Mapping Jurisprudence of WTO Dispute Settlement Body Using Deep Learning

Suyeol Yun [*]

December 14, 2020

Abstract

The world trade organization (WTO) legally regulates the world trade system with its dispute settlement body (DSB). There exists a shared understanding among legal experts about how articles of WTO agreements systematically interact with each other. However, the complexity of the WTO legal framework has constrained many developing countries with limited legal knowledge and resources from fully utilizing the WTO DSB. To address this issue, I propose a new method that summarizes the systematic interactions between articles of WTO agreements. I collected past 20 years of WTO disputes and trained a neural network that mimics the reasoning process of legal experts that determines which articles to cite for given factual description of the dispute. Then I collected all the predictions from the trained neural network and fitted the summarization network using Random Forest. I verified the quality of the fitted network by checking that the network captures the important systematic interactions as explained by the Panel and Appellate Body, two main judicial authorities of the WTO DSB.

[*] Applicant to Ph.D. program of MIT Political Science, Address: 118 Seorim-gil, Sillim-dong, Gwanak-gu, Seoul, 08839. Email: syyun@snu.ac.kr

1 Introduction

The Dispute Settlement Body (DSB) of the World Trade Organization (WTO) deals with trade disputes between WTO members. WTO members file a lawsuit in WTO DSB to claim their impaired benefit related to the WTO agreements as a result of another member's possible illegal trade policy. The judicial body of WTO DSB, *Panel* or *Appellate Body*, then adjudicates the dispute and submits a report in which it expresses its judicial opinion as to whether the challenged trade policy is inconsistent to the rules of the WTO or not (World Trade Organization, 2017).

This process requires enormous legal knowledge and resources because the legal system of WTO is highly complex. This complexity has constrained many developing countries with limited legal knowledge and resources from fully utilizing the WTO DSB (Busch and Shaffer, 2009; Busch and Reinhardt, 2003; SHAFFER, 2006).

To address this issue, I provide a novel method to summarize the network of WTO articles. Currently, understanding of how articles of WTO agreements systematically interact with each other is exclusively shared among legal experts. However, by developing the method that can quantitatively summarize the systematic interaction between articles of WTO, we can lower the cost of understanding the legal system of WTO. This will help resolve the unbalanced legal capacity issue in WTO DSB.

To properly summarize the systematic interactions between articles of WTO agreements, I designed my method based on two following considerations. First, since the legal system of WTO evolves from the way how real-world dispute interacts with the regulatory content of the article of WTO agreement, I considered a way of utilizing two different types of textual data, factual description of the trade dispute and the content of each article of the WTO agreements. Second, since members strategically cite rules of the WTO agreements to encourage the third party participation (Johns and Pelc, 2014) or to reshape the legal precedents(Pelc, 2014; Strezhnev, 2014), I considered a way of generalizing these member-specific strategic citations.

Upon these two considerations, this paper uses deep learning. Deep learning is empirically known as good at extracting information from the textual data. In addition to it, deep learning also generalizes the patterns inside data. Therefore, this paper designs a deep neural network that processes two different types of textual data, description of the dispute and each article content of the WTO agreements. The design mimics the reasoning process of the legal experts, where the experts read the textual description of the dispute and imagine applicable legal articles of the WTO agreements according to its regulatory content.

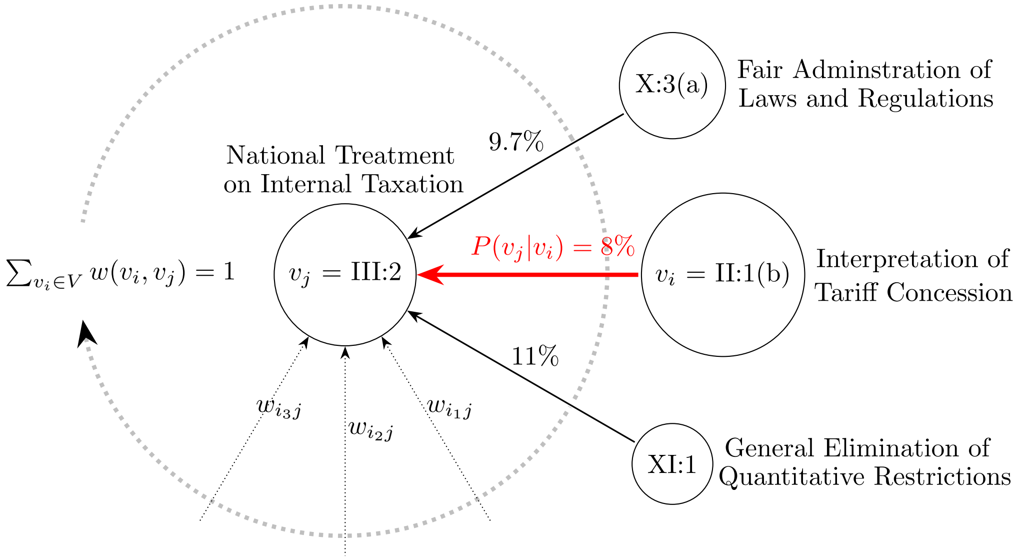
To train this neural network, I collected textual description of trade dispute and articles of the WTO agreement cited for each dispute requested to the WTO DSB from 1995 to 2018. Using this collected data, I trained the neural network by enforcing the neural network to answer correctly whether a given article of the WTO agreements can be cited for the given textual description of trade dispute. After training, I fitted a network that summarizes the systematic interactions between articles of WTO agreements using *Random Forests* (Breiman, 2001; Huynh-Thu et al., 2010). The network is fitted as to best explain the variance of each article's citabilities. Those citabilities are collected from the predictions of the trained deep neural network.

To verify the quality of the fitted network, I compared the fitted network with the jurisprudence of WTO DSB appearing in the Panel and Appellate Body reports. Specifically, I found three major principles of WTO DSB, *Market Access*, *Reciprocity*, and *Non-discrimination*, are clustered in the fitted network. The systematic interactions between articles of WTO agreements are formed as how the Panel and Appellate Body explained in their judicial opinions. As Panel and Appellate Body authoritatively constitute the jurisprudence of WTO DSB, one can conclude that the method qualitatively summarizes the systematic interactions of articles of WTO agreements.

2 Modeling and Formal Definitions

2.1 Network of Articles of the WTO agreements

I define the network of articles of WTO agreements as directed weighted graph $G = (V, \vec{E}, W)$ which is comprised of vertex set V, set of directed edges $\vec{E}$, and edge weight matrix W. I define each legal article of WTO agreement as a vertex, thus $V = \{v \mid v \text{ is a legal article of WTO agreement}\}$. Then I define all ordered pairs of vertices as a set of directed edges $\vec{E}$, thus $\vec{E} = \{(v_i, v_j) \mid (v_i, v_j) \in V \times V)\}$. Finally, I define the edge weight matrix $W = (w(v_i, v_j)) \in [0, 1]^{|V|\times|V|}$ where all incoming edge weights sum up to 1 for all given target vertex v_j, thus $\sum_{v_i \in V} w(v_i, v_j) = 1$. w denotes a map that assigns a weight for each ordered pair of vertices, thus $w : V \times V \to [0, 1]$. I always assign weight 0 for the directed edge comprised of the same vertex, thus $w(v_i, v_i) = 0 \ \forall v_i \in V$. For convenience, I define $w_{ij} = w(v_i, v_j)$.



Fair Adminstration of Laws and Regulations
X:3(a)
9.7%
National Treatment on Internal Taxation
$P(v_j|v_i) = 8\%$
$\sum_{v_i \in V} w(v_i, v_j) = 1$
v_j = III:2
v_i = II:1(b)
Interpretation of Tariff Concession
11%
w_{i_3j}
w_{i_2j}
w_{i_1j}
XI:1
General Elimination of Quantitative Restrictions

(a) **Illustration of $P(v_j \mid v_i)$ where the target article v_j = Article III:2**

> "The dictionary definition of the noun 'excess' is 'the amount by which one number or quantity exceeds another'. More specifically, 'in excess of' means 'more than'. Thus, as a textual matter, a particular number or quantity is 'in excess of' another number or quantity if it is greater, regardless of the extent to which it is greater. ***Looking at the context of Article II:1(b), first sentence, we note that Article III:2, first sentence, of the GATT 1994 is cast in very similar terms and in fact uses the phrase 'in excess of'***:
>
> > *The products of the territory of any contracting party imported into the territory of any other contracting party shall not be subject . . . to internal taxes or other internal charges of any kind in excess of those applied . . . to like domestic products . . .*

(b) **Jurisprudence of Panel in *Russia – Tariff Treatment* case:** Panel explains that the meaning of the term *'in excess of'* in Article II:1(b) clarifies the meaning of the same phrase in Article III:2.

Figure 1: **Modeling of the Network of Articles of WTO agreements**

2.2 Modeling Interaction between Articles of WTO agreements as Conditional Probability

I interpret every directed edge weight $w(v_i, v_j)$ as the conditional probability $P(v_j|v_i) \in [0, 1]$, where the probability represents how probably the source node v_i clarifies the interpretation of the target node v_j comapred to all other source nodes $v \in V \setminus \{v_i, v_j\}$. The articles of WTO agreements interdependently constitute the legal context to clarify the interpretation of other articles as shown in the the Panel report of *Russia-Tariff Treatment* case, as excerpted in Figure 1(b). In *Russia-*

Name of WTO Agreement	Cited Articles
Agreement on Anti-dumping	1, 5.4, 8, 18.1, 18.4
General Agreement on Tariffs and Trade 1994	VI:3, X:3, XXIII:1, VI:2
Agreement on Subsidies and Countervailing Measures	4.10, 7.9, 10, 11.4, 18, 32.1, 32.5
Agreement Establishing the World Trade Organization	XVI:4

Table 1: **Cited articles in *US - Offset (Byrd Amendment)* by complainants**

Tariff Treatment case, the Panel explained that Article II:1(b) clarifies the meaning of the same phrase *'in excess of'* in Article III:2. By modeling this clarification relationship as the directed edge weight w_{ij}, I let the edge weight w_{ij} represent the realtive importance of a source article v_i clarifying the interpretation of the target article v_j. I illustrated this relationship in Figure 1(a).

2.3 Methodological Objective: Finding G^*

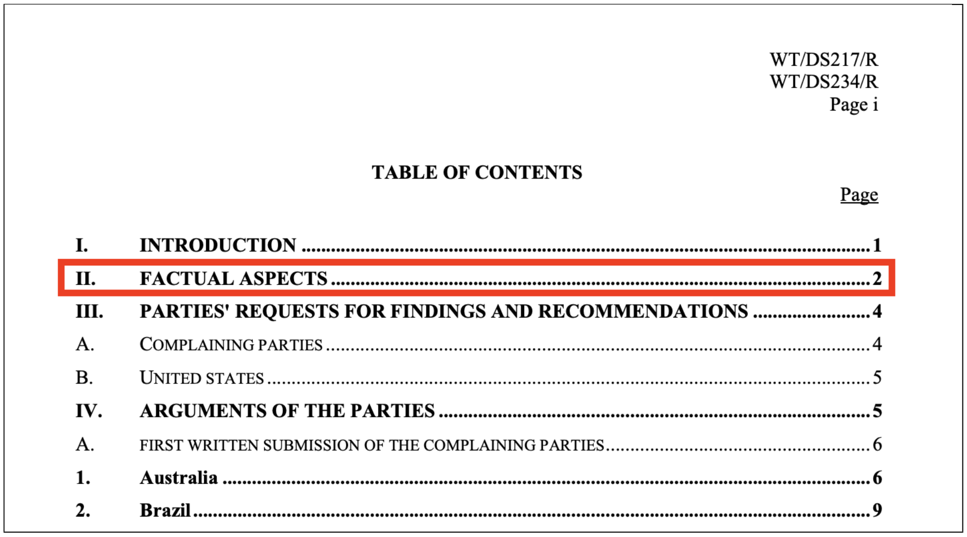
I aim to find $G^* = (V, \vec{E}, W^*)$ where the W^* closely reflect the clarification relationship between articles of WTO agreements as explained by the authoritative judicial bodies of the WTO DSB, Panel and Appellate Body. To find W^*, this paper collected the past 20 years of legal dispute data in WTO DSB. The types and composition of the data collected will be explained in Section 3. Then I design a deep neural network to encode the pattern of interactions of the articles of WTO agreements found in the data. Justification of using deep learning, design and training of deep neural network, and fitting process of W^* using *Random Forest* will be explained in Section 4. Finally in Section 5, I will verify the quality of the fitted G^* by comparing the systematic interaction between articles of WTO agreements found in G^* with the corresponding jurisprudence of the Panel and Appellate Body.

