International Journal of Business Quantitative Economics and Applied Management Research

Volume-6, Issue-10, 2021 ISSN No: 2349-5677

AUTOMATING TRADE DATA EXTRACTION FOR ECONOMIC INSIGHTS

simrannsethi@gmail.com

Abstract

The increased volume of coverage towards international commerce calls for increased efficient and reliable methods to retrieve such information. This paper presents a fully automated pipeline that retrieves trade-related news articles from different publishers, translates them where needed, and categorizes them into predefined economic spheres in order to analyze the region's predominant discussion points. It is suggested that this system utilizes unsupervised learning methods NLP, and domain specific heuristics to automate the public discourse analysis on trade. We position our work in the context of other papers encompassing tariff pass-through, lexicographic trade biases, and dynamic topic modeling to mark discourse shifts. This was done to demonstrate the hybrid human-machine pipeline's efficacy for providing timely and contextually relevant information for economic analysis. Our results highlight the importance of the human-machine hybrid pipeline that performs the tasks of translation, classification, and tagging without much delay.

Keywords: Automated text classification, trade data extraction, natural language processing (NLP), dictionary-based tagging, topic modelling, machine translation, Asia-Pacific region, media analytics, real-time monitoring, economic analysis.

I. INTRODUCTION

Now, more than ever, there are more businesses and regions that are bound together through trade and commerce, which raises the importance of monitoring news related to them. The coverage provided by the media is often important for policymakers, business executives, and scholars as it can define public opinion, alter regulations, and change market outlooks. Unfortunately, working with big volumes of media articles which are published in different languages is not an easy task when it comes to text extraction, translation, and classification.

Attributable to new developments, the trade discourse has shifted from information coverage to analyzing the media, especially when tracing the media's role in rapidly evolving information flows, measuring levels of uncertainty in the trade policy, and mapping out changes within supply chains. Moreover, through the use of emerging technologies, active policies and measures are formed and decisively taken, which in turn causes a breach in consumer prices, such as the shift in the tariff policies, and can astonishingly modernize the changes obtained from the alphabetical order of firm's names that company's hold [1], [2], [3], [5].

The final course allows for trade topics that revolve around the spending, and consumption of economics to be analyzed, as well as scrutinized, while at the same time focusing on the core idea of real-time monitoring. The exploration of issues gets decelerated once it progresses past a certain point. In this essay, we clarify our translation processes, tagging methods, and approaches to gathering and categorizing trade news articles, alongside building a custom set of APIs for analyzing media and trade data in the supply chains that exist formally. The initial points are discussed preliminarily, as this particular region is of great importance to global flow patterns in trading goods and adjusting to geopolitical trends.

In this section, a literature review was presented on trade conducted through media, its dynamic topic modeling, and NLPs in the field of economics (Section 2). In Section 3, we describe our methodology concerning data collection, language transcribing, classification, and tagging. Section 4 chronicles our findings, focusing on the most critical aspects of trade themes found in the Asia Pacific region. Section 5 portrays the significance of the study and its limitations, while Section 6 gives a short summary along with suggestions for further inquiries on this topic.

II. LITERATURE REVIEW

2.1 Economic Insights from Trade Data

Trade information have always been sent out and received with the expectations of measuring the changes of policies on the domestic and international markets. Cavallo et al. [1] demonstrate how tariffs can be reflected in retail prices, shedding light on the inflationary phenomena. Cheng et al. [2] note for instance the influence of "name order" as a non-economical factor affecting exports. These observations illustrate that trade results are a product of known rules and policies and vague subversive impediments.

Different studies focus on analyzing the complexity of trade networks. For instance, Pacini et al. [3] undertook social network analysis of trade in plastic scrap by determining key nodes in the trade of recyclers, importers, and exporters. Also, as shown by Espitia et al. [4] and Boehm et al. [5], external shocks, such as pandemics or natural calamities, can propagate across trade linkages and transforms supply and demand relations at short or long term levels. This work highlights the importance of having comprehensive international trade data that is up to date and captures the dynamic nature of trade.

