2020 Asia–Pacific Statistics Week
A decade of action for the 2030 Agenda: Statistics that leaves no one and nowhere behind
15-19 JUNE 2020 | Bangkok, Thailand

Application of the text mining technique to improve the dataset integration in foreign trade price indexes

Reza Hadizadeh[1]; Saeed Fayyaz[2]; Abbas Moradi[3]

[1] Leader group of PPI, Statistical Center of Iran
[2]Statistician on Labour Force Statistics, Statistical Center of Iran
[3] Expert from Statistical Research and Training Center

Abstract:

Price Indexes are considered as one of the oldest criteria to assess the economic changes. So high important have them in economic planning, it would be practical and helpful for economists to analyse the price indexes. In order to improve the quality of this statistics, both in terms of accuracy and reliability, Statistical Center of Iran (SCI) have been using custom's administrative data source and integration methods with current surveys to produce price indexes of import and export merchandises based on base- year 2012 and international HS classification system since 2015. In this classification, the hierarchy code is given based on ingredients, level of dispensation, application and economic activity type which include 21 main classes, 98 sub-classes, 1241 title and 5113 subtitles respectively. Meanwhile, there are some significant challenges when HS classification is used for producing price indexes of import and export merchandises. Conventionally, Price Indexes has been used for monitoring price fluctuation that fundamentally has different dimensions of statistical variables comprising tariff code, product value on both internal and external currencies, country of origin, transportation type, average currency exchange rate and date of the entrance to the country/destination. Clearly, each 8-digit code with above mentioned attributes covers a group of merchandises which are less in common with others in conventional survey and price index calculation. Practically, real price fluctuations cannot be assessable precisely when the price of some merchandise goes up and some others go down in the same group. This might lead to wrong estimation of real price changes. In this paper, new methods will be proposed so as to reach an efficient and practical solution for precise measurement of merchandises' price in the market. This method, also results in higher quality insurance and public reliance on foreign trade registers. Technically, merchandises' description will be added to HS code to differentiate between items in each group.
Nonetheless, text matching, especially in none English language (Persian), is the unavoidable difficult challenge to deal with. Text mining technique will be applied so as to solve this problem and matching the different merchandises in different datasets resulting in more probable record linkage. In this way, statisticians need modernized knowledge based skills (e.g. text mining technique) to produce more reliable, accurate and timely foreign trade price indexes. The results will be beneficial for other statistical offices producing these official statistics.

Keywords: foreign trade statistics, import and export merchandises, price indexes, record linkage, text mining, accuracy and reliability.

1. Introduction:

Price Indices has been considered as one of the oldest indices for monitoring economic changes and fluctuations. These values imply the price variations on merchandise and services in a predetermined period. Normally, there are four main indices in economics literature including Customer Price Index(CPI), Producer Price Index(PPI), Export Price Index(EPI) and Import Price Index(MPI). In Price Statistics system, EPI and MPI are used to different purposes for example EPI depict the price trend of exported merchandise outside the country borders while MPI focuses on a similar trend for imported merchandise from in a specific period. This statistic plays an important role in foreign trade analysis, expenditure side of National account, trade outputs and National Gross expenditure.

Taking the price of both exported and imported merchandise as a key factor affecting the foreign trends in countries, the exact evaluation of price changes is a necessary. Therefore, precise calculations play a pivotal role in efficient decision making in the trade sector. Technically, price indices for both imported and exported merchandise have been calculated and disseminated based on international recommendations and customs registers by application of value unit measure and HS classification. In this classification, the hierarchy code is given based on ingredients, level of dispensation, application and economic activity type which include 21 main classes, 98 sub-classes, 1241 title and 5113 subtitles respectively.

Value Unit index Method

This method that most countries have been using it, is the most common way of calculation. This method uses the customs registers for both imported and exported merchandise based of values of trades and other supplementary data in customs organization. It is a cost effective and easy to apply method which has its own advocants (Word Bank ,2009). This index is the ratio of the value of a unit in considered period on the reference period.

$$p_u = \left(\frac{\sum_{m=1}^{M} p_m^t q_m^t}{\sum_{m=1}^{M} q_m^t} \right) / \left(\frac{\sum_{n=1}^{N} p_n^0 q_n^0}{\sum_{n=1}^{N} q_n^0} \right) \qquad (1)$$

Considering its advantages into account, the value unit index has its own drawbacks for calculation of both imported and exported merchandise with HS classification as well. These issues cause unforeseen errors leading to reduction of precise calculation this is must for monitoring the changes. In order to calculate the price index of considered merchandise with its related attributes, its value in a specific period (has been fixed during the calculation) divide on the value of the same merchandise in the reference period. So, if this attribute changes, the price variations do not describe the price changes only and the quality of the merchandise include too. This might resulted in bias outputs and do not meet the decision maker's interests and mislead them.

