Optimizing Text Clustering Efficiency through Flexible Latent Dirichlet Allocation Method: Exploring the Impact of Data Features and Threshold Modification

Erzsébet Tóth, and Zoltan Gal

***Abstract*—A parallel corpus comprising Croatian EU legislative documents automatically translated into English spans 28 years and is enriched with metadata, including creation year and hierarchical classifier tags denoting descriptors, document types, and fields. However, nearly two-thirds of the approximately 1.5 thousand texts lack complete metadata, necessitating labor- intensive manual efforts that pose challenges for human administration. This incompleteness issue can be observed in the case of official legal sites functioning as regular service provisioning databases. In response, this paper introduces an artificial cognitive and multilabel classification approach to expedite the tagging process with only a fraction of the manual effort. Leveraging the Latent Dirichlet Allocation (LDA) algorithm, our method assigns field values or tags to incompletely labeled documents. We implement a Flexible LDA variant, incorporating the influence of topics close to the most probable topic, regulated by a relative probability threshold (RPT). We evaluate the LDA prediction's dependence on document prefiltering and RPT values. Furthermore, we investigate the dependence of quantitative linguistic properties on the type and speciality of pre-processing tasks. Our algorithm, built on error-correcting optimizing codes, successfully predicts a mixture of topic probabilities for these legal texts. This prediction is achieved by calculating the Hamming distance of binary feature vectors created using the legal fields of the EUROVOC multilingual thesaurus.**

***Index Terms*—Multilabel classification, Legal text clustering, Latent Dirichlet Allocation, Supervised learning, Artificial Intelligence, Natural Language Processing, Quantitative linguistics.**

I. Introduction

OUR research paper focuses on the cluster analysis of the Croatian and English parallel legal corpus included in the MARCELL (Multilingual Resources for CEF.AT in the Legal Domain) corpus. Previously the MARCELL corpus and its related resources were created in the MARCELL CEF (Connecting Europe Facility) Telecom Action. The CEF Telecom project Multilingual Resources for CEF.AT in the Legal Domain (MARCELL) intends to improve the eTranslation system implemented by the European Commission by providing seven large-scale corpora comprising national legislative documents effective in Poland, Bulgaria, Croatia, Romania, Hungary, Slovakia, and Slovenia [1].

We think that our paper has a close connection with the cognitive infocommunications (CogInfoCom) interdisciplinary research field because one of its main goals is to support the effective interaction between computers and humans and extend human cognitive capabilities with the help of infocommunications devices such an example can be a high-level artificial cognitive capability of a neural network used for text clustering. In addition, CogInfoCom's objective is to provide a systematic view of the co-evolution of human cognitive processes and infocommunication devices [23-25].

Recently there has been a growing interest in quantitative linguistic laws [2] and artificial intelligence used in text clustering. Rijsbergen emphasizes Luhn's work in automatic text analysis which assumed that word frequencies could be used for extracting words and sentences to articulate the content of a document. Let us consider f the word frequency in a particular position of the text and r the rank order of the words (i.e. the order of their frequency of occurrence), then a plot f versus r yields a hyperbolic curve. Besides this, it is a curve illustrating Zipf's law (studied intensively by Stephanie Evert in Nürnberg [21-22]) which declares that the product of word intensity values and their rank order is around constant. Luhn used this law as a null hypothesis to determine the upper and lower cut-offs of the rank order of words. He interpreted the significance of the words as their ability to express the topic of the text. With the help of these arbitrarily specified cut-offs, he omitted insignificant words such as rare and common words from the rank order of the items and in this way, he could identify those important words which describe the content or topic of the text [3].

Highlights of the paper are the following: a) Supervised learning based on the Latent Dirichlet Allocation (LDA) method is applied to 1119 legal EU English texts to classify 392 legal texts without having field attributes; b) The used artificial cognitive and multilabel classification was able to assign fields to the 392 EU legal texts successfully; c) It is proved that the LDA is less sensitive to the cleaning state of the analysed texts. The paper unfolds as follows: Section two provides an overview of related work on the Croatian-English legal corpus. In section three, we detail the applied methodology, delve into data

E. Tóth is with Faculty of Informatics, University of Debrecen, Debrecen, Hungary (e-mail: toth.erzsebet@inf.unideb.hu).

Z. Gal is with Faculty of Informatics, University of Debrecen, Debrecen, Hungary (e-mail: gal.zoltan@inf.unideb.hu)

DOI: 10.36244/ICJ.2024.5.7

processing elements, and engage in a discussion of results using quantitative linguistics approaches. Section four encapsulates conclusions drawn from our findings and outlines potential avenues for the continuation of this research work.

II. Related Works

Numerous scientific papers address challenges in automatically classifying texts based on the inherent features of document context. This section aims to bridge theoretical considerations with practical applications, offering key insights into context interpretation within the legal domain. We address current issues by presenting a non-exhaustive list of notable projects and analysis mechanisms in the legal field.

During the period from 2018 to 2020, the MARCELL project pursued the goal of providing fresh monolingual training material for CEF.AT Neural Machine Translation Services facilitated by the European Commission. This endeavor led to the organization of the introductory workshop for the CEF-project EU Council Presidency Translator in Zagreb in 2020. The workshop discussed the initial outcomes of the Machine Translation (MT) tool developed for English-Croatian and Croatian-English directions by the University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb (Croatia), and Tilde, Riga (Latvia). The participants included professionals in translation and communication from diverse Croatian industry sectors and public authorities. The event also attracted Croatian translators from various EU bodies. Marko Tadić, a member of the MARCELL project and contributor to the development of the MT system, highlighted MARCELL as a significant source of freely available language resources for training similar MT systems used by CEF.AT users and translators [4].

