



# A new knowledge discovery approach for mining business trade barriers

Yazid Bounab<sup>1</sup> · Mourad Oussalah<sup>1</sup>

Received: 7 December 2021 / Revised: 3 March 2022 / Accepted: 10 March 2022 /

Published online: 30 May 2022

© The Author(s) 2022

## Abstract

Cross-border trade barriers introduced by national authorities to protect local business and labor force cause substantial damage to international economical actors. Therefore, identifying such barriers beyond regulator's audit reporting is of paramount importance. This paper contributes towards this goal by proposing a novel approach that uses natural language processing and deep learning method for uncovering Finnish-Russian trade barriers in the fish industry from selected business discussion forums. Especially, the approach makes use i) a three-leg ontology for data collection, ii) a BERT architecture for mapping Onkivisit-Shaw-Kananen trade barrier ontology to negative polarity posts and, iii) a new reverse-engineering clustering approach to identify the causes of individual trade-barrier types. A comparison with official statistical reports has been carried out to identify the salient aspects of trade-barriers that hold regardless of the time difference. The findings reveal the dominance of the Time-length barrier type in the Finnish discussion forum dataset and import vs export tariff discrepancy and product requirement barrier types in the Russian forum dataset. The developed framework can serve as a tool to assist companies or regulators in providing business-related recommendations to overcome the detected trade barriers.

**Keywords** Trade barriers · Sentiment analysis · Natural language processing · BERT · Causality

## 1 Introduction

Trade is a fundamental economic act that involves buying, selling or exchange commodities and services between participants that can be individuals, business entities or national organizations. Foreign trade occurs when these participants belong to different countries

---

Bounab Yazid and Oussalah Mourad contributed equally to this work

✉ Yazid Bounab  
Yazid.Bounab@oulu.fi

✉ Mourad Oussalah  
Mourad.Oussalah@oulu.fi

<sup>1</sup> Faculty of ITEE, CMVS, University of Oulu, Pentti Kaiteran katu 1, 90570, Oulu, 90014, Finland

with different trade policies (Lawrence et al., 2018). In this context, foreign trade can be negatively impacted by various trade-barriers. The latter refers to the set of government-induced policies that put restrictions on international trade (Lawrence et al., 2018) by making the trade either more difficult or expensive through some tariff or regulation like barriers, or even by fully forbidding the entry of some specific goods / services as in case of trade embargo or economical sanctions. Motivations for introducing trade barriers are rooted back to government need to protect infant industries from international competitions and promote local business actors to gain competitiveness or maintain high employment rate. Nevertheless, in the age of internationalization and globalization, economists commonly admit that trade barriers are against free-trade policy advocated by the World Trade Organization (WTO) and undermine investors as well as small- and medium-sized enterprises (SMEs) who seek long term stability, predictability and economic profitability in their international markets. Although, as a result of the General Agreement on Tariffs and Trade (GATT), tariff-barrier is left at quite low levels, the total trade costs remain relatively high (Anderson, 2004). This is due to the growing impact of non-tariff barriers, which, include, among others, product regulation, border inspections, transportation, access to efficient distribution channels, time delay, cultural and language clues, among others. Therefore, timely and efficient identification of such trade-barriers is of paramount importance for both business entities and regulator agencies. Indeed, as pointed out in (Bahja, 2020), this contributes to reinforce E-Governance duties in terms of promoting transparency in international business and supporting local business operators in easing access to international market opportunities (Ghosh, 2009). Moreover, shedding light on trade barriers provides valuable input data to the World Trade Organization (WTO) (Global trade, 2021) that will foster the development of more evidence-based policing. Strictly speaking, national agencies and European Union (EU) institutions maintain a database of trade-barriers that can be made available for their export enterprises. This includes, for instance, the Market Access Database and Access to Market Portals (European Commission. Market Access Database, 2021) in the EU, and the monthly reports of The Office of the United States Trade Representative (USTR) in USA (The Office of the U.S. Trade Representative (USTR). FOREIGN TRADE BARRIERS., 2021). However such databases often contain a high-level description of the trade-barrier according to the agency own categorization when appropriately reported by business actors and lack a fine-grained description that might be of a direct interest to a random investor or SME. Therefore, identifying efficient approaches for uncovering such barriers remain a constant priority for both trade agencies and business actors. This motivates the current paper, which advocates a novel approach based on natural language processing (NLP) and deep learning technology to mine the cross-border trade-barriers between Finland and Russia in fish industry. In essence, this study uses commerce discussions (feedback, questions, and complaints) extracted from Finnish and Russian specialized business discussion forums to uncover trade barriers faced by Finnish and Russian entities (individuals, companies, and government agencies) in fish industry sector. As discussed in (Bahja, 2020), a such study contributes to reinforce E-Governance duties in terms of promoting transparency in international business and supporting local business operators in easing access to international market opportunities. Moreover, shedding light on trade barriers provides valuable input data to the World Trade Organization (WTO) in a way to foster the development of more evidence-based policing.

Overall, this research aims to contribute to the task of automatic identification of trade-barriers from unstructured textual sources. We confine our trade-barrier analysis to the trade-barrier ontology provided by Onkivisit & Shaw (Onkvist & Shaw, 2009), which advocates 11 classes of Tariff and Non-Tariff barriers. We shall refer to this classification as

OS-K trade-barrier ontology. More specifically, this research addresses the following three research questions:

- $RQ_1$ : How to automatically identify the OS-K trade-barriers from unstructured textual inputs?
- $RQ_2$ : How to automatically identify the discourse and causes related to each trade-barrier type?
- $RQ_3$ : How to evaluate the results generated by a such automatic trade-barrier identification system?

We restricted our study to the Finnish-Russian trade-barriers in fish industry according to dataset provided as part of an European CBC Karelia research project. This paper advocates five-fold contributions.

- A comprehensive review of NLP applications in business sector has been conducted, highlighting the key pitfalls in this area to address the problem of automatically identifying trade-barriers from textual sources. This is carried out in the background section of this paper.
- A three-leg ontology, consisting of barrier-business-trade ontology, for querying Finnish and Russian forum dataset has been put forward. This enables us to match user textual inputs to OS-K trade-barrier ontology.
- A new architecture that uses state-of-the-art BERT model and reverse engineering clustering model has been devised to identify the various trade-barrier types as well as their associated causes and discourse in Methodology section. This answers research questions  $RQ_1$  and  $RQ_2$ .
- Exploratory and visualization-based approaches have been conducted to interact with the user and contribute to the evaluation of the findings, which answers research question  $RQ_3$ .
- A comparison of the findings with previously identified Finnish-Russian trade barrier report further sheds light on the statistical assessment of the model and, thereby, contributes to  $RQ_3$  as well.

Section 2 of this paper presents the background information in terms of use of NLP technology in business sector for the purpose of trade barrier identification. It also provides a brief outlook of the adopted OS-K ontology. Section 3 describes the dataset employed in this study. Section 4 details the method and the data pipeline employed to answer the three aforementioned research questions. Results and discussions are reported in Section 5. Trend comparison with economic statistics is carried out in Section 6. Finally, research implications and concluding statements and perspective works are reported in Section 7 and Section 8, respectively.

## 2 Background

It should be noted that surprisingly the problem of automatic identification of trade-barriers from textual sources is much overlooked in both business and information processing research communities where no well cited reference can be identified as far as our investigation is concerned. This is probably due to inherent challenges of NLP tasks involved into this process as well as the scarcity of data and multidisciplinary aspects. Therefore, we first explored the state-of-the-art of NLP research in business sector with the objective of

identifying related research that can contribute to our problem. Then we explicitly provided a background of the *OS-K* trade barrier ontology that we adopted in this paper.