3 Data: Types, Composition and Collection Process

3.1 Overview: How Members Raise Claims in WTO DSB

Members who raise the claim (preferably called *complainant* in WTO DSB) usually cite multiple aritcles of the WTO agreements. This is to cover the complex characteristics of a trade policy that led to the dispute. For example, in the *US - Offset* case, a group of complainants[1] cited articles as shown in Table 1 from the WTO agreements to claim inconsistencies of *Continued Dumping*

[1] Australia, Brazil, Chile, European Communities, India, Indonesia, Japan, Korea and Thailand



WT/DS217/R
WT/DS234/R
Page i

TABLE OF CONTENTS

Page

I. INTRODUCTION **1**
II. FACTUAL ASPECTS **2**
III. PARTIES' REQUESTS FOR FINDINGS AND RECOMMENDATIONS **4**
A. COMPLAINING PARTIES 4
B. UNITED STATES 5
IV. ARGUMENTS OF THE PARTIES **5**
A. FIRST WRITTEN SUBMISSION OF THE COMPLAINING PARTIES 6
1. Australia **6**
2. Brazil **9**

Figure 2: **Table of Contents of Panel Report:** Panel provides factual aspect in the panel report with its page location.

and Subsidy Act of 2000 (CDSOA) of the United States to those cited articles[2].

Upon this understanding of multiple citation, I collected two different types of data. One is textual description of the dispute[3] and the other one is set of articles of the WTO agreements that are cited for each dispute[4]. I will explain each type of data in the following subsections in detail.

3.2 Factual Aspect: Textual Description of the Dispute

Textual description of the dispute is preferably called as *Factual Aspect* in WTO DSB. Since Panel always provide a factual aspect[5] that describes the circumstances of the dispute in each report, I wrote a program that automatically search and collect the panel reports from the WTO official document website[6]. Then I located the factual aspect using the page information from the table of contents in each panel report as shown in Figure 2. I collected the total 143 numbers of different factual aspects . The collected case numbers are listed in Figure 3.

[2] It is worth noting that the WTO agreements comprises many different agreements covering each specific topic in trade such as *Agreement on Anti-dumping, Agreement on Subsidies and Countervailing Measures, Agreement on Agriculture* and so on.

[3] *Check* the CDSOA example at Appendix A.1

[4] *See* Appendix A.3

[5] It's worth noting that Appellate Body doesn't provide any factual aspect because they always use the factual aspect provided by the Panel.

[6] `http://docs.wto.org`

DS 2, 18, 22, 31, 34, 46, 56, 58, 60, 62, 67, 68, 69, 75, 76, 87, 90, 98, 103, 108, 121,
122, 135, 136, 139, 141, 146, 152, 155, 161, 162, 165, 166, 174, 175, 177, 184, 202, 207,
212, 217, 219, 221, 231, 234, 238, 244, 245, 246, 248, 257, 264, 265, 266, 267, 268, 269,
276, 282, 283, 286, 290, 294, 295, 296, 301, 302, 308, 312, 315, 316, 320, 321, 322, 332,
336, 339, 343, 344, 345, 350, 353, 360, 363, 366, 371, 379, 381, 384, 392, 394, 396, 397,
399, 400, 406, 412, 414, 415, 422, 425, 427, 429, 430, 431, 435, 436, 437, 440, 442, 447,
449, 453, 454, 456, 457, 461, 464, 468, 471, 472, 473, 475, 476, 477, 479, 480, 482, 483,
484, 485, 486, 488, 490, 492, 493, 495, 499, 504, 505, 513, 518, 523

Figure 3: **List of the Collected Case numbers:** "DS + number" uniquely identifies each dispute. For example, DS 523 refers to *US — Pipe and Tube Products (Turkey)* where the United States was challenged by Turkey for its possibly inconsistent anti-dumping measure.

3.2.1 Joint Adjudication & Early Settlement

The number 143 may seem small compared to the total 596[7] number of cases that are requested to WTO DSB. This is due to the following two reasons. First, panel jointly adjudicates different cases together if the cases raise the claim toward the same trade policy of the same member state. For example, in *US - Offset (Byrd Amendment)*, panel merged DS217[8] and DS234 together because they were asking the judicial opinion for the same government measure of the United States as shown in Figure 4. This paper selects the smallest case number as a representative number for this case of joint adjudication. For example, DS217 and DS234 share the same panel report then this paper chooses DS217 as a representative number as shown in Figure 3 where the list includes DS217 but not DS234. Second, members sometimes find *mutually agreeable solution* before the panel expresses its judicial opinion by publishing its panel report. Then Panel stops there and no factual aspect is provided. I omitted this kind of *early settled* cases as well.

3.3 Cited Articles: Set of Articles Cited for the Same Dispute

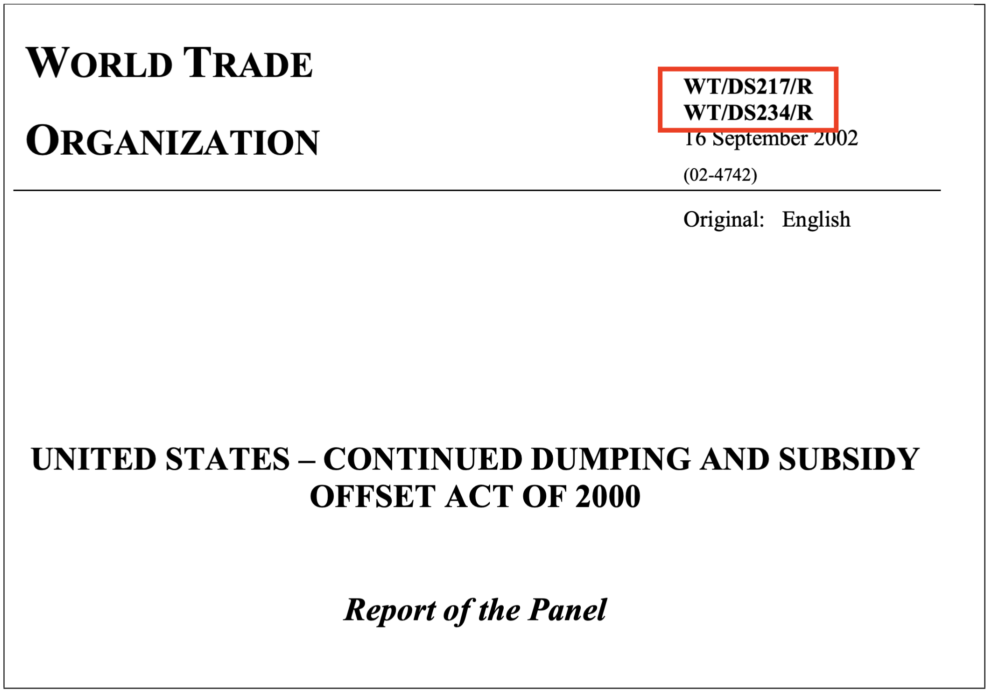
Every legal claim in WTO DSB cites multiple set of articles as shown in Table 1. To collect this set of articles claimed for each dispute, I wrote a program that collects this set of articles cited from the WTO official webpage[9]. The webpage chronologically lists up all dispute cases requested to WTO DSB and the program visits each page and collects the cited articles. Among all the agreements included in the WTO agreements[10], this paper collected articles from the *General*

[7] As of November 1st, 2020.

[8] DS refers to *Dispute Settelement.* DS is the official prefix that indicates the case in WTO DSB.

[9] `https://www.wto.org/english/tratop_e/dispu_e/dispu_status_e.htm`

[10] WTO agreements is comprised of multiple agreements, such as General Agreement on Tariffs and Trade 1994, Agreement on Agriculture, Agreement on the Application of Sanitary and Phytosanitary Measures, Agreement



WORLD TRADE

ORGANIZATION

WT/DS217/R
WT/DS234/R
16 September 2002

(02-4742)

Original: English

UNITED STATES – CONTINUED DUMPING AND SUBSIDY OFFSET ACT OF 2000

Report of the Panel

Figure 4: **Cover of a Panel Report Includes Information about Joint Adjudication:** Panel explicitly marks which different cases are adjudicated together in the cover of the panel report. DS217 and DS234 are handled together in this example.

Agreement on Tariffs and Trade 1994 (GATT 1994) only. This is because articles in GATT 1994 constitutes basic set of trade rules of WTO and other agreements elaborates the articles of GATT 1994 more in detail (World Trade Organization, 1999). For example, the official name of *Agreement on Anti-dumping* is *Agreement on Implementation of* ***Article VI of the GATT 1994*** where the name self-explains that it elaborates on the article VI of GATT 1994. The collected result is listed in the Appendix A.2. Figure 5 lists up 80 different articles of GATT 1994 cited in 143 cases without duplication.

3.3.1 Various Levels of Scope in Cited Articles

As shown in Figure 5, members sometimes cite articles in different levels of scope. For example, For the Article II, member sometimes cites Article II as a whole but sometimes cites Article II:1 or Article II:1(a). This is because two main judicial bodies of WTO DSB, *Panel and Appellate Body*, both constitute its legal precedents citing articles of the WTO agreements in various levels of scope. The various levels of scopes include, *Title, Article, Paragraph, Sentence* or *Term*

on Textiles and Clothing, Agreement on Technical Barriers to Trade, Agreement on Trade-Related Investment Measures, Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade 1994 (antidumping), Agreement on Subsidies and Countervailing Measures, Agreement on Rules of Origin, Agreement on Safeguards and so on.

I, I:1, II, II:1, II:1(a), II:1(b), II:2, II:3, III, III:1, III:2, III:4, III:5, III:7, IV, IX, IX:2, V, V:1, V:2, V:3, V:3(a), V:4, V:5, V:6, V:7, VI, VI:1, VI:1(a), VI:1(b), VI:2, VI:3, VI:5(a), VI:6, VII, VII:1, VII:2, VII:5, VIII, VIII:1, VIII:3, VIII:4, X, X:1, X:2, X:3, X:3(a), XI, XI:1, XIII, XIII:1, XIII:2, XIII:3(b), XIX, XIX:1, XIX:2, XIX:3, XV, XVI, XVI:1, XVI:4, XVII, XVII:1, XVII:1(c), XVIII, XVIII:10, XVIII:11, XX, XXI, XXII, XXII:1, XXIII, XXIII:1, XXIII:1(a), XXIII:1(b), XXIV, XXIV:12, XXIV:5(b), XXIV:6, XXVIII

Figure 5: **Set of articles of GATT 1994 collected and used in this paper:** These articles comprises the node set V and their ordered pairs comprise the edge set $\vec{E}$ as formally defined in Section 2.1

$$\begin{gathered}\text{Let } D \text{ is a set of DS case numbers listed in Figure 3.}\\ \text{Then there exists } c_d = \{v_d \in V \mid \; v_d \text{ is an article cited in the case } d \in D\}\\ \text{where } V \text{ is set of articles listed in Figure 5.}\\ \text{Then define set of cited articles } C = \{c_d \mid d \in D\}\end{gathered}$$

Figure 6: **Formal Definition of Set of Cited Articles:** I formally define a set of cited articles C. The elements of C are listed in Appendix A.2.

