UDC 004.8
DOI: https://doi.org/10.17721/1812-5409.2025/1.21

Olena SKLIARENKO[1], PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0001-6555-1223
e-mail: olena.skliarenko@e-u.edu.ua

Anatolii PASHKO[2], DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0001-6944-8477
e-mail: anatoliipashko@knu.ua

Leonid LYTVYNENKO[1], PhD (Engin.)
ORCID ID: 0000-0002-0828-383X
e-mail: leonid.lytvynenko@e-u.edu.ua

Yanina KOLODINSKA[1], Senior Lecturer
ORCID ID: 0000-0002-3330-7565
e-mail: yanina.kolodinska@e-u.edu.ua

[1]Private Higher Educational Establishment "European University", Kyiv, Ukraine
[2]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

DEEP LEARNING MODELS FOR SOLVING THE PROBLEM OF COMPARING UNSTRUCTURED TEXTUAL INFORMATION

This study is devoted to the application of modern deep learning methods, such as universal multilingual sentence vector representations, for analyzing and comparing the similarity of unstructured text documents. The paper develops an algorithm for determining the similarity of multilingual texts using deep neural networks. The authors propose a supervision-free method for fast analysis and comparison of complex documents on similar topics written in different countries and in different languages. The presented methodology involves pre-processing documents, reformatting them for comparison, and applying deep learning methods to detect textual similarities. The algorithm is tested on the example of a comparative analysis of the similarity of the texts of the Tax Code of Ukraine and the German fiscal code. The Ukrainian fiscal code, presented in HTML format, was processed using the library for analyzing HTML documents and regular expressions, while the German fiscal code in PDF format required special analysis of unstructured content. In the course of the study, to improve the accuracy and overcome the identified limitations of universal multilingual vector representations, the Ukrainian document was translated into English using the Google Translate API. Solving the given task requires deep pre-processing of the data, which was carried out in this study. The article provides a step-by-step algorithm of the proposed methodology, which leads to the formation of a correlation table, which determines the similarity between similar multilingual text documents. Additionally, the authors support every step with their reasoning about how to interpret intermediate results and understand if the method should be adjusted to the need or the problems of the specific documents, like applying machine translation or advanced pre-processing.

The authors propose an algorithm for applying this method to the fiscal codes of European Union member states or other countries. In this research, we identified and described the limitations of this approach, demonstrated the verification of intermediate results, and supplemented the method with additional capabilities to increase its reliability. Recommendations are given on the possibilities of applying the proposed methodology, in particular, the development of algorithms for standardizing certain documents of the European Union.

K e y w o r d s : ***deep learning, universal vector representations of sentences, financial documents, fiscal code, regular expressions.***

AMS 2020 classification: 68T07, 68T50.

Introduction

We focus our research on the comparison of the Ukrainian fiscal code with the German fiscal code. These fiscal codes share common principles but have a very different structure and different sizes – and so making them difficult to compare manually.

As a part of the preparation for the research, we analyzed the frequency of the changes to the Ukrainian fiscal code, and we can see that it has received constant updates and amendments during the last decade shown on Fig. 1 which signifies the importance of our research not only from a scientific perspective, but also from the practical need.

Ukraine is actively moving to the European Union, and an important part of the movement is to harmonize Ukrainian legislative standards with the standards required in the European Union. At the same time every country, while following the general principles, adopts its own legislative documents which are not identical.

Natural language analysis has made many steps forward in recent years, and by applying modern Deep Learning techniques to document analysis and comparison we are offering a method to identify missing common parts. We do not aim to substitute human analysis, but aim to offer a robust, easily verifiable approach that will supplement it – so we're leaving generative models out of the scope of this research.

We're using Multilingual Universal Sentence Embeddings initially introduced by Google (see Basystiuk, & Melnykova, 2022; Shakhovska N., Basystiuk, & Shakhovska K., 2019) – which allows us to analyze texts in different languages.

In recent years, scientists have been using artificial intelligence technologies, in particular deep learning methods, to analyze unstructured data, recognize and compare natural language texts.

For example, works (Golomoziy et al., 2023; Ianevych et al., 2023; Peng et al., 2023) investigated methods for classifying unstructured texts related to the processing of court documents. The transformation of unstructured text data into a statisti,al dataset allowed them to be used for analysis by scientists, journalists, government officials, politicians, and anyone else. This papers provide a literature review summarizing the methods and approaches used in the application of natural language processing (NLP) techniques specifically for the legal domain.