In Iran Customs organization register there is a series of information based on the HS coding system that is called on attribute set. This attribute set include but not limited to tariff code, value in US dollar and Iran's currency (Rial), weight, country of origin, transportation type, data of arrival and average of exchange rate. In this circumstance, for price calculation of each unit of considered merchandise, a fixed value unite can be applied that is similar to average daily price of in deliberated period. But there is still a big challenge that is related to the HS coding system that has been group classified this is in contrast with the definition of price index calculation. For example, HS 71110000 (Base metals, silver or gold, clad with platinum, not further worked than semi- manufacture) includes gold and silver with all side products. If the price index displays a growth in comparison to last period it is not exactly possible to emphasize that which one resulted to this increase, gold or silver? It is also possible that one of them increased while the other decrease. As a result, while the attribute of appearance, time, quality and quantity are not considered in the calculations, the price indices will have errors and biases. Additionally, in time that there are repetitive shocking changes in merchandise' prices due to the economic sanctions, international and national political changes, etc. it is wrongly determined as outliers and omit from the data and caused over/underestimation. To solve and overcome this problem, text-mining technique is offered in this study.

2. Methodology

Text mining is the process of seeking or extracting the useful information from the textual data. It is an exciting research area as it tries to discover knowledge from unstructured texts. It is also known as Text Data Mining (TDM) and knowledge Discovery in Textual Databases (KDT). KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Text mining process is same as data mining, except, the data mining tools are designed to handle structured data whereas text mining can able to handle unstructured or semi-structured data sets such as emails HTML files and full text documents etc. (Gupta et al., 2009). Text Mining is used for finding the new, previously unidentified

information from different written resources. Structured data is data that resides in a fixed field within a record or file. This data is contained in relational databases and spreadsheets. The unstructured data usually refers to information that does not reside in a traditional row-column database and it is the opposite of structured data. Semi-Structured data is the data that is neither raw data, nor typed data in a conventional database system. Text mining is a new area of computer science research that tries to solve the issues that occur in the area of data mining, machine learning, information extraction, natural language processing, information retrieval, knowledge management and classification. Figure 1 gives an overview of the text mining process (Vijayarani et al., 2013).

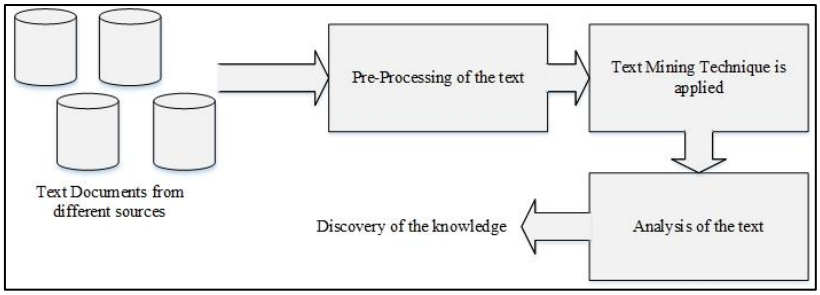

Fig. 1: Text Mining Process

In other to text mining in the database of registered data' customs, it would be necessary to follow a series of steps. first, we descript Harmonized Commodity Description and Coding Systems (HS) that it classifies export and import merchandise. second, we present proposed approach for calculating precise MPI and XPI. we utilize R software and "tm" package for text mining in this paper.
The Harmonized System is an international nomenclature for the classification of products. It allows participating countries to classify traded merchandise on a common basis for customs purposes. At the international level, the Harmonized System (HS) for classifying merchandise is a six-digit code system.

