**Report: How to detect fake online shopping reviews using linguistic features**

**Shiyan Wei**

**300569298**

**LING226 - Special Topic: Introduction to Computational Linguistics (CRN 15285)- 2022**

# **Introduction and Research Questions**

With the COVID breakout, more and more people will use online shopping instead of physical store shopping since it is one of the most effective ways to keep people physically apart and to block the virus from spreading. In most cases, people will check the previous comments before buying online. Reviews are mixed, but some have not come from the actual user and have been made to mislead customers to sell products. To increase sales of their products, some sellers create false customer reviews. However, another incorrect comment was posted by sellers from their competitors in an attempt to reduce their sales. This can lead to distrust in reviews as customers need to find out whether the comments are genuine.  Hence, it is important for customers to identify product reviews before making a purchase, and it is also important for online stores to vet reviews to ensure accuracy and build trust. Using linguistic characters, Python-based methods enable the detection of fake online reviews, allowing us to acquire genuine and useful feedback and avoid purchasing low-quality products.

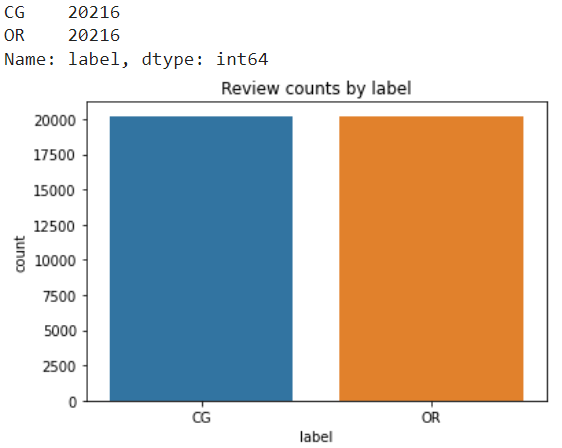
In this project, I will use NLTK to identify the characteristics of fake comments in linguistics by comparing real and fake reviews. NLTK offers us many suitable methods to detect fake online shopping reviews, such as sentiment analysis, and identifying particular words, phrases, or topics that appear only in fake reviews. All these methods will enable us to compare the language structure, complexity, and content of the reviews to detect any irregularities that indicate a fake review. Through this research, I aim to understand better the linguistic characteristics that distinguish between fake and real online shopping reviews. Additionally, I can better understand which techniques are more accurate and less effective by exploring how the fake review detection methods used by NLTK perform compared to the manual processes for identifying fake reviews.

# **Data and description of data**

* 1. Data label and sizes:

This data set contains fake and genuine online shopping reviews, with over 20,000 reviews generated by computers and 20,000 reviews caused by humans combined. Therefore, there will be enough data to distinguish between real and fake reviews, and the results will be more accurate. The data set consists of two labels: OR and CG, where the OR label is for original reviews and the CG label is for computer-generated reviews, and it includes a total of 40,000 reviews.

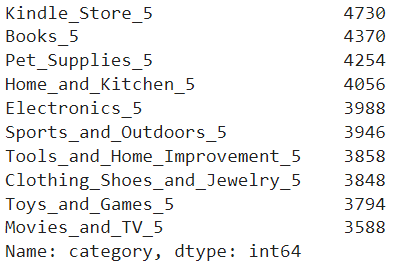
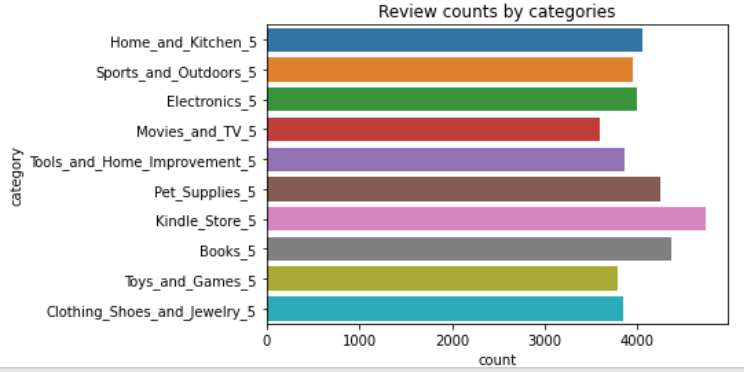
Figure 1:Reviews counts by label



* 1. Data categories:

There are ten categories, including apparel, sports, and home. Each category has a minimum of 3,500 reviews.