ISSN 1812-5409 (Print), ISSN 2218-2055 (Online)

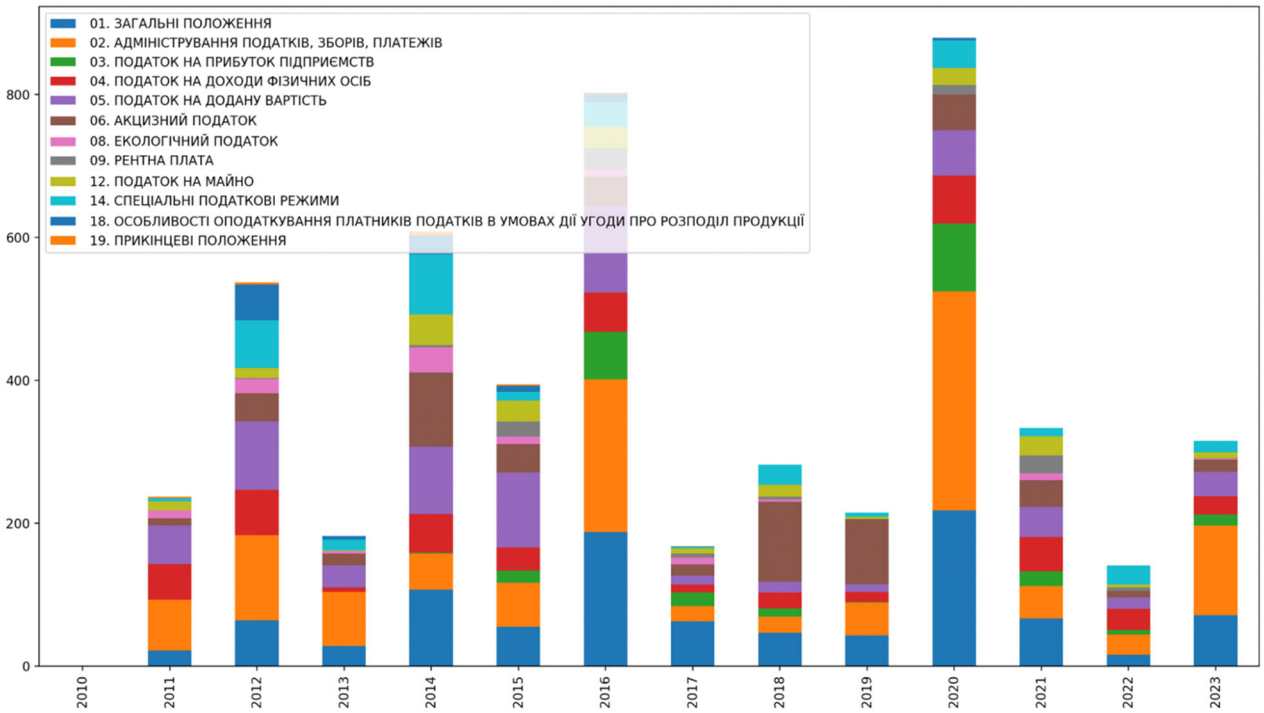

Fig. 1. Frequency of changes to different sections of the Ukrainian fiscal code

It should be mentioned that other researchers' scientific works on similar topics mainly focus on the study of methods and algorithms for classification, translation, and clustering. Unlike other studies, this article is devoted to the development of methods and algorithms for determining the similarity of the content of pro-lingual unstructured text documents.

One of the problems noted by researchers in improving object recognition is the availability of prepared large datasets required for training neural networks. The authors of this article propose a methodology for preparing, processing, and analyzing multilingual similar text documents to determine their similarity in content. The model proposed by the authors was tested by comparing the texts of fiscal codes of different countries. For the sake of purity of the experiment, the authors chose German texts rather than English-language documents, which is a fairly common practice.

1. Methodology

The solution to this problem requires deep pre-processing of relevant information and unstructured data, the so-called sub-deep (in the sense of deep learning) processing.

When processing unstructured data, sub-deep processing includes Understanding, Cleaning, Data Transformation, Feature Engineering, Splitting DataSet, Data Augmentation.

Let us present our methodology in the form of a step-by-step algorithm, considering the case of comparing the similarity of the texts of the fiscal codes of Ukraine and Germany.

Ukrainian and German fiscal codes, in addition to having different structures and sizes are also offered in different formats.

1.1. Data preparation. Ukrainian and German fiscal codes, in addition to having different structures and sizes are also offered in different formats.

