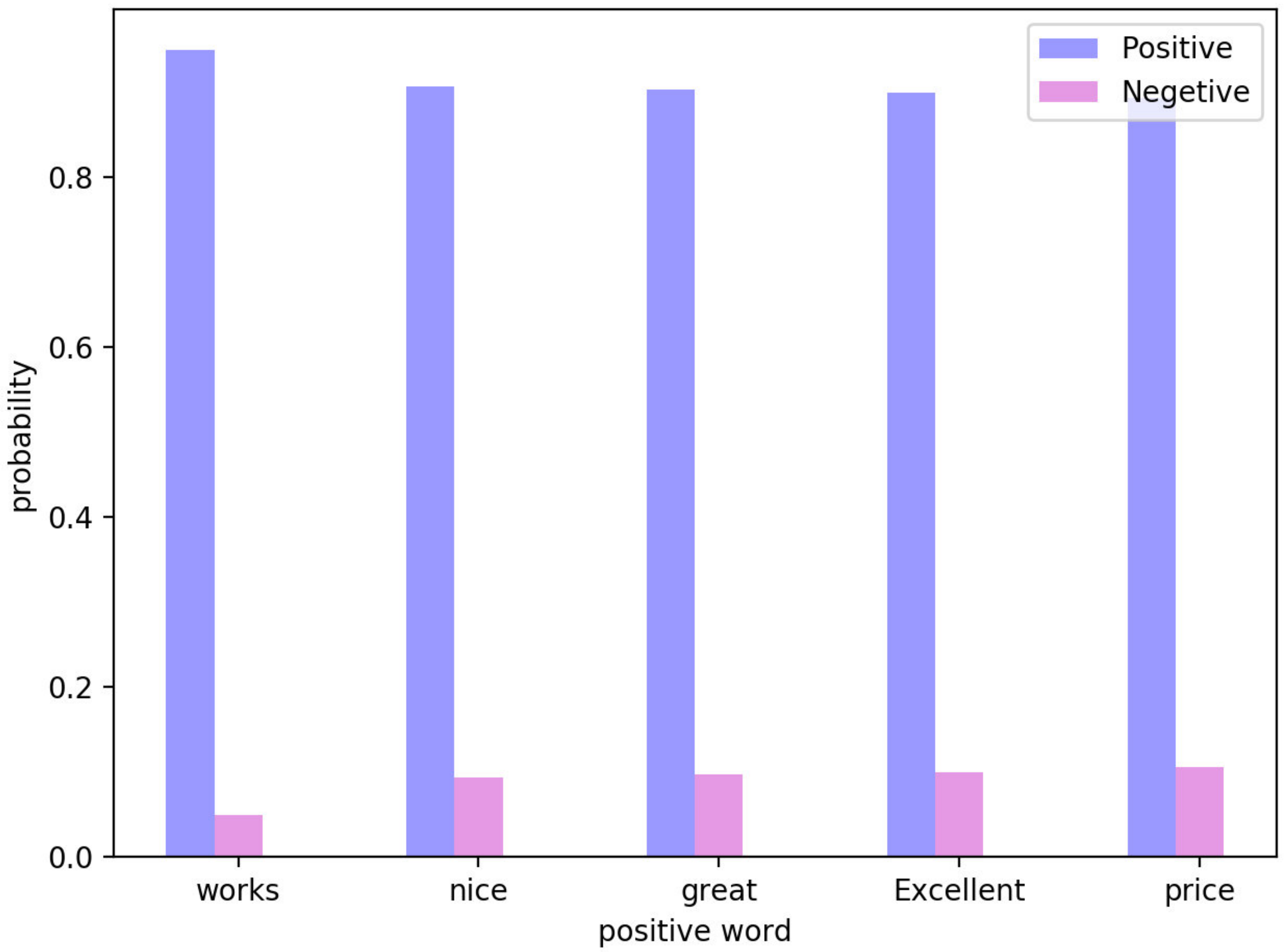


Negative word cloud:

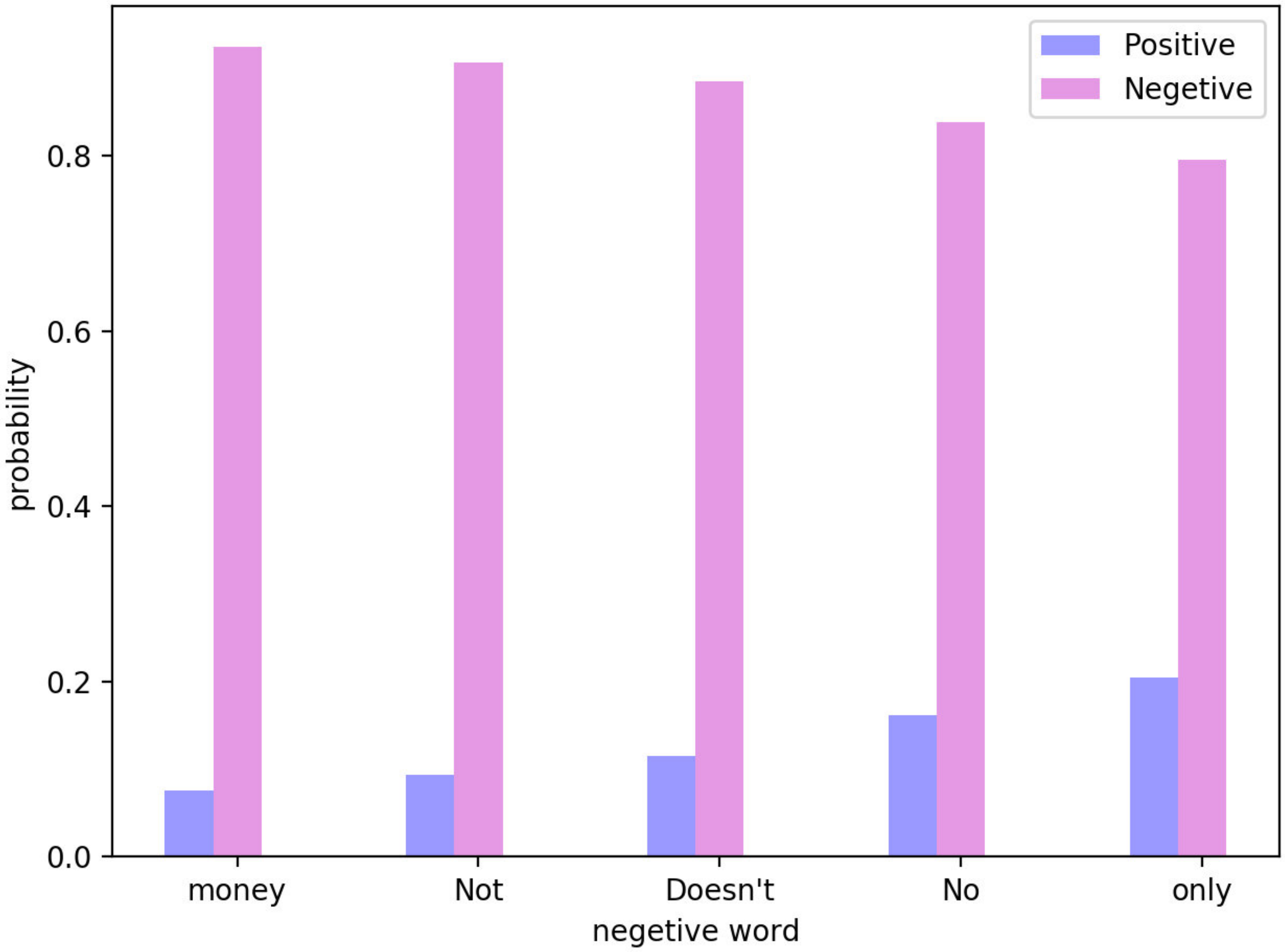


The larger the volume of a word, the more classification effect this word has. According to the above words with good classification effect, we can visualize their corresponding posterior probability.