The HS comprises approximately 5,300 article/product descriptions that appear as headings and subheadings, arranged in 99 chapters, grouped in 21 sections. The six digits can be broken down into three parts. The first two digits (HS-2) identify the chapter the merchandise are classified in, e.g. 09 = Coffee, Tea, Maté and Spices. The next two digits (HS-4) identify groupings within that chapter, e.g. 09.02 = Tea, whther or not eflavored. The next two digits (HS-6) are even more specific, e.g. 09.02.10 Green tea (not fermented)... Up to the HS-6 digit level, all countries classify products in the same. The Harmonized System was introduced in 1988 and has been adopted by most of the countries worldwide. It has undergone several changes in the classification of products. These changes are called revisions and entered into force in 1996, 2002, 2007, 2012 and 2017.

Table 1. HS Codes for different sectors

HS code	Group's Name	**HS code**	Group's Name
01-05	Animal & Animal Products	50-63	Textiles
06-15	Vegetable Products	64-67	Footwear / Headgear
16-24	Foodstuffs	68-71	Stone / Glass
25-27	Mineral Products	72-83	Metals
28-38	Chemicals & Allied Industries	84-85	Machinery / Electrical
39-40	Plastics / Rubbers	86-89	Transportation
41-43	Raw Hides, Skins, Leather, & Furs	90-97	Miscellaneous
44-49	Wood & Wood Products		

As mentioned, each two-digit HS code contains several six-digit/eight-digit HS codes. The Iran's customs register information has a description of the merchandise which include the specifications and attributes of the imported or exported merchandise. For example, in table 2, an example of the code 02041000 with the sample description of the merchandise was presented. In order to calculate price

indices based on 8 HS digit codes, relative prices should be calculated and it would be necessary that each 8-digit code should have identical attributes between two different periods. Taking register data as a data source, there are some remarkable challenges of different type of merchandise as well as redundant characters resulting in low efficient linkage. So, if the linkage is apply based only 8-digit, the calculated indices will mislead and be biased. In this paper we proposed text mining technique for solving this problem.

Table 2. Sub categories for specific HS code

HS code	Descriptions
02041000	The carcass/lamb are left according to the value statement
02041000	The carcass of the remaining meat according to the, declaration of value
02041000	@ carcass of fresh mutton remain s according, to the value statement
02041000	Hot mutton 1- value statement
02041000	The remaining carcass of the sheep according to the declaration of value

Data preprocessing plays a very important role in text mining techniques and applications. It is the first step in the text mining process.

STEP 1
Data preprocessing

- Removing numbers
- Removing punctuation
- Removing stop words
- Removing strip whitespace
- Steaming

Commonly each code's description contains a series of characters includes but not limited to numbers, symbols, low importance signs and redundant spaces. Thus, in order to prepare high quality analysis on theses codes' descriptions it would be necessary to remove these characters (punctuation, numbers, stop words and whitespaces). Last but not least step is Steaming that is done in different ways that are demanding and interested readers can find details in many sources. Steaming is one of the prominent steps in text mining techniques.

STEP 2
Documents similarity

- Jaccard similarity
- Cosine similarity

Text similarity measures play an increasingly important role in text related research and applications in tasks such as information retrieval, text classification, document clustering, topic detection, topic tracking, questions generation, question answering, essay scoring, short answer scoring, machine translation, text summarization and others. Finding similarity between words is a fundamental part of text similarity which is then used as a primary stage for sentence, paragraph and document similarities. Words can be similar in two ways lexically and semantically.

Many methods have been introduced for similarity findings in two different texts. One of the famous methods is the metric distance that exists between two texts. Technically, the R programming uses Jaccard similarity and Cosine similarity methods to find similarities between two texts (Niwattanakul et al., 2013). It is important to note, however, that this is not a proper distance metric in a mathematical sense as it does not have the triangle inequality property and it violates the coincidence axiom (Gunawan et al. ,2018).

It is worth mentioning that unqualified data matching based on merchandise' descriptions is one of the immense challenges for precise calculation of imported and exported price indices. Technically, linkage with the only merchandise' descriptions (tariff code, the value in US dollar and Iran's currency (Rial), weight, country of origin, transportation type, data of arrival and average of the exchange rate) can result in missing some important data, while in price indices calculation the important parts are the specific merchandise' attributes and stability of considered merchandise during the specific period.

In connection with relative price formula, after preprocessing phase, similarity verdict of merchandise' descriptions for current and last month is necessary for the superior quality of prices. In other words, each merchandise's description of 8-digit tariff code in data set (1), current month, should be linked with

similar description in dataset (2), last month. The relative price criteria will be the similarity degree with at least 70% similarity threshold.