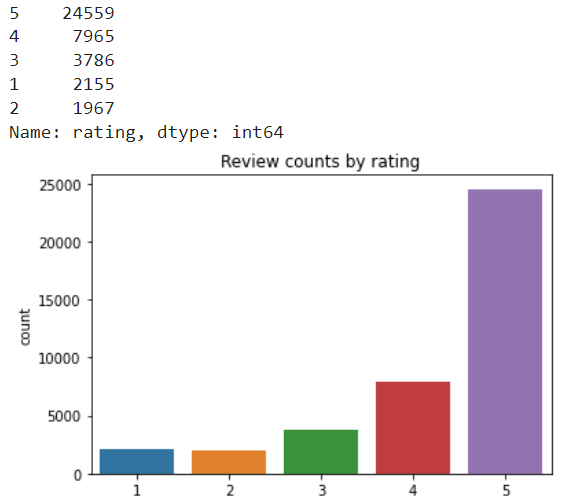
Figure 2: Reviews counts by categories



* 1. Data rating:

Review text data is scored on a scale from 1 to 5. Our data set can view the proportion of different rating levels. A rating of 1 indicates a negative review, while a rating of 5 shows a positive opinion. We can use these ratings to analyse customer sentiment by comparing the distribution of rating levels to determine whether the majority of customers have a favorable or unfavorable opinion.

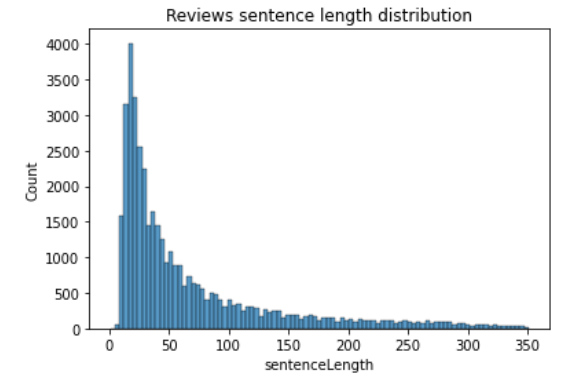
Figure 3: Reviews counts by ratings



* 1. Data type and length of sentences

In addition, since our data set is a CSV file downloaded from OSFHOME, we sorted the length of reviews within this data set. This classification allowed us to comprehend the length and structure of the customer reviews' sentences. We used Pandas to present the data and sort the evaluation's sentence length. The majority of paragraphs in this collection contain fewer than 50 sentences.

Figure 4: Reviews sentence length distribution



# **Explanation of the program**

* 1. Project summarises

The project considers four key steps of data analysis: data collection, data cleaning, data analysis, and interpreting results and reflection. First, we gather the fake review datasets online and import them using Pandas into the collaboratory. Then, we carry out data cleaning and pre-processing to ensure the accuracy of the analysis. We will use tokenisers, regular expressions, and other techniques to clean the data. After the data cleaning and pre-processing, we build models to classify the reviews as genuine or fake. Following that, we apply natural language processing techniques such as word frequency analysis and sentiment analysis to detect anomalies in the data. Finally, we use different classification algorithms to detect fake reviews and interpret the output to arrive at meaningful conclusions.

* 1. Preprocess data in NLTK

Before we can do more analysis, we need to clean up our data. NLTK's regular expressions and tokeniser will help us clean up the current data. For example, removing unnecessary punctuation, special characters, extra spaces, and other elements will help to treat each text equally. After removing all these elements, we can also tokeniser the data and apply lemmatisation or stemming to clean up the dataset further. All of these steps, when followed together, can help us create a more efficient dataset.

* 1. Clarify the test set

For the final test set, I selected 2,547 reviews with sentence lengths under 50 from the Home and Kitchen category. Since most of the comments are in this range, and this trend is clear from the graphs above or check them at the python program, it makes sense to use this data to figure out which statements are honest and which are not.