## 2.1 Natural language processing in business sector

Natural language processing is becoming the new standard for business intelligence where self-service analytics service providers, e.g., IBM Watson, are increasingly adding NLP uses to their platforms to enhance search experience and go beyond string matching. NLP also helped developing enhanced personalized services to account for individual user's profiles generated through continuous monitoring of his purchases and actions as well as capitalizing on user's review and feedback, taking into account the inferred sentiment and rationality as in (Bounab et al., 2020), (Eisenstein, 2019), (Tarnowska & Ras, 2019), (Tarnowska & Ras, 2019), (Tarnowska & Ras, 2021) (Krouka et al., 2021). Besides, with the developments in e-commerce and finance applications, NLP research contributed further in the emergence of sophisticated decision-support and recommendation systems that efficiently interact with customers and provided relevant recommendations to the user (Tarnowska & Ras, 2021). A simple observation of the functioning of the very popular Amazon and Netflix recommendation systems that build on user's profiling, which explores user's past purchases, navigation history, product contents, rating and recent news update, reveals the success of such applications in our daily-life. Chatbots or virtual assistants become part of business ecosystem in many large scale or even SMEs where the interaction with customers is part of their core business. In this context, chatbot is found to substantially increase company initiated communications, mine customers' needs and interests, raise awareness, guide users to other selected service providers and achieve high level of cost effectiveness (Luo et al., 2013). For instance, Public Tableau Ltd. suggests that 30% of customer service positions in the US can be automated using chatbots, resulting in \$ 30 billions annual saving.

Data collection of either structural or unstructured textual inputs is another business application of NLP. It can constitute the core business of the whole company as in AlphaSense (Technologies, 2021). Without claiming an exhaustive coverage of the NLP applications in the business field, a summary of the dominant practices in this area is presented in Table 1. In essence our survey of the existing literature in the field of NLP-based business revealed the following (see Table 1). First, there is a lack of comprehensive and up-to-date survey papers in the field within information processing community as testified by the identified literature, where several key publications were rather published in sparsely forums and communication channels. Second, the dominance of sentiment analysis literature for exploratory-analysis, prediction or even business strategy is surprisingly very striking. From this perspective, identifying trade-barrier can be initiated by further exploring posts with negative polarity, or tracking special vocabulary that translate our interpretation of trade-barrier. Third, there is a recognition that chatbots and conversational assistants will become state-of-the art in the next decade, acting as a cross-edge where many AI technologies ranging from speech recognition to business analytics passing through various components of NLP, deep learning technology, knowledge discovery, dialogue constructs and emotional monitoring can be applied to infer and/or maintain high focused conversations with potential customers. From this perspective, chatbot data can also be utilized to infer insights regarding potential trade-barrier by monitoring utterance conveying user's discomfort and feedback. Fourth, many business related domains are unfortunately left unexplored. Among these domains, one shall mention the automatic identification of trade-barriers for business entities, which motivates our current work in this paper. There are

**Table 1** Applications of NLP in business

Application	Objective	Input	Technique	Limitation
Customer service (Lagi, 2019; Technologies, 2021)	Enhance user's experience	chatbot customer call	SR, IC QA	Need continuous update
Reputation monitoring (Lagi, 2019)	Capture user's feedback	SocMed feedback website	S.A CoreReso	Vulnerable to misinformation SA inherent limitations
Advertisement (Lagi, 2019)	Target advertising	Emails SocMed, E-C, B.D.	Key_M SeD	Privacy issues
Market intelligence (Lagi, 2019)	Business insights	SocMed Emails	Ev_Ext Sen_CI	Missinformation pre-processing
Stock Market (Senyuk, 2021)	Stock pred.	Trading Data	SA, NER DL, RNN	Data relevancy price Volatility
Portfolio optimization (Senyuk, 2021)	profit maximization	Trading data	ML NP statistics	Data coverage stock price volatility
Accounting auditing (Senyuk, 2021)	Detecting anomalies	Transaction records	TC, SA	High positive rate
Risk assessments (Technologies, 2021; Senyuk, 2021)	Creditworthiness of borrowers	loan profile	NA, OCR	Data sparsity
Coaching (Technologies, 2021)	Draw financial plan	Chatbot Customer call	SR, IC QA	Data reliability Trust

SR: Speech recognition, IC: Intent classification, QA: Question Answer, SocMed: Social Media, SA: Sentiment Analysis, CoRe: Coreference resolution, E-C: E-commerce, B-D: Browsing data, Key\_M: Keyword matching, SeD: Sense disambiguation, Ev\_Ext : Event extraction, Sen\_cl : sentence classification, ML : Machine Learning, NP : Non-parametric, TC: Text clustering, OCR: Optical character recognition, NE: Named-entity recognition, DL: Deep Learning

increasing theoretical frameworks from AI, modal reasoning, approximate reasoning, causal reasoning (Morgan & Winship, 2007) that might be utilized for trade barrier mining. Nevertheless, their success in practical NLP applications is rather limited and further research still needed to narrow the gap between theory and practice. Sixth, the challenges in applying NLP technology to business sector cannot be ignored. Indeed, this is rooted back to the challenges of text analysis where inherent ambiguity, dynamic change of contextual discourse, presence of noisy information, use of metaphors and other complex lexical constructs render the application of standard NLP pipeline open to debate (Bahja, 2020; Lagi, 2019; Mitchell et al., 2014; Techlabs, 2018). This calls for subsequent research and exploration.

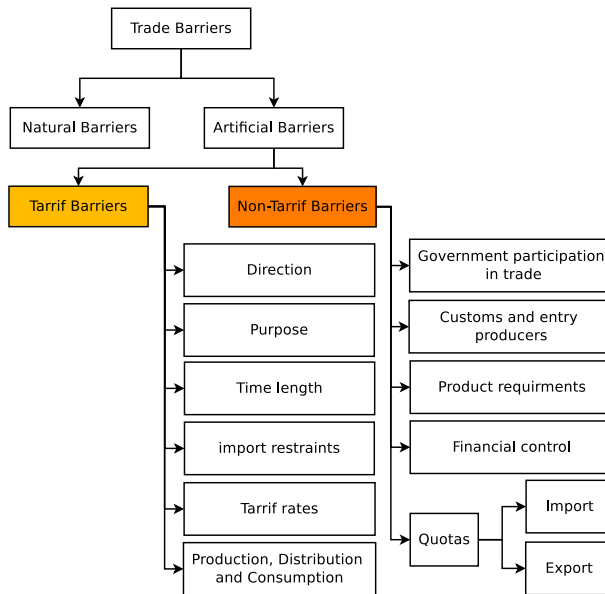
## 2.2 Trade barriers classification

Several trade barriers classifications have been introduced by several scholars and regulators. Grüber and Reiter (Grüber & Reiter, 2021) in their classification emphasized three different forms of trade barriers: Tariff, Non-Tariff and Free-trade. In the same spirit as World Trade Organization (WTO) categorization, Onkivist & Shaw (Onkivist & Shaw, 2009) classified trade barriers into two main classes: Natural trade barriers that deal with non-human causes such as cultural or language cues, and Artificial barriers that encompass Tariff and Non-Tariff barriers. Tariffs barriers involve a financial burden on imported commodities and include six different attributes:

- **Direction (1):** focuses on the discrepancy between import and export movement of goods/services.
- **Purpose (2):** refers to the reason for introducing Tariff extra duties. This is usually related to compensation for local industry or additional tax revenues to protect domestic economy from foreign competitors.
- **Time length (3):** distinguishes Tariff surcharges which are temporary and limited in time, and those which are permanent.
- **import restraints (4):** correspond to special and variable duties associated with frequent changes in world market prices.
- **Tariff rates (5):** correspond to duties that vary per unit of weight, volume, gauge or other indicator of quantity on the good crossing the border. This includes three tax rates: specific, ad valorem and combined.
- **Distribution points (6):** correspond to indirect taxes collected at a critical point of distribution or while a purchase or a consumption occurs.

Non-tariff barriers force other overt and clandestine means to restrict imports and occasionally exports. It includes five different attributes:

- **Government participation in trade (7):** includes administrative guidance to restrict foreign purchases to a certain percentage of local demand, government procurement controlling commercial operations and subsidies involving governmental support domestic producers (Onkivist & Shaw, 2009).
- **Customs and entry procedures (8):** consist of a set of enforcing procedures that add extra difficulty and time to import the goods. This includes preliminary valuation, documentation, license, inspection, and Health and safety regulations.
- **Product requirements (9):** include product standards related to physical characteristics and/or product performance, product testing and packaging, labelling, and national safety regulations prior to allowing the product to market.



**Fig. 1** Onkivisit & Shaw and Kananen (OS-K) Trade Barrier Classification

- **Financial control (10):** involves monetary policies to control capital flow to protect local currency. It includes exchange control, multiple exchange rates, prior import deposits and credit restrictions, and profit remittance restrictions.
- **Quotas (11):** refer to quantity control on imported and exported goods (Onkivist & Shaw, 2009; Schaffer et al., 2014).

