Final Report

1. What is the name of your project? Please write it as a research question and provide a short description.

Project title: "Analysis of the Palestinian-Israeli Conflict from Chinese and English Language Perspectives"

research question:
What are the differences between Chinese and English media's coverage of the Israeli-Palestinian conflict? Do two different languages have an impact on the same thing? For example, I subconsciously feel that Chinese reports will be more conservative and English reports will be more exaggerated.

short description:
This project aims to analyze and compare Chinese and English media reports and perspectives on the Palestinian-Israeli conflict. By crawling and analyzing Chinese and English reporting data on the Palestinian-Israeli conflict in Wikipedia and Google News, the purpose is to reveal reporting biases and differences in opinions under different language backgrounds. The expectation is to use data analysis technology in python as a programming language to specifically implement the entire process and quantify the differences.

2. What data did you collect? How did you collect it? How many data samples did you collect?
a. Specify exact data sources and your approach.
b. Describe what has been changed from your original plan, what challenges you encountered or resolved.

I collected data from Wikipedia and Google News and used the yifan_scrape_wikipedia_page function to scrape content about the Israeli-Palestinian conflict from English and Chinese Wikipedia pages. The yifan_save_news_to_csv function was used to extract news reports about the Israeli-Palestinian conflict from the RSS feed of Google News. The data sample is the entire page content of the Chinese and English entries on the Palestinian-Israeli conflict, as well as 100 Chinese Google news items and 100 English Google news items.

a. The exact data source is:
Wikipedia Palestinian-Israeli Conflict English Version
https://en.wikipedia.org/wiki/Israeli%E2%80%93Palestinian_conflict
Wikipedia Palestinian-Israeli conflict Chinese version
https://zh.wikipedia.org/zh-cn/%E4%BB%A5%E5%B7%B4%E5%86%B2%E7%AA%81",

Google News English version of the Israeli-Palestinian conflict
https://news.google.com/rss/search?q=Israeli-Palestinian+conflict&hl=en-US&gl=US&ceid=US:en
Google News Chinese version of the Israeli-Palestinian conflict
https://news.google.com/rss/search?q=%E5%B7%B4%E4%BB%A5%E5%86%B2%E7%AA%81&hl=zh-CN&gl=CN&ceid=CN:zh-Hans

The method is to write two functions for data acquisition.
yifan_scrape_wikipedia_page function. This function is designed to grab data from a specified Wikipedia page and save it to a CSV file. It is mainly used to collect title and paragraph content on the page. Use requests.get(url) to send an HTTP request to the specified URL to obtain the page content. Verify that the response's status code is 200 to ensure the request was successful. Use BeautifulSoup to parse the response content and find all titles and paragraphs (`h1`, `h2`, `h3`, `h4`, `h5`, `h6`, `p`). Iterate through each found element and extract its type (such as `h1`, `p`, etc.) and text content. Save the extracted data to a CSV file, with each row including the type and content of the element.

yifan_save_news_to_csv function. This function is used to grab news items from the Google News RSS feed and save them in CSV format. It is mainly used to collect news titles, links and publication dates. Use requests.get(rss_url) to send a request to the RSS source to obtain news data. Make sure the response status code is 200, indicating the request was successful. Use xml.etree.ElementTree to parse the XML content of RSS sources and extract news items. For each news item, extract the title, link, and pubDate. Save the extracted news items to a CSV file, with each line containing a news item's title, link, and publication date.

b. In the original plan, I planned to use social media data, specifically comments on Twitter. Because the acquisition of social media data has more complexities and limitations. I turned to data from Wikipedia and Google News. The challenge is that scraping social media data usually requires going through official APIs, which often have strict usage restrictions and quota limits. Obtaining large amounts of data may require special access rights or face cost issues. My solution was to choose more stable data sources and turn to Wikipedia and Google News as data sources, probably because they provide more stable and reliable information.

My data preprocessing code is a text processing pipeline for cleaning, word segmentation, and feature extraction of Chinese and English text. First, it defines a Chinese text cleaning function that retains text with only Chinese characters. Then, there is a word segmentation function that uses the jieba library to segment Chinese and simply split words for English, and removes stop words in both languages. Next, the pipeline processes four different datasets, two in Chinese and two in English, including Google News and Wikipedia content about the Palestinian-Israeli conflict. For each data set, it first reads the file, then cleans the text, performs word

segmentation processing, and finally uses the TF-IDF (term frequency-inverse document frequency) method to convert the text data to facilitate subsequent text analysis or machine learning applications. The processed data is saved to a new CSV file.

