28th Meeting of the Wiesbaden Group on Business Registers

Global Legal Entity Identifier Foundation (GLEIF)
Dominik Jany – Data Scientist
Zornitsa Manolova – Head of Data Quality Management and Data Science

Session No. 6

"New data sources: Opportunity and challenges"

LENU – Using AI for legal form detection

1 Abstract

The Legal Entity Identifier (LEI) is a 20-digit alpha-numeric code based on the ISO 17442 standard. It connects to the key reference information that enables clear and unique identification of legal entities. The Global LEI Repository is the transparency island in a cloudy environment – it provides open, free-of-charge, high-quality legal entity data with global coverage.

There are more than 2.3 Mio active LEIs in the system. The reference data of these LEIs is collected, verified, and managed by a network of LEI issuing organizations across the globe. The embedded global standards in the data format and the established data quality framework ensure consistency and high-quality data among the different organizations and jurisdictions.

Using the LEI data, GLEIF and our partners from Sociovestix Labs, developed a state-of-the-art open-source AI tool. The tool, called LENU (Legal Entity Name Understanding), automatically assigns standardized Entity Legal Form (ELF) codes (ISO 20275) to entities based on legal name and legal jurisdiction only.

As of today, we utilize a wide range of traditional Machine Learning models as well as more advanced Deep Learning (transformer) models that predict the ELF code for any legal name within a given jurisdiction. This enables public and private organizations of any size to start adopting the ISO standard for legal forms by assigning the codes easily and effortlessly.

In our paper, we demonstrate the end-to-end process from analyzing the feasibility of the initial idea and testing of different algorithms from simple string matching to more sophisticated neural networks, to making available the open-source tool for the public. We will also highlight the motivation and benefits of having standardized data as part of an organization's dataset and the use cases we see for the example of standardized legal forms.

2 Introduction

The wide range of entity legal forms existing within and between different jurisdictions has made it challenging for organizations to categorize and structure this information effectively. This task becomes even more difficult due to the similarities in types and textual representation of these legal forms across jurisdictions. For instance, as one can see in the examples (a) and (b) in Table 1, different jurisdictions can have their distinct versions of a "Limited Liability Company" (LLC), where each one operates under its specific legal framework.

It is essential to note that even if multiple jurisdictions, like US-Delaware and US-New York, have LLCs, they are distinct entity forms. The complexity therefore lies in handling the diversity of entity legal forms across jurisdictions and the need to distinguish between seemingly similar legal structures.

Table 1: Examples of inconsistent legal form representations

	Legal name	Jurisdiction	Legal form	ELF code
(a)	Dean Quarry Apartments LLC	US-NY	Limited Liability Company	SDX0
(b)	RUBICON TECHNOLOGY MANAGEMENT L.L.C.	US-DE	Limited Liability Company	HZEH
(c)	LOCKWOOD RIVERFRONT HOTEL, LLC	US-DE	Limited Liability Company	HZEH
(d)	GIANT Weilerswist g21 GmbH	DE	Gesellschaft mit beschränkter Haftung	2HBR
(e)	Selbstfahrer Union G.m.b.H.	DE	Gesellschaft mit beschränkter Haftung	2HBR
(f)	Interproximal AB	SE	Aktiebolag	XJHM
(g)	Konstlist i Heby Aktiebolag	SE	Aktiebolag	XJHM
(h)	Aktiebolaget Clas Grönwalls Lantbrukstjänst ilkividation	SE	Aktiebolag	XJHM
(i)	Infrastrukturentwicklungsgesellschaft Hilden mbH	DE	Gesellschaft mit beschränkter Haftung	2HBR
(j)	Katholische Kirchengemeinde Maria Königin Lingen	DE	Körperschaft des öffentlichen Rechts	SQKS
(k)	むつ小川原風力合同会社	JP	合同会社	7QQ0
(l)	合同会社まつお	JP	合同会社	7QQ0

Furthermore, the representation of legal forms within entity names is influenced by the cultural and linguistic context of the jurisdiction and the specific legislation used to manage registrant information. Expert interviews highlighted that, for example, many entities in France are not legally required to include their legal form in their names. Furthermore, the same entity legal form may be inconsistently represented between various entities, especially across different data sources. These discrepancies often manifest in different punctuation styles, as shown in the examples (d) and (e) in Table 1. Also, inconsistent use of abbreviations as demonstrated by the entities (f), (g) and (h) may cause ambiguity: the legal form "Aktiebolag" can be written out in full at the end of the name or represented by the abbreviation "AB", and it can appear at the beginning of the entity's name as "Aktiebolaget". This issue is also present in Germany. According to the official business register of the German Federal Statistical Office in 2018, of the names of the more than 9,500 entities that are stock corporations, 13% include the term "Aktiengesellschaft" while 87% include the abbreviation "AG". Similar figures can be observed for the almost 700,000 private limited companies (Gesellschaft mit beschränkter Haftung, GmbH) [6].