Ukrainian fiscal code is an HTM document, where different sections are represented as tags with the specific class names:

```
<span class=rvts15>РОЗДІЛ І. ЗАГАЛЬНІ ПОЛОЖЕННЯ</span></p>
<p class=rvps2><a name=n131></a>
<span class=rvts9>Стаття 1.</span> Сфера дії Податкового кодексу України</p>
```

Using the regular expressions, we split the whole document into small segments – one segment is one article of the code with all the related sub articles. Then, using the Python library BeautifulSoup, we extract the text from HTM.

```
from bs4 import BeautifulSoup
def strip_html(html):
  return BeautifulSoup(html, 'html.parser').get_text(separator=' ')
```

We have also excluded a small number of articles which consisted of the tables or other non-textual information.

To parse the PDF file of the German fiscal code we used PyMuPDF library and its "fitz" Python package. This PDF file has a table of contents (ToC) – so we tried to utilize the built-in "get_toc()" function to get the references to different sections of the fiscal code. But the ToC of this PDF file is not a properly structured ToC in terms of the PDF format and only has visual resemblance. So, to properly split the sections of the file we had to write our own ToC parser, and then our own extractor of the fragments from the ToC.

Finally we used the fitz builtin "get_text()" method to get the textual representation of the articles, skipping tables and other non-textual information similar to what we did for the Ukrainian fiscal code.

As a result, we were able to have similarly structured and sized articles of both fiscal codes, which we saved in the JSON format for the convenience of our analysis.

1.2. Research design. One of the popular approaches to analysis of multilingual texts is based on the usage of the Multilingual Universal Sentence Embeddings (Basystiuk, & Melnykova, 2022). We used an open, pre-trained model published by Google, available at the TensorFlow hub (Basystiuk et al., 2021).

As we can see on Kaggle (Boyko et al., 2019; Yang et al., 2019), this model has gained popularity, but to see if it can be applied to our use-case and our data we performed an initial test and checked how it would behave on the samples from our texts by comparing vector distances which we calculated using the embedding from the model with our perception of what articles are close to each other:

```
import tensorflow as tf
import tensorflow_hub as hub
from sklearn.metrics.pairwise import cosine_similarity
embed = hub.load("https://www.kaggle.com/models/google/universal-sentence-encoder/frameworks/TensorFlow2/variations/multilingual/versions/2")

sentences = [
  '1.1. Податковий кодекс України регулює відносини...',
  'A commercial business is an independent...',
  'Платниками податку є: 212.1.1. Особа, постійне представництво...',
  '(1) The Federal Central Tax Office assigns every taxpayer...',
  '(2) For the purposes of this subsection...',
  '...'
]

embeddings = embed(sentences)
similarities = cosine_similarity(embeddings)
```

Finally, to evaluate the approach at scale, we used the heat map visualization from the "seaborn" Python package. We are also excluding points below some threshold to make the visualization less cluttered:

```
import seaborn as sns
import matplotlib.pyplot as plt

similarities = cosine_similarity(embed(de), embed(ua))
df = pd.DataFrame(similarities).round(2)

significant = df[corr>=.30]
plt.figure(figsize=(12,8))
sns.heatmap(significant, cmap="Greens")
```

1.3. Data analysis. After performing analysis and applying the visualization to a subset of the articles we've seen a visualization shown on Fig. 2. This result has immediately raised our concerns.

On the visualization we see recognizable horizontal and vertical lines – a clear indication that some articles of one fiscal code correlate with many articles from other fiscal code and vice versa. It is a clear indication that the model picked up some trivial textual similarities, but not the similarities in meanings.

After reviewing the data again, we realized that our versions of Ukrainian and German fiscal codes also included references and dates of the changes of the codes and by removing them we've seen some improvement shown on Fig. 3. Nevertheless, the pattern of horizontal and vertical lines was still there and we continued analyzing our data.

We were including the name of the article in the text of the article, but because every name of the Ukrainian article included "Стаття" as a part of the name – that could impact the result too. In the end, we decided to remove the article names from the article text, because even if they correlate in different fiscal codes, we're interested in the text of the articles itself, because we want to assess not only the structure but also the meaning of the fiscal codes.

The pattern was still there, so we decided to also remove all the references to other articles of the fiscal codes because they were represented as numbers and were also influencing the similarity score.

Finally, after running our analysis and visualization again, we've seen improvements on the subset data, but when we applied everything to the whole data set we've seen a picture shown on the Fig. 4. There was a clear vertical line on the chart, but after manual review we didn't find the correlation the chart showed us.