Positive word:



Negative word:



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

Review Emotion Quantification

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.

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:

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

Among,

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

$$C = min_i \sum_j P_{ij}$$

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

- 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.

Information Entropy

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:

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

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

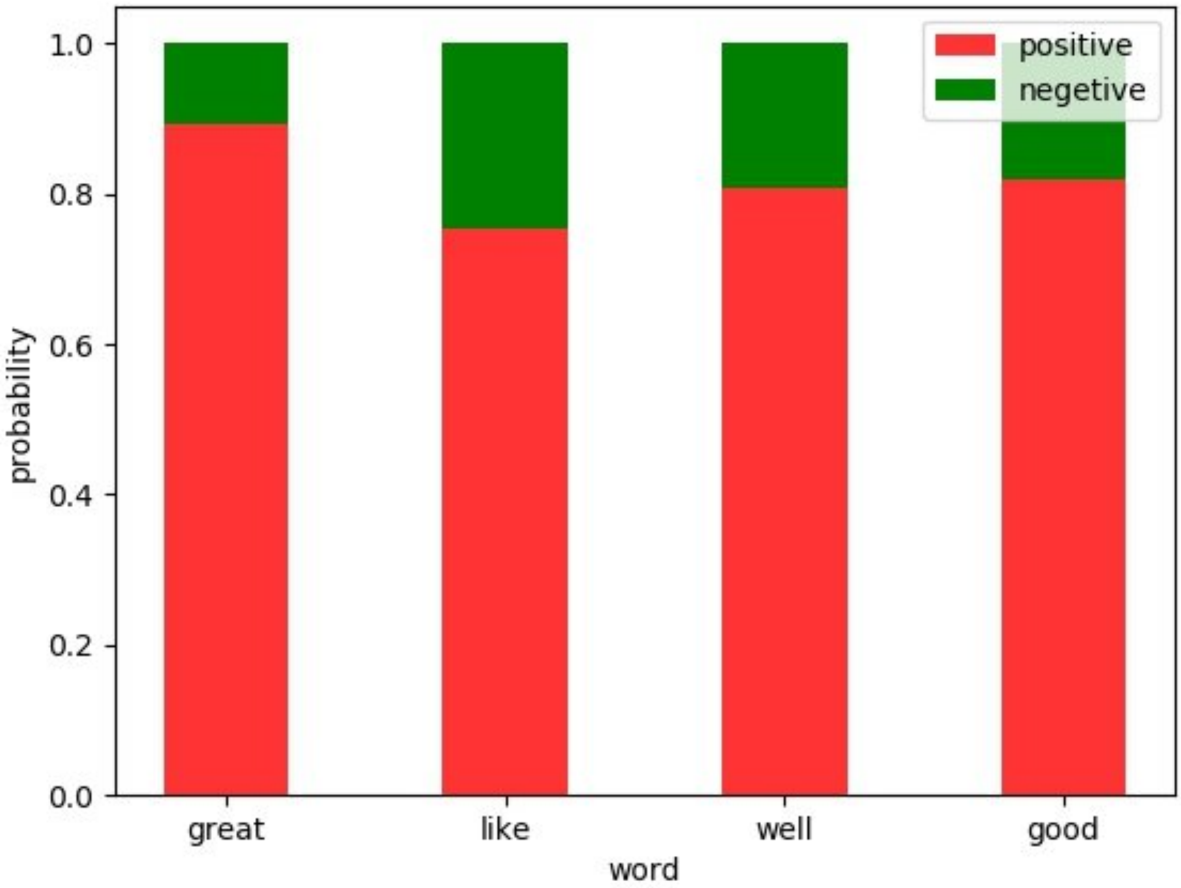