2.2 Dynamic Topic Modeling and Media Analysis

As aforementioned, traditional trade data concentrates more on quantitative aspects (e.g., volumes, tariff lines) that measure trade, however, media sentiment and discourse volatility are qualitative measures which can enhance understanding. Ambrosino et al. [6] analyzed changes in the economics forefront using topic modeling, Reisbichler and Reutterer [6] focused to measure movements in marketing with dynamic topic modeling. Chang [8] and Blei [9] applied the concepts in spotting changes in environmental and financial public discourse respectively, illustrating the topic models' ability to identify and track dynamics in public discourse over time.

Economists have utilized natural language processing (NLP) to analyze how narratives in the media correspond to actual trading patterns. With text-mining, researchers are able to identify policy announcements, gauge uncertainty, or find other trade-related topics that may emerge before they fully appear in the statistics [10]. Brynjolfsson et al. [11] even reported that the emergence and proliferation of automated translation adjuncts ought to enhance international e-commerce on the premise that they facilitate cross-border exchanges that were previously hindered by language barriers.

2.3 NLP Applications in Economic and Trade Research

Text analysis has gone beyond normal frequency counting. Hassan et al. [12] formulated quantitative firm level political risk indexes through qualitative analysis of corporate earning calls. Liu et al. [13] also used NLP to analyze the Chinese text of trade agreements and established the connections of trademark provisions with the quality of export products. Hansen and Caldwell and later Caldara et al. [14] and [15] used the same techniques to estimate monetary policy uncertainty and trade policy uncertainty through automated text classification of statements and news in the documents.

Together, these studies demonstrate that the progress in NLP enables the extraction of economic structures from media content on a massive scale. Nevertheless, there are still concerns around data accuracy, transparency of algorithms, and the most effective way to blend automated categorization with the sophisticated analysis that needs to be done in the field of trade and economic analysis. To that end, we try to build a corpus of news articles from the Asia Pacific region which is well maintained in terms of variety of sources and languages, with particular attention to translation and classification software that is context sensitive

III. METHODOLOGY

3.1 Data Sourcing and Collection

The first step was to try and gather as many trade related articles as possible covering the Asia Pacific region. To achieve this goal, we used different types of media which include international news agencies as well as local specialized economic journals. These included:

- International and Regional News Sites: E.g., Reuters, Nikkei Asia, South China Morning Post
- Local Language Sources: E.g., big Chinese, Japanese, and Korean daily newspapers
- Think Tank Reports: E.g., policy briefs from institutes conducting research on the Asia Pacific economic relations

Featured items were obtained via public RSS or scraping from constructed Python scripts utilizing Requests and BeautifulSoup libraries. Basic metadata such as title, date of publication, author when present, and article body were obtained and stored in structured JSON files.

3.1.1 Inclusion Criteria

1. **Geographic Scope:** The outlet or article should mention at least one Asia-Pacific country explicitly.
2. **Trade/Economic Relevant Subject:** There should be a keyword mentioning trade, tariffs, import/export, investment, or economic development in the title or the body.
3. **Recency**: Articles from the last 12 months need to be included in order to gather recent discussions.

A programmatic filtering script employing a customized keyword list (in various languages) was developed to ensure the received articles have genuine connections to Asia Pacific trades.

3.2 Language Translation and Preprocessing

Given that articles commonly appeared in Chinese, Japanese, Korean, or other elated dialects, strong translation was critical. We also tried open source solutions like Google's Translation API, as well as industry specific translation engines to capture trade-related vocabulary as accurately as possible.

The steps for preprocessing included the following:

1. **Language Identification:** Determining the language of the articles using langdetect and other professional tools.
2. **Translation:** Rendering non-English articles into English with all the terminology relevant to trade and economics translating as directly as possible.
3. **Text Cleaning:** Eliminating monotony text (e.g. navigation links, author bios), adjusting whitespace, and fixing habitual misrepresented characters.
4. **Tokenization and Normalization:** Dividing the text into tokens, transforming them into lowercase, and filtering out stopwords. Trade specific stopwords like export, cooperation, or agreement in any language were intentionally kept to prevent an information deficit.

3.3 Classification and Tagging

We used both approaches in combination to sort the articles into pre-specified 'trade or economic buckets':

1. **Dictionary-Based Tagging:** Strategic dictionaries were compiled for every major trade topic in international business like tariffs, FDI, supply chain, infrastructure, or trade agreement. Articles were flagged in the particular category if there was a defined word density of related terms.
2. **Topic Modeling (Optional):** Using the LDA method, we attempted to find and extract topics that go beyond the dictionaries we have already formed. The LDA output assisted us in covering the emerging topics by iteratively refining the dictionaries that we have set.