So we decided to make a more radical move and used machine translation to bring both fiscal codes to the same language. We used Google Translate API for the translation, it supports Ukrainian and works relatively well. We had to write a script to split articles into the segments to fit Google API limitation on the request length, but other than that we didn't experience problems with the API. Because Google charges money for the API access, we translated the initial, non-cleaned version of the Ukrainian fiscal code so that we're able to repeat any post processing without calling the Google Translation API again. But the tests showed that our initial cleanup made sense for the translated text as well, so we repeated it in full for the translated version.

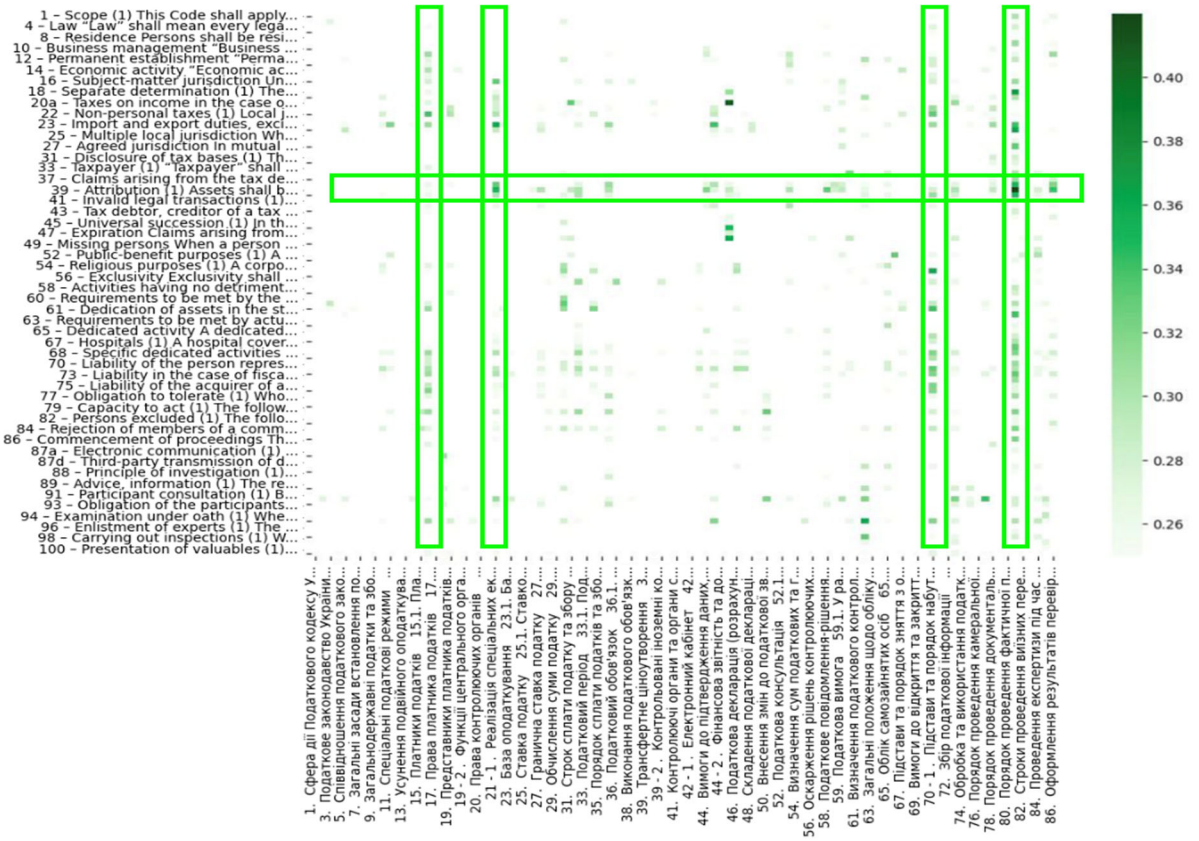

Fig. 2. Cosine similarity of the subset of articles from the Ukrainian and German fiscal codes