Figure 1 shows the above trade-barrier classification, referred **OS-K** trade barrier ontology. This classification is adopted in our study because of its ease, clarity and also due to the fact that it agrees to a large extent to the classification employed by the Finnish Ministry of Foreign Trade <sup>1</sup>

As alternative to the aforementioned classification, one shall mention the classification introduced by USTR, which is yearly updated utilizing inputs from US foreign embassies (The Office of the U.S. Trade Representative (USTR). FOREIGN TRADE BARRIERS., 2021), the European Market Access Database (MADB) (European Commission. Market Access Database. 2021), which provides valuable insights to European exporting SMEs.

### 3 DataSet

In this study, as part of the European Finnish-Russian project, operated by CBC Karelia organization, which aims to foster joint Finnish-Russian SMEs in the fish industry sector, two datasets have been utilized (from Finnish and Russian side). The datasets reflect the wide community interest, including professionals, users and SMEs in fish industry ecosystem (i.e, fish marketing, transportation, production and consumption of fish products) in

<sup>1</sup>[https://um.fi/documents/35732/48132/internationalisation\\_and\\_barriers\\_to\\_trade\\_2013/ae4b492a-3ccc-4489-6b0d-c8101d9281ae?t=1525859635639](https://um.fi/documents/35732/48132/internationalisation_and_barriers_to_trade_2013/ae4b492a-3ccc-4489-6b0d-c8101d9281ae?t=1525859635639)

both Finland and Republic of Karelia. The overall database has been shared under the data-sharing and usage privacy agreement. The database has been elaborated in the first quarter of 2021 and contains historical data as well.

### 3.1 The finnish dataset

We used the Suomi24 forum ([www.suomi24.fi](http://www.suomi24.fi)) because of its popularity and its wide spread coverage including business related discussion. We selected the Seafood section of Suomi24 as it was found to contain interesting fish business related to discussions. We crawled all data linked to seafood thread, which span from 2021 till 2014. This dataset is constituted of 14701 text entries between topic titles and comments written by users in different languages such as Finnish, Swedish, Hebrew, Spanish, and some English. Table 2.(a) shows some statistics related to the Finnish dataset.

### 3.2 The russian dataset

The dataset is generated by crawling 13 different Russian fish farming business forums. The detail of the dataset is summarized in Table 2. The collected dataset contains around 1000 messages from the 375 users registered to the 13 Russian fish-farming online discussion forums, covering 114 topics in overall. The dataset includes time attribute, indicating the date where the post has been posted (from 2010 till 2019). Table 2.(b) shows the list of websites used for building the Russian dataset and their statistics.

## 4 Methodology

### 4.1 Rationality and overall architecture

Given the nature of the input data, a combination of NLP, deep-learning technology, and clustering approach has been adopted to answer research question  $RQ_1$  and  $RQ_2$ . Prior to detailing the technical architecture used to handle the above research questions, a set of rational assumptions have been adopted for this purpose.

*Assumption 1* Trade-barriers are rather associated with negative polarity posts.

The preceding draws on the intuition that when users discuss about the trade-barriers, they often convey negative sentiment posts. Therefore, a sentiment analysis approach can

**Table 2** (a) Finnish Dataset Statistics and (b) Russian Fish farming dataset Statistics

<b>(a) Suomi24, Section Fish and Sea Food</b>	
Number of Topics	1760
Number of Answers	11643
<b>(b) Russian Data Resources Sources</b>	
No. of Sources	13
No. of Sections	232
No. of Topics	3519
No. of Answers	141995
No. of Users	3038

**Russian Websites:** agro-forum, fishretail, agroxxi, forumhouse, apkforum, fermer mirfermera, bio.moy, NN.RU, bioformer, rybovod, farmerforum, rybovodstvo



be used as a preliminary exploration to identify potential trade-barriers from the discussion forums.

*Assumption 2* The association of a single post to trade-barrier types is not mutually exclusive.

This draws on the inherent ambiguity pervading natural language so that it is not fully excluded that a given post will be associated to more than one trade-barrier class among the 11 classes of the (OS-K) ontology.

*Assumption 3* The causes of each trade-barrier type can be revealed from the analysis of the discourse built from the constructed corpus associated to this trade-barrier.

The preceding confines the causality analysis, whose scope often goes beyond to linguistic and semantic aspect to complex philosophical issues, to the output of discourse-based analysis. This motivates a clustering based approach followed by a manual checking step for this purpose. Figure 2 provides a generic pipeline for identifying the various trade-barriers and their causes. The approach highlights at least four distinct phases: Preprocessing, Data collection, Trade-barrier mapping, Causes identification.

### 4.2 Preprocessing

Initially, since the most powerful tools used in NLP are in English, all the posts were translated into English language using Google Translate (GT) application programming interface (API). Although, we acknowledge that such a translation may overlook important messages conveyed by Russian blogs, our initial exploration revealed that this loss is rather marginal and still all named-entities and verbal statements were unaffected by this translation. This also testifies of the efficiency of Google translation, which is widely adopted in

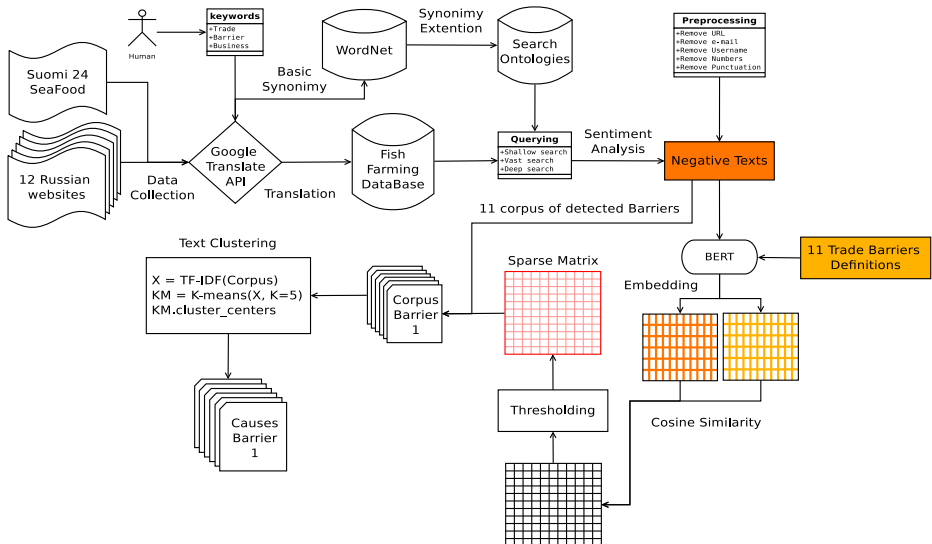


Fig. 2 Generique pipeline

professional forums. Strictly speaking, other empirical studies showed that translation API often fail in case of metaphors (Solutions, 2017), culinary texts (Hasyim et al., 2021), or when attempting to capture writing style and demographic attribute. Nevertheless, these aspects are less apparent in short text messages of our database, which partly explain the efficiency of translation API in our case.

Given the nature of unstructured text collected from the Finnish and Russian forums, which contain several uncommon characters and noisy terms that can deviate the outcomes of the parsers and other NLP modules, preprocessing stage is required. This goes through successive steps to clean out all the noise in the retrieved textual data. Especially, this includes the following:

- Remove URLs and emails;
- Replace combined tokens by separate ones, e.g., “hasn’t” becomes “has not”;
- Remove distracting single quotes;
- Remove punctuation and Numbers;
- Remove extra spaces.

### 4.3 Ontology for data collection

From the raw dataset (Finnish and Russian), there is a need to identify useful subset of data that is relevant to our purpose of mining trade-barrier evidence only. For this purpose, we have created a three-leg ontology to define the keywords needed for querying the database.

- Barrier ontology: represents keywords that define the concept of “Barrier”.
- Business ontology: represents keywords that define the concept of “Business”.
- Trade ontology: represents keywords that define the concepts of “Import” and “Export”.

To build the keyword list associated with each ontology-leg, we follow a three-stage process. Initially, a manual crafting of wording associated with each ontology-leg are drafted. This is followed by an automatic expansion using WordNet synonymy relation and then a purification and manual checking stage is carried out for potential inconsistency detection. Table 3 summarizes each component of this leg-ontology construction and its associated keywords used for querying the database.