Inconsistencies are further introduced by variations in the capitalization of individual characters or the entire name. Many business registers default to representing reference data in uppercase letters, which adds to the complexity, as shown in examples (b) and (c) in Table 1. Also, the intermingling of legal form and name elements as in example (i), the absence of any mention of the legal form like in example (j), or the use of non-Latin characters in examples (k) and (l), exacerbates the challenge of accurately detecting legal forms, in particular if data users lack local domain expertise.

In summary, the significant variability in representation not only presents considerable obstacles for any automated identification approach but also for anyone without a minimum amount of domain knowledge.

In recent years, with the rise of deep learning, and in particular with the Transformer model [13], a new class of neural network language models has surpassed traditional machine learning-based approaches in numerous text classification benchmarks [10]. The following chapters will show the strengths of transformer models in automated legal form detection.

3 Methodology

In this paper, we explore the application of machine learning and deep learning techniques to accurately classify entity legal forms. Our focus lies specifically on utilizing the Entity Legal Form (ELF) code (ISO 20275)[1] as a fundamental basis for classification. The ELF code standard serves as a comprehensive solution for standardized legal form representation and is incorporated in the freely available LEI data, which encompasses over 2.3 million active entities worldwide as of September 2023. Consequently, the LEI data and the ELF code standard provide an optimal data source for training classifiers dedicated to legal form detection.

[1] https://www.iso.org/standard/67462.html

We address the challenge of legal form classification by employing a novel approach using Transformer language models based on the BERT architecture in combination with standardized ELF codes. The results of the Transformer models are compared to a traditional Bag-of-Words setting. To evaluate the performance of Transformer and traditional approaches, a substantial subset of LEI data comprising over 1 million legal entities from 30 different legal jurisdictions is employed as the evaluation dataset. Lastly, we conducted an expert review using 7,256 entities, further corroborating the plausibility of our findings.

The results of our research are available as a Python library on GitHub (https://github.com/Sociovestix/lenu). The associated Transformer models are accessible on HuggingFace (https://huggingface.co/Sociovestix).

3.1 Entity Legal Forms (ELF) code list

The ELF code, established by the International Organization for Standardization (ISO)[1], is a unique 4-digit alpha-numeric code and serves as a comprehensive solution for standardized legal form representation. As of September 2023, there are 3,250 legal forms in 175 jurisdictions worldwide available[2] in version 1.4.1 of the openly accessible ELF code list. GLEIF has been acting as the maintenance agency secretariat of the ELF code list since 2017, regularly introducing new legal forms and jurisdictions. By way of example, Table 2 shows the ELF codes for US-Delaware for entities with RegistrationStatus ISSUED. Each jurisdiction has its unique set of legal forms. Legal forms that appear to be similar across jurisdictions must be treated individually due to differing legislature within each jurisdiction. ELF code 8888 is a so-called reserved code that is used in case no existing specific legal form can be assigned. Notably, the legal forms are not equally distributed within US-Delaware. Similarly imbalanced legal form data can be observed in all jurisdictions. As ELF codes are part of the LEI data, the Global LEI Repository serves as the basis for our work. For the scope of our work, we use the 30 largest jurisdictions by number of entities that have RegistrationStatus ISSUED. This leaves us with a training data set of 1.1 million LEI records.

Using this standardized and up-to-date list of entity legal forms, data users are enabled to uniquely identify the legal forms of entities around the globe without having any local knowledge about language or legislature. This prevents costly legal name analysis. For instance, there are more than 500 legal form checks implemented in Deutsche Bundesbank's Financial Statement Data Pool on approximately 125,000 entities each year [6]. These checks would not be necessary, if all entities were assigned with an ELF code.