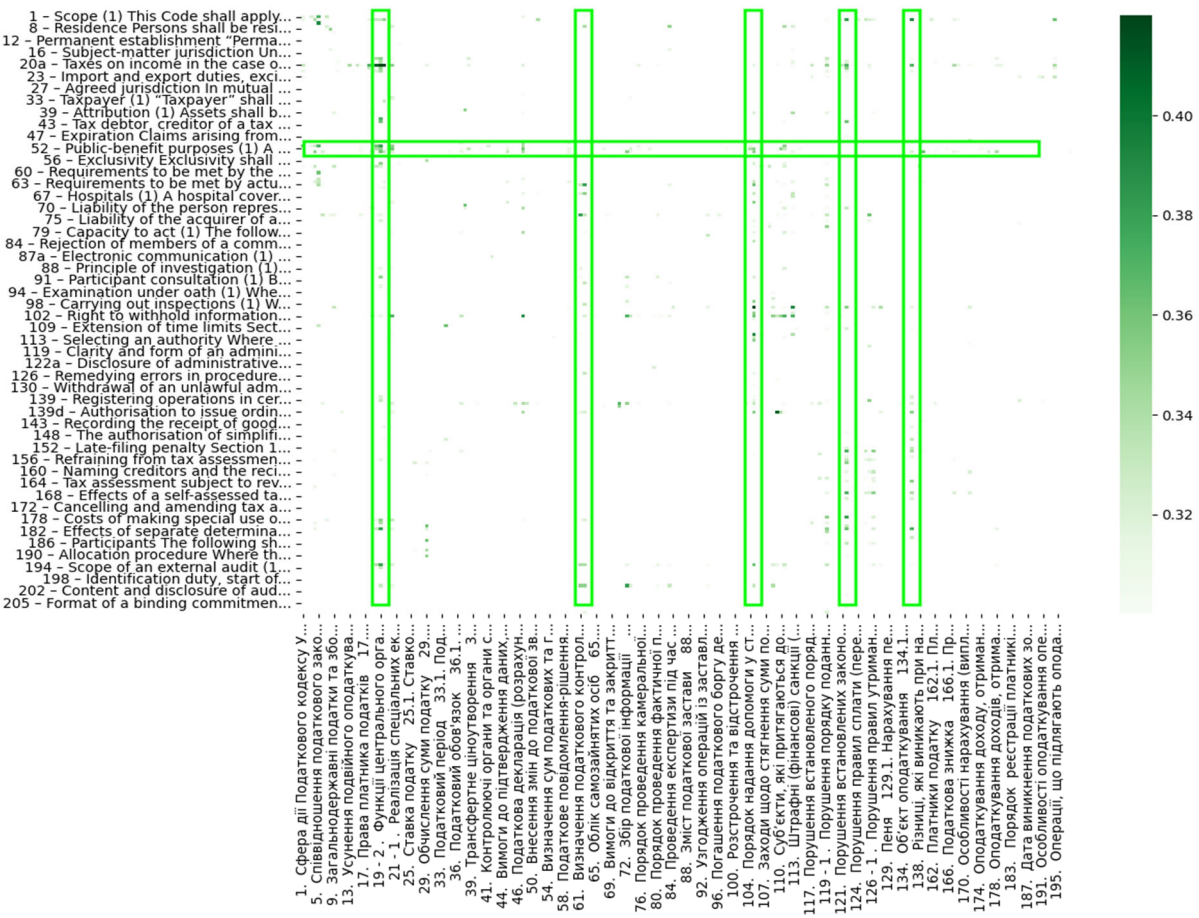

Fig. 3. Cosine similarity on the subset of data after removing references to changes in the Ukrainian and German fiscal codes

Fig. 4. A clear line pattern after cleaning up titles and references to other articles

2. Results. Finally, after applying the translation to English, increasing the correlation threshold

```
significant = df[corr>=.70]
```

and repeating the analysis and visualization, we've seen the picture without line patterns as shown on Fig.5.

To manually asses the results, we're dropping rows and columns that have correlation below our threshold

```
table = significant.dropna(how='all').dropna(axis='columns', how='all').fillna('')
```

and as shown in Table 1. we can see a clear relation of articles, for example, "Local taxes include ..." of the Ukrainian fiscal code correspond very well to the German "Non-personal taxes Local jurisdiction ..." article.

Table 1

Correlation table between articles of the Ukrainian and German fiscal codes

	6 When setting the tax, the followin...	9 Local taxes include:\n \n property t...	11 The Verkhovna Rada of Ukraine esta...	27 The unit of measurement of the tax...	31 Absolute (specific) is the tax rat...	33 Tax relief - provided by the tax a...	40 The duty of the taxpayer to calcul...	42 Fulfillment of the tax obligation ...
1 – Taxes, ancillary tax payments\n "T...						0.72		
16 – Taxes on income in the case of c...								
18 – Non-personal taxes\n Local jurisdi...		0.7	0.7			0.7		
28 – Taxpayer\n "Taxpayer" shall mean a...							0.73	
30 – Claims arising from the tax debto...							0.71	
36 – Tax debtor, creditor of a tax reb...								