**Table 3** Ontologies used for search

Ontology	keywords	No.
Barrier	fence, railing, barricade, hurdle, bar, blockade, curb roadblock, obstacle, obstruction, stumbling block impediment, hindrance, roadblock, barrier, limit, restriction	18
Business	work, line of work, occupation, profession, career, employment, job, position, vocation, calling, field, sphere, trade, métier, craft, biz, racket, game, trade, trading, commerce, dealing, traffic, merchandising, dealings, transactions, negotiations, affair, matter, thing, case, circumstance, situation, event, incident, firm, happening, occurrence, episode, company, concern enterprise, venture, organization, operation, corporation, undertaking, office, agency, franchise, practice, outfit, responsibility, duty, function, obligation, worry, problem, beeswax, bailiwick, patronage, clientele, line, business organization business concern, byplay, commercial enterprise, business enterprise	70
Trade	import, imported, importation, export, exported, exportation	6

We have devised a three-level query system to search for all possible matches within the Russian/Finnish database.

1. **The shallow search:** performs the keyword search operation within the titles (threads' names of the discussion) only of the documents.
2. **The vast search:** extends the keyword search operation to both title and main text body of the topic.
3. **The deep search:** performs the keyword search operation on title, main text body, and replies/answers of each topic.

Algorithm 1 describes the different steps used for querying the retrieved datasets utilizing the predefined ontology-legs. In practice, we have only used the “vast” and “deep-search” modes with the three-leg ontology mentioned above, where 84 keywords have been employed. This generated 14701 output files for the Russian dataset and 3636 output files for the Finnish dataset, which are then used as input to our reasoning pipeline detailed in the next section. We shall denote by  $\omega$  the set of posts resulting from the application of this search operation.

---

**Algorithm 1** Querying\_the\_database(Dataset, Level).

---

```

1: Texts ← []
2: for Ontology in Ontologies do
3:   for keyword in Ontology do
4:     switch Level do
5:       case shallow – search
6:         for title in Dataset do
7:           if keyword in title then
8:             Texts.Add(title)
9:           end if
10:        end for
11:       case vast – search
12:         for (title, main_text) in Dataset do
13:           if keyword in title then
14:             Texts.Add(title)
15:           else if keyword in main_text then
16:             Texts.Add(main_text)
17:           end if
18:         end for
19:       case deep – search
20:         for replies in Dataset do
21:           if keyword in replies then
22:             Texts.Add(replies)
23:           end if
24:         end for
25:     end for
26:   end for
27: return Texts

```

---

#### 4.4 Trade-barrier identification

As pointed out in the previous section, the newly extracted subset of dataset, which best accommodates the three-legs ontology (Barrier, Business and Import & Export) that encapsulates the concept of cross-border trade-barrier is now utilized for subsequent reasoning. For this purpose, in light of Assumption 1 and Assumption 2 endorsed in this study, the task of identification trade-barrier type undergoes three distinct steps:

#### 4.4.1 Sentiment analysis

Given the assumption that barriers are negatively perceived by the users, sentiment analysis has been used to identify negative posts from the subset dataset  $\Omega$  generated by the application of the above three-leg ontology. We used SentiStrength package (Thelwall et al., 2010), which is found to have human-level accuracy for short social web text, to identify sentiment polarity of each post. It reports two sentiment strengths  $E_+$  for negative sentiment and  $E_+$  for positive sentiment in the scale (-4 to +4). Therefore, each  $i^{th}$  post  $p_i$  of  $\zeta$  such as  $E_+(p_i) + E_-(p_i) \leq -1$  is considered to bear a negative polarity, and will thereby be included in the newly created subset, denoted  $\zeta$ , where  $\zeta \subseteq \Omega$ . In other words,  $\zeta$  contains only posts that more likely include discussions directly linked to trade-barriers. The next step concerns the identification of the type of the trade barrier associated to each post in  $\zeta$ .

#### 4.4.2 Trade-barrier type matching process

To uncover the type of trade-barrier associated to a given post, a deep learning-based approach has been designed. This consists of encoding the definition of each of the 11 trade-barrier types as well as each post in  $\zeta$  using the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2018) released by the Google research group in 2018. Figure 3 presents a high level graphical illustration of this process where pre-trained BERT model is employed.

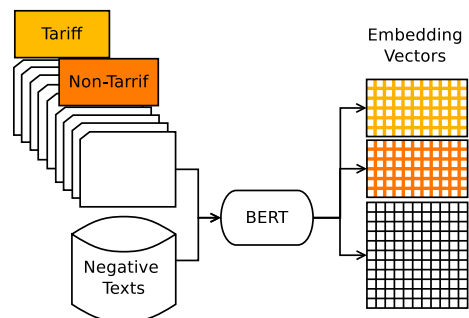
We mainly use BERT as an encoder for text representation, which yields sentence vector representation for each trade-barrier type definition, say,  $H_i$ ,  $i=1$  to  $c$ , and each post  $p_i$ ,  $i=1$  to  $M$ , where  $c$  and  $M$  stand for the total number trade-barriers types (which is 11 as per the OS-K employed ontology) and the number of posts in  $\zeta$ , respectively. More specifically, we have:

$$H_i^s = BERT(H_i), \quad i = 1 \text{ to } c \quad (1)$$

$$p_j^s = BERT(p_j), \quad j = 1 \text{ to } M \quad (2)$$

where  $H_i^s$  and  $p_j^s$  are the vector representation of the  $i^{th}$  trade-barrier type and  $j^{th}$  post in  $\zeta$ , respectively, corresponding to their associated BERT embedding. The use of BERT embedding instead of Glove (Pennington et al., 2014) or word2vec embeddings (Mikolov et al., 2013) is motivated by several grounds. First, the BERT model builds on the state-of-the-art Transformer architecture in NLP, trained over 2.5 billions of words and uses bi-directional learning to gain context of words from both left to right context and right to left context simultaneously, which offers a substantial ability to account for the context and various linguistic constructs of the input word sequence (Vaswani et al., 2017; Renjith et al., 2020).

**Fig. 3** Trade barriers embedding



This intuitively offers an edge over any semantic similarity like analysis. Second, the configuration of BERT architecture enables relatively large word-sequence to be treated as a single input, which captures all the information conveyed by this ordering word-sequence. This contrasts with word2vec or Glove embedding whose model assigns a separate embedding vector to each individual word of the sequence, which intuitively leaves the ordering information of the sequence fully ignored. Third, the model offers reasonable flexibility in fine-tuning the model hyper-parameters without important complexity burden. Now given that both the trade-barrier type  $H_i$  and post  $p_j$  are represented in real-valued vector representation of same size, the matching between the two entities can be assessed using standard techniques well established within the scope of information retrieval (IR) field (Salton & Harman, 2003) which covers, among others, the process of matching between a given query and a set of documents. In this respect, in line with vector-based information retrieval methods, we use a simple inner product or a cosine similarity to assess the similarity between  $H_i$  and post  $p_j$ . More specifically, the similarity between  $H_i$  and  $p_j$  is calculated as:

$$Sim(H_i, p_j) = Cosine(H_i^s, p_j^s) = \left( \frac{H_i^s \bullet p_j^s}{\|H_i^s\| \|p_j^s\|} \right), i = 1 \text{ to } c; j = 1 \text{ to } M \quad (3)$$

The application of Eq. 3 over all posts and trade-barrier types yields a similarity matrix  $\mathbf{M}$  with size  $c \times M$  and whose values lie in the unit interval.

Next, in order to find out whether a given post can be assigned to a given trade-barrier type, often thresholding-like approach is employed. For this purpose, we adopt a two-stage strategy.