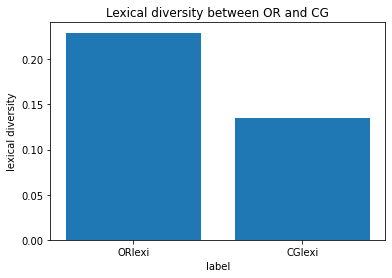
* 1. Analysis methods

The implementation code will be written in Python and executed in collaboratory; the following analysis methods will be used to identify the linguistic characteristics of fake reviews.

* + 1. Lexical Diversity
    2. Frequency Distribution Analysis
    3. Sentiment Analysis
    4. Latent Semantic Analysis
    5. More…

# **Result and Discussion**

* 1. Lexical diversity
     1. Comparing the number of words used in reviews written by humans and computers in the sample dataset shows that human-written reviews have more words than computer-written reviews.
     2. The original has a score of 0.225 for lexical diversity, while the version made by a computer only has a score of 0.134.

Figure 5: Lexical diversity between OR and CG

* 1. Frequency distribution
     1. Quality words like "great" and "nice" are among the top 20 most frequent words in both genuine and fake reviews. This indicates that our evaluation texts may contain more positive language. Figures 6 and 7 show the first frequency words of human-created and computer-generated reviews.

Figure 6: The graph of the first 20 frequency words in the original reviews

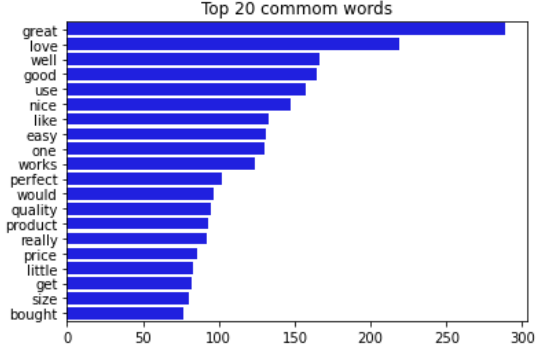
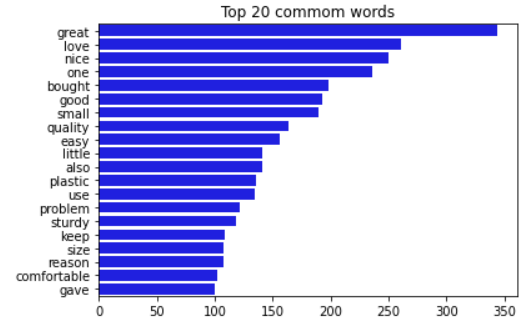
****

Figure 7: The graph of the first 20 frequency word in the computer-generated reviews



* 1. Word Cloud
     1. By comparing the frequency words, we can see that the frequency words in both real and fake reviews are very similar and that the word "great" appears frequently. However, words that describe specific household and kitchen items, such as "knife", is rare. This may indicate that computer-generated reviews focus more on general sentiments than detailed descriptions of the product, as they need to be more capable of accurately capturing the product's specific characteristics.
     2. These observations are in Figures 8, 9, and 10.

Figure 8: word cloud for test set



Figure 9: word cloud for human-created reviews



Figure 10: word cloud for computer-generated reviews



* 1. Sentiment analysis
     1. Sentiment By Rating Of reviews

Identify the distinctions between rating and sentiment. The graph below shows that most reviews are positive, whether generated by humans or computers.

The rating scale is: 4 or 5 is positive, 3 is neutral, and 1 or 2 is negative.