[2] https://www.gleif.org/en/about-lei/code-lists/iso-20275-entity-legal-forms-code-list

Table 2: Legal forms and their ELF codes in US-Delaware for ISSUED LEI records

Legal form name	ELF code	# entities
Commercial Bank	9ASJ	2
Corporation	XTIQ	5,379
Limited Liability Company	HZEH	30,553
Limited Liability Limited Partnership	TGMR	44
Limited Liability Partnership	1HXP	64
Limited Partnership	T91T	9,707
Non-deposit Trust Company	MIPY	2
Partnership	QF4W	16
Savings Bank	JU79	0
Statutory Trust	4FSX	1,266
Unincorporated Nonprofit Association	12N6	1
Legal form not yet in code list	8888	7,118

3.2 Rule-based approach

For the rule-based approach, we applied two methods:

a) For all LEIs, select the ELF code that appears most frequently in the respective jurisdiction.

b) For those jurisdictions in which the ELF code list contains abbreviations, assign the corresponding legal form if the abbreviation is part of the legal name.

These two approaches served as an initial sanity check and benchmark, whether any sophisticated machine learning or deep learning approach is necessary.

3.3 Traditional machine learning

Our machine learning baseline approach to legal form classification follows a traditional text classification pipeline setup. The setup consists of the pre-processing of input text, feature selection and training of a classifier.

The **pre-processing** "cleans up" the input text to achieve a minimum degree of harmonization, which in turn eases subsequent processing and aims to enhance classification performance. For this, we compared two pre-processing approaches:

a) We transform each input name string to lower-case letters, ensuring a fundamental degree of harmonization of the input names. This is the default setup.

b) We adopt a set of harmonization rules by following ideas for record linkage as presented in [8] and [9]. This includes (1) converting the string to lower-case, (2) replacing diacritics, (3) replacing multi-spaces, (4) removing double quotation marks, (5) replacing trailing non-alphanumeric characters, (6) correcting commas and periods, (7) applying purge rules and (8) replacing multi-spaces. Our purge step removes special characters like "-", "(", ")", ";", "/", "," with simple white spaces and converts " & " and " + " to " and ". We denote this setup with "+ prep".

In the feature selection step, we transform each legal name into their Bag-of-Words representation, due to its simplicity for classification purposes [1]. For this, we split the (pre-processed) legal name strings at their white spaces into their set of words. This kind of representation disregards any information about the position of words in the legal name.

For **classification,** we explore four methods. The first to mention is the Complement Naive Bayes (CNB) algorithm, which is an adaptation of the standard Multinomial Naive Bayes algorithm. This algorithm is particularly suited for imbalanced data sets [11]. Secondly, we apply a Decision Tree classifier (DT), which divides the underlying data space with the use of different text features [1]. Thirdly, we use Random Forest (RF) [2], which - as an ensemble method - fits a number of decision tree classifiers on sub-samples of the data and uses averaging to improve predictive accuracy and avoid over-fitting. Lastly, we use Support Vector Machines (SVM), which partition the data by using non-linear delineations between the different classes [1]. For all presented traditional classifiers, we apply the implementations of the Python machine learning library scikit-learn[3] in its version 1.3.0. We do not modify any of the libraries' default parameters.

3.4 Transformer models

In this study, we apply Bidirectional Encoder Representations from Transformers (BERT) [3] to the entity legal form classification problem. More precisely, we evaluate several variants of pre-trained BERT models. Language models come along with pre-trained tokenizers specifically tied to the respective model. BERT uses WordPiece, which tokenizes the input string into sub-word units, resulting in a vocabulary size of 30,000 tokens. By using sub-word units, it can handle rare words, which is an important feature when working with legal names. Sub-word units also allow for a good balance between the flexibility of single characters and the efficiency of full words [14]. We omitted any custom pre-processing or tokenization in favor of following the end-to-end processing as defined by the BERT variants.

In contrast to the Bag-of-Words model, BERT is a sequence model in which the word order within a given text is taken into account. The output of the model is an embedding vector that captures the whole sequence of tokens, including positional context information. When fine-tuning the model for classification, the sequence embedding serves as input to a classification head, which is usually a

[3] https://scikit-learn.org

single layer of randomly initialized weights. During training, not only the classification head but also the pre-trained weights within the model are tuned to the task.