- *Acceptance threshold* It corresponds to the minimum similarity value below which the agreement between a given post and a given barrier type is deemed void. In other words, if there is no posts  $p$  such that  $Sim(p, H_k)$  is beyond the *acceptance-threshold*, then the barrier type  $H_k$  is considered not reflected in the existing pool of posts. We empirically set the value of *acceptance-threshold* to 0.5.
- *High-Frequency-based threshold* After the initial screening of the similarity values, which discards those posts whose similarity value is below the *acceptance-threshold*, this phase further scrutinizes the restricted range of permissible similarity values of each trade-barrier type to select a final threshold beyond which the post is deemed to be assigned to the underlining barrier class. For this purpose, we proceed in two phases. First, for each column of the similarity matrix  $\mathbf{M}$ , corresponding to a given trade-barrier type, we create a subdivision of the minimum and maximum values of the similarity values. Second, the bin value, which contains the largest number of items and relatively high similarity value is selected as a final threshold  $\epsilon$ . The rational behind doing so is driven by our desire to ensure some flexibility in terms of post-barrier type association in order to ensure the existing of a sufficient number of posts in each trade-barrier type. This contrasts with standard approach often adopted in information retrieval literature where the maximum similarity value is often adopted as a criterion for the association. On the other hand, this is also motivated by the subsequent task of the cause investigation where one desires to have a corpus associated to each trade-barrier type for both Finnish and Russian dataset. More formally, the corpus  $\mathcal{U}_i$  of each trade barrier type  $H_i$  is defined as:

$$\mathcal{U}_i = \{p_j : Sim(p_j^s, H_i^s) \geq \epsilon\} \quad i = 1 \text{ to } c \quad (4)$$

For a detailed implementation and the employed trade-barrier definitions, one shall refer the reader to the Github repository of the project.<sup>2</sup>

#### 4.5 Trade-barrier cause identification

This subsection aims to comprehend the reasons behind each trade barrier type in both Finnish and Russian corpora. For this purpose, a causality analysis has been carried out. Although, the field of causality detection from textual is an active research field in both computational linguistics and computer science communities where several linguistic rule-based, supervised and unsupervised machine learning methods as part of competitions and challenges, e.g., SemEval2007, SemEval2011 (Keith et al., 2020). Nevertheless, applying most state-of-the-art approaches to our case is rather challenging for at least two reasons. First, the quasi-majority of the state-of-the-art machine-learning approaches are domain-specific. Therefore, attempting to apply them to our dataset will be intuitive of little relevance. Second, the quasi-majority of causality literature instead focused on identifying the cause-effect relationship. We are only interested in comprehending the causality aspect without paying importance to the “effect” parameter. We, therefore, designed a semi-supervised approach to handle the causality aspect. This consists of at least five phases:

- For both Finnish and Russian, we take the set of posts associated to each barrier type as an input; namely,  $\mathcal{U}_i$ , for  $i=1$  to  $c$ .
- We use the list of tokens formed by all posts in  $\mathcal{U}_i$ ,  $i=1$  to  $c$ , to construct a TF-IDF vector representation of each post  $p_j$  in  $\mathcal{U}_i$ .
- We perform a k-means clustering (Rong & Liu, 2020) of the posts  $p_j$  in  $\mathcal{U}_i$  using their associated TF-IDF vector representation. Then, we report the centroids  $L_k$  of each cluster.
- We perform the reverse-engineering operation from TF-IDF to textual representation to transform each centroid  $L_k$  into its textual representation, say,  $T_k$ .
- We hypothesize that the content of the textual centroids bears to a large extent the required information to comprehend the cause of the barrier type occurrence.
- We interact with the operator that reads the content of the centroid for each barrier type for subsequent reflection, if any.

The preceding allows us to extract more detailed reflections about the trade barriers’ content, revealing and summarising the causes of the presence of the detected trade barriers. Figure 4 summarizes the five steps used for Trade Barrier Causes Identification.

##### 4.5.1 Evaluation

In the absence of ground truth information that would allow us to perform more quantitative analysis of the usual information-retrieval adopted performance metrics, and in order to address  $RQ_3$ , we mainly relied on two main approaches for this purpose:

- We performed exploratory and visualization-based analysis as a way to seek the relevance of the outcomes.
- We performed a brief comparison analysis with previously available reports regarding the Finnish-Russian trade barrier.

<sup>2</sup>[https://github.com/bounabyazid/Bert\\_Trade\\_Barriers](https://github.com/bounabyazid/Bert_Trade_Barriers)

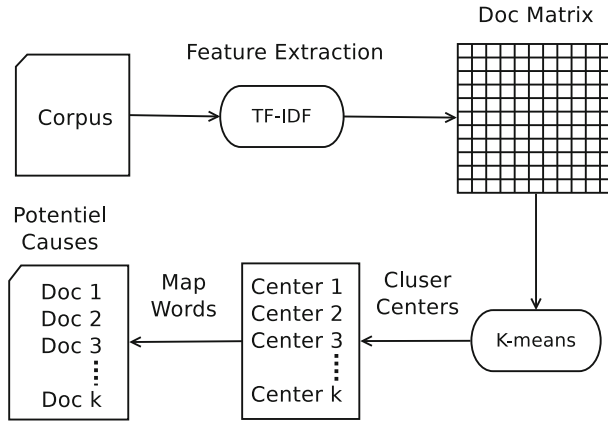


Fig. 4 Text clustering

## 5 Results and discussions

### 5.1 Exploratory analysis

Initially, we performed an exploratory analysis to comprehend the Finnish and Russian overall corpora content using WordCloud visualization, which helps identify its core contents in terms of word frequency and aesthetic features, providing general insights on the most discussed topics. Figure 5 (a) and (b) show the WordCloud representation of Finnish and Russian Corpora, respectively. From Fig 5 (a), one may notice that the Finnish dataset is rather dominated by fish and related seafood (food, eat, salmon), food taste, pricing, and various fish recipes. This indicates the tendency of the users in this forum to act as fish consumers instead of international business dealers. However, discussions related to Russian imported fish and Russian fish recipes are highly present as well.

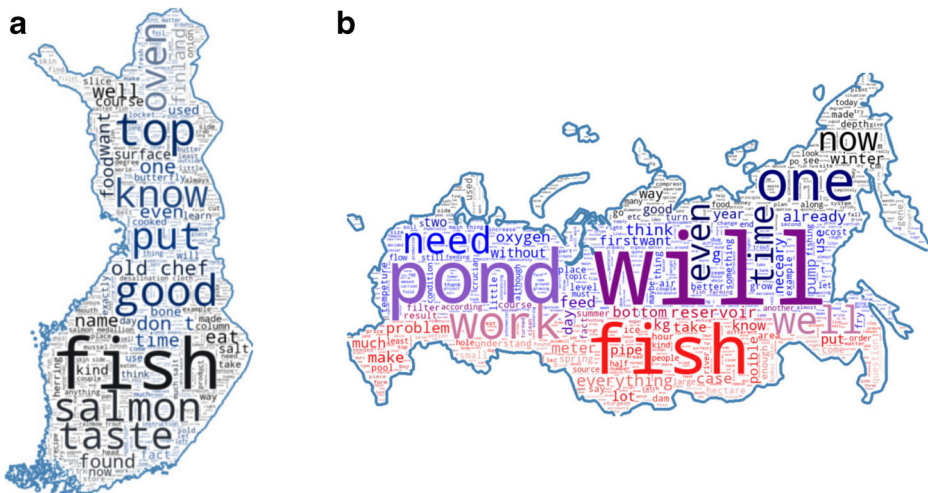


Fig. 5 Most frequent words in (a) Finnish Suomi24 Forum and (b) Russia Forums

It is noticeable from WordCloud visualization that this is not enough to reveal insights in terms of trade-barrier identification and answer our research questions. This motivates the use of other tools such as sentiment analysis, similarity calculus and clustering analysis to help grasp other hibernated aspects in both datasets.

## 5.2 Sentiment analysis results

Initially, we want to quantify the extent of positive and negative sentiment in the initial Finnish and Russian dataset using SentiStrength sentiment analysis method. Table 4 summarizes the results of Finnish and Russian sentiment scores in terms of number of posts in each sentiment polarity.

It is noticeable that negativity is the most dominant sentiment polarity for the Finnish side, followed by the neutrality and positiveness polarities. In contrast, neutrality appears to be the dominant sentiment polarity for the Russian side, followed by positiveness and negativity. Since the core methodology of our model builds on the handling of posts with negative polarity, we can therefore notice that the amount of such datasets is subsequently higher in the case of the Finnish dataset compared to the Russian dataset.

## 5.3 Trade-barrier identification results

Critical in implementing the trade-barrier type identification is determining the high-frequency-based threshold  $\epsilon$  in expression (4). In order to highlight the convenience of our choice, Fig. 6 shows the result of assigning individual (negative) posts to each tariff and non-tariff barrier type for the Finnish and the Russian dataset, respectively. In the chart related to the Finnish dataset, there is a high discrepancy in values of tariff barriers compared to non-tariff ones. In contrast, this trend is no longer valid for the Russian dataset.

