WTO: An NLP analysis of the state of trade multilateralism

Th. Warin, PhD - Bullet points for conversation

2024-01-07

Introduction

- Our research question is:
  "How can Natural Language Processing (NLP) techniques be applied to analyze WTO documents and communications to understand the evolving dynamics and challenges in trade multilateralism?"

Introduction

1. Application of NLP Techniques: For a detailed exploration of how these techniques can be employed to analyze textual data.
2. Understanding Trade Multilateralism: The question aims to gain insights into the complex nature of trade multilateralism.
3. Evolving Dynamics and Challenges: It hints at not only understanding the current state but also **identifying trends, changes, and challenges in trade multilateralism**.

Introduction

- Since the establishment of WTO in 1995, the landscape of international trade has undergone **transformative changes**.
- These changes have been propelled by rapid technological advancements, which has facilitated the **creation of global value chains** (Baldwin in 2016).
- However, the multilateral rules have struggled to keep up with these rapid developments. The **Doha Round negotiations** have been in a stalemate since 2001, highlighting the risk of these rules becoming obsolete. In response, states are increasingly resorting to PTAs to navigate the complexities of 21st-century commerce.

Literature Review

- Originally, when PTAs were fewer and mainly focused on tariff reductions, their impact was gauged by their mere existence or through typologies that categorized them based on the level of economic integration, ranging from simple free trade agreements to comprehensive economic unions (Baier et al., 2014).
- These efforts have significantly enhanced the understanding of PTAs.

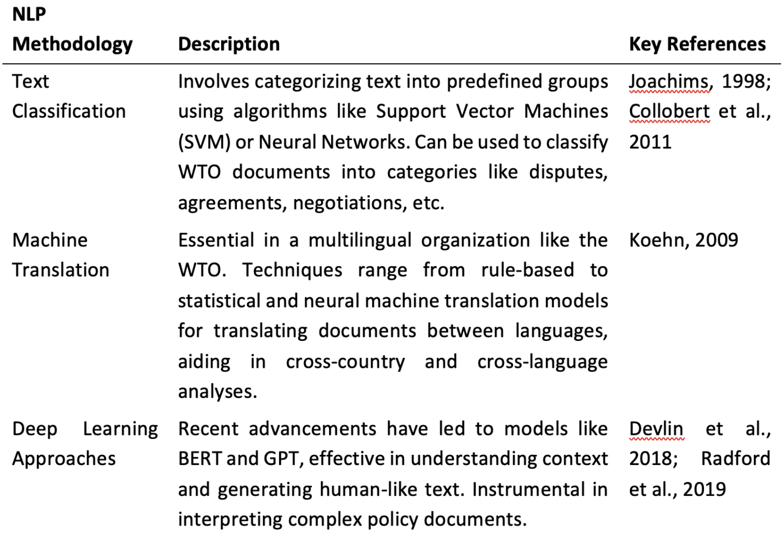
Literature Review

- The **availability of text corpora** allows for the application of methods like textual similarity providing insights into the extent of textual overlap between PTAs.
- Machine learning techniques are applied to leverage the high dimensionality of textual data for **prediction and classification tasks** (Gentzkow et al. 2017).
- Some studies have employed text-as-data methods, particularly **textual similarity**, to explore subsets of the PTA universe (Allee & Lugg, 2016).

Literature Review

- The application of NLP in this field is **not just a technological advancement** but also a methodological shift, allowing for more nuanced and comprehensive analysis of complex trade negotiations and agreements (Griffiths and Tenenbaum, 2004).
- By applying NLP techniques to WTO documents, we aim to **uncover patterns, trends, and insights**.
- This approach can provide a deeper understanding of the state of trade multilateralism, the challenges faced by the WTO, and the potential future directions of global trade policies (Lazer et al., 2009).