Due to its availability and strong performance, a zoo of pre-trained and fine-tuned variations of BERT has emerged, in particular with the goal of covering specific languages. For each jurisdiction, we evaluate a different set of BERT variants, mainly driven by the official language(s) within the respective jurisdiction. For instance, in US jurisdictions we test the standard BERT base variants trained on an English corpus, whereas in non-English jurisdictions we test language-specific variants. As an exception to this, we test on all jurisdictions - except for Japan - the multilingual version of BERT[4], which is pre-trained on 102 different languages. By doing so we anticipate catching language-specific intricacies in legal names more efficiently.

Another important aspect is the "casing" of the models. Some BERT variants are trained as "cased" models, whereas others are trained as "uncased". In “cased” models, the text is tokenized as is, including any capitalized letters. In “uncased” models, the text is converted to lowercased before training. We test both, the cased and the uncased versions, where available. In addition to language-specific models, we evaluate FinBERT[5, 15], which has been specifically pre-trained for the financial domain on corporate 10-K & 10-Q ports, as well as earning call transcripts and analyst reports.

Regardless of the specific model variation or underlying jurisdiction data, we optimize each model for 5 epochs with an AdamW optimizer as described in [7], with a learning rate gamma = 0.00002 and a weight decay of lambda = 0.01.

4 Results

Table 3 shows the results as a comparison of Traditional and Transformer-based models. We evaluate our models in terms of F1 score to account for precision and recall simultaneously. Additionally, we consider the Macro F1 Score (F1-M), which is useful for multiclass classification problems exhibiting imbalanced classes. Each model has been cross-validated with five stratified, non-overlapping folds. The scores are computed on the concatenated predictions of all folds.

Generally, the classification performance varies significantly between jurisdictions, which highlights the uniqueness of each jurisdiction in terms of their unique characteristics that may become challenging for the classification task. These challenges include the varying number of samples per legal form within a given jurisdiction which leads to an imbalanced distribution of legal forms. Additionally, there are jurisdiction-specific intricacies that have an impact on the quality of the data and therefore influence the performance. Notably, Belgium (BE) and France (FR) exhibit significantly lower scores than other jurisdictions. In Belgium and France we generally observe that the majority of entities do not carry any legal form information within their legal names. Furthermore, France exhibits an exceptionally high number of 165 ELF codes as target classes that are unequally distributed among the entities.

[4] https:/huggingface.co/bert-base-multilingual-uncased

Table 3: Comparison of traditional machine learning approach and Transformer models. For traditional models the best performing model is shown for both F1 and F1 Macro (F1-M) score separately. For Transformers, the best BERT variant is selected solely by F1 score

LEI data			Traditional				Transformer		
Jurisdiction	# entities	# ELF	Best variant by F1		Best variant by F1-M		Best variant by F1	F1	F1-M
DE	135,079	31	RF + prep	0.9537	DT + prep	0.5906	bert-base-german-uncased	**0.9616**	**0.6174**
IT	104,968	50	SVC + prep	0.899	DT + prep	**0.3218**	bert-base-italian-uncased	**0.901**	0.3121
NL	89,748	20	RF + prep	0.9812	RF + prep	0.7529	bert-base-multilingual-uncased	**0.9847**	**0.7676**
IN	87,491	35	SVC + prep	0.8845	DT	**0.4872**	bert-base-multilingual-uncased	**0.8862**	0.4705
ES	84,231	41	RF + prep	0.9491	DT + prep	**0.5219**	bert-base-multilingual-uncased	**0.9505**	0.5191
GB	74,847	29	SVC + prep	0.9666	DT + prep	**0.4081**	bert-base-uncased	**0.969**	0.4047
FR	59,973	165	SVC + prep	**0.5769**	DT	**0.189**	bert-base-multilingual-cased	0.571	0.1107
DK	56,226	22	RF + prep	0.9349	RF + prep	0.587	danish-bert-botxo	**0.9444**	**0.5941**
US-DE	54,156	12	SVC	0.9871	RF + prep	**0.6094**	finbert-pretrain	**0.9878**	0.5719
SE	48,083	18	RF + prep	0.9789	RF + prep	0.5424	bert-base-multilingual-uncased	**0.9854**	**0.5647**
FI	35,587	52	RF + prep	0.9839	DT + prep	0.5618	bert-base-finnish-uncased-v1	**0.9858**	**0.5978**
LU	33,683	28	SVC	0.8565	DT + prep	**0.4279**	bert-base-multilingual-uncased	**0.8761**	0.3817
NO	32,996	27	RF + prep	0.9888	DT + prep	**0.6815**	bert-base-multilingual-uncased	**0.991**	0.5942
AT	24,433	21	RF + prep	0.9411	DT + prep	0.5496	bert-base-german-uncased	**0.9635**	**0.6001**