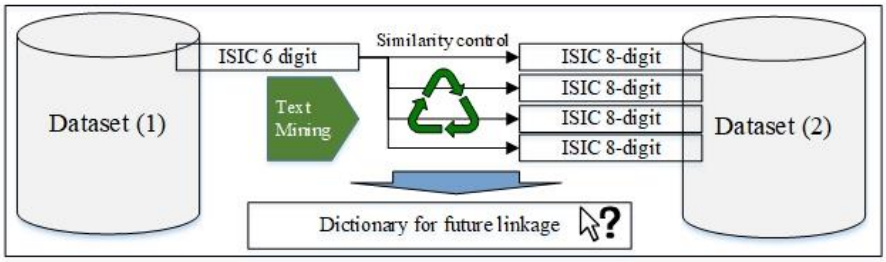

Fig. 2: Linkage process with text mining in price indices

STEP 3 Library making	•Labeling •Identify Code

Following the similarity determination, a label was appointed to each of tariff codes based on the keyboard's in description and the number of repetitions in the database. These labels however can be a 4-digit number which can be attached to previous 8-digit tariff codes. The new codes had 12-digit that would be more beneficial for the next linkage.

3. Result

The process of price indices calculation for imported and exported merchandise in Iran's NSO based on Customs registered data have 2 important phases. These two phases are 1) preparation and investigation and 2) relative prices and indices calculation. Mostly, receiving data from custom is not well structured as statisticians need and it should be purified and outlays must be removed. Subsequently, relative prices and price indices for 8-digit, 2-digit, merchandise' group and total price of both imported and exported merchandise. Calculation of relative prices (the ratio of current price to last month price) should be used keys and group attributes mentioned before. Regarding huge amount of daily data, more than 400,000 records, it would be impossible to eye-control and review the data and a systematic mechanism should be developed to prevent outliers. There is no doubt that outliers in data sets after the initial phase are seen due to a variety of merchandises in 8-digit codes and it would be making linkage with descriptions because the descriptions are same apparently but there are different merchandises in practice.

Lastly, text mining was the solution that was used by authors to overcome the problem. These attempts resulted in the reformation of the current calculation process and the text-mining phase was added.

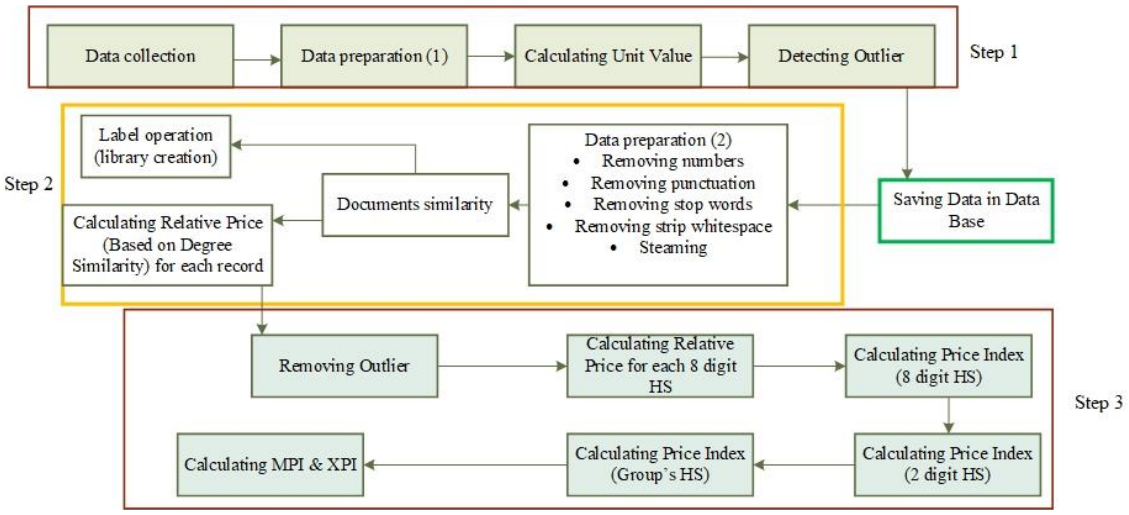

Fig. 3: Different steps of innovative linkage

A practical sample of applying text mining technique in phase 2 of indices calculation process was shown in below table.