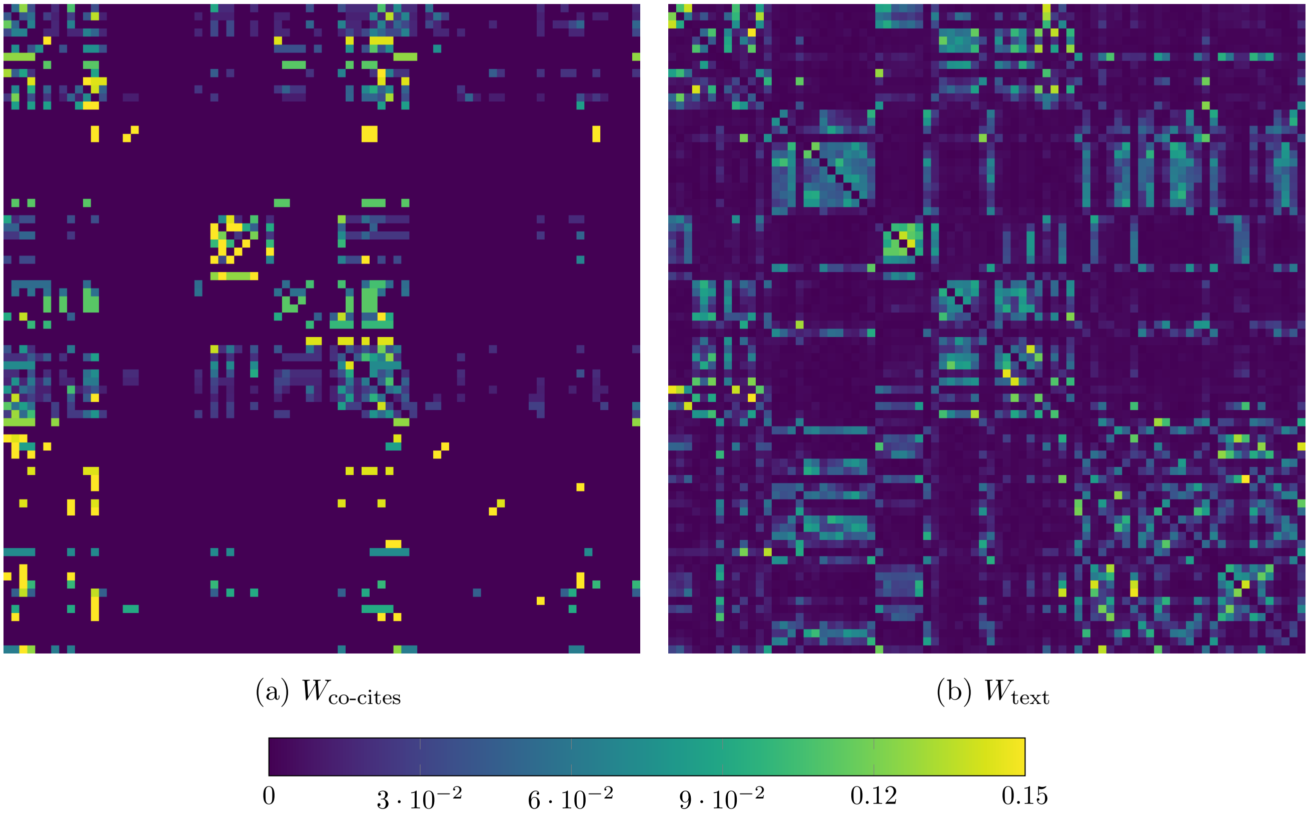
as shown in Appendix A.3. Following this practice, members also cite articles in different levels of scope to make their legal claim fit and valid according to the general jurisprudence of WTO DSB.

4 Methodology: Considerations and Development

This section introduces two main considerations to design the method to qualitatively fit G^*. I justify the use of the deep learning upon those considerations. Then I explain the structure of the deep neural network and its training process. After training of the deep neural network, I explain the process of fitting the network G^* using *Random Forest* (Breiman, 2001; Huynh-Thu et al., 2010).

4.1 Two Main Considerations: Utilizing Information in Textual Data & Generalization of Member-Specific Citation

I considered two main points to determine its method to qualitatively fit the edge weight matrix W^* defined in Section 2.1. One is to utilize the information in textual data, the factual aspect



(a) $W_{\text{co-cites}}$

(b) W_{text}

Figure 7: **Heatmap of Two Different Edge Weight Matrices:** Above two subfigures visualizes two different *edge weight matrices* $W_{\text{co-cites}}$ and W_{text}. One can check that $W_{\text{co-cites}}$ is sparser than W_{text}. It means $W_{\text{co-cites}}$ captures fairly less interactions compared to W_{text}

and the legal article as exemplified in Appendix A.1 and Appendix A.4 respectively. The other one is about how to generalize member-specific citation patterns. Since members of the WTO strategically cite the articles of WTO agreement expecting different outcomes according to each member's specific interest (Johns and Pelc, 2014; Pelc, 2014; Strezhnev, 2014), I select a method that can generalize these member-specific citation patterns.

4.1.1 Importance of Using Textual Information

I emphasizes the necessity of using textual information to qualitatively fit the edge weight matrix W^* as defined in Section 2.1. Rather than using the textual data, one can simply model the co-citation matrix as W^*, which counts the co-occurrences of each article with other articles. However, it simply allocates a large edge weight for frequently cited articles and fails to explain how articles systematically interact with other articles. This failure is mainly due to the insufficient information in the co-citation matrix. Members cite the articles of the WTO agreements based on the complex characteristics of the trade policy that led to the dispute. However, the co-citation pattern omits this contextual information . To emphasize the necessity of using the textual information, I

$$\text{Let } \delta_{ij}^{d} \text{ is defined to be 1 if } \{(v_i, v_j) \mid v_i, v_j \in V \text{ and } i \neq j\} \subset c_{d \in D} \ \text{ else } 0$$

$$\text{where } V, D \text{ and } c_d \text{ is defined as in Figure 6.}$$

$$\text{Then let } \textit{co-citation matrix } M = (m_{ij}) \in \mathbb{N}^{|V| \times |V|} \text{ s.t. } m_{ij} = \sum_{d \in D} \sum_{i,j \in V} \delta_{ij}^{d}$$

(a) **Formal Definition of Co-citation Matrix**

	I	**I:1**	**II**	**II:1**	⋯
I	**0**	**3**	**7**	**2**	
I:1	**3**	**0**	**3**	**4**	
II	**7**	**3**	**0**	**4**	
II:1	**2**	**4**	**4**	**0**	
⋮					

(b) **Illustration of Co-citation Matrix**

Figure 8: **Formal Definition and Illustration of Co-citation Matrix:** This paper defines co-citation matrix M as subfigure (a) and it's illustrated as subfigure (b) using the paper's dataset. Note that the co-citation matrix is *symmetric*, $m_{ij} = m_{ji}\ \forall i, j \in V$.

prepared two different matrices $W_{\text{co-cites}}$ and W_{text} that are both following the definition of *edge weight matrix* W in Section 2.1. $W_{\text{co-cites}}$ is calculated using the co-citation pattern between the articles of the WTO agreements as formally defined in Figure 9. W_{text} is the one fitted using the textual information and the way how it's fitted will be explained at the following bodies of this section, in particular in Section 4.3.2. Two heatmaps visualized in Figure 7 shows how sparse the $W_{\text{co-cites}}$ is compared to the W_{text}. This sparsity indirectly refers to the insufficient information to qualitatively map the jurisprudences of WTO DSB. In contrast with it, if we fit the *edge weight matrix* W using the textual information, we get a more dense and informative matrix as visualized in Figure 7(b). Upon this observation, this paper justifies the use of a deep neural network to process information embedded in the text description of the dispute and regulatory content of the articles. This is because deep neural network is known to effectively extract information from the textual data to perform various tasks such as text classification (Minaee et al., 2020), text summarization (Magdum and Rathi, 2020) and text generation (Guo et al., 2017).

$$\text{For given } M \text{ defined in Figure 8(a),}$$
$$\text{let } \textit{normalized co-citation matrix } N = (n_{ij}) \in \mathbb{R}^{|V| \times |V|} \text{ s.t. } n_{ij} = \frac{m_{ij}}{\sum_{j \in V} m_{ij}}$$

(a) **Formal Definition of Normalized Co-citation Matrix**

	I	**I:1**	**II**	**II:1**	**···**
I	**0**	**0.053**	**0.125**	**0.035**	
I:1	**0.040**	**0**	**0.04**	**0.054**	
II	**0.114**	**0.049**	**0**	**0.065**	
II:1	**0.032**	**0.065**	**0.065**	**0**	
⋮					

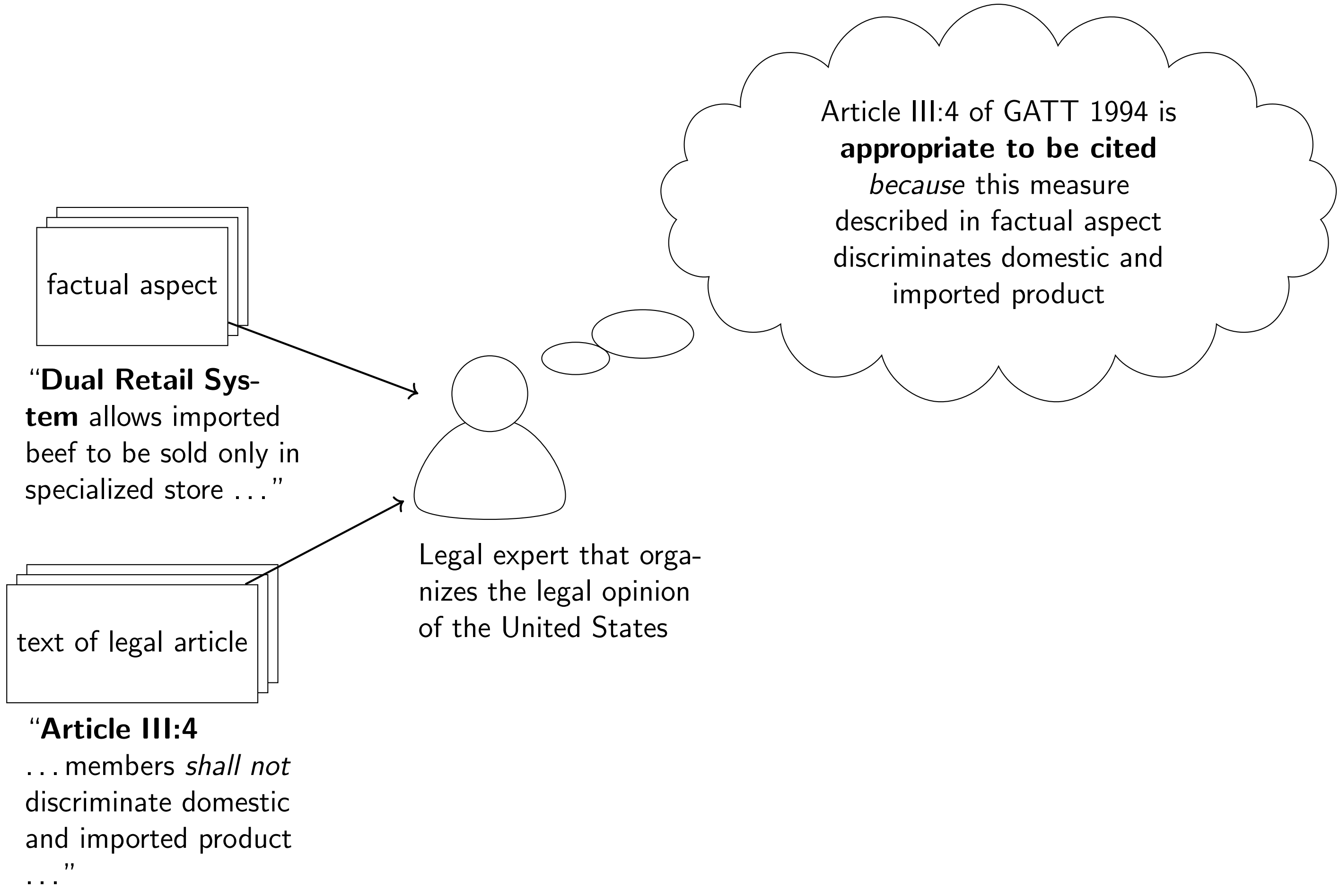
$\sum_{j \in V} n_{ij} = 1$

(b) **Illustration of Noramlized Co-citation Matrix**

Figure 9: **Formal Definition and Illustration of Normalized Co-citation Matrix:** This paper defines normalized co-citation matrix N of M as subfigure (a) and it's illustrated as subfigure (b) using the paper's dataset. Note that normalized co-citation matrix is no more *symmetric*, $n_{ij} \neq n_{ji}\ \forall i, j \in V$. This definition is prepared to fit the definition of co-citation matrix to that of W as defined in Section 2.1.

4.1.2 Generalization of Each Member's Strategic Citation

This paper aims to map the regulatory system of WTO DSB in a form of *directed weighted graph G* as defined in Section 2.1. To achieve this objective, we need to fit W to generalize member specific strategic citation behavior. Deep neural network is known as good at generalization despite its large capacity (Neyshabur et al., 2017), possible instability of training algorithm (Charles and Papailiopoulos, 2017), nonrobustness (Zahavy et al., 2017), and sharp minima (Dinh et al., 2017). Therefore, this paper trains a deep neural network without any member specific information, such as geolocation, GDP or specialized industry. By training a deep neural network only using the text description of the trade dispute and the legal text of the cited articles, this paper expects the fitted $G^* = (V, \vec{E}, W^*)$ can show interactions between articles of the WTO agreements without being biased to member-specific citation patterns.



factual aspect
“**Dual Retail System** allows imported beef to be sold only in specialized store . . . ”
text of legal article
“**Article III:4**
. . . members *shall not* discriminate domestic and imported product . . . ”
Legal expert that organizes the legal opinion of the United States
Article III:4 of GATT 1994 is **appropriate to be cited** *because* this measure described in factual aspect discriminates domestic and imported product

Figure 10: **Visualization of How Member Cites in WTO DSB (Citable Case):** With two different contexts, factual aspect and text of legal articles, member judges whether the given legal article is appropriate to be cited or not.