LEI data			Traditional				Transformer		
Jurisdiction	# entities	# ELF	Best variant by F1		Best variant by F1-M		Best variant by F1	F1	F1-M
BE	23,969	41	SVC + prep	0.5089	DT	**0.1444**	bert-base-multilingual-uncased	**0.5344**	0.1391
KY	20,541	13	RF + prep	**0.728**	DT + prep	**0.4627**	bert-base-multilingual-uncased	0.7108	0.3844
PL	20,173	36	DT + prep	**0.9898**	DT + prep	**0.6252**	bert-base-polish-uncased-v1	0.9879	0.5355
AU	15,350	13	SVC + prep	**0.8887**	DT + prep	**0.3301**	bert-base-multilingual-uncased	0.8861	0.3198
IE	15,294	19	RF	0.9189	DT	**0.5116**	bert-base-uncased	**0.9251**	0.4569
VG	15,086	9	SVC + prep	**0.8663**	DT	**0.2696**	bert-base-multilingual-uncased	0.8374	0.1622
CZ	14,477	52	RF + prep	0.9893	RF + prep	**0.4355**	bert-base-multilingual-uncased	**0.9908**	0.3824
EE	13,824	13	RF + prep	0.9954	RF + prep	0.6291	bert-base-multilingual-uncased	**0.9965**	**0.6329**
CH	13,742	28	RF + prep	0.9211	RF + prep	**0.4066**	bert-base-multilingual-uncased	**0.9367**	0.3902
HU	10,041	33	RF + prep	**0.9326**	DT	**0.5791**	bert-base-multilingual-uncased	0.9265	0.4511
JP	9,690	12	RF + prep	0.8968	DT + prep	0.2598	bert-base-japanese	**0.9828**	**0.44**
LI	9,458	13	CNB + prep	0.9522	CNB + prep	**0.7708**	bert-base-multilingual-uncased	**0.9525**	0.6616
US-MA	6,987	13	RF	**0.9548**	DT	**0.5107**	bert-base-multilingual-uncased	0.9501	0.4969
PT	6,427	20	RF + prep	**0.9129**	RF + prep	**0.295**	bert-base-multilingual-uncased	0.9088	0.2566
US-CA	6,176	14	RF + prep	0.9362	RF + prep	**0.4067**	bert-base-uncased	**0.9399**	0.3896
US-NY	4,836	10	RF + prep	0.952	DT + prep	0.4998	bert-base-uncased	**0.9582**	**0.525**

4.1 Traditional machine learning and rule-based approach

The **rule-based approach** was discarded in the early stages of this project. The mean accuracy over all jurisdictions for assigning the most frequent ELF code per jurisdiction to all LEI records is 59%. When using the legal form abbreviations of the ELF code list, we reached a mean accuracy of 63% over all jurisdictions in which legal form abbreviations are present. Both values indicate that traditional machine learning and transformer models are clearly outperforming simple rule-based solutions.

Considering **traditional machine learning** methods, Random Forest with pre-processing performs best in 17 jurisdictions for the F1 score, while for the Macro F1 score, a Decision Tree Classifier with pre-processing yields the best results in 13 jurisdictions, followed by RF + prep in 9 jurisdictions.

Regarding the impact of adding pre-processing to the pipeline, we evaluated for each classification method the number of instances in which the pre-processing enabled pipeline outperforms its simpler counterpart. Encompassing F1 and F1-Macro scores, we generate 60 evaluations for 30 jurisdictions. Incorporating the pre-processing was superior to omitting it in 32 (CNB + prep), 45 (DT + prep), 48 (RF + prep), and 46 (SVM + prep) cases respectively. This strongly suggests that the presented pre-processing is indeed supportive for the task within the traditional setup.