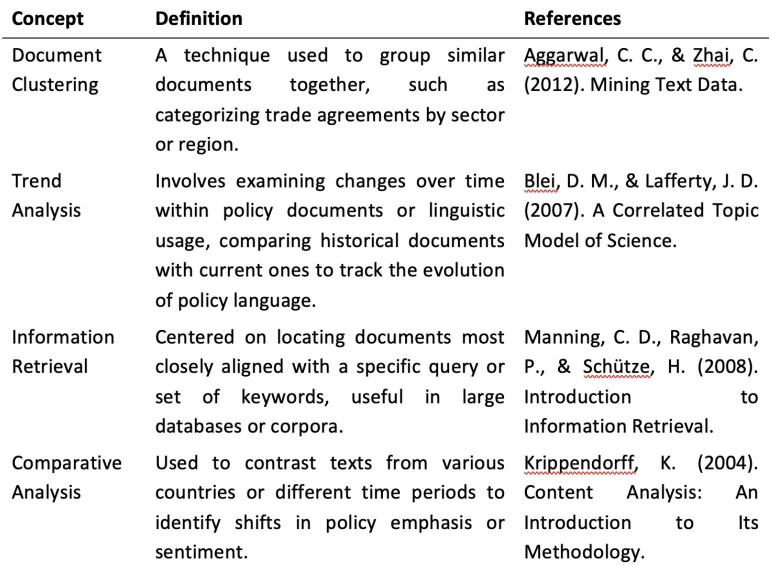
NLP Methodology	Description	Key References
Text Preprocessing	Involves cleaning and preparing text data. Key tasks include tokenization (breaking text into words or tokens), removing stop words, stemming (reducing words to their base form), and lemmatization (ensuring the root word belongs to the language).	Manning and Schütze, 1999; Bird et al., 2009
Sentiment Analysis	Analyzes text to determine the sentiment expressed within it (positive, negative, or neutral). Useful in assessing the tone and stance of policy documents or public statements in the context of trade policies.	Pang and Lee, 2008
Topic Modeling	Techniques like Latent Dirichlet Allocation (LDA) are used to discover abstract topics within a collection of documents. Can reveal underlying themes in trade negotiations or agreements.	Blei et al., 2003
Named Entity Recognition (NER)	Identifies and classifies key information in text into predefined categories (e.g., names, locations, monetary values). Useful for extracting specific data from trade agreements or legal documents.	Nadeau & Sekine, 2007



NLP Methodology	Description	Key References
Text Classification	Involves categorizing text into predefined groups using algorithms like Support Vector Machines (SVM) or Neural Networks. Can be used to classify WTO documents into categories like disputes, agreements, negotiations, etc.	Joachims, 1998; Collobert et al., 2011
Machine Translation	Essential in a multilingual organization like the WTO. Techniques range from rule-based to statistical and neural machine translation models for translating documents between languages, aiding in cross-country and cross-language analyses.	Koehn, 2009
Deep Learning Approaches	Recent advancements have led to models like BERT and GPT, effective in understanding context and generating human-like text. Instrumental in interpreting complex policy documents.	Devlin et al., 2018; Radford et al., 2019

Methodology	Definition	Equation/Representation
Cosine Similarity	Measures the cosine of the angle between two vectors in a multi-dimensional space, commonly used in text analysis to compare document content similarity.	$\text{Cosine Similarity} = \frac{A \cdot B}{\parallel A \parallel \parallel B \parallel}$
Jaccard Similarity	Compares two sets by dividing the size of their intersection by the size of their union. Useful for comparing sets of words or tokens in documents.	$\text{Jaccard Similarity} = \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$
Euclidean Distance	Measures the 'distance' between two points (or documents) in a vector space, typically used in clustering algorithms.	$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$

Methodology	**Definition**	**Equation/Representation**
Levenshtein Distance	Calculates the minimum number of single-character edits (insertions, deletions, substitutions) required to change one string into another.	No simple equation; algorithmic computation.
TF-IDF (Term Frequency-Inverse Document Frequency)	A statistical measure used to evaluate the importance of a word to a document in a collection or corpus. It increases with the number of times a word appears in a document but is offset by the frequency of the word in the corpus.	$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$
Latent Dirichlet Allocation (LDA)	A generative statistical model that allows sets of observations to be explained by unobserved groups, to discover abstract topics within documents.	Based on probabilistic graphical models; no single equation.