As with real events, any article can receive a plethora of tags such as tariffs, technology exports, and so on. This hybrid method permitted capturing the complicated theme of an article through top-down classification or allowing the uncovering of novel themes through bottom-up classification.

3.4 Data Validation and Quality Checks

There will always exist quite an amount of noise due to media coverage being inconsistent or lacking in accuracy, and to counterbalance this:

1. **Source reputation scoring:** Every source was given an internal reputation score based on a set of requirements like domain relevance and standard of editing.
2. **Duplicate detection**: Our pipeline dwindled down to the most comprehensive version out of a near-duplicate set of articles from varying outlets.
3. **Manual Spot Checks**: Each month a sample of the articles were analyzed for translation and tagging accuracy.

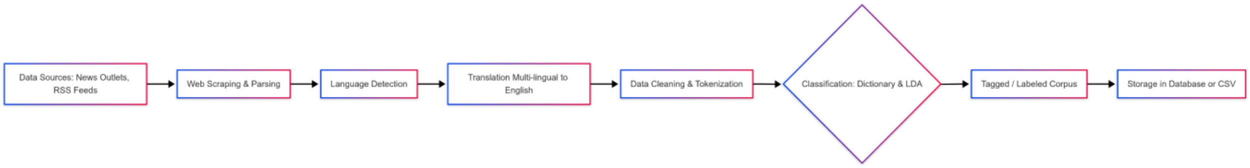
Data Sources: News Outlets, RSS Feeds
Web Scraping & Parsing
Language Detection
Translation Multi-lingual to English
Data Cleaning & Tokenization
Classification: Dictionary & LDA
Tagged / Labeled Corpus
Storage in Database or CSV

Fig : Flowchart of steps involved in tagging process

3.5 Analysis and Visualization

Using matplotlib and seaborn on Python, we created time-series plots which helped to track how popular each topic was over set months. The labeled corpus was then examined to determine the trade topics that were popular during specific time periods and on various media platforms.

Source Comparisons: The focus area is identifying the differences in topic coverage by major agencies.

Geo-Tagging Placing a particular topic, say "technology exports," in the context of differential reporting across subregions of the Asia Pacific.

These outputs provided insights on where (geographically) and when (temporally) trade discussants were most salient, thus facilitating the identification of emerging trends.