3. What kind of analysis and visualizations did you do?
a. What analysis techniques did you use, and what are your findings?
1. Keyword and word frequency analysis: Extract keywords and statistical word frequencies of Chinese texts through the jieba library. For English texts, use the TF-IDF method to extract keywords and statistical word frequencies.

2. Time series analysis: Based on the publication date of news articles, analyze the distribution of data in different months.

3. Text similarity analysis: Use TF-IDF vector and cosine similarity method to calculate the similarity between texts.

4. Sentiment analysis: Use the SnowNLP library to perform sentiment analysis on Chinese texts, and use the TextBlob library to perform sentiment analysis on English texts to evaluate the emotional tendency of the text.

5. Topic modeling analysis: Use the LDA (Latent Dirichlet Allocation) model to perform topic modeling analysis on the text data to identify the main topics in the text.

6. Language style analysis: Analyze the vocabulary richness, average sentence length, most common words and part-of-speech distribution of the text to evaluate the language style of the text.

During the entire analysis process, various natural language processing tools and methods were used, such as jieba, TextBlob, SnowNLP, gensim, NLTK, etc. The data processed included Chinese and English news articles and Wikipedia content.

Here are my findings based on experimental results.
1. Keyword and word frequency analysis: This part reveals the most frequently occurring and important words in the text. These keywords and word frequency data can help us quickly understand the topics and focus of each data set. According to the experimental results, Chinese and English are similar

2. Time series analysis: By analyzing the number of articles in different time periods (for example, by month), the time trend of news reports or article releases can be revealed. We found that the attention in Chinese and English is different. Chinese first increased and then decreased, while English continued to increase.

3. Text similarity analysis: This analysis helps identify the degree of similarity

between texts. High similarity between texts may indicate that they discuss similar topics or events, while low similarity may imply differences or diversity in content. The similarity between Chinese and English is very high.

4. Sentiment analysis: Sentiment analysis results show the emotional tendency in the text, such as positive, negative or neutral. This can help to understand the emotional tendency of news reports or articles. According to experiments, we were surprised to find that both Chinese and English are very objective, and the attitudes tend to be neutral.

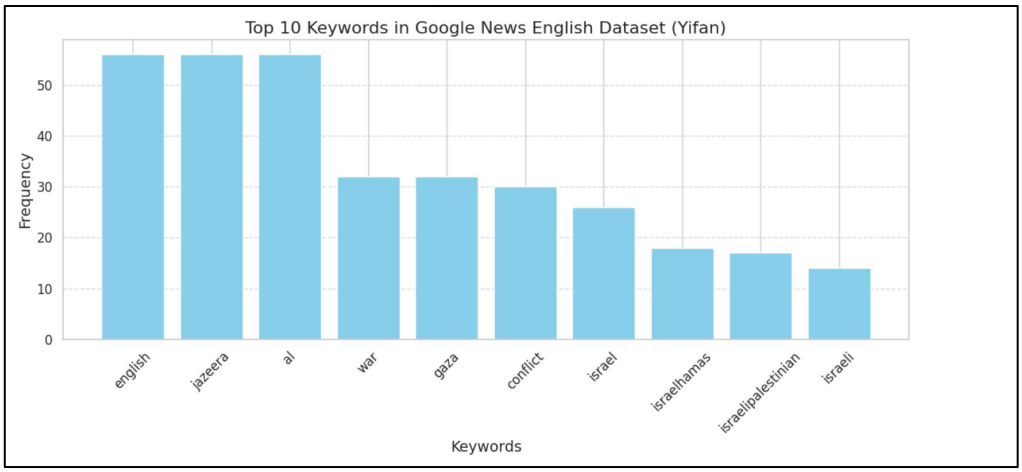
5. Topic modeling analysis: The topics identified through the LDA model can reveal the main topics or discussion points in the text. This type of analysis is particularly useful for understanding the distribution of topics in large text collections, such as news collections or long-form articles. We found that the themes in Chinese and English are also similar.

6. Language Style Analysis: This analysis reveals the lexical diversity, sentence structure complexity, and grammatical features of the text. For example, high vocabulary richness may indicate that the author uses a wide range of vocabulary, while average sentence length and part-of-speech distribution may reflect the author's writing style and the complexity of the text. The language style of English on Wikipedia is richer than that of Chinese. I think it is because more people are sharing English. As for Google News, the richness of Chinese language style is greater than that of English.

c. Describe the figures you made. Explain its setup, meaning of each element.