The EU Council Presidency Translator toolkit, an MT service developed for the 2020 Croatia's Presidency of the Council of the European Union, operates as a multilingual communication tool, facilitating instant translation of texts, documents, and websites between Croatian and English. Utilizing neural networks enhanced with artificial intelligence and machine learning, this toolkit is accessible online or through SDL Trados Studio with a plug-in, powered by the CEF eTranslation platform [5][6].

In the realm of multilabel classification, a noteworthy contribution is found in the paper [18]. The authors propose a novel method enhancing multilabel classifier performance by considering label correlations. The paper provides a comprehensive description of the classifier chains method, comparing it with various multilabel classification methods across diverse datasets. Experimental results demonstrate the superiority of the classifier chains method in multiple metrics, including precision, recall, and F1-score.

Another paper introduces Label-Specific Feature Learning (LSFL), a method involving the acquisition of label-specific feature vectors capturing the characteristics of each label [19]. These feature vectors are integrated with input feature vectors, forming a new representation for each instance, subsequently used to train a multilabel classifier. LSFL, based on regularized matrix factorization, optimizes the input feature matrix and label-specific feature matrix, preventing overfitting and encouraging sparsity. The LSFL method, as reported, outperforms several state-of-the-art multilabel classification methods across various evaluation metrics.

A widely-cited paper reviews diverse approaches to multilabel learning, encompassing problem transformation methods, algorithm adaptation methods, and ensemble methods [20]. The authors discuss evaluation measures, including precision, recall, F1-score, and Hamming loss, presenting experimental results on benchmark datasets to compare algorithm performance based on these measures.

III. Applied Methodology, Data Processing and Discussion

In the next subsections, we describe the input data set applied in the multilabel classification process [7][8]. Topic discovery with the Latent Dirichlet Allocation method will be detailed in [9][10][11]. Interpretation of the results will be based on quantitative linguistic approaches [12][13].

A. Description of the Data

Upon Croatia's accession to the EU in 2013, the country initiated the translation of national legislative documents between Croatian and English, resulting in the creation of the Croatian-English Parallel Corpus of Croatian National Legislation spanning texts from 1990 to 2019, totaling approximately 1,800 documents. Notably, earlier Croatian legal documents were monolingual, and the English translations commenced in PDF format in the late 1990s.

However, the extraction of text from various PDF files posed challenges, impacting the quality of the automatic extraction process. To address this, the researchers employed sentence splitting and alignment using LF-aligner [14], an open-source tool utilizing HunAlign in the background [15]. To ensure high-quality alignment, proofreading was manually conducted for all 1,816 documents, resulting in accurately aligned TMX files. This meticulously curated parallel corpus serves as a valuable resource for the noiseless training of Neural Machine Translation (NMT) systems. The corpus encompasses 396,984 token units, with 14.4 million and 17.7 million tokens in Croatian and English, respectively.

In general, the 1816 source documents collected from 1991 to 2018 (28 years) have a header with a specific structure and a body with Croatian and corresponding English split sentences. The header consisted of the document type (noted T attribute), year of creation (noted Y attribute), EUROVOC descriptors (noted D attribute) and field (noted F attribute) elements. The number of unique entities of the document type, year, descriptor and field elements is $|T| = 11, |Y| = 28, |D| = 1393$ and |F| = 22, where $|X|$ represents the cardinality of the document subset with attribute $X \in \{T, Y, D, F\}$. In this data set, there are 1585 documents having type and year in the header (noted TY attributes). A subset of 1511 documents has a type, year in the header and sentences in the body (noted $TY{:}S$ attributes). Only 1119 documents have type, year, field in the header and sentences in the body (noted $TYF{:}S$ attributes). In this way, the number of documents without field is $1511 - 1119 = 392$, called a set of unlabelled documents [16]. In our analysis, the document ID range is 1…1511 in chronological order.

TABLE I.
LIST OF DOCUMENT TYPES (T).

ID	Type	ID	Type
1	Constitution	7	Order
2	Constitutional act	8	Ordinance
3	Decision	9	Other
4	Instructions	10	Regulation
5	Law	11	Standing orders/rules of procedure
6	Legal code		

Table 1 provides a list of distinct document types, aligning with the CELEX identification system of the EU. The CELEX database stands out for its expansive coverage and advanced search functionalities, making it an invaluable resource for legal research and analysis. Users can conduct searches based on document type, date, subject matter, and keywords, as well as utilize specific CELEX numbers assigned to each document.

Table 2 summarizes the unique fields within the legal corpus, with each field serving as a label derived from the EUROVOC multilingual, hierarchical thesaurus developed by the Publications Office of the EU. The collection comprises a total of 22 fields, representing various subject areas. This diversity enables users to navigate and search for terms and concepts pertinent to specific domains within different official information systems.

TABLE II.
LIST OF DOCUMENT FIELDS (F).

ID	Field
1	Agriculture, forestry, fishery
2	Communication and information technology
3	Construction and city planning
4	Culture and cultural property
5	Defence, internal affairs and national security
6	Economy, trade and commerce
7	Education and sports
8	Energy production
9	Environment and natural heritage
10	Finance, budget and monetary affairs
11	Health care
12	Industry and technology
13	Information, media, documentation, statistic
14	International relations and cooperation
15	Labour, employment and pension scheme
16	Law and the judiciary
17	Politics and public authority
18	Science and research
19	Social activities and human rights
20	Social care
21	Tourism and tourist activities
22	Traffic and traffic infrastructure

The distribution of the 1511 documents with type, year and sentences is represented on the left side of Fig. 1. according to the analysed years. The histogram of the number of documents vs. years is illustrated on the right side of Fig. 1.