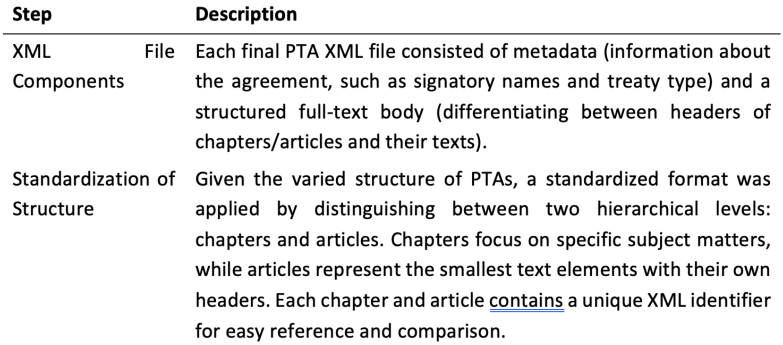


Concept	Definition	References
Document Clustering	A technique used to group similar documents together, such as categorizing trade agreements by sector or region.	Aggarwal, C. C., & Zhai, C. (2012). Mining Text Data.
Trend Analysis	Involves examining changes over time within policy documents or linguistic usage, comparing historical documents with current ones to track the evolution of policy language.	Blei, D. M., & Lafferty, J. D. (2007). A Correlated Topic Model of Science.
Information Retrieval	Centered on locating documents most closely aligned with a specific query or set of keywords, useful in large databases or corpora.	Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval.
Comparative Analysis	Used to contrast texts from various countries or different time periods to identify shifts in policy emphasis or sentiment.	Krippendorff, K. (2004). Content Analysis: An Introduction to Its Methodology.

Case Study

- Let's use Alschner et al. (2018) for a case study: Data represents a total of **448 PTAs**.
- These agreements were signed between the years **1948 and 2015**.
- Within this collection, the majority of the treaties, numbering 423, are available in English, with an additional 23 in Spanish and two in French.

Step	Description
Collection from WTO RTA System	Full texts were collected from the WTO Regional Trade Agreements Information System, supplemented by manual searches for broken links.
Format Transformation	Original treaties in varying formats were transformed into a unified, machine-readable XML format. This included correcting metadata errors and converting scanned documents to digitized text.
Preprocessing Steps	Two key preprocessing steps were undertaken: 1) Removal of annexes and schedules to focus on the main agreement body. 2) Using structural information from PDFs and manual work to segment each treaty into a structured format.
Segmentation	Each treaty was segmented into a hierarchical structure, enabling the differentiation of various structural text elements, such as chapters, articles, headers, and full texts.



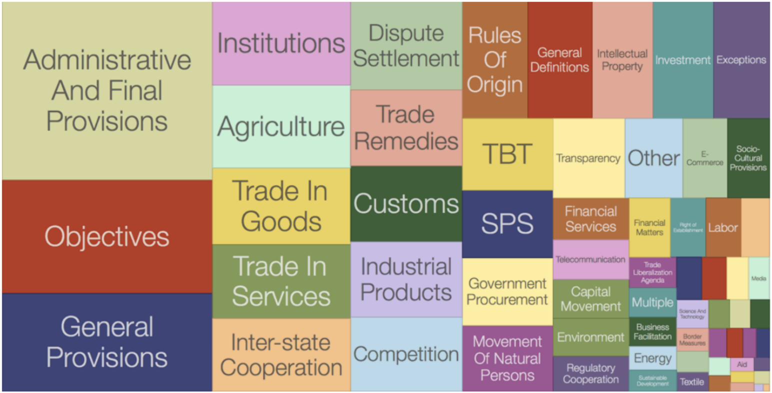
Step	Description
XML File Components	Each final PTA XML file consisted of metadata (information about the agreement, such as signatory names and treaty type) and a structured full-text body (differentiating between headers of chapters/articles and their texts).
Standardization of Structure	Given the varied structure of PTAs, a standardized format was applied by distinguishing between two hierarchical levels: chapters and articles. Chapters focus on specific subject matters, while articles represent the smallest text elements with their own headers. Each chapter and article contains a unique XML identifier for easy reference and comparison.

Case Study

- In the 1950s, treaties averaged around **5,000 words**, but by the 2010s, this figure had escalated to over **50,000 words**.
- These agreements have continued to increase in number, whereas the growth of Customs Unions has plateaued, and Goods FTAs have only seen modest expansion.

