

Unraveling India-China Trade Relations from an NLP Perspective

Harika Konada, Ramesh Chandra, Varun Balle

Abstract—This research employs a multi-faceted approach, integrating Natural Language Processing (NLP) and Machine Learning (ML) techniques, to analyze and summarize articles on India-China trade relations. The study includes data collection from diverse news sources, text data preprocessing, topic modeling, abstract summarization, and sentiment analysis.

The first hypothesis posits that both abstractive and extractive summarization sentiment should be similar, but our investigation reveals nuanced differences. The sentiment analysis leverages the Longformer model for a comprehensive evaluation. The findings aim to enhance understanding and insight into the subtle aspects of sentiment across different sections of the articles.

Index Terms- *Abstractive Summarization, Extractive Summarization, Latent Dirichlet Allocation(LDA), Seq2Seq, longFormer, Topic Modelling*

I. INTRODUCTION

The intricate relationship between India and China, two of the world's most populous nations, extends beyond their shared borders, permeating the global economic and geopolitical landscape. Their trade ties, characterized by both cooperation and competition, have far-reaching implications for the world stage.

The sheer volume and diversity of data surrounding India-China trade relations present a challenge for traditional analysis methods. To effectively navigate this intricate web of information, the adoption of advanced technologies such as Natural Language Processing (NLP) and Machine Learning (ML) has become essential. NLP, with its ability to comprehend and interpret human language, offers a powerful tool for extracting meaningful insights from vast troves of textual data. By combining the strengths of NLP and ML, a hybrid approach emerges, capable of dissecting articles on India-China trade relations with unprecedented precision and depth.

This approach encompasses three key NLP areas: summarization, topic modeling, and sentiment analysis.

Summarization involves automatically condensing lengthy articles into concise summaries, capturing the essence of the content without sacrificing crucial details. Topic modeling, a sophisticated NLP technique, uncovers the underlying thematic structure of a collection of documents by identifying recurring topics and their relationships. Sentiment analysis delves deeper, gauging the emotional tone and opinions expressed in trade-

related discourse. This nuanced understanding of sentiment is crucial for comprehending the evolving nature of India-China trade narratives, revealing shifts in public opinion, policy stances, and international perceptions.

This paper introduces a hybrid approach, leveraging NLP and ML to dissect articles on trade relations. The objective is to harness these technologies for summarization, topic modeling, and sentiment analysis. Understanding sentiment nuances becomes crucial for comprehending the evolving nature of India-China trade narratives.

II. LITERATURE REVIEW

The literature review delves into existing studies on India-China trade and text analysis. It identifies gaps in the literature that the current analysis aims to fill. Recent literature emphasizes the growing significance of NLP and ML in scrutinizing complex datasets, especially in the domain of sentiment analysis. Studies highlight the role of abstractive summarization and topic modeling in extracting meaningful insights from textual data. The integration of transformer models, such as Longformer, signifies the evolving landscape of sentiment analysis in capturing document-level sentiments. [3] Conducted extensive research to find articles that would be suitable for our goals of understanding nuances and emotions. They picked the topic of India-China trade relations as their topic, as it is close and highly relevant to today's global economic trends. [2] Performed data preprocessing and extracted text using the requests library from the article URLs. [2] did necessary preprocessing to remove non-alphanumeric characters and spaces. [1] A non-parametric test (Mann-Whitney U test) was introduced which is broadly used for comparing groups of normal and non-normal data, analyzing ordinal, and exploring differences. [2] Deployed abstractive summarization with the T5 model which facilitates efficient information retrieval by identifying the most relevant and informative parts of large document collections. [1] Utilized Latent Dirichlet Allocation (LDA), in conjunction with extractive summarization, which is a generative probabilistic model for topic modeling. [3] Undertook emotional analysis of the sentiment obtained from the article by dividing each article into chunks of smaller topics. [1] [2] Deployed Longformer Tokenize, Longformer-For-Sequence Classification, and pipeline to perform sentiment analysis for the entire text.