4.2 Design of Deep Neural Network

Upon the justification of using deep neural network with above two main considerations in Section 4.1, I explain the design process of the deep neural network that can encode the pattern of citations in WTO DSB. Since the citations are performed upon the general understanding over jurisprudences of WTO DSB, encoding of those citation patterns could reflect the jurisprudences of WTO DSB.

4.2.1 Design Input/Output of Deep Neural Network: by Analogy with How Member Cites in WTO DSB

A rule of thumb to design input and output of deep neural network is to mimic how humans do for a given task. Therefore I present a visualization of how legal experts of WTO agreements determine whether to cite a legal article of WTO agreements with an example of *Korea - Beef* case

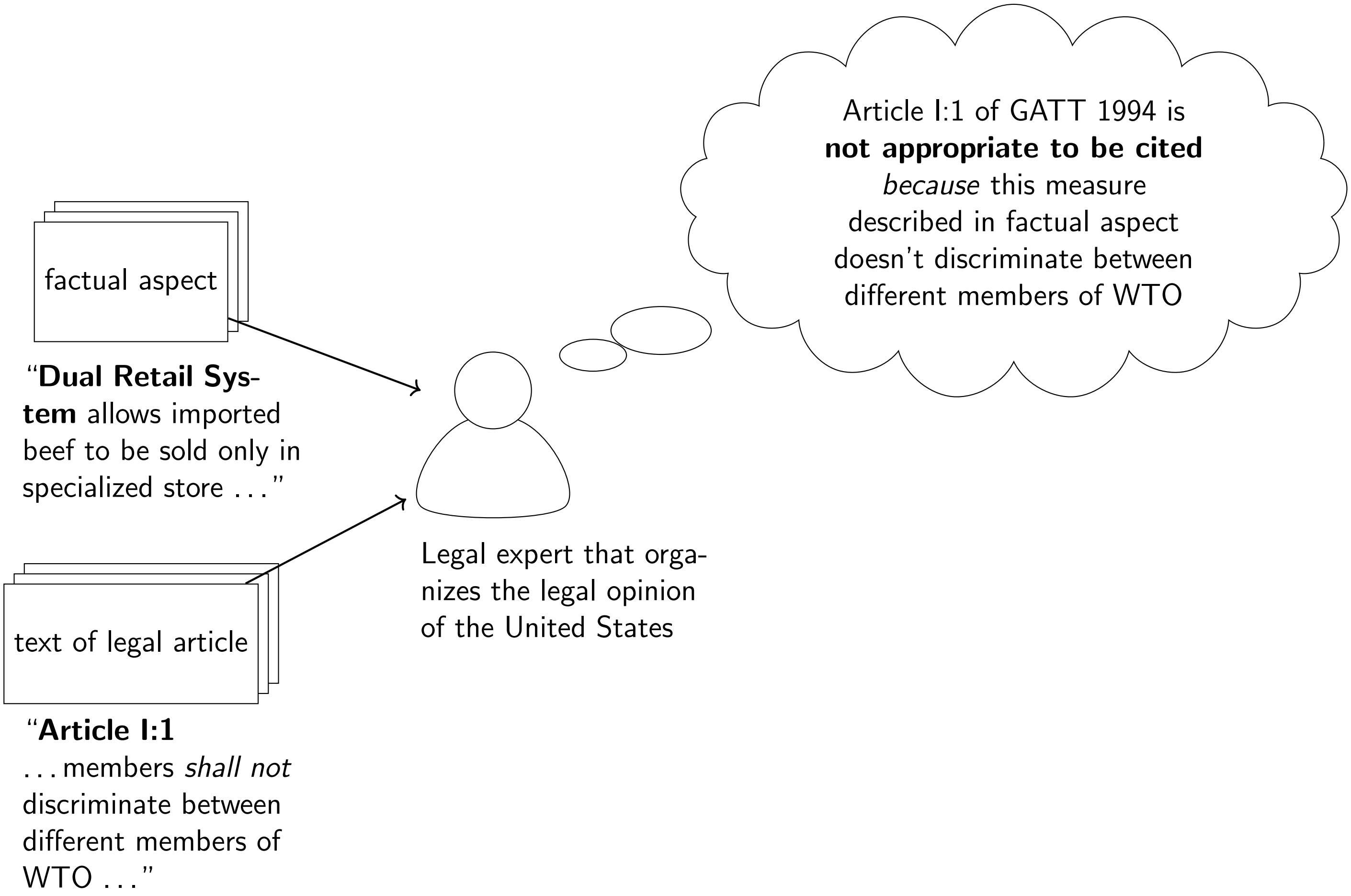

Article I:1 of GATT 1994 is **not appropriate to be cited** *because* this measure described in factual aspect doesn't discriminate between different members of WTO
factual aspect
"**Dual Retail System** allows imported beef to be sold only in specialized store . . . "
Legal expert that organizes the legal opinion of the United States
text of legal article
"**Article I:1** . . . members *shall not* discriminate between different members of WTO . . . "

Figure 11: **Visualization of How Member Cites in WTO DSB (Non-Citable Case):** With two different contexts, factual aspect and text of legal articles, member judges whether the given legal article is appropriate to be cited or not.

(Figure 10 and Figure 11). In this case, the United States raised a claim relating to the *Dual-retail system* maintained by South Korea. In *Dual-retail system*, South Korea maintained two distinct retail systems for imported and domestic beef. There existed stores specialized for imported beef and they can sell only imported beef and cannot sell domestic (Korean) beef. U.S. claimed that the *Dual-retail system* is inconsistent to the Article III:4 (National Treatment) of GATT 1994 because *Dual-retail system* discriminates between domestic and imported beef. A measure that discriminates domestic and imported products falls under the scope of the Article III:4 of the GATT 1994, which states the principle of *National Treatment* that prohibits the discrimination between imported and domestic product. However, U.S. didn't cite the Article I:1 of GATT 1994 that prohibits the discrimination between members of WTO because *Dual-retail system* didn't discriminate against the United States from other countries who export beef to South Korea, such as Argentina, Australia, etc.

We can understand that there exists a shared understanding over jurisprudences of WTO DSB

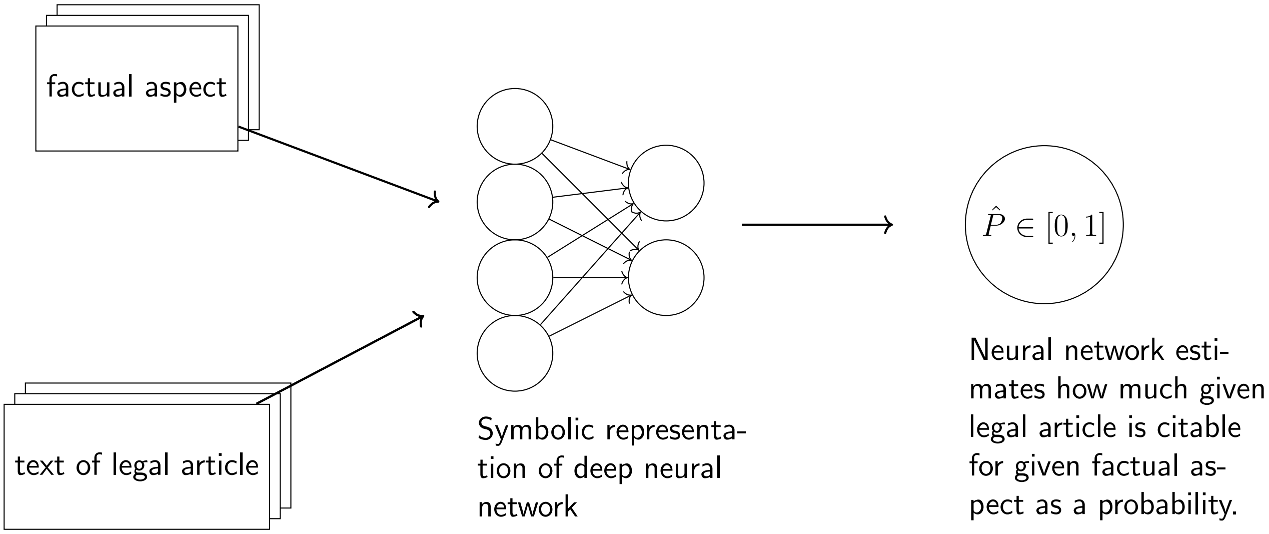

factual aspect
text of legal article
Symbolic representation of deep neural network
$\hat{P} \in [0,1]$
Neural network estimates how much given legal article is citable for given factual aspect as a probability.

Figure 12: **Design of Training Framework of Deep Neural Network:** I designed a training framework of deep neural network by analogy with how member cites in WTO DSB as visualized in Figure 10 and Figure 11.

among legal experts of the WTO agreements. They follow up new cases and study jurisprudence stated in the *Panel* or *Appellate Body* reports. Then they organize a legal argument by citing certain article(s) upon this shared understanding for given possible inconsistent measures claimed by a member of WTO.

To mimic this reasoning process, I designed the input and output of the deep neural network as illustrated in Figure 12 and formally defined in Figure 13. The neural network is designed to estimate the citability for a given pair of a factual aspect and text of a legal article. By iteratively training the neural network with data explained in Section 3, I expect the neural network can learn a shared understanding over jurisprudences of WTO DSB closely to that of legal experts. The detailed structure of neural network and training schemes will be explained in the later subsections.

For given V, D defined in Figure 5 and Figure6,

let $E = \{e \mid e$ is an english word or special charcter$\}$ and

$n_{\text{factual}}, n_{\text{article}} \in \mathbb{N}$ represents *max token length* of factual aspect and legal article respectively.

Then define $T = \{t_d \mid t_d = (e_1, e_2, \ldots, e_{n_{\text{factual}}})$ *s.t.* $d \in D$ and $e_{i \leq n_{\text{factual}}} \in E\}$

where t_d represents a factual aspect of one of DS cases listed in Figure 3.

Also define $A = \{a_v \mid a_v = (e_1, e_2, \ldots, e_{n_{\text{article}}})$ *s.t.* $v \in V$ and $e_{i \leq n_{\text{article}}} \in E\}$

where a_v represents texts of one of a legal article listed in Figure 5.

Now defines a deep neural network f with a set of parameters θ

$$f_\theta : T \times A \to [0, 1]$$

Figure 13: **Formal Definition of Input/Output of Deep Neural Network**: T and A represents a set of "documents" that are factual aspects and text of legal articles respectively. The term "document" refer to a *tuple* of English words or special characters like $(e_1, e_2, \ldots, e_{n_{\max}})$. Then a pair of two documents, factual aspect and legal article, is fed into the neural network f and returns a probability that represents how much the given article is citable for the given factual aspect.

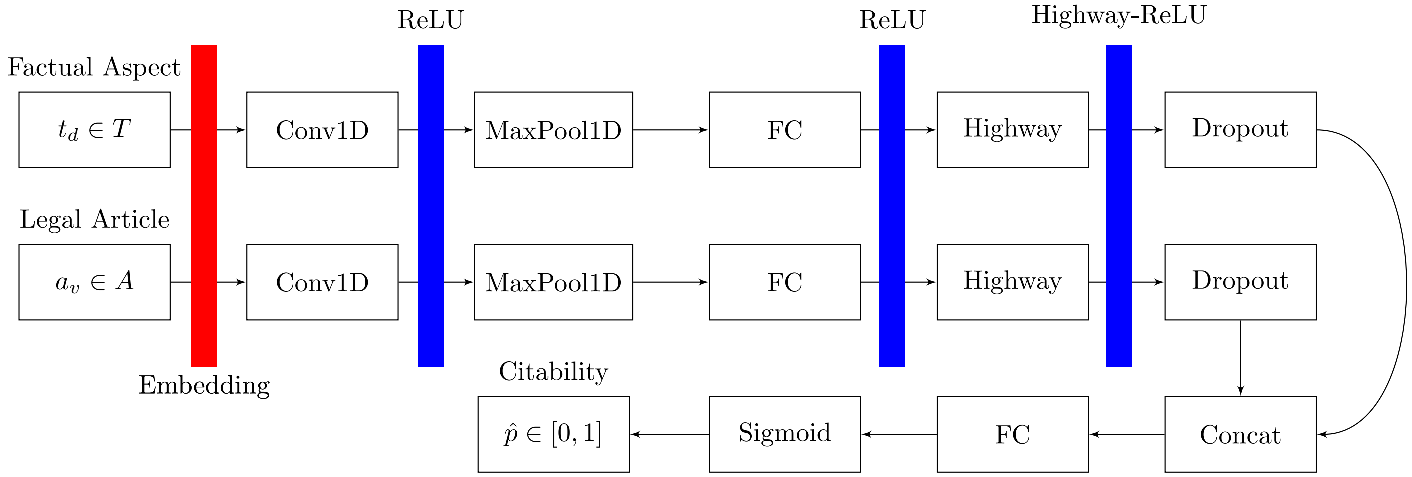


Figure 14: **Layers of Deep Neural Network:** The term "Layer" refers to a mathematical operation. How layers are stacked in which order determines the "structure" of deep neural networks. This figure illustrates the structure of the deep neural network that this paper used. The figure follows the notations defined in Figure 13.