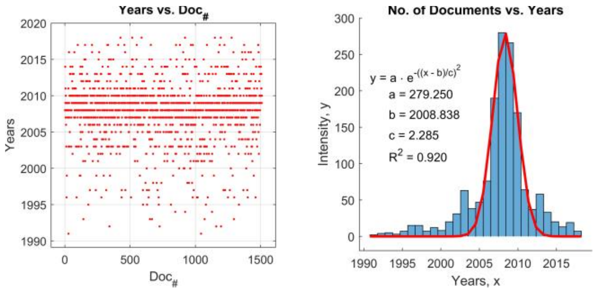

Fig. 1. Years vs. DocID (left); No. of documents vs. years (right)

The largest number of documents was created around the year 2008. The fitting curve of the intensity versus years has the following equation:

$$Intensity = a \cdot e^{-[(year - b)/c]^2} \quad (1)$$

Values of the parameter triplet (a, b, c) are $(279.25, 2008.84, 2.28)$ with 92% coefficient of determination. The independent variable takes values in the following interval: $year \in \{1991, \dots, 2018\}$.

In Fig. 2, the left side presents an overview of the potential number of descriptors per document yearly, while the right side illustrates the distribution of descriptors per document across the years. Notably, the number of descriptors per labeled document remains consistently below a dozen for the examined years. However, when considering both labeled and unlabeled documents, the average number of descriptors is less than four. This difference arises due to a substantial portion of documents in the English-Croatian legal corpus lacking descriptors.

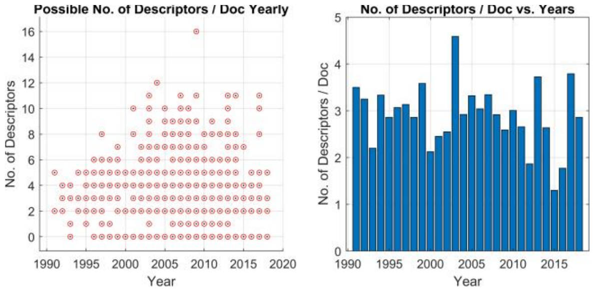

Fig. 2. Possible No. of descriptors/doc yearly (left);
No. of descriptors/doc vs years (right).

Note that no more than 16 descriptors were assigned to any of the documents belonging to the 28 years of the survey (see Fig. 3 left). Most of the documents, ~25% have just one descriptor and ~70% of them have between 3 and 7 descriptors (see Fig. 3 right).

Because the number of descriptors is very high ($|D| = 1393$), this attribute was not considered in this multilabel analysis [17]. In contrast to descriptors, the number of fields is just 22 creating the possibility to use them as a multilabel binary vector of 22 dimensions.

We define a topic of the document by the field pattern of its feature vector. The majority of documents exhibit one of two fields, but a significant portion of legal texts lacks fields altogether (refer to Fig. 4, left).

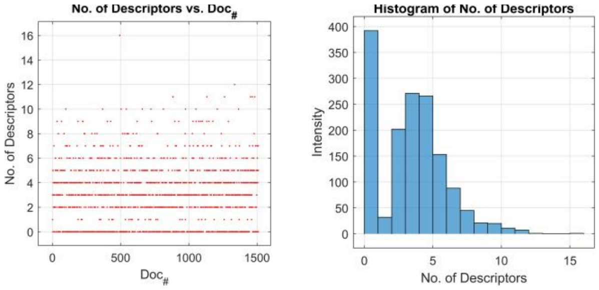

Fig. 3. No. of descriptors vs. docID (left);
Histogram of No. of descriptors (right).

Overall, the average number of fields per document is predominantly less than one during the examined years (refer to Fig. 4, right).

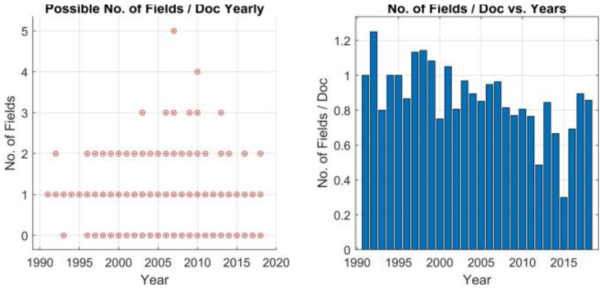

Fig. 4. Possible No. of fields/doc yearly (left);
No. of fields/doc vs. years (right).

This reflects that the EUROVOC thesaurus was not exploited sufficiently for assigning labels to these documents. A maximum of five fields are assigned to the investigated texts and the majority (~66%) of them have just one field. The number of unlabelled items is 392 (see Fig. 5).

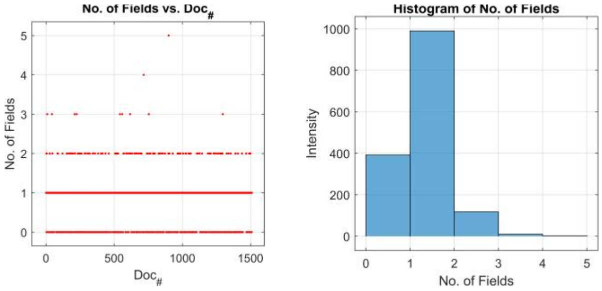

Fig. 5. No of fields vs. docID (left);
Histogram of No. of fields (right).

We observed that most of the items have around 100 sentences and the longest item has less than 5000 sentences (see Fig. 6 left). The histogram of the sentences conforms to an exponential function:

$$Intensity \;=\; a \cdot e^{-b \cdot year} \tag{2}$$

Values of the parameter triplet (a, b) are $(1316.183, -0.007)$ with 99% coefficient of determination. The independent variable takes values in the following interval: $year \in \{1991, \dots, 2018\}$ (see Fig. 6 right).