III. METHODOLOGY

The research begins with data collection from diverse news articles on the India-China trade. Subsequent preprocessing involves cleaning and vectorizing text data. Topic modeling utilizing LDA provides a structural framework for analyzing the articles. Abstractive summarization, powered by the T5 transformer model, condenses the textual information. Sentiment analysis, both chunk-wise and on the entire corpus, uses the Longformer model. The methodology ensures a holistic examination of sentiment dynamics.

A. Data Collection and Preprocessing

The data collection and preprocessing procedure involved gathering information from specific URLs related to the India-China trade and subsequently refining the obtained textual data. The data collection began with the utilization of the `get_article_text` function, which utilized the requests library to fetch the HTML content from predefined URLs.

The *BeautifulSoup* library was then employed to parse the HTML structure and extract the text from paragraphs within each article. This ensured that the relevant textual content was retrieved from the specified URLs. Following data collection, the preprocessing phase was initiated using the `preprocess_text` function. This function systematically cleaned each article by removing non-alphanumeric characters and eliminating extra spaces. The cleaning process involved regular expressions to substitute non-alphanumeric characters with spaces and to replace consecutive spaces with a single space.

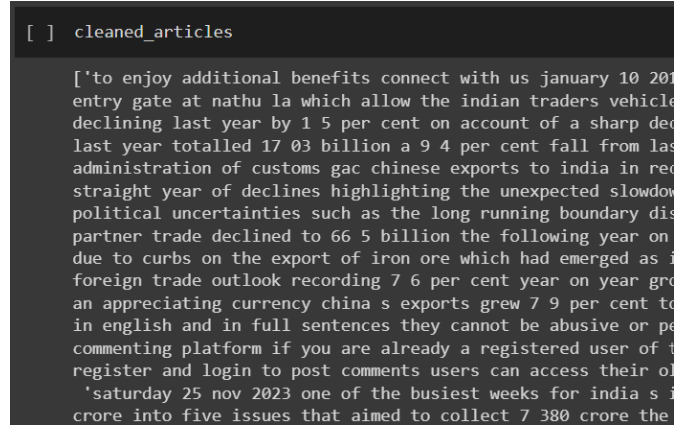
Additionally, the text was converted to lowercase, and leading/trailing whitespaces were removed. These preprocessing steps were crucial to ensure the integrity and uniformity of the text data, eliminating potential inconsistencies introduced by irrelevant characters or spaces. This process resulted in a set of cleaned articles, stored in the `cleaned_articles` list, which served as the foundation for subsequent analyses and model applications.

Fig. 1. Text before cleaning



```
articles
["To enjoy additional benefits CONNECT WITH US January 10, 2013, with two-way trade declining last year by 1.5 per cent... Indian exports to China last year totalled $ 17.03 billion... by the Chinese General Administration of Customs (GAC). China's annual figures marked the second straight year of declines, drivers of a relationship amid political uncertainties such as the long running boundary dispute between the two countries. India's largest trading partner. Trade declined to $ 17.03 billion from $ 17.03 billion in 2012. The fall in exports was largely due to curbs on the export of iron ore which had emerged as India's largest export. The data showed an overall recovery in China's foreign trade out of the world. Despite the challenge from grim global demand and an appreciating currency china's exports grew 7.9 per cent to $ 17.03 billion in full sentences they cannot be abusive or pe... commenting platform if you are already a registered user of the platform. register and login to post comments users can access their old comments. 'saturday 25 nov 2023 one of the busiest weeks for india's foreign trade. crore into five issues that aimed to collect 7 380 crore the
```