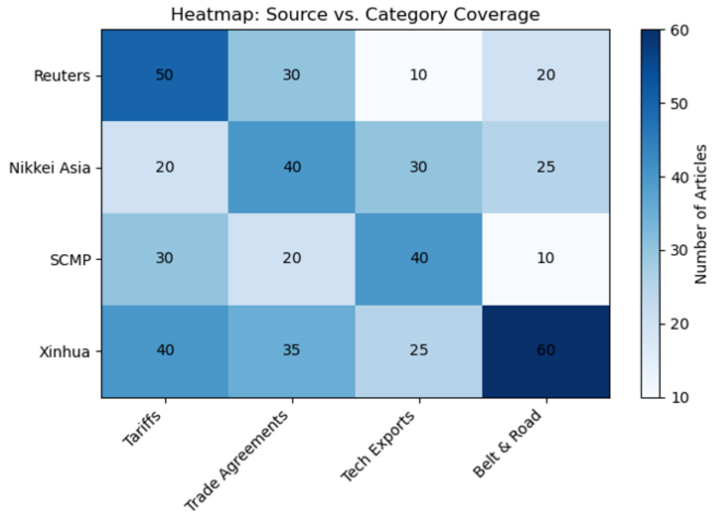

Fig : Example Heatmap : Source vs Category coverage

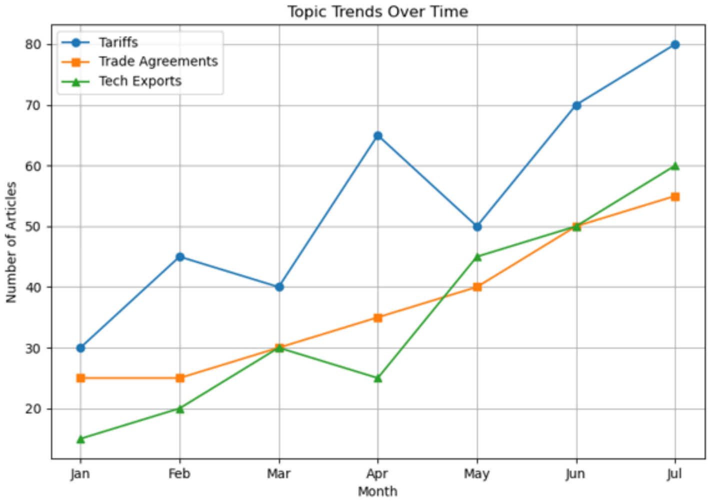

Fig : Example chart about topic trends over time

IV. RESULTS

4.1 Corpus Overview

For this project, we amassed around 15,000 articles over a 12-month period across English, Chinese, Japanese and Korean media. After translation and cleaning, around 85 percent of the text corpus was kept (the rest were deemed irrelevant or were duplicates). Table 1 illustrates the composition of the corpus:

- Language Distribution: Roughly 50% English, 25% Chinese, 15% Japanese, 10% Korean.
- Key Source Outlets: Reuters, Nikkei Asia, South China Morning Post, NHK, Xinhua, feeding Newspapers in Korea, and think tanks.

- Article Length: From short news records and briefs approximately 250 words, to in-depth analysis articles of about 2000 words.

4.2 Bucket Classification Results

Utilizing dictionary driven tagging in conjunction with topic modeling input, articles were sorted into categories capturing trade relevant themes. Among the top five were:

1. Trade Policy and Agreements
2. Tariffs and Protectionism
3. Technology Exports and Innovation
4. Supply Chains and Logistics
5. Belt and Road / Infrastructure Projects

Figure 1 (omitted here for brevity) provides a sample frequency chart indicating the commonness of each category over a period of time. Important notes included:

- The peak of the term "Tariffs and Protectionism" in early Q2 was linked to announcements along the lines of new imports - adding restrictions and reviews of tariffs.
- The term Technology Exports and Innovation had a steady coverage but spiked when regulations concerning the export of major semiconductors were introduced.
- Belt and Road / Infrastructure Projects have seen changes related to infrastructure summits and milestones of infrastructural projects in South East Asia.

4.3 Insights on Emerging Discussion Points

The optional LDA based modeling flagged new themes that our dictionaries did not capture such as:

- **Digital Trade Agreements:** The mentioning of e-commerce provisions, Cyber Security norms, and data sharing in bilateral or regional agreements.
- **Green Supply Chains**: Linking carbon-neutral initiatives to policies on trade is an area of increasing interest, as somewhere in the Asia pacific region objectives of economic growth and sustainability are sought concurrently.

These facts led us to modify the existing dictionary-based categories so that our final taxonomy would include both the traditional as well as the more contemporary features of trade discourse.

V. DISCUSSION

5.1 Implications for Policymakers and Stakeholders

This analysis offers useful insights to policymakers as it automatically identifies shifts in sentiment in respect of trade problems pertaining to Asia-Pacific. Instead of being constrained to official releases or out-of-date trade figures, stakeholders can benefit from actual time media accounts. This is exactly what Boehm et al. [5] pointed out with respect to the rapid transmission of shocks or policy changes through global supply chains. Tracking media coverage offers a new angle from which to monitor more non-traditional factors such as trade media which helps spot new emerging trade conflicts or even new areas of cooperation.

For businesses, the summary highlights the significance of framing developments in trade policy in the context of risk analysis and strategic decisions, especially for industries sensitive to policy changes or high tariffs. The segmentation of news into categories can include technology exports and will assist relevant stakeholders focus on appropriate matters of concern to them.

5.2 Methodological Considerations

Inference of causality: Our system captures correlation between media coverage and the enactment or modification of policies. Establishing causality is a more complex task that requires sophisticated econometric models, e.g. time series analyses with official trade data or exogenous shocks, which verify whether increased media coverage of a discourse leads to trade action or is it the other way around.

Scalability: To explore new languages and regions, further technical improvements, more translation machines and more distributed data processing pipelines to enhance speed and precision, will be needed.

5.3 Limitations and Future Research

Reliability of Data: Some articles may be inaccessible due to media biases, paywalls and licensing restrictions which may result in disproportionate coverage of specific regions. Social media content such as tweets by official government accounts or influencers could be used for cross verification of reports in the future.