4.2.2 Structure of Deep Neural Network

An efficient way to architect the structure of deep neural network is to use the human analogy as I did in subsection 4.2.1. A legal expert reads the text line by line and creates several local understandings. Then he/she merges those local understandings into a global level to summarize an essential information to determine whether the given article is citable or not for the given factual description of the dispute.

To analogy this process, I borrowed *1-Dimensional Text Convolutional Neural Network* (TextCNN) from Kim (2014). This is because of the following two reasons. First, convolutional neural networks are known as good at learning how to integrate local and global features to perform classification tasks (Lawrence et al., 1997; Nahid and Kong, 2017). Second, Kim (2014) has implemented this convolutional neural network into the text domain and has shown its high performance on classification.

Figure 14 demonstrates the entire structure of the deep neural network that is used in this paper. General flow is to return the citability of an legal article that is fed into the neural network with a factual aspect. Each block in Figure 14 represents a set of unique mathematical operations. We prefer to call those blocks as "layer" and each layer has its desired role regarding how to process the information for which purpose. I will explain each block's role and composition in the following subsections.

4.2.2.1 Inputs: a Pair of Documents, Factual Aspect and Legal Article

I defined the neural network in terms of its input and outputs in Figure 13. First we need to *tokenize* the text of factual aspect and legal article. The term "*tokenize*" refer to the process of decomposing the text into the sequence of words or special characters. I used the *off-the-shelf* tokenizer provided by *Spacy* API[11]. Since the neural network is represented as a *function*, it must hold a predefined input and output dimension. I checked the *max token length* over all tokenized results for factual aspects and legal articles. The *max token length* was $35,842$ and $20,158$ for factual aspects and legal articles respectively. These numbers correspond to n_{factual} and n_{article} in Figure 13. Then I *padded* a special token [**PAD**] at the tail in case the token length of a factual aspect or legal article is shorter than those *max token lengths*.

It's worth noting that the field of deep learning preferably calls this ordered set of tokens as '*Document*'. Therefore I prepare documents from the raw data collected in Section 3 at this stage and move on to the next step, *Embedding Layer*.

4.2.2.2 Embeddiing Layer: From Documents to Numerical Vectors

Since the deep neural network is comprised of mathematical operations, we need to transform the word tokens into a form of numerical vectors. This process can be conducted with a single *Embedding Layer*. This layer is defined as $|\textit{Size of Dictionary}| \times k$ matrix where this matrix works as a dictionary for the neural network. Neural network refers to this matrix to find the meaning of a token in a form of k-dimensional vector and updates its value while it's trained to reflect domain specific meaning of each token.

For example, WTO DSB prefers to use the word *inconsistent* rather than the word *breach* to refer to the illegality of a member's trade policy. This kind of domain specific information will be stored in this matrix as a distribution of those numerical vectors in this layer.

I used *Google News Word2Vec*[12] to initialize this embedding layer. This is an open-source pre-trained vector provided by Google. It contains 300 dimensional English word vectors for 3 million unique words. Because of the GPU memory limit, I set 400,000 as a maximum number of word vectors to read form the Google Word2Vec and this number corresponds to the $|\textit{Size of Dictionary}|$.

Figure 15 represents the "output" of the embedding layer for given document that is tokenized from the example in Appendix A.1. The neural network find correspondent k-dimensional vectors

[11] https://spacy.io/api/tokenizer

[12] https://code.google.com/archive/p/word2vec/

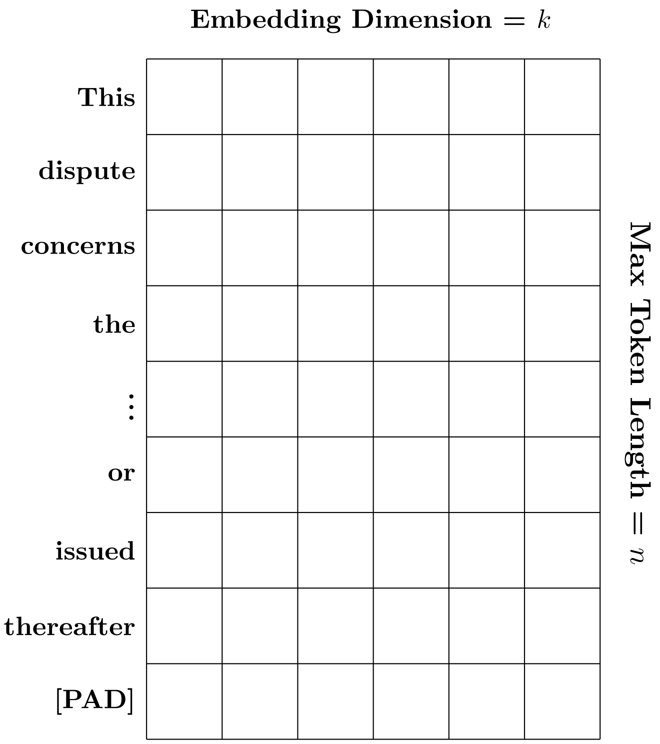


Figure 15: **Output of Embedding Layer:** This figure illustrates an output of embedding layer. Embedding layer maps each token in the input document to the specified numerical vector in dimension k.

for given ordered tokens in the document and returns (n_{factual}, k) or (n_{article}, k) size of matrix for given two different types of inputs. This matrix is fed into the next layer, *Conv1D*.

4.2.2.3 Conv1D: Capturing the Local Features

This subsection explains the how convolution filters runs over the (n_{factual}, k) or (n_{article}, k) size of matrix that are passed from the previous layer. I define this output matrix from the embedding layer more formally as below following the notation used in Kim (2014)

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_{\max}$$

where $\oplus$ represents concatenation and

$x_i \in \mathbb{R}^k$ represents the i-th embedded token from the document.

Let $x_{i:i+j}$ refer to the concatenation of the embeddings$x_i, x_{i+1}, ..., x_{i+j}$

Then a convolution filter $w \in \mathbb{R}^{h \times k}$ is simply defined as (h, k) size of matrix where h represents the filter size and k represents the embedding dimension which is same to that of the embedding layer. Then the convolution is defined as $w \cdot x_{i:i+h-1} + b$ where $b \in \mathbb{R}$ is a bias term. I illustrated

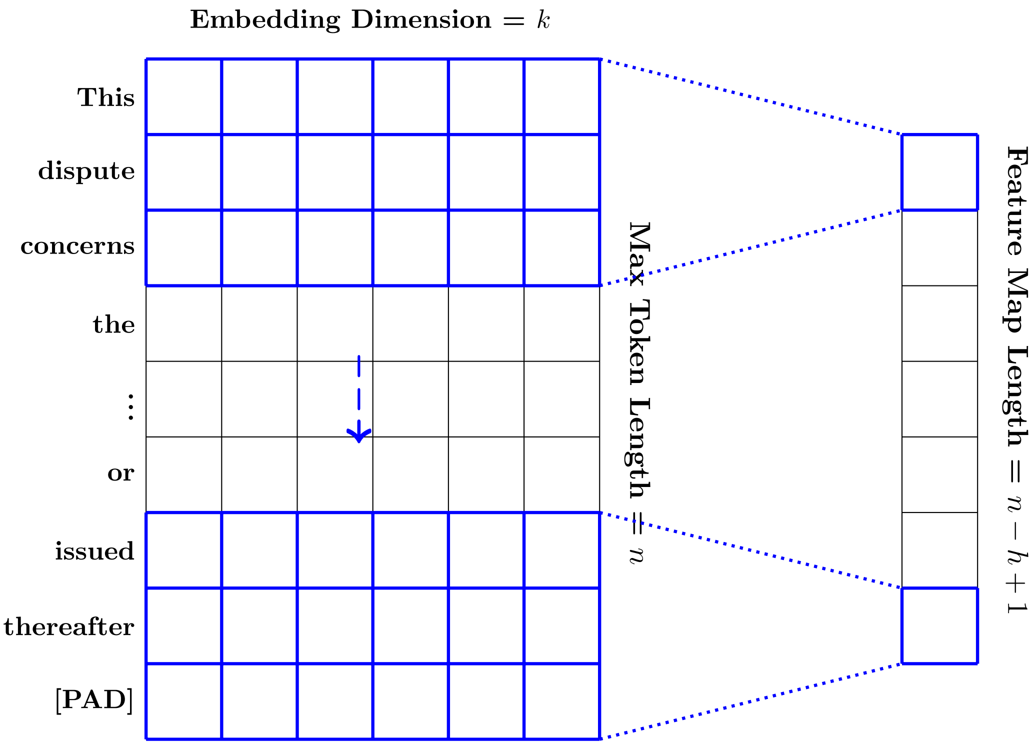


Figure 16: **Conv1D:** h-sized filter runs over the $n \times k$ embedding matrix and produces $(n-h+1)$ size of feature map.

this convolution operation in Figure 16 where the blue box refers to the convolution filter w with the filter size 3 and it runs over the each set of 3 embeddings as increasing the i of x_i one by one. This will eventually generate a feature vector c which is a size of $n - h + 1$.

Each convolution filter w learns how to summarize information *locally* for the region of filter size. For example, a convolution filter w holds its unique way of understanding the meaning of the ***"This dispute concerns"*** in Figure 16. This convolution filter w generalizes its own way of understanding 3 token embeddings together over all $n-h+1$ number sets of 3 token embeddings.

I prepared 128 different convolution filters $\{w\}$ so that each of them can hold their unique way of comprising *local understandings* where the size of locality corresponds to its filter size. Moreover, I also prepared three different filter sizes 3, 4 and 5. This means we will eventually get $(n-2, 128), (n-3, 128)$ and $(n-4, 128)$ size of vectors by the *Conv1D* operation.

It's important to note that I introduced *non-linearity* using **ReLU** ($ReLU(x) = max(0, x)$) after each convolution operation. This means output of convolution operation $w \cdot x_{i:i+h-1} + b$ becomes 0 in case the value is smaller than 0. This sequential layering of linear (such as convolution operation) and non-linear operations (such as ReLU) lets the model to be able to encode more complex patterns like citation patterns appearing upon the jurisprudence of WTO DSB.