Case Study

- Another measure of a treaty's scope is the number of chapters or articles it contains. Within our corpus, 319 out of 447 PTA texts adhere to a **chapter structure**. The remaining 29 percent are structured differently, mainly divided into **articles without chapters**. Focusing on comprehensiveness, our analysis includes the 39,978 articles contained within the corpus.
- When these articles are categorized by PTA types, the differences in scope between different types of agreements become evident. For instance, Goods & Services FTAs typically include a substantially higher number of articles, with a **median of 194 provisions**.



Source: Alschner et al. (2018)

Figure 1: Heatmap of tokens by categoy. Source: Alschner et al. 2018

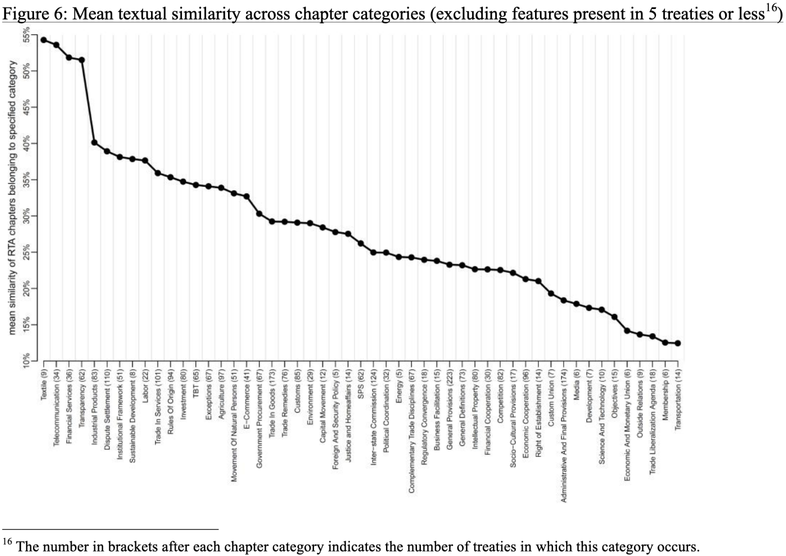


Figure 6: Mean textual similarity across chapter categories (excluding features present in 5 treaties or less[16])

[16] The number in brackets after each chapter category indicates the number of treaties in which this category occurs.

Figure 2: Mean Textual Similarity. Source: Alschner et al. 2018

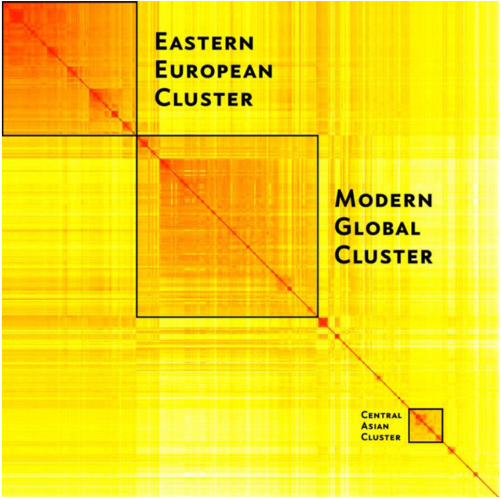


Figure 3: Clusters. Source: Alschner et al. 2018

Discussion

How can we use this kind of unstructured data?

- The emergence and proliferation of PTAs do indeed raise questions about the potential impact on multilateralism.
- However, given the observed clustering and similarities among these agreements, we are not completely moving towards the opposite direction: a world based solely on bilateralism.

Conclusion

- The observed clustering and similarities among PTAs suggest a trend towards regional multilateralism, where regional blocs are increasingly shaping trade norms and practices. This development poses a challenge to traditional multilateral institutions like the WTO, urging them to adapt and evolve to maintain their relevance in a rapidly changing global economy.

Conclusion

1. Application of NLP Techniques: For a detailed exploration of how these techniques can be employed to analyze textual data.
2. Understanding Trade Multilateralism: The question aims to gain insights into the complex nature of trade multilateralism.
3. Evolving Dynamics and Challenges: It hints at not only understanding the current state but also **identifying trends, changes, and challenges in trade multilateralism**.

Conclusion

Thank you!