Fig. 2. Text after cleaning



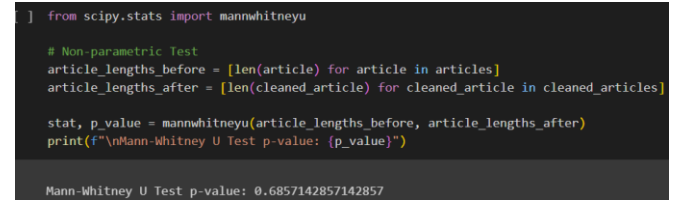
```
cleaned_articles
['to enjoy additional benefits connect with us january 10 2013, with two-way trade declining last year by 1 5 per cent on account of a sharp decline in india's annual figures marked the second straight year of declines, drivers of a relationship amid political uncertainties such as the long running boundary dispute between the two countries. india's largest trading partner. trade declined to $ 17.03 billion from $ 17.03 billion in 2012. the fall in exports was largely due to curbs on the export of iron ore which had emerged as india's largest export. the data showed an overall recovery in china's foreign trade out of the world. despite the challenge from grim global demand and an appreciating currency china's exports grew 7 9 per cent to $ 17.03 billion in full sentences they cannot be abusive or pe... commenting platform if you are already a registered user of the platform. register and login to post comments users can access their old comments. 'saturday 25 nov 2023 one of the busiest weeks for india's foreign trade. crore into five issues that aimed to collect 7 380 crore the
```

B. Non-parametric test

A non-parametric test was deployed as we decided to do a statistical test that does not make any assumptions about the underlying distribution of the data. A non-parametric test is often used when the researcher does not know the underlying distribution of the data. Some of the most widely used non-parametric tests are the Mann-Whitney U test, the Wilcoxon signed-rank test, the Krsukal-Wallis test, the Friedman test, and the Chi-square test.

C. Mann-Whitney U test

Fig. 3. Mann-Whitney U test



```
from scipy.stats import mannwhitneyu

# Non-parametric Test
article_lengths_before = [len(article) for article in articles]
article_lengths_after = [len(cleaned_article) for cleaned_article in cleaned_articles]

stat, p_value = mannwhitneyu(article_lengths_before, article_lengths_after)
print(f"\nMann-Whitney U Test p-value: {p_value}")

Mann-Whitney U Test p-value: 0.6857142857142857
```

We chose the Mann-Whitney U test for our evaluation and the length of articles before and after cleaning were the target groups. The p-value obtained from the Mann-Whitney U test was 0.6857. The p-value is a probability value that helps us decide about the null hypothesis. The p-value can be interpreted as follows:

Null Hypothesis (H0): There is no significant difference between the lengths of the articles before and after cleaning.

Alternative Hypothesis (H1): There is a significant difference between the lengths of the articles before and after cleaning.

Interpretation: If the p-value is less than the significance level (commonly set at 0.05), we reject the null hypothesis, suggesting that there is a significant difference. If the p-value is greater than the significance level, we fail to reject the null hypothesis, indicating that there is not enough evidence to suggest a significant difference.

The interpretation of the p-value determines whether there is a significant difference in lengths.

The resulting p-value of 0.6857 indicates no significant difference.

D. Topic Modelling with Latent Dirichlet Allocation (LDA)

In this phase, we created a corpus with the *cleaned_articles* and vectorized it with the *CountVectorizer*. We also employed LDA in this step to determine the topics in the articles.

LDA is a generative probabilistic model for topic modeling, a technique used to uncover the underlying thematic structure of a collection of documents. LDA assumes that each document is a mixture of topics, and each topic is characterized by a distribution over words. By analyzing the word distribution of each document, LDA can identify the topics that are most relevant to each document and the relationships between those topics.

The topics found from the article after LDA are the following:

Fig. 4. Topics

```
Topics:
Topic 1: global, 10, national, affordable, strategies
Topic 2: trade, china, billion, india, comments
Topic 3: prime, year, offer, et, exclusively
Topic 4: global, 10, national, affordable, strategies
```

E. Seq2Seq Model

Seq2seq models are a type of neural network architecture specifically designed for handling sequence-to-sequence tasks. They consist of two main components: an encoder and a decoder. Seq2seq models are particularly well-suited for NLP tasks that involve transforming one type of text into another type of text. They are effective for tasks such as machine translation, text summarization, and question answering. Seq2seq models can be used to generate concise summaries of lengthy documents. They can capture the main points of the original text while preserving the core meaning.

We created a tokenizer with the *AutoTokenizer* class of the Hugging Face Transformers library, which simplified the process of using it with the pre-trained T5 base model, a large language model (LLM) trained on a massive dataset of text and code.