$$H(X) = -\sum_{i=1}^{n} p_i log p_i$$

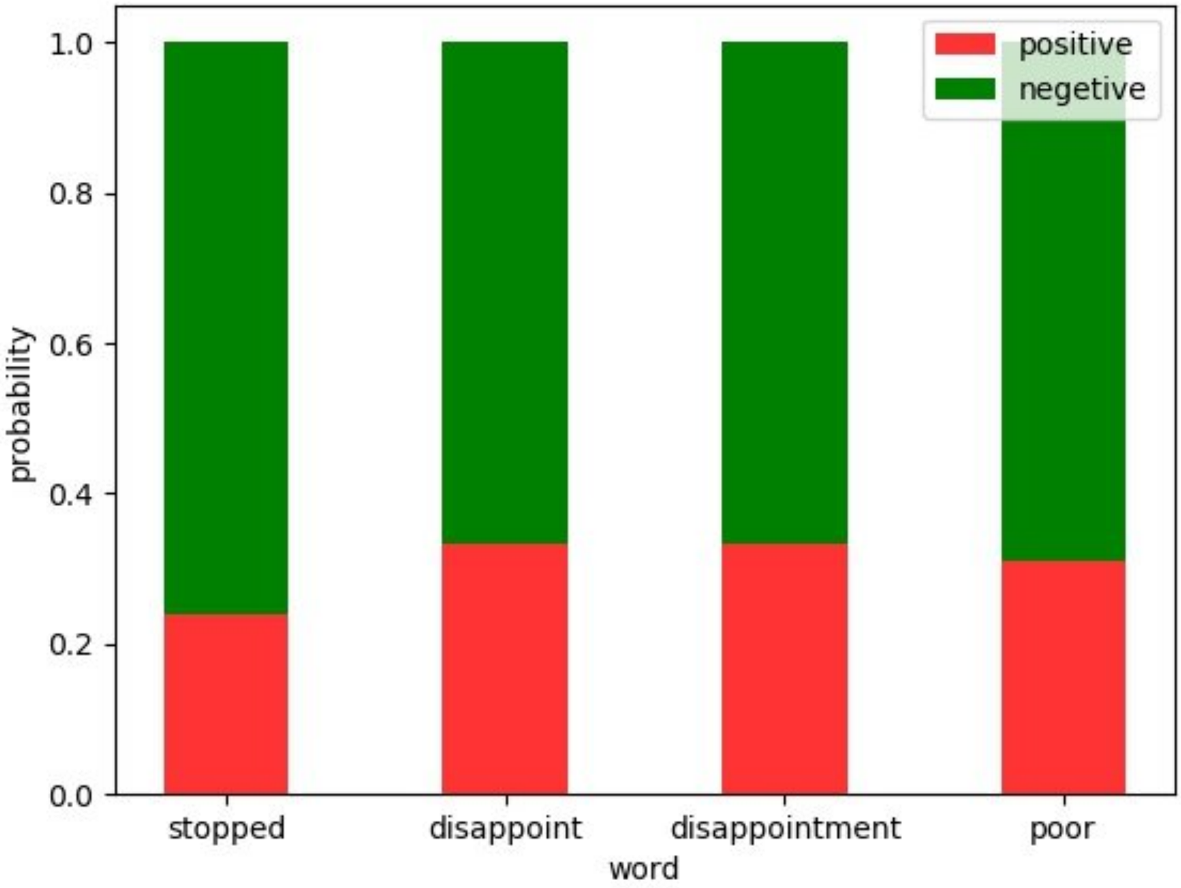
The larger the value of the information entropy, the greater the uncertainty of the random variable. The amount of information contained in a random variable is directly related to its uncertainty. For example, we need to know a lot of information if we want to figure out something that is very ambiguous 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.

Relationship Between Specific Quality Descriptions and Rating Levels

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.

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:





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.