There was a high number of sentences created from 2007 to 2012 (see Fig. 7 left). The largest item has less than 500 and the shortest one has at least 30 sentences, respectively. Note that from 2007 to 2021 a higher amount of documents was created resulting in a lower number of sentences per document (see Fig. 7 right).

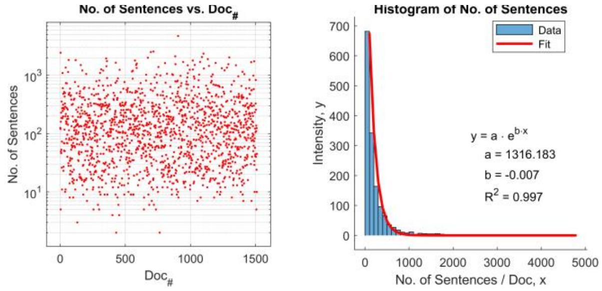

Fig. 6. No of sentences vs. docID (left);
Histogram of No. of sentences (right).

The longest documents were created in the years of 1995 and 2013.

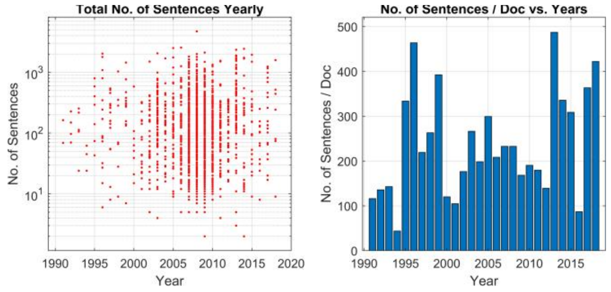

Fig. 7. Total No of sentences yearly (left);
No. of sentences/doc vs. years (right).

The histogram of the sorted unique descriptors follows an exponential equation:

$$Intensity \;=\; a \cdot e^{-b \cdot x^{c}} \tag{3}$$

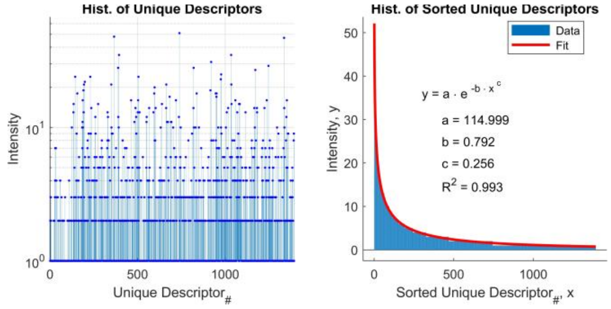

Fig. 8. Histogram of unique doc types (left);
Histogram of sorted unique descriptors (right).

Values of the parameter triplet (a, b, c) are $(114.99, 0.79, 0.26)$ with 99% coefficient of determination. The intensity of the unique descriptors covers the range of 1 and 50 and they have specific occurrences in the corpus (see Fig. 8 left). The sorted unique descriptor (independent variable), x takes values in the following interval: $x \in \{1, \dots, 1393\}$ (see Fig. 8 right).

Most of the documents conform to type 8 (Ordinance), followed by type 5 (Law) and type 3 (Decision) of Table I. (see Fig. 9 left). The first two most frequent fields of the documents listed in Table II. belong to the "10: Finance, budget and monetary affairs" and "1: Agriculture, forestry, fishery" domains having ratio of approximately 30% and 20%, respectively (see Fig. 9 right).

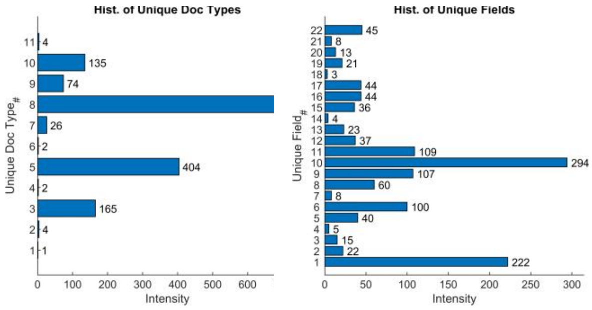

Fig. 9. Histogram of unique doc types (left);
Histogram of unique fields (right).

The smallest intensity of the fields was “18: Science and research”, “14: International relations and cooperation” and “4: Culture and cultural property”. In the next section, we highlight the main aspects of the LDA method and its application in our data set.

B. Topic Discovery with Latent Dirichlet Allocation Algorithm

LDA operates as a hierarchical Bayesian model featuring three levels. The items within the collection set constitute a blend of topic probabilities, where each topic represents an infinite amalgamation of the basic set of topic probabilities. In the context of a document, the likelihood of a topic imparts characteristics to the text corpus. The Hamming distance, a straightforward and intuitive distance metric, finds utility in various applications, including clustering, classification, and information retrieval. It serves to compare document similarity based on labels, contexts, or themes.

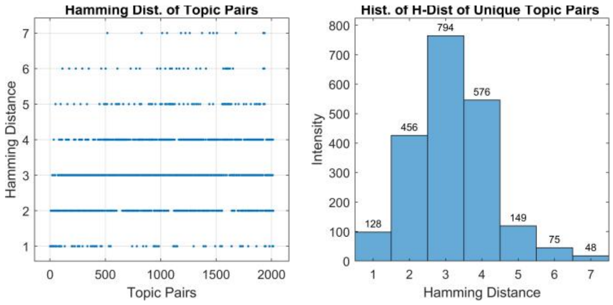

Fig. 10. Hamming distance of topic pairs (left);
Histogram of Hamming distance of unique topic pairs (right).