For the tariff barriers, it is notable that most Finnish companies suffer from the length of time their exported commodities last before they reach the final consumer with 14.74%. On the other hand, one can see a significant presence of barriers linked to rates, imports restraints, and distribution points with an average of 35%. This can be explained by the unusual increase of direct taxes imposed on the imported/exported commodities that do not follow the changes in world market prices and fair competition. Moreover, it can also be related to the indirect taxes collected at different points of distribution. Furthermore, Fig. 7 shows a shallow presence of direction-barrier with 5.49%, possibly due to WTO constraints and lack of active professional import/export in the forum.

Furthermore, the proportions of the non-tariff barriers are close to each other, with a level of 10.12% for government participation in a trade associated with administrative guidance, government procurement and state trading, and Subsidies, besides the financial control with 8.67%, which may include monetary policies to control capital flow. Customs and entry procedures with product requirements also showed a high occurrence of 8.09%, which involve eligibility in terms of paperwork and worthiness of the product. Overall, the “time length” is the most frequent one in tariff barriers, while “the government participation in trade” is the most frequent non-tariff barrier.

**Table 4** Summary of sentiment polarity analysis

Dataset	Positive	Neutral	Negative
Finnish Data	4200	5112	5320
Russian Data	9888	4060	1620



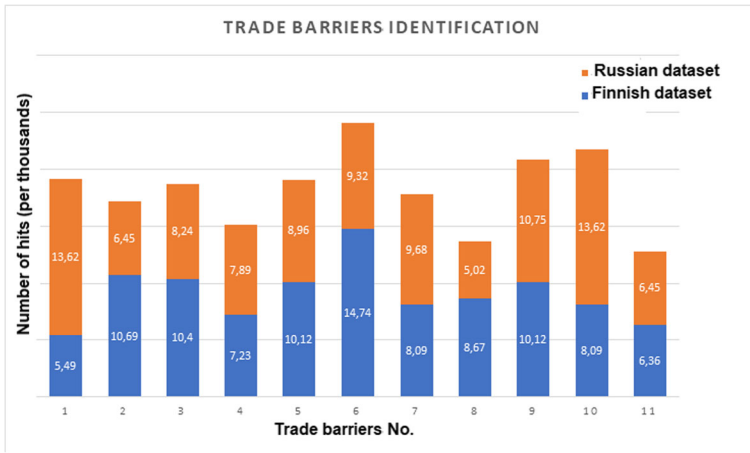


Fig. 6 Detected trade barriers in the Finnish dataset

On the other hand, Fig. 7 shows that in the tariff part, most of the Russian companies suffer from the direction tariff of the imported or exported products with 13.62%, followed by time-length and rate-change barrier with 9.32% and 10.12%, respectively. On the non-tariff part, it is clear that product requirements are the most critical barrier with 13.62%, followed by government participation in a trade with 10.75% and customs and entry procedures 9.68%. Finally, the quotas with 6.45% and the financial control with 5.02% were the lowest present barriers. Overall, the direction barrier is the most frequent one in tariff barriers, while the product requirement is the dominant one in non-tariff barriers.

### 5.4 Causes identification result

We applied the K-means clustering algorithm on each encoded corpus associated with each barrier type into five clusters  $K = 5$  with seven keywords only. Since the preliminary exploratory analysis of the influence of individual K-values, while considering the interpretation aspect, a unique centroid (K=1) will fail to grasp the diversity aspect that may involve any rational explanation of the cause. Unlike a large K-value renders the tasks of finding

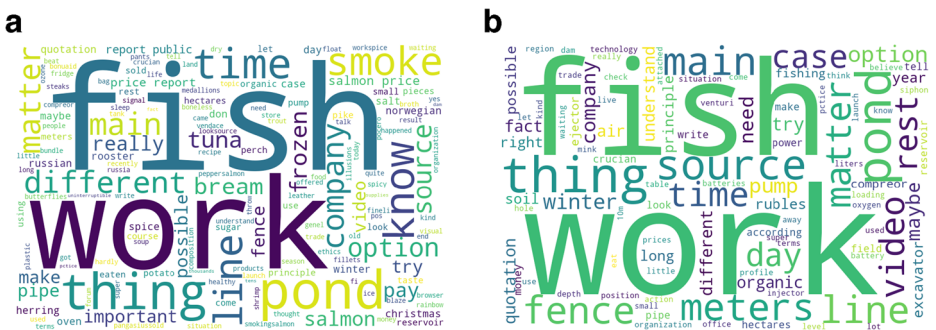


Fig. 7 (a) Most frequent words in Finnish trade barriers causes, (b) Most frequent words in Russian trade barriers causes

appropriate interpretation rather tricky. In essence, we extracted the centers of each cluster after reverse engineering from TF-IDF representation to token-level representation while removing unnecessary words to understand the scope of the remaining words used for each trade barrier type. Finally, we aggregate the remaining words to build a bag of words associated with each barrier type to capture its associated discourse as a way to reveal the causes of such a barrier. In this regard, the result shows the following:

- *Direction barrier*: seem to be caused by importing Norwegian salmon, bream and frozen Russian tuna in Christmas season with different prices (bream, frozen, tuna, Russian, salmon, Norwegian, price, different).
- *Distribution points*: seem to be caused by imposing different prices on salmon, Norwegian fish pike and Russian tuna at different import lines (Russian, Norwegian, fish pike, salmon, price, tuna, report, public).
- *Import restraints*: seem to be caused by imposing different prices on the Russian's imported fish products which reflect on the seafood prices (salmon, steaks, tuna, frozen, bream, herring, smoke, important products).
- *Purpose*: seems to be linked to the discrepancy in fish product selling points and imposing tariff rates at different points (spice, salmon, price, pay, frozen, herring, rainbow trout, tuna).
- *Tariff Rates*: caused by imposing different tariffs leading to different prices of both fresh and frozen Norwegian fish in different seasons, particularly at Christmas (frozen, salmon, prices, Norwegian, shrimp, trout, Christmas).
- *Time length*: seems linked to difficulty to accommodate extended period of Christmas discount on Russian and Norwegian fish products where this is no a such extended discount on the imported products.
- *Customs and entry procedures*: linked to the understanding of the working procedures of buying Russian fresh and frozen tuna and time requirement.
- *Financial control*: discusses about the vendace through applying different prices for buying Russian fresh and frozen fish products (salmon, price, fish, rainbow, products, Russians, frozen).
- *Government participation in trade*: discusses the official trade of fish policies (offered, common, official, trade, species, salmon).
- *Product requirements*: discusses the quality standards required for trading fish products such as frozen or fresh or ever half frozen (Russian, fresh, half frozen, frozen, products).
- *Quotas*: linked to the limited quantity of rainbow trout and salmon sold in Finland (tuna, fish, sold, year, Finland).

Similarly, interpreting the Russian barriers causes, one may notice the following:

- *Direction barrier*: seems to be caused by the tariff discrepancy of the imported vs exported fish and fish farming equipment to make fish ponds (rubles, pump, pond, battery, tell price, fish source).
- *Distribution points*: caused by the variation in import/export costs of fish products as well as the additional costs incurred in winter season (company, aquarium, fish, eat, case winter, possible, come, costs).
- *Import restraints*: linked to import restriction on special pipes, which are needed in construction of artificial small ponds for fish farming and their equipment (pipe, small ponds).
- *Purpose*: concerns local industry protection by slowing down the import process (fish pond, day rest, need right).