Figure 11: Sentiment scores in human-created reviews by rating

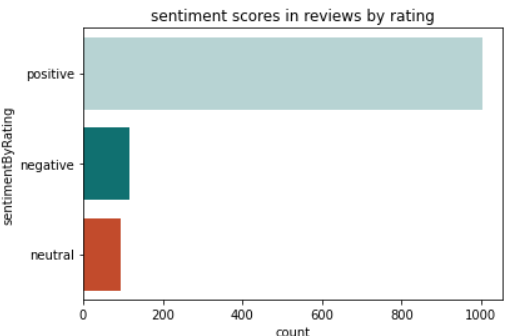
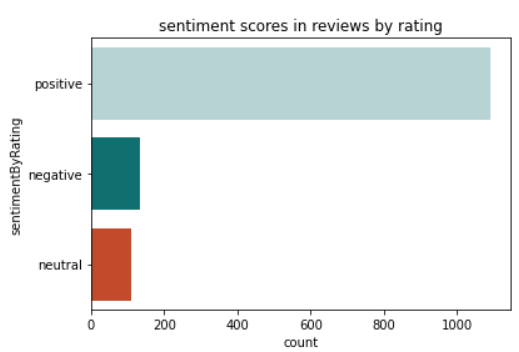


Figure 12: Sentiment scores in computer-generated reviews by rating



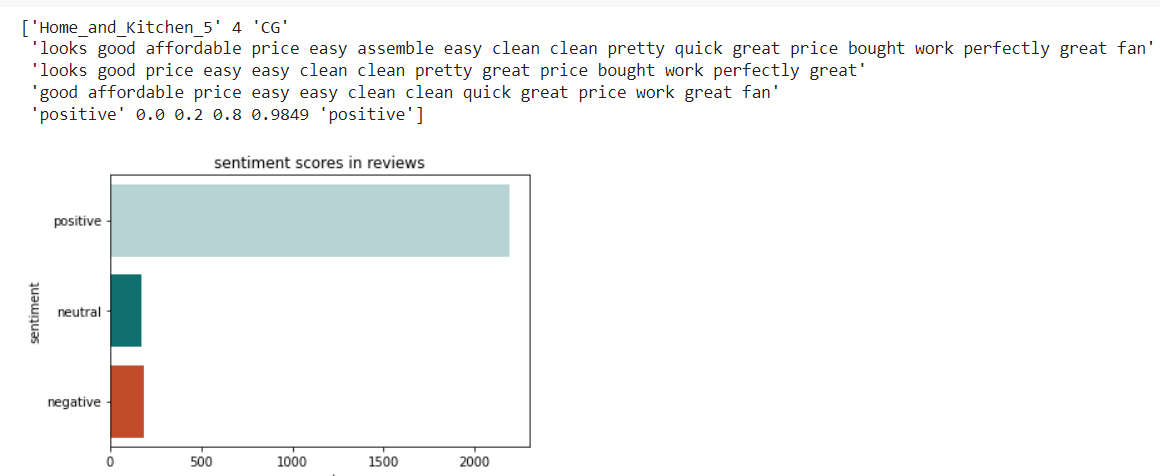
* + 1. VADER Lexicon

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based tool for figuring out the polarity of sentiment in text data.

Using VADER to analyse review sentiment to calculate the polarity score, the sentiment distribution is similar to the previous rating sentiment distribution. The majority of reviews are positive.

No matter if the reviews are real or not, we can assume that most of the products are good.

Figure 13: Sentiment score in all reviews by VADER



* + 1. Article length and polarity

When the text length of human-created reviews is compared to that of computer-generated reviews, the text length of human-created reviews is longer. This suggests that when people write reviews, they are more likely to provide details, which could lead to a more accurate product review. (Figure 14). However, the trends of sentiment polarity indicate that there is no significant difference between real and fake reviews in terms of sentiment. This implies that the increased detail and length of a human-generated review do not necessarily lead to an increase in polarity, and may indicate that people are more likely to provide more information than computers. (Figure 15, 16). This suggests that the additional details provided by human-generated reviews do not necessarily increase the accuracy of reviews, as the sentiment polarity remains the same regardless of the length or details.