Our approach involved utilizing the Hamming distance between topic pairs and their corresponding histogram to assess the dissimilarity of label assignments across the texts (refer to Fig. 10). The maximum distance observed between topics is seven, with the majority of topic pairs positioned at a distance of three. The overall distribution converges towards a Gaussian bell shape (refer to Fig. 10, right).

Feature binary vectors of two sets of documents with cardinality 30 are represented in Fig. 11, respectively. The square means bit one in the corresponding field position of the vector. Document IDs increase from bottom to top.

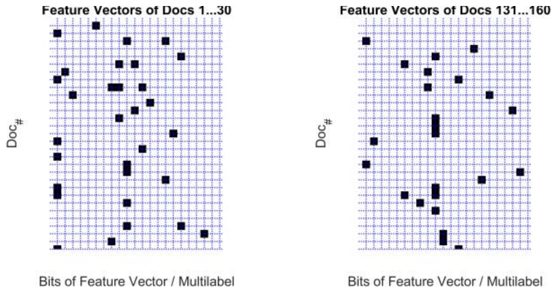

Fig. 11. Feature vectors of docs 1...30 (left);
Feature vectors of docs 131...160 (right).

The horizontal pattern refers to the binary feature vector representing the topic details of a specific document. It's worth noting that a small number of documents in both sets depicted in Fig. 11 lack a pattern, indicating the absence of fields in these texts. Our objective was to predict the missing fields in these texts by leveraging the meaning of sentences through the application of the LDA method.

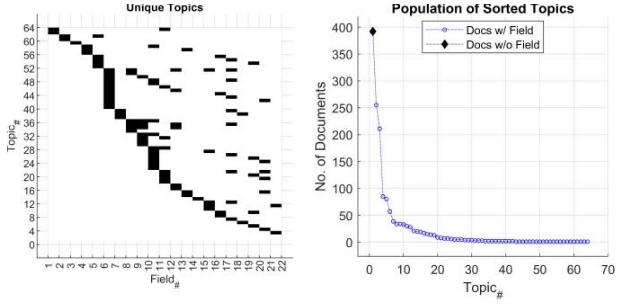

Fig. 12. Unique topics of the document corpus (left);
Population of the sorted topics (right).

Within the analysed corpus, feature vectors exhibit distinct bit patterns corresponding to a total of 64 topics (refer to Fig. 12). A binary vector featuring homogeneous 0-s is termed a pseudo-topic, serving to represent documents devoid of fields. The 392 documents associated with pseudo-topics are illustrated in the right figure using a diamond shape.

Fig. 13 (left) showcases a toplist of tokens derived from both original and cleaned texts on a log-log scale. Notably, both intensities exhibit well-approximated straight lines, suggesting power functions in the linear scale. The scatterplot of the intensity of the cleaned and original tokens shows linear dependencies in two intervals (see Fig. 13 right).

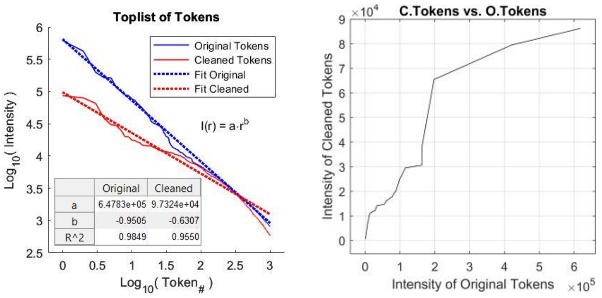

Fig. 13. Toplist of tokens (left);
Cleaned Tokens vs. Original Tokens (right)

This indicates that the cleaning ratio of the tokens is constant for the majority of the large documents. The slope of the original tokens is close to -1, corresponding to Zipf's law. However, the tokens of the cleaned texts have an exponent of -0.63 proving unconformity to Zipf's law. A possible explanation for this unconformity may be due to the cleaning task executed on the corpus. The cleaning process of the texts and usage of working objects are listed in Table III.

Tasks executed in each of the cleaning steps are as follows. 1: Convert string into tokens; 2: Add part of speech details; 3: Details of Original Tokens; 4: Reduce each token to stem; 5: Remove stopwords; 6: Remove Short (less than 3 characters) and Long Words (greater than 15 characters); 7: Details of Stem Tokens; 8: Create a bag-of-words; 9: Create a bag-of-words of cleaned docs; 10: Create TopBag of Original Tokens; 11: Create TopBag of Cleaned Tokens. Variable noTokens has a value of 1000 because we consider only the first 1000 most frequent tokens as significant.

TABLE III.
LIST OF DATA PROCESSING STEPS.

StepID	Function
1	docs = TokenizedDocument(texData)
2	docs1 = AddPartOfSpeechDetails(docs)
3	tokenDetailsOrig = TokenDetails(docs1)
4	docs2 = NormalizeWords(docs1)
5	docs3 = RemoveStopWords(docs2)
6	docsClean = RemoveShortWords(docs3, 2, 15)
7	tokenDetailsStem = TokenDetails(docsClean)
8	bag = BagOfWords(docs)
9	bagClean = BagOfWords(docsClean)
10	topBag = Topkwords(bag, noTokens)
11	topBagClean = topkwords(bagClean, noTokens)

Applying LDA for topic prediction on labeled documents yields a list of potential topics quantified by their respective probabilities of belonging. Word clouds are subsequently generated by arranging tokens in descending order of occurrence within the text. These tokens represent reduced stems of words (such as "servic," "articl," "manufactur," "measure"), with some exceptions (like "electron," "croatian," "conform," "limit").

Prediction examples of four selected texts (doc ID = 25, 26, 33, 34) in decreasing order of the belonging probability are shown in Fig. 14. right side.

We observed that some documents are unique, but others have a few dominant probabilities.