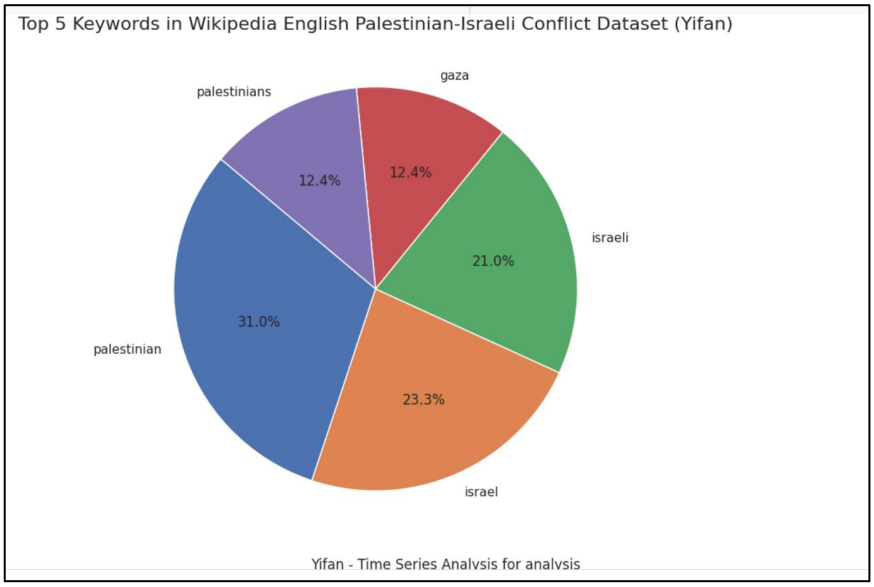
1. Visualization of keyword and word frequency analysis:

Bar chart: Shows the top 10 keywords and their frequencies in the Google News English data set. The bar chart uses sky blue bars to represent the frequency of each keyword. The x-axis is the keyword and the y-axis is the frequency. This helps to quickly identify the most common topics in the dataset.



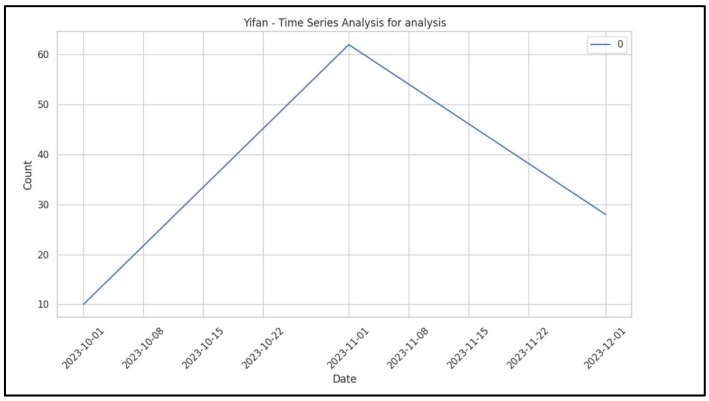
Word Cloud Chart: A word cloud was generated for the Wikipedia English Sino-Pakistani Conflict Dataset. Word cloud graphs generate visualizations of words based on their frequency, with more frequent words appearing larger in the graph. This chart form provides a visual representation of the main topics in the data set.

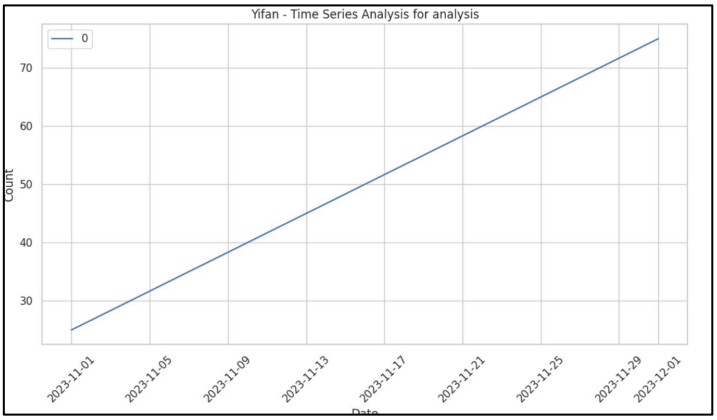
Pie chart: Also for the Wikipedia English China-Pakistan conflict data set, the frequency distribution of the top five keywords is shown. Through pie slices of different sizes and percentage labels, you can visually see the proportion of each keyword in the whole.