Figure 14: Distribution of article length

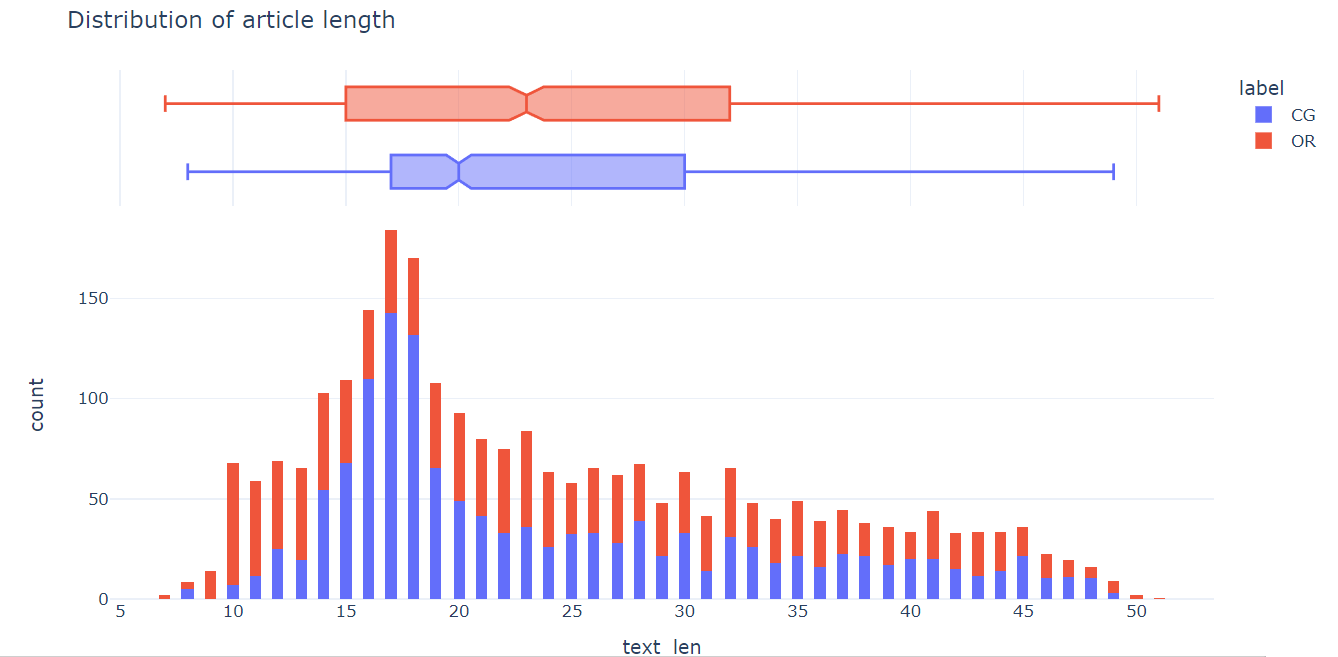


Figure 15: Polarity by label

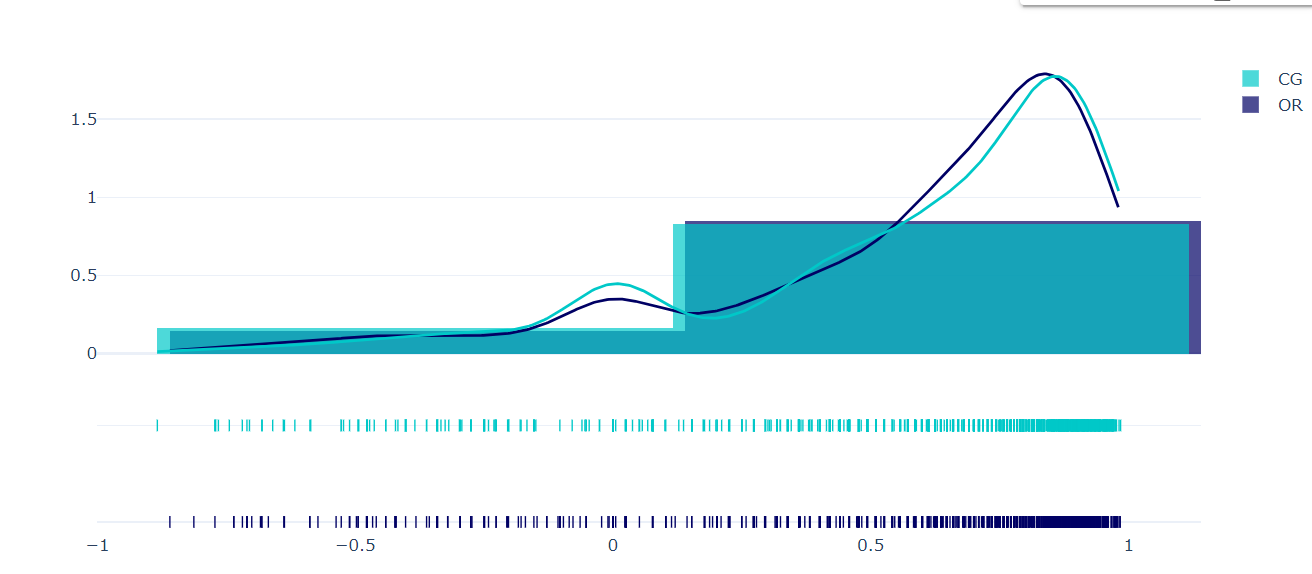
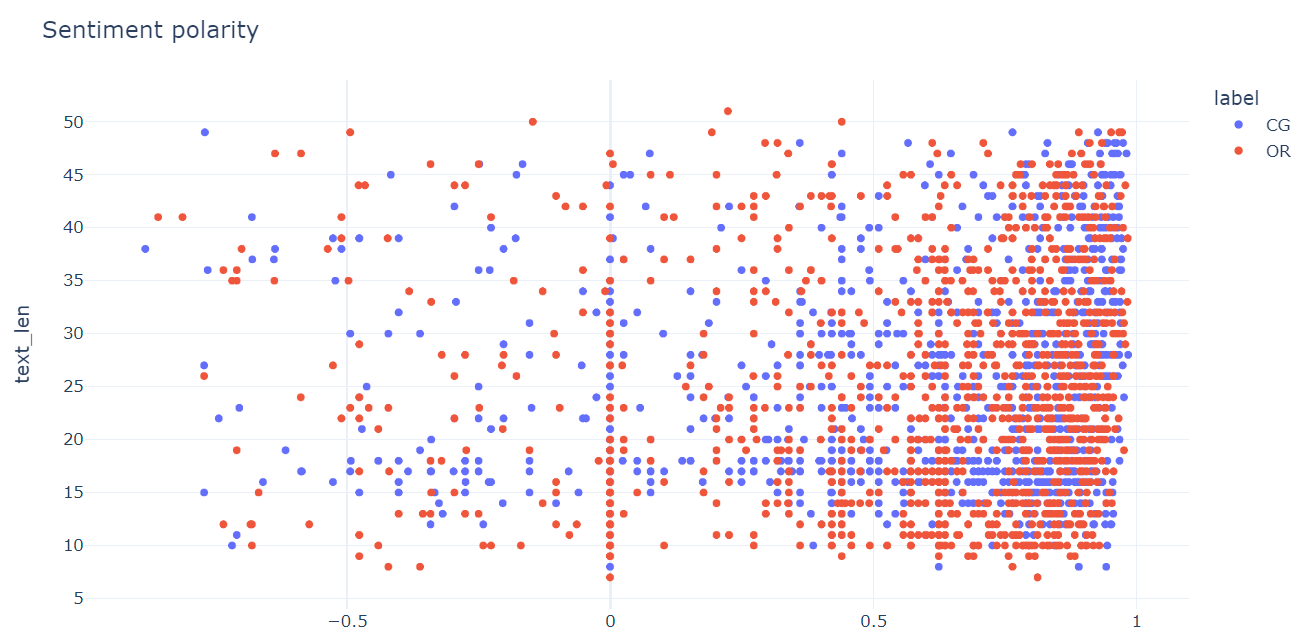


Figure 16: Sentiment Polarity Scatter



* 1. Bigram

Find out which words are often used together by using Bigram to figure out what the data set is about.