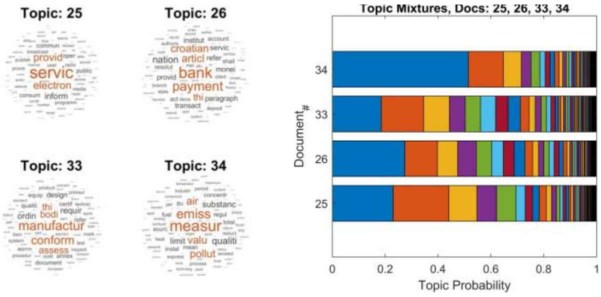

Fig. 14. Sample of topic word clouds of a text quartet (docs 25, 26, 33, 34) (left); Topic mixtures of the text quartet (right).

The latter documents weaken the goodness of the topic prediction.

C. *Impact of the similarity probabilities on the LDA*

To consider significant topics not just the absolute first candidate we introduced a parameter called topic Relative Probability Threshold (selection of significance threshold), $RPTh$. This parameter is used to binary classify topics into significant and non-significant groups of the normalized topic probabilities. Note that when $RPTh = 100\%$ we have the classical LDA algorithm with only one solution, which means that each of the selected four texts has only one dominant topic.

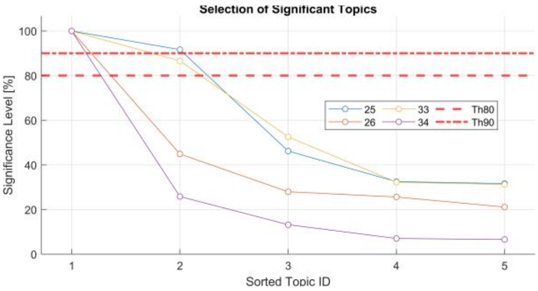

Fig. 15. Selection of significant topics of the text quartet (docs 25, 26, 33, 34).

In our case, this means selection of Topic34, Topic48, Topic2 and Topic 62 for the texts 25, 26, 33 and 34, respectively. For $RPTh \in (0,1)$ the extended LDA may have more than one topic proposal for close probability values. For $RPTh = 0$, the LDA gives all 63 combinations as a proposal of the multilabel classification problem, implying non-usability in practice.

In the case of the significance threshold of the LDA having value $RPTh80 = 80\%$, the estimation of the text quartet is shown in Table IV. Two of the documents (25, 33) have more than one estimation and others are identified by one estimation (see Fig. 15). Note that in this randomly selected text quartet, the first and second topic estimation for the same document differs strongly. The first estimation for text 33 is Topic2 ("Traffic and traffic infrastructure"). The second estimation is Topic43 ("Economy, trade and commerce", "Politics and public authority") which is quite different from Topic2. If the significance threshold value is set to $RPTh90 = 90\%$, just one text (25) has more than one multilabel estimation and text 33 gets only one dominant topic (multilabel2). This behaviour of the modified LDA mechanism proves the strong dependence of the estimation decision on the value of the significance threshold.

TABLE IV.
LIST OF SIGNIFICANT TOPICS OF THE TEXT QUARTET (DOCS 25, 26, 33, 34), THRESHOLD = 80%.

Doc ID	Significant Topic IDs
25	1st estimation : Topic34 (Field8, Field9, Field12) 2nd estimation: Topic61 (Field2, Field5)
26	1st estimation: Topic48 (Field6, Field10, Field17)
33	1st estimation: Topic2 (Field22) 2nd estimation: Topic43 (Field6, Field17)
34	1st estimation: Topic62 (Field1)

This dependence requires other deeper studies in this area. We note that if $RPTh80 \in (53, 86)$ then the text quartet estimation is the same. This indicates large intervals to have the same topic prediction of the text.

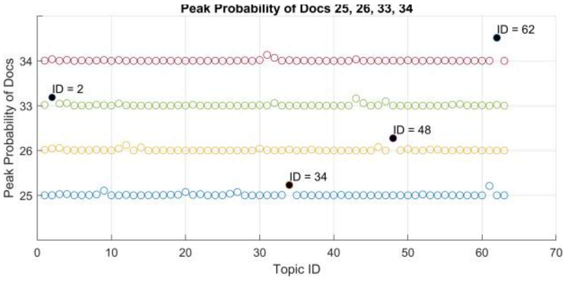

Fig. 16. Topic probability of the text quartet, Rigid LDA (docs 25, 26, 33, 34). The index of the peak is the ID of the topic.

For threshold $RPTh100$ only the largest probability counts in the decision. We name this classical case Rigid LDA (R-LDA). When the threshold is less than 100% we call it Flexible LDA (F-LDA) and the impact of it is explained in subsection E of this paper (see Fig. 16).

D. Impact of the text preprocessing on the LDA

We applied the algorithm to both the original and cleaned versions of the texts to assess the performance of the LDA method. Successfully, all 392 documents lacking fields were categorized into topics, with one or more fields being attached to each text.

Fig. 17 illustrates the topic assignment for the 392 documents in question. Each circle in the figure corresponds precisely to one topic assigned to the respective document. The chosen topic represents a list of individual fields that aligns with the feature vector patterns shown in Fig. 12 on the left. We observed a similar pattern in the allocation of predicted topics to documents categorized under pseudo-topics in both cleaned and uncleaned cases (refer to Fig. 17, left and right).

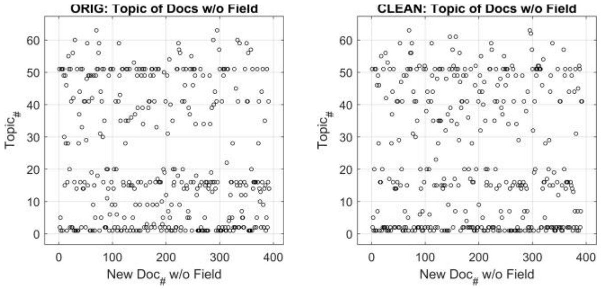

Fig. 17. Topics of Docs w/o Field (left);
Topics of Clean Docs w/o Field (right).