Table 2. Sub categories for specific HS code

HS code	Descriptions	Identify Code
02041000	carcass lamb left accord value statement	0111
02041000	carcass remain meat accord declaration value	0112
02041000	carcass fresh mutton remain accord value statement	0113
02041000	Hot mutton accord value statement	0114
02041000	remain carcass sheep accord declaration value	0115

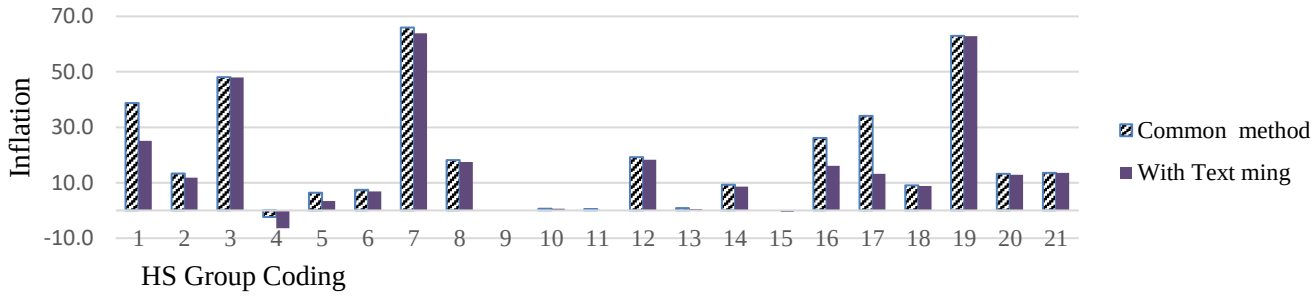

Fig. 4: Inflation rate for imported merchandise in Quarter 3 2o19 in Iran

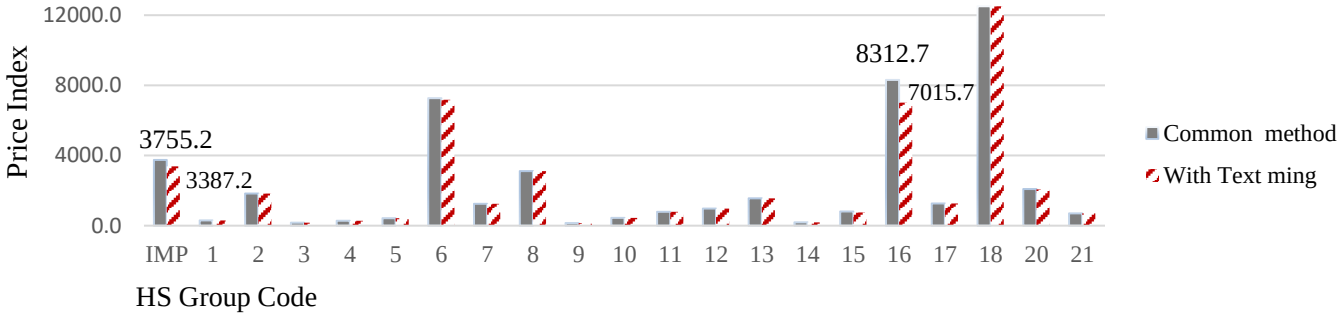

Fig. 5: Inflation rate for imported merchandise in Quarter 3 2o19 in Iran

4. Discussion, Conclusion and Recommendations:

Average price report has been under attention of statistical users as merchandise price indices. This managerial report has been prepared based on 8-digit HS codes. In many cases that the price fluctuation for each tariff code of merchandise group is not in a harmony, the average price in not precise and may be biased. In these cases, however text mining can be considered as a last resort for solving the problem. Text mining can cluster similar merchandise with a high level of similarities resulted in high quality average prices for each merchandise group.

References:

D. Gunawan et al. (2018),' The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents', Journal of Physics Conference Series 978(1):012120

S. Vijayarani et al, (2013) International Journal of Computer Science & Communication Networks, Vol 5(1),7-16.

S. Niwattanakul et al. (2013) Using of Jaccard Coefficient for Keywords Similarity, Proceedings of the International Multi Conference of Engineers and Computer Scientists 2013 Vol I, IMECS, March 13 - 15, 2013, Hong Kong.

Word Bank (2009),'Export and Import Price Index Manual'.

Vishal Gupta and Gurpreet S. Lehal, (2009)," A Survey of Text Mining Techniques and Applications", Journal of emerging technologies in web intelligence, Vol. 1, No. 1.