4.2 Traditional approach vs. Transformer models

When comparing Transformer models with the traditional approach, we observe that Transformers outperform the traditional pipeline in 22 out of 30 cases in the F1 score. On the Macro F1 score, we find that the Transformers still outperform in 9 cases. Its strongest competitor is the Decision Tree - with and without pre-processing. The Transformer is beaten by the Decision Tree 15 times on the F1-Macro score. This is due to a tendency of the Transformer to be beneficial for the majority classes while being comparably less accurate when classifying weakly represented classes with few samples.

4.3 Transformer models

Language and jurisdiction-specific characteristics prove to have a major influence on the models' performance. The multilingual BERT version performs best in 18 jurisdictions. This is mostly due to the fact that in many cases we did not find any suitable language-specific models for fine-tuning. Also, for jurisdictions like Switzerland and Luxembourg, which have multiple administrative languages, or Liechtenstein, which exhibits a large variety of international names, using a fine-tuned multilingual BERT version performed better than a fine-tuned German BERT.

Also, the multilingual model for example slightly outperformed BERTje (F1 0.9834, F1-M 0.7582) in the Netherlands. For some jurisdictions though, we were able to find language-specific models that outperformed the multilingual version. In the case of Japan, the Transformer clearly outperformed the Traditional approaches. One reason is that no delimiters in the Japanese language exist to tokenize at [4], which causes problems with the Bag-of-Words approach. The Japanese Bert developed by Tohoku University, though, uses a custom Japanese Tokenizer, and therefore achieves a high F1 score of 0.9828.

Regarding the cased and uncased variants of the models, we observe that throughout all jurisdictions, the uncased variants perform better than the cased variants. This may be attributed to the large number of entities whose legal name is given in upper-case letters. The portion of upper-cased legal names varies across all jurisdictions. Based on our research this is linked to the representation of legal names in the local business registers and authoritative sources, which in some jurisdictions by default capture legal names in upper-case letters. A cased model, which relies on the input text using upper and lower case characters, might be negatively affected by fully upper-cased entity names, whereas an uncased variant might be more robust against such deficiencies in the data as it does not infer any meaning to capitalization of specific tokens.

For the subset of mainly English-speaking jurisdictions (GB, US-DE, US-MA, US-CA, US-NY, AU, IE, VG, KY) we evaluated FinBERT. Even though the performance was mostly close, FinBERT, in general, was not able to outperform the standard bert-base and multilingual models. The only exception is US-Delaware, for which it was on par with bert-base-uncased in the F1 score, and even outperformed it reaching an F1-M score of 0.5719 vs. 0.5248 for bert-base-uncased.

Table 4: Token sequence analysis

	Entity A	Entity B
Token	**Attribution**	**Attribution**
[CLS]	0	0
unsere	0.11	
kinder	0.12	
,	0.05	
unsere	0.1	
zukunft	0.17	
-	0.08	
stiftung	**0.82**	
der	**0.43**	
volksbank	0.01	0.09
oden	0.06	0.06
##wald	0.18	0.17
eg	0.18	**0.98**

Non-obvious domain expertise is necessary if the legal form is only implicitly given by the entity name. It is not always mandatory for entities to include legal form information within a legal name. However, presuming sufficient domain knowledge, the legal form may be determined from non-obvious name characteristics. For instance, the legal entity "Langholtgaard" is registered with ELF code FUKI ("Enkeltmandsvirksomhed") and the entity's legal name indicates that it is a farm. In Denmark, the legal form of "Enkeltmandsvirksomhed" (English: "sole proprietor") is predominantly used for farms and therefore, the model assumes that Danish farms predominantly have the ELF code FUKI. Rule-based approaches are unlikely to cover such context information. Transformers, however, can generally capture these patterns within the given data and thus acquire non-obvious domain expertise.

To illustrate this point, let's consider two German entities: Entity A has the legal name "Unsere Kinder, unsere Zukunft – Stiftung der Volksbank Odenwald eG", while entity B has the name "Volksbank Odenwald eG". Both names exhibit the abbreviation "eG", which represents the legal form "eingetragene Genossenschaft" (ELF code AZFE). This would be correct for entity B, however, entity A represents a "Stiftung" (ELF code V2YH) (English: "foundation") that belongs to entity B. The Bag-of-Words approach is unable to classify the ELF code in this scenario, as it generally does not consider the

sequence of tokens and therefore assigns the same relevance to "Stiftung" and "eG". Also, we claim that any simple rule-based approach will fail to predict the correct ELF code in this scenario. The Transformer, however, predicts the correct ELF code for both legal names. Using the Python library transformers-interpret[5], which utilizes integrated gradients [12] we are able to compute and visualize in Error! Reference source not found. the relevance attribution of each token to the Transformer's entity legal form prediction. Please note that the legal name of entity B can be understood as a substring of entity A's legal name.