2. Visualization of time series analysis:

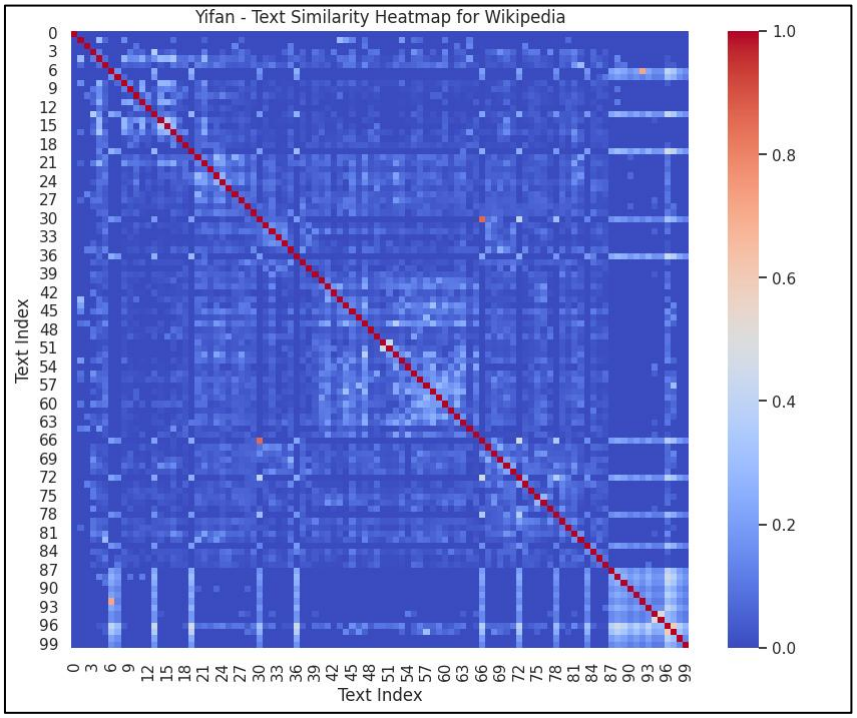
Line chart: shows the trend of the number of articles published in the Google News Chinese and English data sets over time. The x-axis of the line chart is time (divided by month), and the y-axis is the number of articles, which is helpful for observing article publishing trends in different time periods.





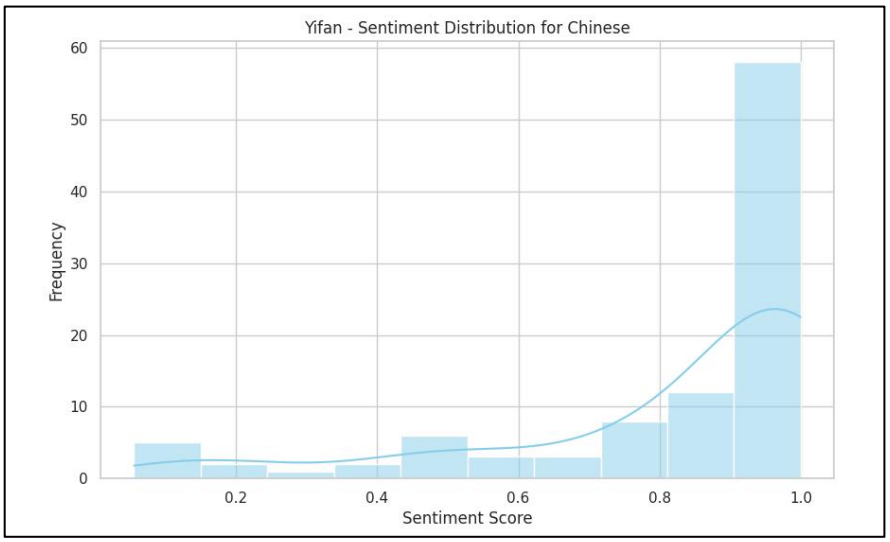
3. Visualization of text similarity analysis:

Heat map: shows the similarity scores between different texts. Each grid in the heat map represents the similarity between the two texts, and the warmer the color, the higher the similarity. This helps to quickly identify similar or related text in the dataset.