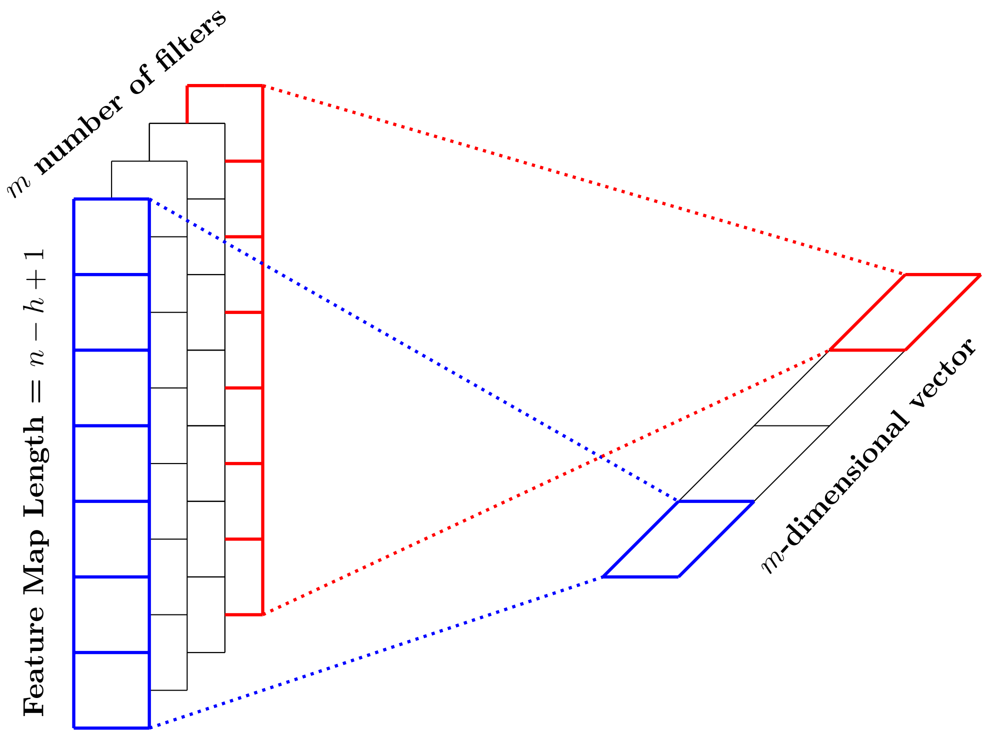


Figure 17: **MaxPool1D:** Filter out max value for all m number of feature map outputs from m different convolution filters. MaxPool1D produces a m dimensional vector as an output for collection of those filtered max values.

4.2.2.4 MaxPool1D: Choose the Most Prominent Feature

Our general goal is to estimate the citability $\hat{p} \in [0, 1]$ as closely to that of real world citation patterns. Therefore we need an efficient *down-sampling* process to sample the large dimensional features into smaller dimensions. With a convolution, *Max-pooling* is preferred to use for this purpose. *Max-pooling* simply selects the largest values among $(n - h + 1)$ size of vector where each of dimension is calculated by the same convolution filter w, i.e., $ReLU(w \cdot x + b)$. Figure 17 illustrates this process. I used 128 ($= m$ in Figure 17) number of filters for each filter size, thus I got 3 different 128-dimensional vectors as a result of *MaxPool1D*.

4.2.2.5 FC: Enlarge the Capacity

Using the $(3, 128)$ size of feature map from the *Max-pooling*, I flatten those three (128) dimensional feature maps into a (384) dimensional vector. To increase the model capacity in terms of size of parameters and non-linearity, I introduced a upscaling *Fully Connected Layer* (FC) that is defined as $W_{fc} \cdot x + b_{fc}$ where $W_{fc} \in \mathbb{R}^{1024 \times 384}$ is a linear map that increases the feature dimension, x is the flattened (384) size of feature vector and $b_{fc} \in \mathbb{R}^{1024}$ is a bias term. Then by applying **ReLU** again to the output of this *Fully Connected Layer*, I introduced another *non-linearity.* This is to let the model to introduce more complexity in terms of nonlinearity without losing important features from this non-linear operation by introducing upscaled dimension to the feature map. This operation returns a (1024) size of feature vector.

4.2.2.6 Highway: Intorducing another Non-linearity while Preventing the *Vanishing Gradients*

I implemented *Highway network* (Srivastava, Greff and Schmidhuber, 2015) to add more non-linearity. Adding non-linearities increases the chance that the neural network can approximate more complex patterns in the data, however, as the network gets deeper *vanishing gradients* problem arises. Deep neural network is trained by updating its parameter weight proportional to the partial derivative of the loss function with respect to the current parameter (*See* Algorithm 1) and computes this gradient by chain rules. Therefore, as the network gets deeper in terms of non-linearity, the value of gradient tends to become *vanishingly small* compared to that of the front layer.

Highway network prevents this *vanishing gradient problem* by introducing an additional parameter that learns the *adequate amount of non-linearity.* The highway network is formally defined as following.

$$Y = H(W_H \cdot x + b_H) \cdot T(W_T \cdot x + b_T) + x \cdot (1 - T(W_T \cdot x + b_T))$$

$$\text{where } H \text{ is ReLU, } T \text{ is Sigmoid}$$

$$(W_H \in \mathbb{R}^{1024\times 1024}, W_T \in \mathbb{R}^{1024\times 1}, b_H \in \mathbb{R}^{1024} \text{ and } b_T \in \mathbb{R} \text{ in this paper's setting})$$

By defining W_T a linear map that returns 1-dimensional output, the sigmoid of its output $T(W_T \cdot x + b_T)$ is confined in $[0, 1]$. Therefore, the network gets to introduce $T(W_T \cdot x + b_T)$ amount of ReLU to $W_H \cdot x + b_H$ where x is (1024) shape of feature vector from the previous layer.

4.2.2.7 Dropout: Ensure more Generalizability

Highway network returns (1024) shape of feature vector and I apply *dropout* operation (Srivastava et al., 2014) with the drop rate 0.5 on this feature vector. Dropout randomly sets the value of each dimension as 0 or scales up by $1/(1 - \text{drop rate})$ thus dimension-wise sum of the feature vector is unchanged. Dropout is one of the most widely accepted regularization techniques which prevents overfitting and helps the model to achieve its generalizability.

4.2.2.8 Concat: Feature Map from Factual Asepct and Legal Content Meets

We have run through the two same processes of Embedding, Conv1D, MaxPool1D, FC, Highway and Dropout for two different types of data, *Factual Aspect* and *Legal Article*. This generates two (1024) sizes of feature vectors. We simply concatenate those two and generate a (2048) size of feature vector.

4.2.2.9 FC: Generates Logit

Each feature vector before concatenation in the previous layer corresponds to the understanding of the neural network for each type of data. Now we simply concatenated those two and reduce the size to 1 by applying *Fully Connected Layer* with the liner map $W_{final} \in \mathbb{R}^{2048\times 1}$ with bias $b_{final} \in R$. This generates a scalar and we consider it as a logit, $\log \frac{\hat{p}}{1-\hat{p}}$

4.2.2.10 Sigmoid: Generates Citability

By applying *Sigmoid* to this size 1 logit, we get $\hat{p} \in [0, 1]$.

4.2.3 Train of Deep Neural Network

I have total 11,440 number of data instances that is calculated by $|D| \times |V|$ where $|D| = 143$ and $|V| = 80$. Before training, I randomly split the entire dataset $T \times A$ into train and test data in 8:2 ratio. The number of split results is 9,153 for train data and 2,287 for test data. I used train data only to train the neural network and used test data to check the trained model's performance. By measuring a performance metric on the inference results of the test data, one can check the generalizability of the trained neural network regarding how well the trained neural network performs over the data that it has never seen before.

The term "training" refers to adjust a set of parameters θ that constitutes mathematical operations defined in layers illustrated in Figure 14 and explained in the following subsections. The procedure to fit those set of parameters is described in Algorithm 1.

I fitted the nerual network f_θ using *weighted cross entropy loss* as formally defined in Algorithm 1. I used *cross entropy loss* because it measures how much the probability distribution projected by the trained model f_θ deviates from the true distribution which represents a shared understanding regarding the citability of certain legal article for given textual description of the case among the group of legal experts of WTO agreements.

Algorithm 1: Steps to Train Neural Network f_θ

```
Input: neural network $f_\theta$, train dataset $T_{train} \times A_{train}$ and set of cited article $C$ defined
       in Figure 6
Output: fitted set of parameter $\theta^*$
Let weight of binary cross entropy $w = 26.303$
    $\textit{weighted binary cross entropy loss } L(y, \hat{p}) = w * y \log \hat{p} + (1 - y) \log (1 - \hat{p})$
    Epochs $e \in \mathbb{N}$
    Learning Rate $\alpha \in \mathbb{R}$
for ( $i = 1,\ i{+}{+},\ \textit{while } i \leq e$ ) {
    for ( $(t_d, a_v) \in T_{train} \times A_{train}$ ) {
        $y \leftarrow 1$ if $v \in c_d$ else 0
        $\theta \leftarrow \theta - \alpha * \nabla_\theta L(y, f_\theta(t_d, a_v))$
return $f_\theta^*$
```

It's worth noting that I "weighted" *cross entropy loss*. This is because our dataset is highly imbalanced in terms of citability. Among all 11,440 data instances, I have only 435 data instances where the given article a_v is actually cited for the given case description t_d. This is only 3.802%. This is because a case tends to have only 3 - 4 articles cited on average among 80 different articles in V. Therefore, I adopted a weight of 26.303 which is inverse of the 3.802% to penalize the neural network with higher loss in case the network fails to cite correctly for the positive case where $y = 1$ in Algorithm 1.

Epochs e and learning rate α in Algorithm 1 is a hyperparameter whose value shall be determined before training. First, one epoch refers to one cycle that the neural network trains over the entire training data once and I set e as 15. Therefore, this neural network sees each training data instance 15 times. Second, the learning rate is about how much we are adjusting the model parameters with respect to the loss gradient. I trained the neural network with learning rate of 0.01 but decays with the rate of 0.95 for seeing every 5000 data instances because gradual decay of learning rate is known as good at preventing the training from being stuck in the local minima.

4.2.4 Training Result: AUC-ROC as Performance Metric

Train loss and test loss converged to around 2 and 1.25 after epoch of 10 respectively. However one needs a well-defined performance metric that can measure how well a model performs for a given task. Our task can be understood as a classification problem where the model decides a pair of factual aspects and the legal article falls into the citable case or not. Therefore I used the AUC-ROC metric to measure the model performance.

The AUC-ROC measures model performance at various threshold settings. Since the model accuracy varies upon the threshold selected, one needs to consider every case of threshold together to measure the model's classification power conclusively. ROC is measured with a given threshold $t \in [0, 1]$ and defined as follows.

$$\begin{aligned} \text{ROC}(t) &= \frac{\text{TPR (True Positive Rate)}}{\text{FPR (False Positive Rate)}} \\ \text{where TPR} &= \frac{\text{TP (True Positive)}}{\text{TP (True Positive)} + \text{FN (False Negative)}} \\ \text{FPR} &= \frac{\text{FP (False Positive)}}{\text{FP (False Positive)} + \text{TN (True Negative)}} \end{aligned}$$

Then AUC-ROC measures the "area" under the ROC curve where the curve represents ROC value at every threshold of t. The maximum value of AUC-ROC is 1 and its baseline is 0.5 where the model randomly estimates the case as citable or not. Our model's AUC-ROC converged to around 0.85 after the epoch of 5. It means that the model encoded the pattern inside data and predicts better than the naive baseline. However, we need more substantial analysis of what the value 0.85 means. I will generate a network of articles using this trained neural network at the following subsections and will empirically interpret the generated network with my background knowledge about jurisprudences of WTO DSB in Section 5.

4.3 Fitting $G^* = (V, E, W^*)$ using Random Forests

This subsection explains how I found the best set of directed edge weights W^* that closely maps the shared understanding about jurisprudences of WTO DSB among legal experts in a form of *directed weighted graph* as defined in Figure **??** and illustrated in Figure 1 using the trained neural network f_{θ^*}.

4.3.1 Definition of W^*: Best Set of Directed Edge Weights

Let f_{θ^*} represent the trained neural network that is equipped with an optimized set of parameters θ^*. Then we can construct a prediction matrix $P = (f_{\theta^*}(t_d, a_v)) \in [0, 1]^{|D| \times |V|}$ by collecting all predictions $f_{\theta^*}(t_d, a_v)$ from the trained neural network f_{θ^*} using the all pairs of data $(t_d, a_v) \in T \times A$ as illustrated in Figure 18.

Upon the assumption that the trained neural network f_{θ^*} effectively encodes a shared understanding about jurisprudences of WTO DSB among legal experts of WTO agreements, our task is to find a set of directed edge weights $W^* = \{w^*_{ij} \mid w^*_{ij} = w^*(v_i, v_j)\ s.t.\ w^*(v_j, v_j) = 0 \text{ and } \sum_{v_i \in V} w^*(v_i, v_j) = 1\ \forall v_j\}$ by exploiting the information encoded in the prediction matrix

$P = (f_{\theta^*}(t_d, a_v)) \in [0,1]^{|D| \times |V|}$. This set of directed edge weights $W^* = \{w^*_{ij}\}$ shall represent a set of conditional probability $P^*(v_j|v_i)$ how probably a source node v_i clarifies the meaning of the target node v_j compared to other source nodes closely to a shared understanding about jurisprudences of WTO DSB among legal experts.