For entity name A, the transformer attributes the highest scores to the tokens "stiftung" and "der", which represent the phrase "foundation of", resulting in a correct classification of the legal form. For entity name B, the highest attribution is correctly associated with the token "eg". This demonstrates how the transformer model is able to correctly predict the legal form for both entities due to its ability to consider the positioning and sequence of individual tokens.

Table 5: Requested reference data updates

Jurisdiction	Requested updates	Entities updated	Acceptance rate
DE	3,282	3,224	98.24%
ES	1,284	1,273	99.14%
DK	750	317	42.26%
GB	547	420	76.78%
LU	395	253	64.05%
CH	324	169	52.15%
US-DE	332	183	55.12%
NL	163	106	65.03%
NO	118	87	73.72%
SE	61	56	91.80%

An **expert review** in ten jurisdictions with models having an above-average F1 score was carried out to verify the plausibility of our results. In these jurisdictions, we examine entities, for which the models predicted an ELF code different from the recorded ELF code present in the LEI data at that time. For this task, we utilized GLEIF's publicly accessible Challenge Facility (https://www.gleif.org/en/lei-data/gleif-data-quality-management/challenge-lei-data). Using this facility, we requested updates to the LEI data based on the legal forms suggested by our models. The suggested legal forms were then reviewed by the respective local LEI issuing organizations, which are considered experts in entity data management within their accredited jurisdictions. A detailed overview of the accepted requests is shown in Table 5.

Upon review by the LEI issuing organizations, we found that many of the model's predictions were indeed correct, especially for instances with a high prediction probability value. In total, we requested updates for 7,256 entities in the ten selected jurisdictions, and 6,088 of these were updated based on the proposed entity legal form, resulting in an overall acceptance rate of 83.9%. The high acceptance rate proves that our models produce plausible legal form classifications.

[5] https://github.com/cdpierse/transformers-interpret

Moreover, this indicates that the vast majority of LEI data accurately reflects the correct legal form information and therefore serves as a reliable source for training.

4.4 Open-source command line tool: LENU: Legal Entity Name Understanding

The results of our project are publicly available as an open-source library called LENU (Legal Entity Name Understanding - https://github.com/Sociovestix/lenu). The library is freely available and is distributed under Creative Commons Zero 1.0 (CC0-1.0) Universal license[6]. This tool allows users to predict the legal form for any legal name within a given jurisdiction. The user can either use the Naïve Bayes classifier or the Transformer models. We provide pre-trained transformer models on Huggingface for selected jurisdictions. Users can also train their own models and change the source code according to their needs. Notably, our tool returns confidence values for each output so that users receive an indication whether the predicted legal form should be subject of manual review. Figure 1 shows an example of the command line tool in action.