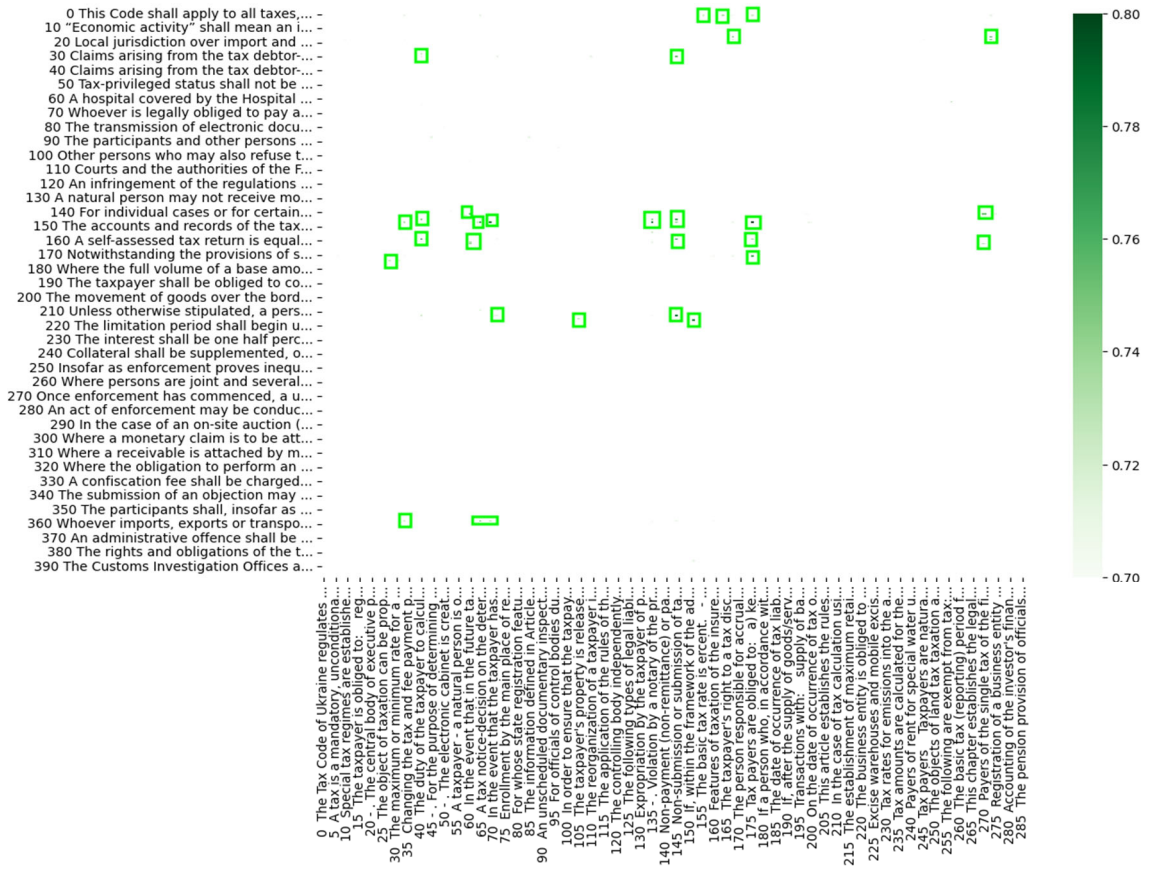

Fig. 5. No line patterns, only individual dots after translating both fiscal codes to English

Collecting all the steps we can summarize our method as following:

1. Analyze document formats and structures. Select appropriate method to convert it to data frames – for example, html and pdf processing
2. Decide what is considered a unit in every document – single article of the tax code in our case
3. Choose embeddings and metric, test it on samples from both documents – for example, Universal Sentence Embeddings with cosine metric
4. Analyze and visualize correlations, cleanup, iterate. Our research showed some unexpected correlations, which we were able to explain and eliminate with the further cleanup of our data
5. Finally, prepare the correlation table which shows us the similarities between the documents

While the proposed approach allows us to find the relation between the most similar articles of the two documents in the straightforward way, it can also be extended to highlight the most distant parts of the documents suggesting places where two fiscal codes differ a lot.