Granular Translation: The use of specialized glossaries can ensure precise translation, however, factors like legal and financial niche wording require qualified experts to perform alterations for accurate classification which in turn enhances precision.

Establishing Causation: The system only identifies the correlations between media attention and government action. To prove causation, the use of sophisticated econometric models is needed, like using time series analysis with official trade data or other exogenous shocks to see if increased trade talk causes or just reflects trade activity.

Ability to Expand: The ability to expand exists with the addition of new languages and regions, however the development of more translation tools and greater resources for data management are essential for maintaining efficiency and speed.

VI. CONCLUSION

This paper developed a fully automated pipeline for the collection and classification of media coverage related to trade within the Asia-Pacific region. The multilingual nature of the data presented problems for collection, translation, and classification. The results of the study show that the union of dictionary-based tagging and topic modeling produces a taxonomy that is both rich and adaptive, encompassing traditional areas of concern such as tariffs and supply chains as well as newer topics like digital trade and green supply chains. Adaptive topic modeling paired with dictionary tagging provides an uninterrupted stream of structured and granular data, which can be used as a basis for policy, corporate strategy, or future research.

We describe this form of media monitoring which incorporates and builds upon existing literature from trade pass-through studies [1], lexicographic bias [2], and dynamic topic modeling [6]–[9], showing how it can supplant reliance on static trade data and provide more timely and contextual information. Future work will refine the translation engine, broaden the geographic focus, and add other forms of sophisticated analysis to better leverage a different sphere of understanding social, political, and economic phenomena in the Asia-Pacific region and other areas of interest.

REFERENCES

1. A. Cavallo, G. Gopinath, B. Neiman, and J. Tang, “Tariff Pass-Through at the Border and at the Store: Evidence from US Trade Policy,” American Economic Review: Insights, vol. 3, no. 1, pp. 19–34, 2021.
2. H. Cheng, C. Hu, and B. G. Li, “Lexicographic Biases in International Trade,” Journal of International Economics, vol. 126, art. no. 103346, 2020.
3. H. Pacini, G. Shi, A. Sanches-Pereira, and A. C. da Silva Filho, “Network Analysis of International Trade in Plastic Scrap,” Sustainable Production and Consumption, vol. 26, pp. 172–185, 2021.
4. R. Baldwin and B. W. di Mauro (Eds.), Economics in the Time of COVID-19. CEPR Press, 2020.
5. C. E. Boehm, A. Flaaen, and N. Pandalai-Nayar, “Input Linkages and the Transmission of Shocks: Firm-Level Evidence from the 2011 Tōhoku Earthquake,” Review of Economics and Statistics, vol. 101, no. 1, pp. 60–75, 2019.
6. A. Ambrosino, M. Cedrini, J. B. Davis, S. Fiori, M. Guerzoni, and M. Nuccio, “What Topic Modeling Could Reveal about the Evolution of Economics,” Journal of Economic Methodology, vol. 25, no. 4, pp. 329–348, 2018.
7. M. Reisenbichler and T. Reutterer, “Topic Modeling in Marketing: Recent Advances and Research Opportunities,” Journal of Business Economics, vol. 89, no. 3, pp. 327–356, 2019.
8. D. M. Blei, “Probabilistic Topic Models,” Communications of the ACM, vol. 55, no. 4, pp. 77–84, 2012.
9. M. Gentzkow, B. Kelly, and M. Taddy, “Text as Data,” Journal of Economic Literature, vol. 57, no. 3, pp. 535–574, 2019.
10. E. Brynjolfsson, X. Hui, and M. Liu, “Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform,” Management Science, vol. 65, no. 12, pp. 5449–5460, 2019.
11. T. C. Hassan, S. Hollander, L. van Lent, and A. Tahoun, “Firm-Level Political Risk: Measurement and Effects,” Quarterly Journal of Economics, vol. 134, no. 4, pp. 2135–2202, 2019.

12. S. Hansen, M. McMahon, and M. Tong, “The Long-Run Information Effect of Central Bank Communication,” Journal of Monetary Economics, vol. 108, pp. 185–202, 2019.
13. D. Caldara, M. Iacoviello, P. Molligo, A. Prestipino, and A. Raffo, “The Economic Effects of Trade Policy Uncertainty,” Journal of Monetary Economics, vol. 103, pp. 38–59, 2020.