To quantify dissimilarity, the Hamming distance between topics predicted for original and cleaned documents was assessed, revealing a range of [0, 7] (see Fig. 18). Large Hamming distances (greater than 5) occur very rarely. The majority of the distances are zero with the following mean, standard deviation and skewness values: $(\mu, \sigma, \gamma) = (1.059, 1.597, 1.453)$.

Our findings indicate that LDA exhibits a reduced sensitivity to the execution or omission of the cleaning task as a pre-processing step. To validate the effectiveness of this artificial cognitive capability-based method, we conducted manual testing by randomly selecting a few unlabelled legal texts. We then compared the automatic labelling performed by the LDA method to the actual categorization decided by a human evaluator.

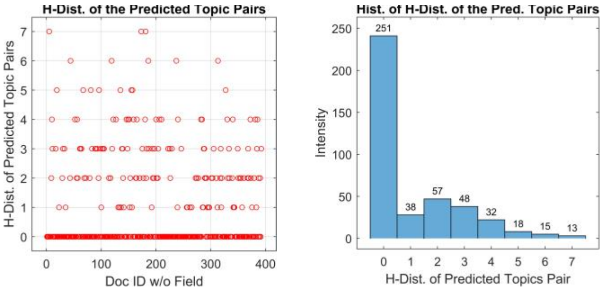

Fig. 18. Hamming distance of the topic prediction by Rigid LDA for cleaned and original documents (left);
Histogram of Hamming distance of pairs of topics (right).

Despite a small sample size of fewer than ten entities which have a large deviation in their text length, we observed a correct matching ratio. The efficiency of the method was further demonstrated by the time it took for manual label assignment by the human evaluator. The manual assignment of one label consumed approximately one hour, considering the inherently time-consuming nature of reading and interpreting the text by a human.

E. Impact of the threshold value on the LDA

We extended the classical LDA by considering not only the largest probability value of the topics, but even others having a relative value close to it determined by the relative probability threshold $RPThx$ (%), $x \in [1, 100]$.

We assume that making decisions based on the topic group of top probabilities offers a more accurate prediction of individual labels compared to relying solely on the selected unique topic given by the top 1 topic probability. This pattern of individual labels constitutes a multilabel pattern, but it may differ from the list of topics determined by the teaching data set. This property has a greater impact when the top 1 topic has a low probability and is close to other topics.

We determined the number of significant topics for each unlabelled document. Mean, standard deviation and coefficient of variation metrics of the number of the significant topics vs. threshold $RPThx$ in the case of 392 unlabelled documents are presented in Fig. 19 (left) and in Fig. 18 (right), respectively.

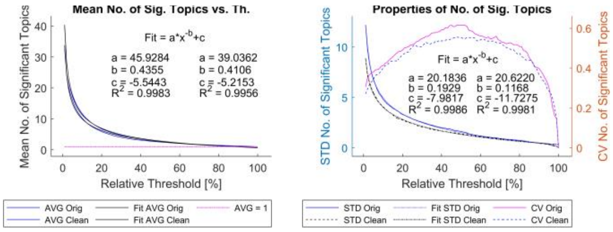

Fig. 19. Dependence of the number of significant topics vs. relative probability threshold (cleaned documents case): Mean (left); Standard deviation and coefficient of variation (right). In both figures, the first and second columns of the fit parameters (a, b, c) belong to the original and cleaned documents, respectively.

For both the original and cleaned unlabelled documents these metrics have a power-law dependence on the threshold value, demonstrating a remarkably high coefficient of determination ($R^2 > 99.5\%$). Fit parameter triplets (a, b, c) are provided in Fig. 18. Dependence of the coefficient of variation (CV) on the threshold $RPThx$ for both original and cleaned unlabelled documents is depicted in Fig. 18 on the right. We can observe that these curves are approximatively symmetric to the vertical axis at $RPTh50$, with a maximum value of 0.6. The impact of the text cleaning can be seen on the smoothness of the CV curves (refer to Fig. 18 on the right). The more the corpus is cleaned, the smoother the CV curve becomes.

IV. Conclusions

In the paper we exhibit automatic labelling methods based on multilabel analysis. The Latent Dirichlet Allocation method can be successfully applied to classify legal texts with multiple fields. Quantitative properties of the documents are influenced by the cleaning and normalizing of the texts. These pre-processing tasks will result in non-conformity to Zipf's law, because the intensity of words instead of inverse proportionality (exponent value = -1) of the rank order becomes a power function with exponent value = -0.63 in the case of the legal corpus. The LDA is less sensitive to the pre-processing of the texts by cleaning. The method in some situations offers the classification of the text to more than one similarly dominant topic. These similar topic pairs have low Hamming distance, causing the ambiguity of making labelling decisions. We implemented the Flexible LDA with enhanced properties, based on a single parameter to increase the efficacy of the labelling. More research is needed to analyse the relationship between the sensitivity of the LDA method and the state of text cleaning.

Acknowledgment

This work has been supported by the QoS-HPC-IoT Laboratory and project TKP2021-NKTA of the University of Debrecen, Hungary. Project no. TKP2021-NKTA-34 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the TKP2021-NKTA funding scheme.

References

[1] T. Váradi et al., "The MARCELL legislative corpus," in Proc. of the 12th Conference on Language Resources and Evaluation. LREC 2020. Marseille, France: European Language Resources Association, 2020, pp. 3761–3768.