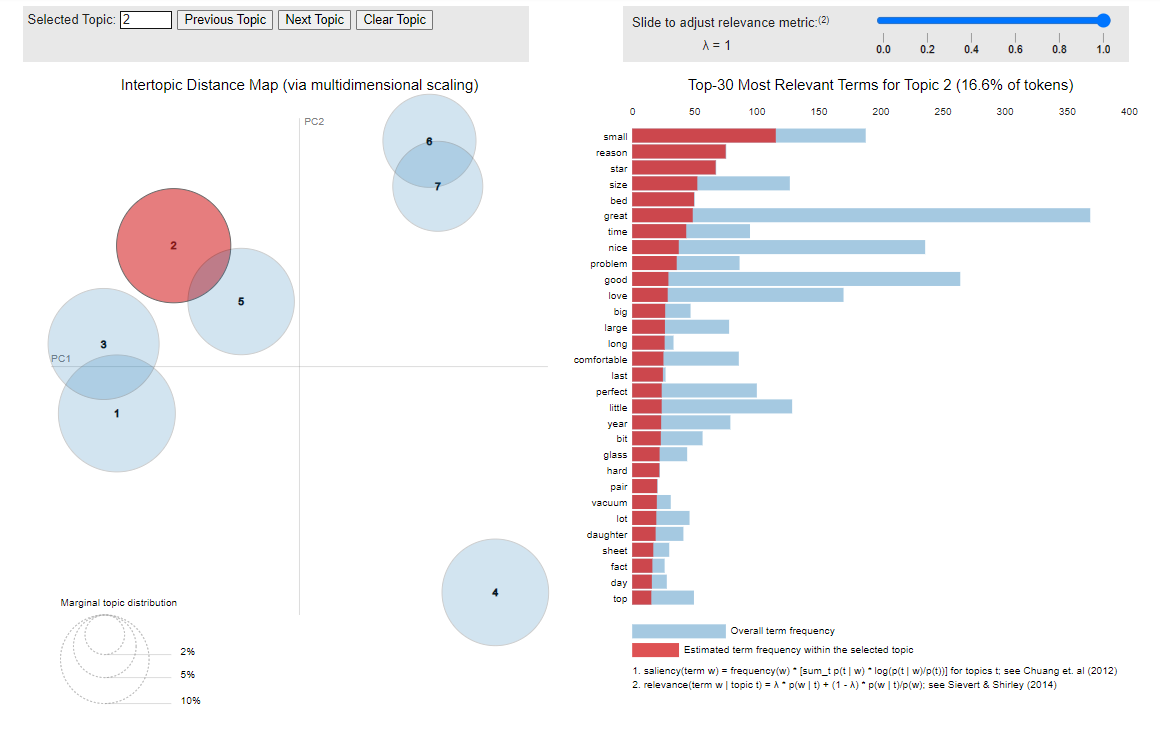
Comparing the human-written reviews to those generated by a computer reveals that the original reviews focused more on the product's usability, as there are many possible combinations for phrases such as "works great." On the other hand, computer reviews are more concerned with why the product should be purchased and its appearance. The use collocations less frequently in computer-generated reviews. However, no phrase matches the product description in the home and kitchen category, and genuine and fake reviews may not be related to the product in any category, regardless of whether it is a real or fake review.

* 1. Latent Semantic Analysis (LSA)

By applying text lemmatization to the reviews, it can reduce the noise and identify meaningful relationships among them

LSA (Latent Semantic Analysis) can be used to further analyze the relationships between reviews and product descriptions in the home and kitchen category to find a theme that most closely relates to the product. This method can help identify relationships between reviews and the product in any category, whether real or fake. Topic 2 might, for example, discuss the size of a product, and the word small appears several times.

Figure 17: LSA in topic 2



# **Reflection**

Using NLTK and Python, we discovered that there are linguistic differences between true and false reviews, such as the fact that true reviews have longer texts than fake reviews and that the topics of discussion in true reviews tend to be more specific than those of fake reviews. However, to ensure the accuracy of our results, we still need additional resources to support our outcome. First, we need to find out whether there was selection bias in the original data collection for our dataset and whether this could have affected our results. Second, it is impossible to determine when the product was reviewed or when the review was generated, which could have a significant impact on the similarity of the review. Third, the test set was chosen based on what I knew, which may have affected the final results. Lastly, we need to improve some parts of the technology we use. For instance, consider conducting additional research on the distinction in sentiment analysis between genuine and fabricated positive reviews. Alternatively, we can compare the similarities between real and fake reviews. This could lead to a better understanding of the differences in sentiment analysis, which could be used to improve our algorithms.