Discussion and conclusions

During our research, we were convinced once again of the importance of the well cleaned and prepared data (Lakhno et al., 2022; Savchenko et al., 2024; Boyko, 2021; Black, Kueper, & Williamson, 2023) – we had to parse different formats of data, replace irrelevant references, article numbers and other non-textual information. While we share our approach to data cleanup and normalization in the Data Cleanup section of this article, it might become less relevant for other types of documents and might limit the ability to replicate our results without the adaptation of the cleanup techniques.

We have also met the limitations of the available Multilingual Universal Sentence Embeddings model. While the model doesn't officially list Ukrainian support, our initial tests gave us a lot of hope that it will work without the translation. At the same time, the usage of translation opens access to a big variety of models and embeddings to evaluate similarity.

Ability to self-test and visualize an intermediate result was an important part of the research, because it allowed to iterate ideas, ensure the data cleanup was performed well, and the limitations of the selected model were addressed by introduction of the translation step.

As a result, we have offered a method of legislative documents analysis applied to the Ukrainian Fiscal Code. In the research we analyzed similarities with the German fiscal code, offering a detailed step-by-step replication guide to apply this method to the fiscal codes of other EU member states or other countries.

The analysis allows us to identify the positive aspects and shortcomings of the Fiscal Code of Ukraine and to develop expert recommendations for bringing the Code in line with European standards.

The continuation of the research initiated in this paper is the development of algorithms for standardizing or bringing to a common denominator the regulatory framework of the European Union, which is quite relevant given the processes of European integration and globalization that are currently being actively developed and implemented.

Authors' contribution: Olena Skliarenko – statement of the task, conceptualization, methodology, intelligence analysis, writing – viewing and editing; Anatolii Pashko – statement of the task, conceptualization, methodology, intelligence analysis, writing – viewing and editing; Leonid Lytvynenko – statement of the task, conceptualization, methodology, software, spelling – original draft; Yanina Kolodinska – data validation, formal analysis, spelling – original draft.

References

Basystiuk O., & Melnykova N. (2022). Multimodal speech recognition based on audio and text data. *Bulletin of Khmelnytsky National University: Technical Sciences*, 313(5), 22–25 [in Ukrainian]. [Басистюк, О., & Мельникова, Н. (2022). Мультимодальне розпізнавання мовлення на основі звукових і текстових даних]. https://www.doi.org/10.31891/2307-5732-2022-313-5-22-25

Basystiuk O., Shakhovska N., Bilynska V., Syvokon O., Shamuratov O., & Kuchkovskiy V. (2021) The Developing of the System for Autimatic Audio to Text Conversion. *IT&AS'2021: Symposium on Information Technologies & Applied Sciences, March 5–6, 2021*, (pp. 1–8). Bratislava, Slovak Republic. https://ceur-ws.org/Vol-2824/paper1.pdf

Black, J. E., Kueper, J. K., Williamson, T. S. (2023). An introduction to machine learning for classification and prediction. *Family Practice*, 40(1), 200–204.

Boyko, N. (2021). Research into machine learning algorithms for the construction of mathematical models of multimodal data classification problems. *Computational Problems of Electrical Engineering*. 11, 1-11.

Boyko, N., Mochurad, L., Parpan, U., & Basystiuk, O. (2019). Usage of Machine-based Translation Methods for Analyzing Open Data in Legal Cases. In *Proc. of the Intl Workshop on Cyber Hygiene (CybHyg-2019) co-located with 1st International Conference on Cyber Hygiene and Conflict Management in Global Information Networks* (p. 328–338). Kyiv, Ukraine. https://ceur-ws.org/Vol-2654/paper26.pdf

Golomoziy, V., Mishura, Yu., Izarova, I., & Ianevych, T. (2023). Processing Big Data of Court Decisions. *Baltic Journal of Modern Computing*, 11. https://doi.org/10.22364/bjmc.2023.11.4.04

Ianevych, T., Golomoziy, V., Mishura, Yu., & Izarova, I. (2023). Comparison of 2D convolutions and dense neural networks for natural language processing models with multi-sentence input. *Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics*, 20–29. https://doi.org/10.17721/1812-5409.2023/2.3

Lakhno, V. A., Kasatkin, D. Y., Skliarenko, O. V., & Kolodinska, Y. O. (2022). Modeling and Optimization of Discrete Evolutionary Systems of İnformation Security Management in a Random Environment. *Machine Learning and Autonomous Systems. Smart Innovation, Systems and Technologies,* 269, 9–22. Springer, Singapore. https://doi.org/10.1007/978-981-16-7996-4_2