To perform this task, this paper adopts a machine learning technique called *Random Forest* that can rank input features, $\{v_i \mid v_i \in V \setminus \{v_j\}\}$, in terms of relative importance to explain the variance of output variables, $\mathrm{Var}(\{f_{\theta^*}(t_d, a_{v_j}) \mid d \in D\})$ for a given target article $v_j \in V$. The step-by-step algorithm of *Random Forest* will be explained in the next subsection.

4.3.2 Random Forest on Prediction Matrix P: Finding W^*

The prediction matrix P illustrated in Figure 18(b) includes a co-citation pattern between articles of WTO agreements. For example, we can see several highlighted bands in Figure 18(b) and it tells us there exists some co-citation patterns inside the matrix.

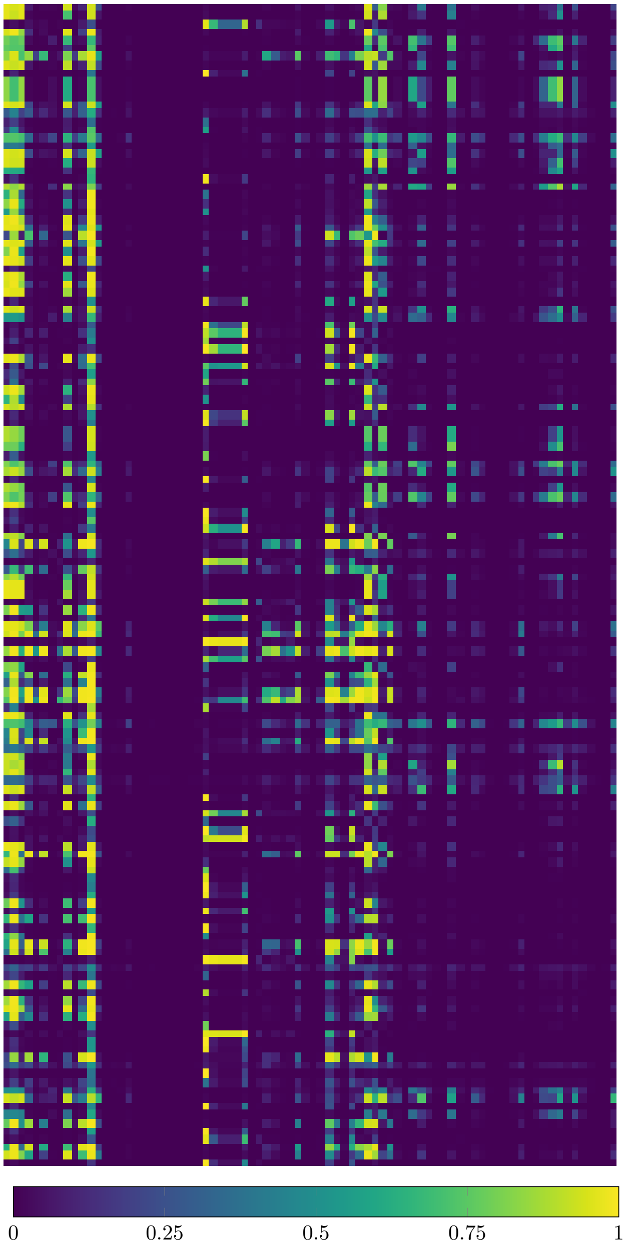
To materialize this co-citation pattern in a form of *directed edge weights* W as defined in Figure **??**, I used a machine learning technique called *Random Forest* (RF). RF is an ensemble method that averages out multiple decision trees.

A decision tree consists of several nodes where each node splits the all observations into two with an inequality criterion of an input feature. This split reduces the variance of the output variables gradually as the tree grows. In our case, for a given target article $v_j \in V$, the rest of the articles $v_i \in V \setminus \{v_j\}$ becomes the input features. For example, we have 143 observations in Figure 18(b) where the number equals that of rows. Then we grow a tree where each node of tree splits the observations according to its own criterion whether the value of that input features greater (or lesser) than certain value. By doing so, variance of output variables in 143 observations keep reduced following down the tree. Then we can collect and assign the amount of variance reduced by each input feature to w_{ij} by interpreting this variance reduction as how much The source article v_i clarifies the interpretation of the target article v_j. I normalize the variance of output variables before constructing the tree, thus total variance reduction sums up to 1. Therefore $W = (w_{ij})$ fits to its definition in Figure **??**.

There are three different aspects that distinguish RF from a single decision tree. First, as noted earlier, RF requires to average out multiple decision trees. I ensembled 1,000 decision trees and averaged out all w_{ij} generated from each decision tree. Second, RF requires *bagging* that random samples observations with replacement before constructing a tree. This is to avoid overfitting by letting each tree being trained on different parts of the same dataset. I sampled 143 observations

	I	I:1	II	II:1	···
DS2	**0.950**	**0.933**	**0.946**	**0.068**	
DS18	**0.950**	**0.912**	**0.947**	**0.013**	
DS22	**0.070**	**0.037**	**0.003**	**0.042**	
DS31	**0.933**	**0.967**	**0.835**	**0.135**	
⋮					
DS505					
DS513					
DS518					
DS523					

(a) **Illustration of Prediction Matrix:** By defining row as a list of DS case numbers and column as legal articles, we can create $|D| \times |V|$ matrix that includes $f_{\theta^*}(t_d, a_v)$ for each cell.



(b) **Prediction Matrix:** This heatmap represents the P from the actual data and its predictions from f_{θ^*}

Figure 18: **Illustration of Prediction Matrix:** I collected all the predictions from the trained neural network f_{θ^*} and constructed prediction matrix P

with replacement for each construction of decision tree. Third, RF requires to compare the result of random subset of input features to split. I random sampled $\sqrt{|V|-1}$ number of input features at each split because Huynh-Thu et al. (2010) reported high performance of use of this parameter for solving regression problem with random forest that is same to our setting.

All the process of finding W^* from the prediction matrix P is formally defined in Algorithm 2 and final output of W^* is visualized as the heatmap W_{text} in Figure 7(b).

Algorithm 2: Random Forest to Find W^*

```
Input: Prediction Matrix $P = (p_{dv}) \in [0,1]^{|D|\times|V|}$ s.t. $p_{dv} = f_{\theta^*}(t_d, a_v)$
Output: $W^* = (w^*_{ij}) \in [0,1]^{|V|\times|V|}$
Let number of features $n = |V|$ ,
     number of obseravations $o = |D|$ and
     number of trees to ensemble $m \in \mathbb{N}$
for ( $v_j \in V$ ) {
    X ← $\{x_d \mid x_d = (p_{dv_1},\ p_{dv_2},\ \ldots,\ p_{dv_n})\ s.t.\ v \in V \setminus \{v_j\}$ and $d \in D\}$
    Y ← $\{y_d \mid y_d = p_{dv_j}\ /\sigma(p_{v_j})\ s.t.\ d \in D$ and $\sigma(p_{v_j})$ is a standard deviation of
    $\{p_{dv_j} \mid d \in D\}\}$
    for ( $k \in \{1, 2, \ldots, m\}$ ) {
        1. $S = \{(x_d, y_d)\} \leftarrow$ Random sample $o$ number of $(x_d, y_d)$ from $X \times Y$ with
          replacement. Then let $X_d$ notate set of all sampled $x_d$.
        2. Construct a decision tree $T_k : X_d \to \mathbb{R}$ where
        $T_k = \{N \mid N = (v_i, b, N_p, \hat{y})$ represents a decision node where
        $b, \hat{y} \in \mathbb{R}, v_i \in V \setminus \{v_j\}$ and $(v_i, b)$ splits $S_N \subset S$ that reached the node $N$
        into $S_{N_{true}}$ and $S_{N_{false}}$ with a criterion $p_{dv_i} \geq b$ with a parent node $N_p \in T_k$
        if $N$ is not a root node. Define $N_p = \emptyset$ if N is a root node. $\hat{y}$ represents
        the node's estimate for given input $x_d$ and $\hat{y} = \frac{1}{|S_N|}\sum_{(x_d,y_d)\in S_N} y_d$.
        $v_i$ and $b$ is $\emptyset$ if $N$ has no child nodes. $(v_i, b)$ at each node N is determined
        among a random sampled $\sqrt{|V|-1}$ number of $v_i$ from $V$ that minimizes MSE
          $\frac{1}{|S_{N_p}|}\sum_{(x_d,y_d)\in S_{N_p}} (y_d - T_k(x_d))^2$. Splitting that generates child node stops when
          $S_N = 1$ at each node N.}
        3. for ( $N \in T_i$ ) {
            if $v_i$ of $N \neq \emptyset$ then
                $I^k_{v_i \to v_j}(N) \leftarrow I^k_{v_i \to v_j}(N) + (\mathrm{Var}(S_N) - \mathrm{Var}(S_{N_{true}}) - \mathrm{Var}(S_{N_{false}}))$ where
                  Var(·) is the variance of the output variable $y_d$ in each subset $S_N, S_{N_{true}}$
                  and $S_{N_{false}}$
            end
        }
    }
}
then $w^*_{ij} \leftarrow \frac{1}{m}\sum_{k\in\{1,2,\ldots m\}} I^k_{v_i \to v_j}(N)$
return $W^* = (w^*_{ij}) \in [0,1]^{|V|\times|V|}$
```

5 Empirical Findings

This section verifies how well the fitted network $G^* = (V, E, W^*)$ aligns with the jurisprudences of the *Panel* or *Appellate Body* of WTO DSB. Since these two judicial bodies of WTO DSB authoritatively opinionate how the regulatory system of WTO DSB systematically organized, this section will validate the quality of the fitted network G^* by introducing three different sub-networks of the fitted network G^* where each sub-network shows how articles of WTO agreements cooperatively achieves important principles of WTO and regulates specific trade issues.

References

Breiman, Leo. 2001. *Random Forests*.

Busch, Marc and Eric Reinhardt. 2003. "Developing Countries and General Agreement on Tariffs and Trade/World Trade Organization Dispute Settlement." *Journal of World Trade* 37:719–735.

Busch, Reinhardt, Eric and Gregory Shaffer. 2009. "Does legal capacity matter? A survey of WTO Members." *World Trade Review* 8(4):559–577.

Charles, Zachary and Dimitris Papailiopoulos. 2017. "Stability and Generalization of Learning Algorithms that Converge to Global Optima.".

Dinh, Laurent, Razvan Pascanu, Samy Bengio and Yoshua Bengio. 2017. "Sharp Minima Can Generalize For Deep Nets.".

Guo, Jiaxian, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu and Jun Wang. 2017. "Long Text Generation via Adversarial Training with Leaked Information.".

Huynh-Thu, Vân Anh, Alexandre Irrthum, Louis Wehenkel and Pierre Geurts. 2010. "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods." *PLOS ONE* 5(9):1–10.

Johns, Leslie and Krzysztof J. Pelc. 2014. "Who Gets to Be In the Room? Manipulating Participation in WTO Disputes." *International Organization* 68(3):663–699.

Kim, Yoon. 2014. "Convolutional Neural Networks for Sentence Classification." *CoRR* abs/1408.5882.

Lawrence, S., C. L. Giles, Ah Chung Tsoi and A. D. Back. 1997. "Face recognition: a convolutional neural-network approach." *IEEE Transactions on Neural Networks* 8(1):98–113.

Magdum, P. G. and Sheetal Rathi. 2020. A Survey on Deep Learning-Based Automatic Text Summarization Models. In *Advances in Artificial Intelligence and Data Engineering*, ed. Niranjan N. Chiplunkar and Takanori Fukao. Singapore: Springer Singapore pp. 377–392.

Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu and Jianfeng Gao. 2020. "Deep Learning Based Text Classification: A Comprehensive Review.".

Nahid, A. and Y. Kong. 2017. Local and Global Feature Utilization for Breast Image Classification by Convolutional Neural Network. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA).* pp. 1–6.

Neyshabur, Behnam, Srinadh Bhojanapalli, David McAllester and Nathan Srebro. 2017. "Exploring Generalization in Deep Learning.".

Pelc, Krzysztof J. 2014. "The Politics of Precedent in International Law: A Social Network Application." *The American Political Science Review* 108(3):547–564.

SHAFFER, GREGORY. 2006. "The challenges of WTO law: strategies for developing country adaptation." *World Trade Review* 5(2):177–198.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15(56):1929–1958.

Srivastava, Rupesh Kumar, Klaus Greff and Jürgen Schmidhuber. 2015. "Highway Networks." *CoRR* abs/1505.00387.

Strezhnev, BuenoAnton. 2014. "Using Latent Space Models to Study International Legal Precedent: An Application to the WTO Dispute Settlement Body." *American Political Science Association 2014 Annual Meeting* .