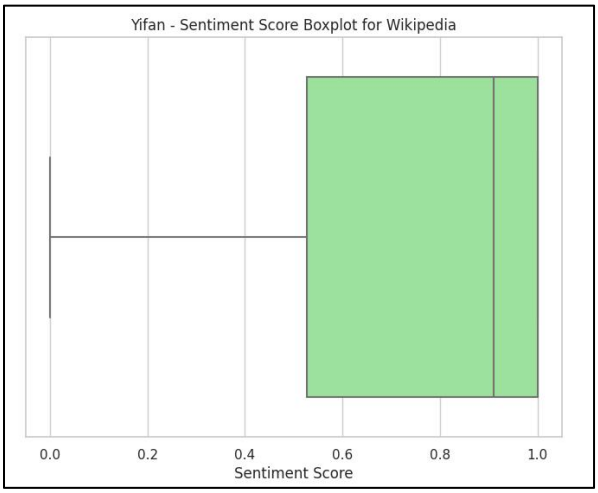


4. Visualization of sentiment analysis:

Histogram: shows the distribution of sentiment scores. The x-axis of the histogram is the emotion score, and the y-axis is the frequency. Through the graph, you can observe the overall distribution of emotional tendencies in the data set.



Boxplot: Shows a statistical summary of sentiment scores, including median, interquartile range, etc. Boxplots provide more detailed information about the distribution of sentiment scores in a dataset.



For reasons of space, I only show some images here. For the six analysis methods, I chose four of them for visual display, which I think is enough to explain the experimental results.

d. Describe your observations and conclusion.

1. Keyword and word frequency analysis: This part reveals the most frequently occurring and important words in the text. Experimental results show that Chinese and English datasets show similarities in terms of keywords and word frequency, which may mean that texts in different languages tend to use similar vocabulary when dealing with the same topic.

2. Time series analysis: By analyzing the number of articles in different time periods, the time trend of news reports or article releases is revealed. The experimental results show that the attention of the Chinese data set first increased and then decreased, while the attention of the English data set continued to grow, which may reflect readers' interests in different language environments or different stages of event development.

3. Text similarity analysis: This analysis helps identify the degree of similarity between texts. The experimental results show that the similarity between Chinese and English texts is very high, which may mean that even in different language environments, the texts still maintain a certain consistency when discussing similar topics or events.

4. Sentiment analysis: Sentiment analysis results show the emotional tendency in the text. The experimental results unexpectedly showed that both Chinese and English texts showed a more objective and neutral attitude, which may indicate that news reports strive to maintain a certain degree of objectivity and neutrality when conveying information. Chinese is more positive than English on Wikipedia, but the degree of positivity is not high.

5. Topic modeling analysis: The topics identified by the LDA model reveal the main topics in the text. Experimental results show that the themes of the Chinese and English datasets are similar, which may reflect common global or cross-cultural concerns.

6. Language style analysis: This analysis reveals the lexical richness, sentence structure complexity and other characteristics of the text. The experimental results show that the English text on Wikipedia is richer in language style than the Chinese text, which may be due to more people sharing and editing the English Wikipedia; while for Google News, the language style richness of Chinese is higher than that of English , may reflect differences in news language styles.

d. Describe the impact of your findings.
Conducting these analyzes on Chinese and English texts, especially comparisons without translating Chinese into English, is meaningful because it helps us understand text characteristics and differences in different language environments. Such comparisons can reveal how information is expressed and focused in different cultural contexts. However, experimental results may be limited by the richness of the dataset. In order to obtain a more accurate and comprehensive comparison, the data set may need to be enhanced and expanded to include more diverse and comprehensive text content.

These findings have important implications for understanding the characteristics and challenges of cross-cultural communication, multilingual content creation, and

multilingual news reporting. They can also provide valuable insights for designing more effective cross-lingual information retrieval and data analysis tools. At the same time, this also prompts us to pay attention to the richness and diversity of texts when processing and analyzing multilingual data to avoid bias and misunderstanding.

4. Future Work

a. Given more time, what direction would you take to improve your project?

1. Expand and diversify the data set: Increase the size and diversity of the data set to include different types of news sources, social media content, forum discussions, etc. to gain a more comprehensive perspective. This allows for a better understanding and comparison of text characteristics in different linguistic and cultural contexts.

2. In-depth semantic analysis: Use more advanced natural language processing technologies, such as BERT, GPT and other deep learning models, to conduct deeper semantic understanding and analysis. This can help capture the meaning of text more accurately, especially when dealing with complex or implicit language.

3. Refinement and optimization of machine learning models: Continuously adjust and optimize the machine learning models used to improve the accuracy and reliability of analysis. For example, fine-tuning sentiment analysis models to better adapt to different cultural and linguistic contexts.