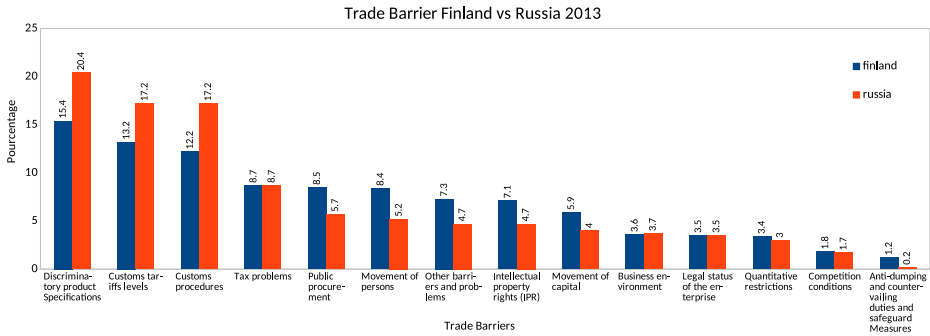
- *Tariff Rates*: linked to the currency fluctuation at various seasons on prices of fish pond and pump in winter as well as the main imported unit (rubles, fish pond, pump, winter, season, main source).
- *Time length*: linked to the time length the amount of (tariff) lasts when importing fish pond, pump and other equipment (fish pond, pump, need, money, rest).
- *Customs and entry procedures*: concerns the understanding of the working procedures in fish farming company and the time needed for import/export goods (work, company, fish, pond, understand, write, time).
- *Financial control*: related to rate discrepancy for buying tolls for building fish bond faced by companies and fish breeding (company, fish, fishing, injector, pump, money).
- *Government participation in trade*: concerns the different instruments supported by the government to help elevate this business area through different support mechanisms (organization, pump, pond, fish, pipe, hole, fence, batteries).
- *Product requirements*: concerns company's needed documentation as well as fish feed quality standards to ensure either selling their fish products or building their fish ponds (right, need, work, time, company, fish, pond, farm, fence, excavator, organic).
- *Quotas*: caused by the limited quantity of fish pond in the winter each year, which reflect on fish production (fish pond, winter, year, quotation).

The overall summary of all the causes in terms of WordCloud representation is shown in Fig. 7 (a) and (b) for Finnish and Russian corpus dataset, respectively.

## 6 Trend comparison with economics statistics

In this section, we want to ascertain how realistic our findings are, compared with the Finnish-Russian official trade barriers statistical reports. Therefore, We used the report from the Finnish trade ministry (Nieminen et al., 2014) because of its detailed classification and ease link with OS-K ontology. It is worth pointing out the following. First, there is a critical time gap between the official report (2013) and the current study, which only marginally includes some historical data dating back to the reporting period, as forum data are overwhelmingly dominated by recent posts up to 2-3 years. The purpose of this analysis is to comprehend the trend in trade-barrier occurrence and disappearance rather than to perform a one-to-one correspondence between official statistics and our forum-based analysis. In addition, tackling identified trade barriers by a given country may take several years. It acknowledged that several identified trade barriers by WTO, especially when dealing with access to the Russian market, were well documented by WTO decades ago but are still subject to Russian-west intense economic tensions. This provides ground for this comparative and trend analysis. Finally, one shall note that in (Nieminen et al., 2014) report that the food industry (including the fish sector) represents only 5.9% of the total barriers in Russia. Therefore extra caution is needed when analyzing such results. Figure 8 illustrates the Finnish-Russian various cross-border trade barriers as reported (Nieminen et al., 2014).

In this plot, a slightly different employed categorization compared to OS-K ontology. Therefore, to ease this comparison, a mapping between this categorization and OS-K. Table 5 shows each report's categorization class and its potential trade-barrier types in OS-K. Because of the insufficient reason in statistics and the absence of any further prior knowledge, the proportion of the report's categorization class is equally split among all its



**Fig. 8** The official reported trade barriers between Russia and Finland in 2013

indicated counterparts in the OS-K ontology. After such adjustment step, it becomes possible to represent the report’s findings in OS-K ontology and derive a direct comparison with forum-based analysis results.

More specifically, let  $C_j$  be the proportion of the  $j^{th}$  in the statistical report. Let  $\iota(j)$  be the set containing OS-K categories that can be equally be mapped to the  $j^{th}$  report categorization. Then the OS-K correspondence is constructed according to expression Eq. 5, where the sum is carried out over all proportions of report’s categories that contain the underlined OS-K class in their mapping as per Table 5<sup>3</sup>:

$$OSK_i = \sum_j^N \frac{C_j}{\|\iota(j)\|} \quad (5)$$

Next, to normalize each proportion score, we multiply it by 100 divided by the sum of all proportions. Figure 9 shows the normalized scores of trade-barriers occurrence and their comparison with forum-data analysis results.

Figure 9 shows a clear discrepancy score of the barrier type proportion between the statistical report and the analysis outcome. Nevertheless, some overall trends can be highlighted.

- From the Russian perspective, there is a rough indication that Product specification and Direction (discrepancy between import and export tariff) barriers achieve close scores between statistical report and forum analysis outcomes.
- Similarly, from the Finnish perspective, Import restraints, Purpose and to some extent government participation in trade are the three main trade-barrier types that show better agreement between statistical report and forum-content analysis.
- Finnish and Russian official statistics have almost equally reported several trade barriers (Direction, Product requirement, Purpose).
- Looking at the highest statistical reported barriers, one recognizes an agreement between Russian and Finnish data where Customs and entry procedure class takes the most significant share. In the second position, Rates class is shared between Russian and Finnish outcomes as well.
- Comparing the proportion of the Tariff and Non-Tariff barriers, we report a close distribution between Russian and Finnish forum data analysis outputs, where (55%, 45%)

<sup>3</sup>By assuming  $C_j=0$ , then ratio in expression 5 vanishes, to avoid singularity, N can be taken equal to the total number of categories in the statistical report, which is 14.

**Table 5** Official vs OS-K trade barriers

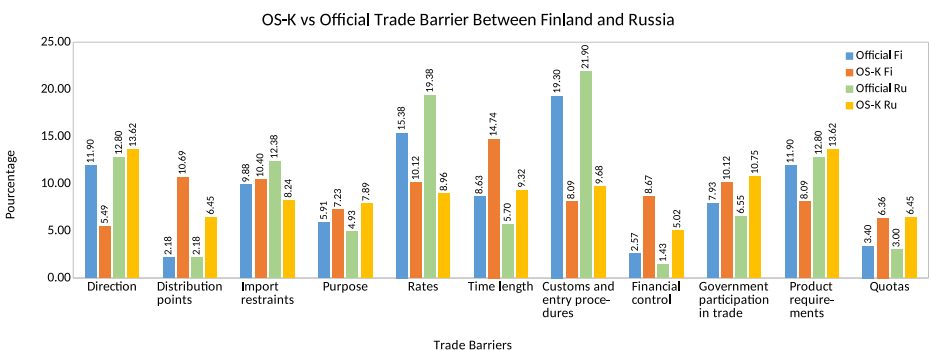
Official Barriers	OS-K Barriers
Discriminatory product specifications	Product requirements, Import restraints.
Customs tariffs levels	Tariff rates.
Customs procedures	Customs and entry procedures.
Tax problems	Tariff rates, Import restraints, Purpose, Distribution points.
Public procurement	Direction, Purpose, Government participation in trade.
Movement of persons	Direction, Product requirements.
Other barriers and problems	Time length
Intellectual property rights (IPR)	Customs and entry procedures.
Movement of capital	Financial control, Time length, Product requirements.
Business environment	Government participation in trade.
Legal status of the enterprise	Customs and entry procedures.
Quantitative restrictions	Quotas.
Competition conditions	Purpose, Government participation in trade.
Anti-dumping and countervailing duties and safeguard Measures	Financial control, Government participation in trade.

and (58%, 42%) are approximately assigned Tariff vs Non-Tariff scores in case of Finnish and Russian data, respectively. This surprisingly closely compares to official report finding as well where the distribution (54%, 48%), (57%, 43%), for Finnish and Russian cases, respectively.

Therefore, in overall, the comparison between forum-based analysis and statistical report findings reveals salient characteristics, confirming the timely endurance of trade-barrier, especially with political tension.

## 7 Implications

This research has twofold research implication: scientific implication and business implication.



**Fig. 9** The OS-K vs the official reported trade barriers between Russia and Finland

**Scientific implication** of this research concerns the contribution to the NLP-based research, which, we demonstrated that it is still in its infancy phase to answer our research questions. In overall, i) a new ontology for fish business data collection, ii) a new deep-learning model that utilizes the state-of-the-art BERT architecture to match barrier-type definitions and negative polarity posts, and, iii) a new reverse-engineering based clustering method for causality identification have been devised and successfully implemented. Due to its intuitive ground, sound interpretation, theoretical foundation and open-source implementation, we expect the platform to easily accommodate other business sectors and find its way to a broad variety of other business applications.