The tokenizer is responsible for converting raw text into numerical representations that can be processed by the seq2seq model.

F. Abstractive Summarization

Abstractive summarization generates concise summaries that may contain rephrased sentences that are not present in the original text. It aims to capture the core meaning and context of the input. In our case, the abstractive summary of the articles emphasizes the decline in Indian exports to China.

The generated seq2seq model in the previous step is used to generate abstractive summaries.

Fig. 5. Abstractive Summaries

```
Summary for Article 1:
the gac said comments share read later a file picture of entry gate at nathu la which allow the indian traders vehicles enter china new trade figures released in beijing on friday showed indian exports to china last year totalled 17 03 billion a 9 4 per cent fall from last year chinese exports to india grew 7 9 per cent to 2 21 trillion the gac said comments share read later a file picture of entry gate at nathu la which allow the indian traders

Summary for Article 2:
a client requested an arrangement of locally grown nargis flowers daffodils and tulips csam minister of state for electronics and it rajeev chandrasekhar said friday wedding planner mukta Kapoor recently had a client request an arrangement of locally grown nargis flowers daffodils and tulips csam minister of state for electronics and it rajeev chandrasekhar said friday
```

G. Extractive Summarization

Using LDA, extractive summarization selects and combines existing sentences from the original text to form a summary. The LDA algorithm assigns scores to sentences based on their topic dominance. In the extractive summary for articles, the sentences were selected based on their relevance to trade deficit figures and relations between India and China.

Considering our hypothesis, where we mentioned nuanced differences in sentiment between abstractive and extractive summarization, this study provided evidence that abstractive summarization provides more flexibility in capturing subtle nuances in sentiment.

Fig. 6. Extractive Summaries

```
Article 1:
to enjoy additional benefits connect with us january 10 2014 07 19 pm updated november 16 2021 09 26 pm ist beijing comments share read later a file picture of entry gate at nathu la which allow the indian traders vehicles enter china india's trade deficit with china reached a record 31 4 billion in 2013 with two way trade declining last year by 1 5 per cent on account of a sharp decline in indian exports new trade figures released in beijing on friday showed indian exports to china last year totalled 17 03 billion a 9 4 per cent fall from last year out of 65 47 total bilateral trade according to figures released by the chinese general administration of customs gac chinese exports to india in recent years largely comprised of machinery were up 1 6 per cent friday's annual figures marked the second straight year of declines highlighting the unexpected slowdown in rapidly growing trade ties that came to be seen as one of the key drivers of a relationship amid political uncertainties such as the long running boundary dispute bilateral trade reached a record 74 billion in 2011 when china became india's largest trading partner trade declined to 66 5 billion the following year on account of the global slowdown and a 20 per cent drop in indian exports the fall in exports was largely due to curbs on the export of iron ore which had emerged as india's single biggest export to resource hungry china friday's data showed an overall recovery in china's foreign trade outlook recording 7 6 per cent year on year growth although missing the government's 8 per cent target despite the challenge from grim global demand and an appreciating currency china's exports grew 7 9 per cent to 2 21 trillion the gac said comments share exports trade balance imports back to top comments have to be in english and in full sentences they cannot be abusive or personal please abide by our community guidelines for posting your comments we have migrated to a new commenting platform if you are already a registered user of the hindu and logged in you may continue to engage with our articles if you do not have an account please register and login to post comments users can access their older comments by logging into their accounts on vuuikle
```

Article 2:

saturday 25 nov 2023 one of the busiest weeks for india s initial public offer i po market ended on a euphoric note with investors pouring an unprecedented 2 6 1 akh crore into five issues that aimed to collect 7 380 crore the government has advised social media and internet intermediaries to align the terms of service o n their platforms within the next seven days to alert users about the consequenc es of creating uploading and sharing prohibited information including deepfake c ontent or child sexual abuse material csam minister of state for electronics and it rajeev chandrasekhar said friday wedding planner mukta Kapoor recently had a client request an arrangement of locally grown nargis flowers daffodils and she s not the only one there s a growing trend of indians opting for domestic produ ce as opposed to imported blooms india s increasing cultivation of exotic flower s such as orchids carnations and tulips is meeting the rising demand regulations are also keeping overseas supplies in check download the economic times news ap p to get daily market updates live business news unwinable wars and the huge to ll it takes on economic humanitarian and military domains sbi bajaj fin axis ban k face the heat of unsecured lending norms what should investors do next tax tro ubles swiggy and zomato grapple with freshly served gst notice on delivery fee a s us eases sanctions on venezuela india finds its next russia to buy discounted oil youtube ustads and insta gurus novice stock traders turn to social media for guidance narayana hrudayalaya s shares are on a growth beat what s driving the hospital chain s optimism kerala four dead dozens injured in stampede tunnel resc

H. Emotion Analysis

We created a sentiment-analysis pipeline for the *cleaned_articles*. The initially considered 4 articles were divided into smaller portions, named *chunks*, depending on the length of the articles.

Article 1 - 5 chunks

Article 2 - 11 chunks

Article 3 - 2 chunks

Article 4 - 5 chunks

The sentiment of each chunk was analyzed alongside confidence values. The provided output shows the sentiment analysis results for each article broken down into chunks. Each chunk is labeled as either positive, negative, or neutral, accompanied by a confidence score, indicating the model's certainty in its prediction.

Fig. 7. Emotion Article 1

Emotions for Article 1:

Chunk 1: NEGATIVE with confidence 0.9872323274612427
 Chunk 2: NEGATIVE with confidence 0.9961349964141846
 Chunk 3: POSITIVE with confidence 0.9075860381126404
 Chunk 4: NEGATIVE with confidence 0.9919589757919312
 Chunk 5: NEGATIVE with confidence 0.9746127724647522

For Article 1, the sentiment fluctuates across chunks, with the initial segments expressing a strong negative sentiment, followed by a positive tone in chunk 3. However, subsequent chunks revert to a negative sentiment. The varying emotions suggest a mixed perspective or narrative in Article 1, with moments of negativity interspersed with a positive section.

Fig. 8. Emotion Article 2

Emotions for Article 2:

Chunk 1: NEGATIVE with confidence 0.9666159152984619
 Chunk 2: POSITIVE with confidence 0.9543963074684143
 Chunk 3: NEGATIVE with confidence 0.9927135109901428
 Chunk 4: POSITIVE with confidence 0.8304776549339294
 Chunk 5: NEGATIVE with confidence 0.9847776889801025
 Chunk 6: NEGATIVE with confidence 0.9987768530845642
 Chunk 7: NEGATIVE with confidence 0.9984546899795532
 Chunk 8: NEGATIVE with confidence 0.9913115501403809
 Chunk 9: NEGATIVE with confidence 0.9810521602630615
 Chunk 10: NEGATIVE with confidence 0.9853445887565613
 Chunk 11: POSITIVE with confidence 0.8221780061721802

Article 2 exhibits a similar pattern, alternating negative and positive sentiment across its chunks. The sentiment changes are notable, indicating potential shifts in the subject matter or the author's stance. Interestingly, chunk 11 introduces a positive sentiment, providing a nuanced emotional context to the article.

Fig. 9. Emotion Article 3

Emotions for Article 3:

Chunk 1: POSITIVE with confidence 0.9988841414451599
 Chunk 2: POSITIVE with confidence 0.9955568909645081

Article 3 maintains a consistently positive sentiment throughout its chunks, suggesting an optimistic or affirmative tone in the content. This stability in positive emotions indicates a cohesive and positive discussion narrative in Article 3.

Fig. 10. Emotion Article 4

Emotions for Article 4:

Chunk 1: NEGATIVE with confidence 0.9930965304374695
 Chunk 2: NEGATIVE with confidence 0.9752012491226196
 Chunk 3: POSITIVE with confidence 0.7095745801925659
 Chunk 4: POSITIVE with confidence 0.5535345673561096
 Chunk 5: NEGATIVE with confidence 0.9944536089897156

In contrast, Article 4 presents a more complex emotional trajectory. The initial chunks convey a strong negative sentiment, which gradually shifts as the article progresses. The latter chunks introduce positive sentiments, indicating a potential evolution or change in the narrative tone within Article 4.