World Trade Organization. 1999. *WTO Agreements Series.* Number no. 2 *in* "WTO Agreements Series" World Trade Organization.

World Trade Organization. 2017. *A Handbook on the WTO Dispute Settlement System.* A WTO Secretariat publication Cambridge University Press.

Zahavy, Tom, Bingyi Kang, Alex Sivak, Jiashi Feng, Huan Xu and Shie Mannor. 2017. "Ensemble Robustness and Generalization of Stochastic Deep Learning Algorithms.".

Appendix A Examples of the Data Collected

A.1 Factual Aspect Example

Excerpt below is from the panel report for the *US - Offset Act (Byrd Amendment)* [13] case.

> **II. FACTUAL ASPECTS**
>
> 2.1 This dispute concerns the Continued Dumping and Subsidy Offset Act of 2000 (the "CDSOA" or the "Offset Act"), which was enacted on 28 October 2000 as part of the Agriculture, Rural Development, Food and Drug Administration and Related Agencies Appropriations Act, 2001.1 The CDSOA amends Title VII of the Tariff Act of 1930 by adding a new section 754 entitled Continued Dumping and Subsidy Offset. Regulations prescribing administrative procedures under the Act were brought into effect on September 21, 2001.
>
> 2.2 The CDSOA provides that :
> " Duties assessed pursuant to a countervailing duty order, an anti-dumping duty order, or a finding under the Antidumping Act of 1921 shall be distributed on an annual basis under this section to the affected domestic producers for qualifying expenditures. Such distribution shall be known as "the continued dumping and subsidy offset". "
>
> 2.3 The term "affected domestic producers" means :
> " a manufacturer, producer, farmer, rancher, or worker representative (including associations of such persons) that –
>
> (A) was a petitioner or interested party in support of the petition with respect to which an anti-dumping duty order, a finding under the Antidumping Act of 1921, or a countervailing duty order has been entered, and

[13] Panel Report, United States — Continued Dumping and Subsidy Offset Act of 2000, WTO Doc. WT/DS217/R (adopted Jan. 27, 2003).

(B) remains in operation.

Companies, business, or persons that have ceased the production of the product covered by the order or finding or who have been acquired by a company or business that is related to a company that opposed the investigation shall not be an affected domestic producer. ”

2.4 In turn, the term “qualifying expenditure” is defined by the CDSOA as “expenditure[s] incurred after the issuance of the anti-dumping duty finding or order or countervailing duty order in any of the following categories: “
(A) Manufacturing facilities.
(B) Equipment.
(C) Research and development.
(D) Personnel training.
(E) Acquisition of technology.
(F) Health care benefits to employees paid for by the employer.
(G) Pension benefits to employees paid for by the employer.
(H) Environmental equipment, training or technology.
(I) Acquisition of raw materials and other inputs.
(J) Working capital or other funds needed to maintain production.” ”

2.5 The CDSOA provides that the Commissioner of Customs shall establish in the Treasury of the United States a special account with respect to each order or finding[8] and deposit into such account all the duties assessed under that Order.[9] The Commissioner of Customs shall distribute all funds (including all interest earned on the funds) from the assessed duties received in the preceding fiscal year to affected domestic producers based on a certification by the affected domestic producer that he is eligible to receive the distribution and desires to receive a distribution for qualifying expenditures incurred since the issuance of the order or finding.[10] Funds deposited in each special account during each fiscal year are to be distributed no later than 60 days after the beginning of the following fiscal year.[11] The CDSOA and regulations prescribe that (1) if the total amount of the certified net claims filed by affected domestic producers does not exceed the amount of the offset available, the certified net claim for each affected domestic producer will be paid in full, and (2) if the certified net claims exceed the amount available, the offset will be made on a pro rata basis based on each affected domestic producer's total certified claim.

2.6 Special accounts are to be terminated after "(A) the order or finding with respect to which the account was established has terminated; (B) all entries relating to the order or finding are liquidated and duties assessed collected; (C) the Commissioner has provided notice and a final opportunity to obtain distribution pursuant to subsection (c); and (D) 90 days has elapsed from the date of the notice described in subparagraph (C)." All amounts that remain unclaimed in the Account are to be permanently deposited into the general fund in the US Treasury.12

2.7 The CDSOA applies with respect to all anti-dumping and countervailing duty assessments made on or after 1 October 200013 pursuant to an anti-dumping order or a countervailing order or a finding under the Antidumping Act of 1921 in effect on 1 January 1999 or issued thereafter. [END]

A.2 Collected Cited Articles for 143 WTO DSB Cases

DS refers to *Dispute Settelement* and this notation is officially adopted by WTO DSB.
WTO DSB identifies each dispute with a unique number for each case such as DS2 and DS18.

Case Number	**Cited Articles (GATT 1994)**
DS 2	I, III, XXII:1
DS 18	XI, XIII
DS 22	VI:3, VI:6
DS 31	III, XI
DS 34	XI, XIII, XXIV
DS 46	XVI
DS 56	II, VII, VIII, X
DS 58	I, XI, XIII, XX

DS 60	VI
DS 62	II
DS 67	II, XXIII, XXIII:1
DS 68	II, XXII:1, XXIII:1
DS 69	II, III, X, XIII, XXVIII
DS 75	III:2
DS 76	XI
DS 87	III:2
DS 90	XI:1, XIII, XVIII:11
DS 98	XIX
DS 103	X, XI, XIII
DS 108	III:4, XVI
DS 121	XIX
DS 122	VI
DS 135	III, XI, XXIII, XXIII:1(b)
DS 136	III:4, VI
DS 139	I:1, III:4, XXIV
DS 141	I, VI
DS 146	III, XI
DS 152	I, II, III, VIII, XI
DS 155	III:2, X:3(a), XI:1
DS 161	II, III, X, XI, XVII
DS 162	III, III:4, VI, XI
DS 165	I, II, VIII, XI
DS 166	I, XIX
DS 174	I, III:4
DS 175	III, III:4, XI, XI:1
DS 177	I, II, XIX
DS 184	VI, X
DS 202	I, XIII, XIX

DS 207	II, XIX:1
DS 212	VI:3
DS 217	VI:2, VI:3, X:3, XXIII:1
DS 219	I, VI
DS 221	VI, VI:2, VI:3, VI:6
DS 231	I, III, XI:1
DS 234	VI, VI:2, VI:3, X, X:3, XXIII:1
DS 238	XIX:1
DS 244	VI, X
DS 245	XI
DS 246	I:1
DS 248	I:1, XIII, XIX:1
DS 257	VI, VI:3, X:3
DS 264	VI, X:3
DS 265	III:4, XVI
DS 266	III:4, XVI
DS 267	III:4, XVI
DS 268	VI, X
DS 269	II, II:1, XXIII, XXIII:1, XXVIII
DS 276	III, III:4, XVII, XVII:1
DS 282	VI, X
DS 283	III:4
DS 286	II, XXII
DS 290	I, I:1, III, III:4
DS 294	VI
DS 295	VI, VI:2
DS 296	VI:3, X:3
DS 301	I:1, III:4, XXIII:1
DS 302	II:1, III:2, III:4, X:1, X:3, X:3(a), XI:1, XV
DS 308	III

DS 312	VI:1, VI:2(a), VI:2(b), VI:6
DS 315	X:1, X:3
DS 316	III:4, XVI:1, XXIII:1
DS 320	I, II
DS 321	I, II
DS 322	VI, VI:1, VI:2(a)
DS 332	I:1, III:4, XI:1, XIII:1
DS 336	VI:3, X:3
DS 339	II:1, III:1, III:2, III:4, III:5, XI, XIII:1
DS 343	I:1, II, II:1, III, VI, VI:2, X:3(a), XI:1, XIII:1, XX
DS 344	VI, VI:1, VI:2
DS 345	I, II, II:1, VI, VI:2, VI:3, X, X:1, X:2, XI, XIII
DS 350	VI:1, VI:2
DS 353	III:4
DS 360	II:1, III:2, III:4
DS 363	III:4, XI:1
DS 366	I:1, II:1, III:2, V:6, VII, VII:1, X:3, X:3(a), XI, XIII:1
DS 371	II:1(b), II:3, III:2, III:4, VII:1, VII:2, VII:5, X:1, X:3, X:3(a)
DS 379	I, VI
DS 381	I, III
DS 384	III:4, IX, IX:2, X:3, X:3(a), XXIII:1(b)
DS 392	I:1, XI:1
DS 394	VIII, VIII:1, VIII:4, X, X:1, X:3, XI, XI:1
DS 396	III:1, III:2
DS 397	I, I:1, VI:1, X:3(a)
DS 399	I:1, II, XIX
DS 400	I:1, III:4, XI:1, XXIII:1(b)
DS 406	III:4, XX, XXIII:1(a)
DS 412	III:4, III:5, XXIII:1
DS 414	VI

DS 415	I:1, II:1, XIX:1, XIX:2
DS 422	VI:1, VI:2(a), VI:2(b)
DS 425	VI:1, VI:6
DS 427	VI, VI:3
DS 429	VI:1, VI:2, VI:2(a), X
DS 430	I, XI
DS 431	VII, VIII, X, X:3(a), XI, XI:1
DS 435	III:4
DS 436	I, VI
DS 437	VI, XXIII
DS 440	VI
DS 442	VI, X:3(a)
DS 447	I:1, III:4, XI:1
DS 449	VI, X
DS 453	I:1, III:2, III:4, XI:1
DS 454	VI
DS 456	III:4
DS 457	II:1(a), II:1(b), X:1, X:3(a), XI, XI:1
DS 461	II:1, II:1(b), VIII:1, X:3(a)
DS 464	VI, VI:1, VI:2, VI:3
DS 468	II:1(b), XIX:1
DS 471	VI:2
DS 472	I:1, II:1(b), III:2, III:4, III:5
DS 473	VI:2
DS 475	I:1, III:4, XI:1
DS 476	I, III, X, XI
DS 477	III:4, X:1, XI:1
DS 479	VI
DS 480	VI, VI:1, VI:2
DS 482	VI

DS 483	VI
DS 484	III:4, X:1, X:3, XI:1
DS 485	II:1(a), II:1(b), VII
DS 486	VI
DS 488	I, X:3
DS 490	I:1, XIX:1, XIX:2
DS 492	I, I:1, II, II:1, II:2, XIII, XIII:1, XIII:2, XXVIII
DS 493	VI
DS 495	XXIII:1
DS 499	I:1, III:4, X:3(a), XI:1, XIII:1
DS 504	VI
DS 505	VI:3
DS 513	I:1, X:1, X:2, X:3(a), XI:1
DS 518	I:1, II:1(b), XI:1, XIX:1
DS 523	VI:3

A.3 Various Levels of Scope Adopted by Panel and Appellate Body

Scope	**Quote**	**Source**
Title	“As the ***title*** **of Article 21 makes clear**, the task of panels forms part of the process of the ‘Surveillance of Implementation of . . . ”	Appellate Body Report, *US – Shrimp (Malaysia)*, paras. 86-87.
Article	“The sequence of steps indicated above in the analysis of a claim of justification under **Article XX** reflects, not inadvertence or random choice, . . . ”	Appellate Body Report, *US – Shrimp (Malaysia)*, paras. 119-120.

Paragraph	“The verb ‘may’ in **Article VI:2** of the GATT 1994 is, in our opinion, properly understood as giving Members a choice . . . ”	Appellate Body Report, *US – 1916 Act*, paras. 116.
Sentence	“The customary rules of interpretation of public international law as required by **the first sentence of Article 17.6(ii) of the Anti-Dumping Agreement**, do not admit of another interpretation . . . ”	Appellate Body Report, *US – Zeroing (EC)*, paras. 132-133.
Term	“ . . . **The term ‘commerce’** is defined as referring broadly to the exchange of goods, . . . ”	Appellate Body Report, *Colombia – Textiles*, para. 5.34.

A.4 Example of Article of the WTO agreements

Article I

General Most-Favoured-Nation Treatment

1. With respect to customs duties and charges of any kind imposed on or in connection with importation or exportation or imposed on the international transfer of payments for imports or exports, and with respect to the method of levying such duties and charges, and with respect to all rules and formalities in connection with importation and exportation, and with respect to all matters referred to in paragraphs 2 and 4 of Article III, any advantage, favour, privilege or immunity granted by any contracting party to any product originating in or destined for any other country shall be accorded immediately and unconditionally to the like product originating in or destined for the territories of all other contracting parties...