[2] *Encyclopedia entry: laws in quantitative linguistics*. [Online]. Available: http://lql.uni-trier.de/ (2023-05-06)

[3] van C. J. Rijsbergen, Information Retrieval. 2nd ed. London: Butterworths, 1979. [Online]. Available: (2023-05-03) http://www.dcs.gla.ac.uk/Keith/Preface.html

[4] *Marcell Project Homegape*. [Online]. Available: (2023-02-22) https://marcell-project.eu/

[5] *EU Council Presidency Translator Workshop*. [Online]. Available: https://hr.presidencymt.eu//workshop (2023-02-23)

[6] M. Mohri, A. Rostamizadeh, and A. Talwalker, Foundations of machine learning. 2nd ed. Cambridge, MA: The MIT Press, 2018.

[7] J. Nam et al., "Large-scale multi-label text classification — revisiting neural networks," in T. Calders, F. Esposito, E. Hüllermeier, R. Meo Eds. Machine Learning and Knowledge Discovery in Databases ECML PKDD 2014, LNCS, vol. 8725, Heidelberg: Springer, 2014, pp. 437–452. DOI: 10.1007/978-3-662-44851-9-28

[8] J. Liu et al., "Deep learning for extreme multi-label text classification," in SIGIR'17: Proc. of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2017, pp. 115–124. DOI: 10.1145/3077136.3080834

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* vol. 3, no. 5, pp. 993–1022, 2003.

[10] M. D. Hoffmann, D. M. Blei, and F. Bach, "Online learning for Latent Dirichlet Allocation," in J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta Eds. Advances in Neural Information Processing Systems 2010, NIPS, vol. 23, Curran Associates, Inc., 2010, pp. 1–9.

[11] K. R. Canini, L. Shi, and T. L. Griffiths, "Online inference of topics with Latent Dirichlet Allocation," in Proc. of the International Conference on Artificial Intelligence and Statistics 2009, JMLR, vol. 5, Clearwater Beach, Florida: MLR Press, 2009, pp. 65–72.

[12] K. Johnson, Quantitative methods in linguistics. Malden, MA: Blackwell Publishing, 2008.

[13] S. Th. Gries, Quantitative corpus linguistics with R: a practical introduction. 2nd ed. New York: Routledge, 2016.

[14] *Homepage of the LF Aligner translator.* [Online]. Available: https://sourceforge.net/projects/aligner/ (2023-02-22)

[15] D. Varga et al., "Parallel corpora for medium density languages," in G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, N. Nikolov Eds. Proc. of the RANLP 2005 (Recent Advances in Natural Language Processing), Borovets, Bulgaria, 2005, pp. 590–596.

[16] S. Duan, S. Chang, and J. C. Príncipe, "Labels, information, and computation: efficient learning using sufficient labels," *Journal of Machine Learning Research* vol. 24, pp. 1–35, 2023.

[17] G. Tsoumakas, and I. Katakis, "Multi-label classification: an overview," *International Journal of Data Warehousing and Mining* vol. 3, no. 3, pp. 1–13, 2007.

[18] J. Read et al., "Classifier chains for multi-label classification," *Machine Learning* vol. 85, pp. 333–359, 2011.

[19] M.-L. Zhang and L. Wu, "Lift: Multi-Label Learning with Label- Specific Features," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 107–120, 1 Jan. 2015, DOI: 10.1109/TPAMI.2014.2339815.

[20] M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms, " in I*EEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014, DOI: 10.1109/TKDE.2013.39.

[21] S. Hartman, "Derivational morphology in flux: a case study of word-formation change in German," *Cognitive Linguistics*, vol. 29, no. 1, pp. 77–119. 2018 DOI: 10.1515/cog-2016-0146

[22] M. Baroni and Stefan Evert, "Testing the extrapolation quality of word frequency models," *Proc. of Corpus Linguistics 2005*. [Online]. Available: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/Lexiconodf/EvertBaroni2005.pdf (2023-07-12)

[23] P. Baranyi and Á. Csapó, "Definiton and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica,* vol. 9, no. 1, pp. 67–83, 2012.

[24] P. Baranyi, Á. Csapó, and G. Sallai, Cognitive infocommunications (CoginfoCom). Springer, 2015.

[25] I. Horváth et al., "Definiton, background and research perspectives behind 'Cognitive Aspects of Virtual Reality' (cVR)," *Infocommunications Journal,* Special issue: Internet of Digital & Cognitive realities, pp. 9–14, 2023. DOI: 10.36244/ICJ.2023.SI-IODCR.2.

Erzsébet Tóth is an assistant professor at the Faculty of Informatics, University of Debrecen, Hungary. She had a degree in English language and literature and library and information science in 1995. In 2008 she obtained her Ph.D degree in information science and technology. In the doctoral dissertation she investigated the evaluation of search engine performance. She is involved in a virtual library project that focuses on the presentation of digitized library content in three-dimensional space, and she studies the enhanced possibilities of the English language teaching and learning in virtual learning environment. Since 2022 she has been an IEEE member.

Zoltan Gal is currently a full professor at the Faculty of Informatics, University of Debrecen, Hungary. He earned MSc in electrical engineering and computer science from the Technical University of Timisoara, Romania and PhD in informatics sciences from the University of Debrecen. His scientific interest is focused on distributed processing and communication systems, sensor technologies and services in the Internet of Things.

He was the CIO of his institute for 20 years and developed the university-level metropolitan area highs-speed data network and services with over 10k Internet nodes. He is Cisco Certified Network Professional lecturer since 1999 and taught over five hundred network professionals in the field. Starting in 2015 he is head of the Centre of High-Performance Computing at his university. He is an IEEE member since 1996 and published over one hundred twenty scientific conferences and journal papers: He supervises his own R&D&I project called QoS-wwHPC-IoT Laboratory.