*Business implications* of this research lies in the critical importance of trade-barrier identification for new investors and small-to-medium scale enterprises where it is commonly acknowledged that the official economical statistical and insurance review report are lacking many unforseen and under-reported genuine barriers, which for various reasons can be overlooked by insurer and audit experts. Such a relevance is also rooted back to the importance of learning from peers in business sectors, where feedback from local SMEs is of paramount importance to either encourage or discourage other economical actors to pursue cross-border business adventures and internationalization. This provides regulators and national agencies with useful insights to reshape policy documents and engage in a more evidence-based trade negotiation. For instance, high and unfair competition and rigid regulations can lead to the bankrupt of the SMEs and economic long run failure (York, 2018). On the other hand, with the dominance of the mutual distrust and political escalation as in recent Russian-European trade relations, the reliance on official reporting for trade-barrier analysis is likely to be biased. Therefore, the contribution of any alternative evidence based collection will be of paramount importance for both economical operators and national agencies.

## 8 Conclusion

This paper introduces an efficient machine-learning-NLP approach for automatic trade type identification from business forums, contributing towards populating NLP technology in business intelligence mining. The architecture pipeline includes three key components. First, a three-leg ontology has been devised to collect relevant data from Finnish and Russian fish business forums. Second, a BERT architecture is used to match each trade-barrier type's definition in the (OS-K) ontology, distinguishing eleven Tariff and Non-Tariff barrier types to negative polarity posts of the discussion forums. An automated thresholding similarity-based matching technique is then employed to quantify the extent of matching between each negative post with all trade-barrier types. Third, a TF-IDF vectorization and k-means clustering reverse-engineering based method is used to grasp the cause(s) behind each trade-barrier occurrence. Visualization using WordCloud and trend comparison with official statistics from Finnish Trade Ministry are employed to demonstrate the feasibility and technical soundness of the proposal, enabling us to identify latent characteristics of trade-barriers that goes beyond time frame constraint. The proposal paves the way for further investigations of other business-related intelligence that can be employed from the overwhelming increasing user-generated content from specialized web forums or blogs. The analysis of the findings suggests the dominance of the Time-length barrier type in the Finnish discussion forum dataset and import vs export tariff discrepancy and product requirement barrier types in the Russian forum dataset. Such a finding overlaps to a large

extent with official statistics where Non-Tariff related barriers, especially product specifications, take the highest share among the reported barrier types. The study fills the gap of the urgent need to sustain mutual trust in cross-border collaborations. The findings can also be cast into a generic framework of the recommender-system umbrella. Indeed, the platform and results can assist companies or regulators in providing business-related recommendations to overcome the detected trade barriers regarding the target country trading policies for instance.

**Funding** Open Access funding provided by University of Oulu including Oulu University Hospital. This work is partly supported by the EU CBC Karelia on IoT and Business Creation, as well as the H2020 YougRes project (Ref. 823701), which are gratefully acknowledged.

**Data Availability** All data generated by this work are available in the supplementary file of this submission and Github project provided in this paper.

**Code Availability** The source code is available in the Github account pointed out in the paper.

## Declarations

**Conflict of Interests** The authors have no conflicts of interest to declare that are relevant to the content of this article.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Anderson, J. E. (2004). E.v.W.: Trade costs. *Journal of Economic Literature*, 42(3), 691–751.
- Bahja, M. (2020). *Natural language processing applications in business*. Singapore: E-Business IntechOpen. [10.5772/intechopen.92203](https://doi.org/10.5772/intechopen.92203).
- Bounab, Y., Oussalah, M., & Ferdenache, A. (2020). Reconciling image captioning and user's comments for urban tourism. In *IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1–6).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
- Eisenstein, J. (2019). *Introduction to Natural Language Processing*. MIT press, USA.
- European Commission. Market Access Database. (2021). <https://madb.europa.eu/madb/>. Accessed: 2021-07-01.
- Ghosh, S. (2009). Application of natural language processing (nlp) techniques in e-governance. In *E-Government Development and Diffusion: Inhibitors and Facilitators of Digital Democracy* (pp. 122–132). IGI Global India.
- Global trade, WTO. (2021). <https://www.wto.org/index.htm>. Accessed: 2021-07-01.
- Grübler, J., & Reiter, O. (2021). Characterising non-tariff trade policy. *Economic Analysis and Policy*, 71, 138–163.
- Hasyim, M., Yolanda Latjuba, Akhmar, A., Kaharuddin, A. M., Jihad, K., & Saleh, N. (2021). Human-robots and google translate: A case study of translation accuracy in translating french-indonesian culinary texts. *Turkish Journal of Computer and Mathematics Education*.



- Keith, K. A., Jensen, D., & O'Connor, B. (2020). Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. In *Proceedings of the 58<sup>th</sup> ACL. ACL. New York*.
- Krouka, M., Elgabli, A., Issaid, C. B., & Bennis, M. (2021). Energy-Efficient Model Compression and Splitting for Collaborative Inference Over Time-Varying Channels.
- Lagi, M. (2019). Natural Language Processing – Business Applications. <https://emerj.com/ai-sector-overviews/natural-language-processing-business-applications/>. Accessed: 2021-07-01.
- Lawrence, J., Gitman, C., McDaniel, A.-S., Monique, R., Koffel, L., Talsma, B., & Hyatt, J.C. (2018). Introduction to Business. Rice University, USA.
- Luo, X., Tong, S., Fang, Z., & Qu, Z. (2013). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Sciences*, 38, 937–947.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Mitchell, C. M., Evanini, K., & Zechner, K. (2014). A triologue-based spoken dialogue system for assessment of english language learners. In *Proceedings International Workshop on Spoken Dialogue Systems*.
- Morgan, S., & Winship, C. (2007). *Counterfactuals and Causal Inference*. London: Cambridge University.
- Nieminen, M., Siikaluoma, H., Koskela, A., & Vilhunen, L. (2014). Internationalisation and Barriers to Trade in 2013, Finland (2014). Finland Ministry of Trade, Finland.
- Onkvist, S., & Shaw, J. J. (2009). *International Marketing. Strategy and Theory*. New York: Routledge.
- Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).
- Renjith, S., Sreekumar, A., & Jathavedan, M. (2020). An extensive study on the evolution of context-aware personalized travel recommender systems. *Information Processing & Management*, 57(1), 102078. 10.1016/j.ipm.2019.10207.
- Rong, Y., & Liu, Y. (2020). Staged text clustering algorithm based on k-means and hierarchical agglomeration clustering. In *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)* (pp. 124–127).
- Salton, G., & Harman, D. (2003). Information retrieval. In *Encyclopedia of Computer Science* (pp. 858–863).
- Schaffer, R., Agusti, F., & Dhooge, L.J. (2014). *International Business Law and Its Environment*. London: Cengage Learning.
- Senyuk, L. (2021). 5 Natural Language Processing (NLP) Applications In Finance. <https://www.avenga.com/magazine/nlp-finance-applications/>. Accessed: 2021-07-01.
- Solutions, A. C. (2017). The Limitations of a Translation. <https://acedtranslations.com/the-limitations-of-a-translation/>.
- Tarnowska, K. A., & Ras, Z. W. (2019). Sentiment analysis of customer data. In *Web Intelligence*, (Vol. 17 pp. 343–363). IOS Press.
- Tarnowska, K. A., & Ras, Z. (2021). Nlp-based customer loyalty improvement recommender system (clirs2). *Big Data and Cognitive Computing* 5, 1–4.
- Techlabs, M. (2018). Re-Modelling the Hospitality Industry with Artificial Intelligence, Predictive Analytics & NLP. <https://chatbotsmagazine.com/re-modelling-the-hospitality-industry-with-artificial-intelligence-predictive-analytics-nlp-e875fff604b8>. Accessed: 2021-07-01.
- Technologies, D. O. (2021). These popular NLP applications are changing the face of Finance. <https://www.day1tech.com/these-popular-nlp-applications-are-changing-the-face-of-finance/>. Accessed: 2021-07-01.
- The Office of the U.S. Trade Representative (USTR). FOREIGN TRADE BARRIERS. (2021). <https://ustr.gov/search?q=trade+barriers>. Accessed: 2021-07-01.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61, 2544–2558.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, (Vol. 5998 p. 6008).
- York, E. (2018). The Impact of Trade and Tariffs on the United States. <https://taxfoundation.org/impact-of-tariffs-free-trade/>.