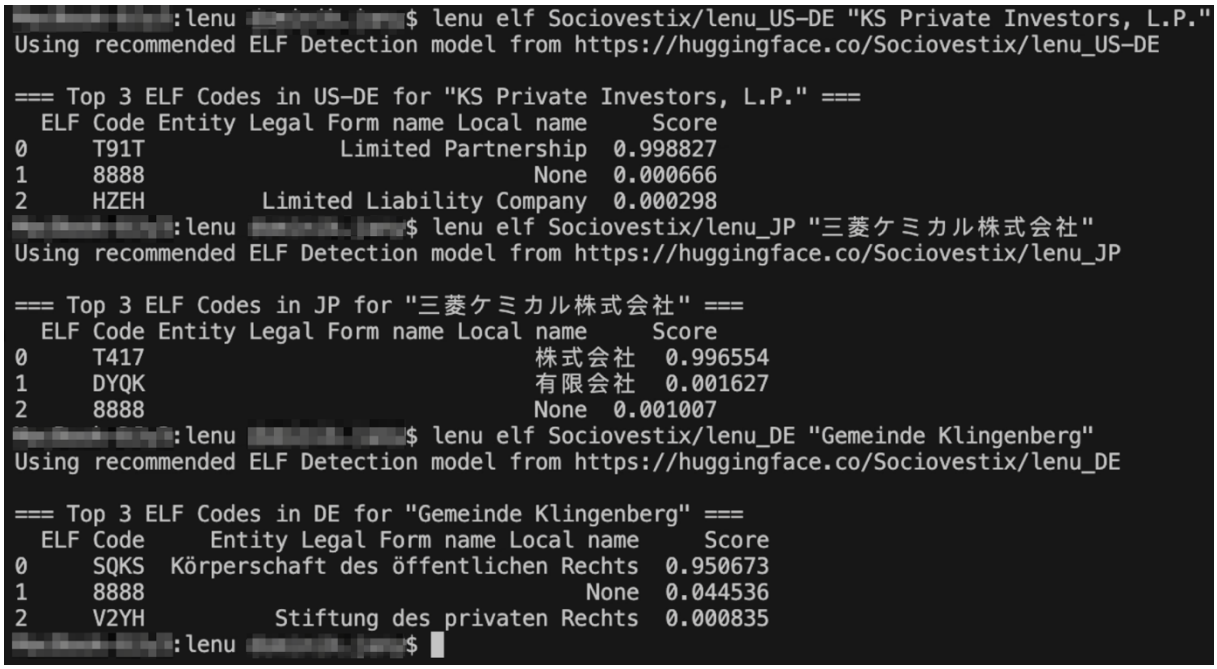

Figure 1: Open Source Command line tool

5 Conclusion

Our study successfully applied Transformer-based language models for classifying entity legal forms from raw legal entity names within their respective jurisdictions. Utilizing various BERT variants, we compared their performance against traditional baselines on a substantial subset of LEI data, covering over 1.1 million legal entities from 30 jurisdictions. Our findings reveal that the presented

[6] https://creativecommons.org/publicdomain/zero/1.0/

models can effectively learn statistical relationships that prove useful, especially when legal entity names lack explicit representation of the corresponding legal form. Rule-based and traditional machine learning approaches on the other hand require significant effort for pre-processing the legal names. Oftentimes the preprocessing mandates considerable domain knowledge to account for various languages and jurisdiction-specific specialties in terms of legal form representation. By leveraging pre-trained BERT models, the Transformer models elegantly solve the language barrier problem and autonomously recognize implicit legal form representations. More specifically, the ability of large language models to capture the sequential nature of legal names also proved advantageous, as it enhances accuracy in handling cases, where traditional methods may struggle.

The expert review process played a critical role in validating the reliability of our models. We sought confirmation from local experts in ten selected jurisdictions, and the vast majority of the proposed legal forms were confirmed. This clear indication from the experts reinforces the trustworthiness and accuracy of our legal form classifications.

The LEI data and ELF codes played a crucial role in our study, providing valuable ground truth labels for legal form classification. We believe that broader adoption of the ELF code standard will significantly enhance transparency while improving data integration tasks in various domains. By making our open-source library freely accessible to the public, we want to facilitate the adoption of the ELF codes to entities worldwide. We invite all stakeholders to use it for entity legal form classification.

References

[1] Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. Mining text data (2012), 163–222.

[2] Leo Breiman. 2001. Random forests. Machine learning 45 (2001), 5–32.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[4] Teruo Hirabayashi, Kanako Komiya, Masayuki Asahara, and Hiroyuki Shinnou. 2020. Composing word vectors for japanese compound words using bilingual word embeddings. In Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation. 404–410.

[5] Allen H Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research 40, 2 (2023), 806–841.

[6] François Laurent. 2021. The benefits of the Legal Entity Identifier for monitoring systemic risk. ESRB Occasional Paper Series 18 (2021).

[7] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[8] Tom Magerman, Bart Van Looy, and Xiaoyan Song. 2006. Data production methods for harmonized patent statistics: Patentee name harmonization. (2006).

[9] Matteo Magnani and Danilo Montesi. 2007. A study on company name matching for database integration. Technical Report UBLCS-07-15, Department of Computer Science University of Bologna (2007).

[10] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. ACM computing surveys (CSUR) 54, 3 (2021), 1–40.

[11] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of the 20th international conference on machine learning (ICML-03). 616–623.

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In International conference on machine learning. PMLR, 3319– 3328.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017).

[14] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016).

[15] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097 (2020).