In summary, the sentiment analysis highlights the emotional dynamics within each article, showcasing the variation and nuances in the expression of sentiments across different sections or themes. These emotional insights can provide a deeper understanding of the subjective tones present in the articles, aiding in the interpretation of the overall sentiment conveyed by each piece.

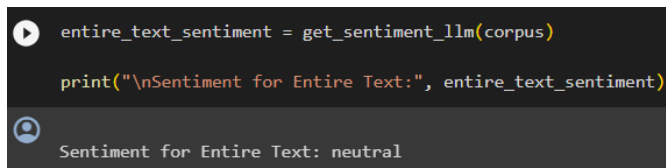
I. Sentiment Analysis Using LongFormer

In this step, we assessed the sentiment of the entire text using *LongFormer* for the seq2seq model. *LongFormer* is a pre-trained transformer model for long documents that was integrated into Hugging Face in 2020.

Input text is tokenized using the LongFormer tokenizer which is then converted to PyTorch tensors. These tensors are used as input for the LongFormer Model and a set of logits are generated which represent the confidence in each sentiment label. The sentiment label is predicted by selecting the class with the highest logit value. The possible

sentiment labels are "positive," "negative," and "neutral."

Fig. 11. Sentiment for the entire text



```
entire_text_sentiment = get_sentiment_llm(corpus)

print("\nSentiment for Entire Text:", entire_text_sentiment)
```

Sentiment for Entire Text: neutral

The output "Sentiment for Entire Text: neutral" indicates that, according to the LongFormer sentiment analysis, the overall sentiment of the entire text (corpus) is considered neutral.

J. OUTCOMES

The research outputs highlight the effectiveness of NLP and ML techniques in uncovering insights from India-China trade articles. Distinct topics, nuanced sentiment changes, and abstractive summaries contribute to a comprehensive understanding of the narratives. The neutral sentiment of the entire corpus suggests a balanced tone in the overall coverage.

K. CONCLUSION

The study successfully integrates diverse techniques to analyze India-China trade articles. The nuanced sentiment analysis provides a deeper layer of understanding, while abstractive summarization condenses complex information. The neutral sentiment across the corpus indicates a balanced portrayal. Further optimization and exploration of advanced models may enhance the depth of analysis in future research.

REFERENCES

- (2022, November 15). India's October trade deficit widens to \$26.91 billion as exports decline 17%. The Economic Times. <https://economictimes.indiatimes.com/news/economy/indicators/indias-october-trade-deficit-widens-to-26-91-billion/articleshow/95529500.cms?from=mdr>
- (2023, January 14). India's trade deficit with China hits \$100bn for first time. Times Of India. <https://timesofindia.indiatimes.com/business/india-business/indias-trade-deficit-with-china-hits-100bn-for-first-time/articleshow/96979850.cms>
- KRISHNAN, A. (2014, January 10). India-China trade: Record \$ 31 bn deficit in 2013. THE HINDU. <https://www.thehindu.com/business/indiachina-trade-record-31-bn-deficit-in-2013/article5562569.ece>
- KRISHNAN, A. (2023, January 13). India's imports from China reach record high in 2022, trade deficit surges beyond \$100 billion. THE HINDU. <https://www.thehindu.com/news/international/indias-imports-from-china-reach-record-high-in-2022-trade-deficit-surges-beyond-100-billion/article66372861.ece>

ACKNOWLEDGMENT

Firstly, we would like to acknowledge Dr. Tony Diana for his exceptional guidance and unwavering support throughout this research endeavor. His consistent encouragement and constructive feedback have been instrumental in shaping our ideas and refining the quality of this paper. We are immensely grateful for his mentorship and the time he has dedicated to us.

Secondly, we express our gratitude to our UMBC GSA Writing Coach, Ben Marino for his support in writing this paper.

Lastly, the authors express their sincere gratitude to the various libraries, models, and frameworks that were instrumental in conducting this research. We extend our deepest appreciation to the research community for their invaluable insights and contributions.