Peng, Li, Xi, Rao, Jennifer, Blase et al. (2021). *CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks.* arXiv. https://doi.org/10.48550/arXiv.1904.09483

Savchenko, Y., Yarovyi, R., Kolodinska, Y., Levchenko, S. (2024). The role of artificial intelligence and machine learning under martial law. *Measuring and computing devices in technological processes,* 2, 213–216. [in Ukrainian] https://doi.org/10.31891/2219-9365-2024-78-25

Shakhovska, N., Basystiuk, O., Shakhovska, K. (2019). Development of the speech-to-text chatbot interface based on Google API. *CEUR Workshop Proceedings,* vol. 2386, 212–221.

Yang, Y., Cer, D., Ahmad, A. et al. (2019). *Multilingual Universal Sentence Encoder for Semantic Retrieval*. arXiv. https://doi.org/10.48550/arXiv.1907.04307

Отримано редакцією журналу / Received: 05.09.24
Прорецензовано / Revised: 27.11.24
Схвалено до друку / Accepted: 07.03.25

Олена СКЛЯРЕНКО[1], канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0001-6555-1223
e-mail: olena.skliarenko@e-u.edu.ua

Анатолій ПАШКО[2], д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0001-6944-8477
e-mail: anatoliipashko@knu.ua

Леонід ЛИТВИНЕНКО[1], канд. техн. наук
ORCID ID: 0000-0002-0828-383X
e-mail: leonid.lytvynenko@e-u.edu.ua

Яніна КОЛОДІНСЬКА[1], ст. викл.
ORCID ID: 0000-0002-3330-7565
e-mail: yanina.kolodinska@e-u.edu.ua

[1]Приватний вищий навчальний заклад "Європейський університет", Київ, Україна
[2]Київський національний університет імені Тараса Шевченка, Київ, Україна

МОДЕЛІ ГЛИБОКОГО НАВЧАННЯ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧІ ПОРІВНЯННЯ НЕСТРУКТУРОВАНОЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Присвячено застосуванню сучасних методів глибокого навчання, таких як універсальні багатомовні векторні представлення речень, для аналізу та порівняння схожості змісту неструктурованих текстових документів. Розроблено алгоритм визначення подібності багатомовних текстів за допомогою глибоких нейронних мереж. Авторами запропоновано безсупервізійний метод для швидкого аналізу та порівняння складних документів подібної тематики, написаних у різних країнах та різними мовами. Представлена методологія передбачає попереднє оброблення документів, їхнє переформатування для порівняння та застосування методів глибокого навчання для виявлення текстової схожості. Алгоритм відпрацьовано на прикладі здійснення порівняльного аналізу схожості за змістом текстів Податкового кодексу України з німецьким фіскальним кодексом. Український фіскальний код, поданий у форматі HTML, було оброблено за допомогою бібліотеки аналізу HTML документів і регулярних виразів, тоді як німецький фіскальний код у форматі PDF потребував спеціального аналізу неструктурованого змісту. У процесі дослідження для підвищення точності й подолання виявлених обмежень універсальних багатомовних векторних представленнь український документ було перекладено англійською мовою за допомогою Google Translate API. Розв'язання поставленої задачі потребує глибокого попереднього оброблення даних, що було здійснено у пропонованому дослідженні. Наведено також покроковий алгоритм представленої методології, що приводить до формування кореляційної таблиці, за якою визначається схожість між подібними різномовними текстовими документами. Крім того, автори доповнюють кожен крок своїми міркуваннями про те, як інтерпретувати проміжні результати та розуміти, чи варто коригувати метод відповідно до потреб або проблем конкретних документів, як-от застосування машинного перекладу чи розширеного попереднього оброблення.

Авторами запропоновано алгоритм для застосування зазначеного методу до фіскальних кодексів країн – членів Європейського Союзу або інших країн. У дослідженні виявлені й описані обмеження цього підходу, а також продемонстровано проведення перевірки проміжних результатів та здійснено доповнення методу додатковими можливостями задля підвищення його надійності. Крім цього, надано рекомендації щодо можливостей застосування запропонованої методології, зокрема розроблення алгоритмів для стандартизації певних документів країн Європейського Союзу.

Ключові слова: ***глибоке навчання, універсальні векторні представлення речень, фінансові документи, фіскальний код, регулярні вирази.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.