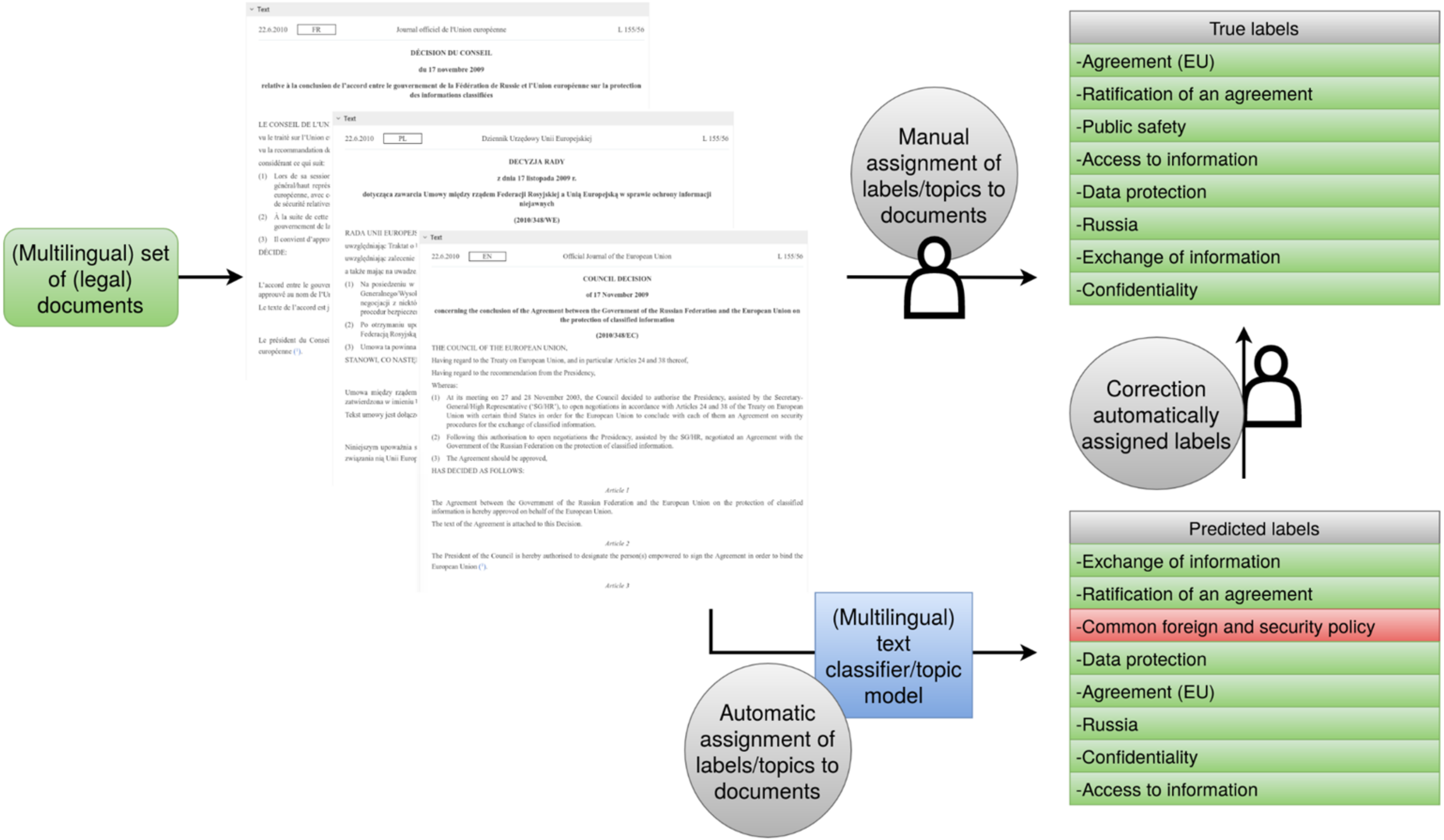
Multilingual text classification of EUR-Lex documents

Arne Defauw

CROSSLANG
TRANSLATION AUTOMATION

European Language
Resource Coordination
Connecting Europe Facility



(Multilingual) set of (legal) documents

Text

22.6.2010 FR Journal officiel de l'Union européenne L 155/56

DÉCISION DU CONSEIL

du 17 novembre 2009

relative à la conclusion de l'accord entre le gouvernement de la Fédération de Russie et l'Union européenne sur la protection des informations classifiées

LE CONSEIL DE L'UN
vu le traité sur l'Union e
vu la recommandation d
considérant ce qui suit:
(1) Lors de sa sessio
général/haut représ
européenne, avec c
de sécurité relative
(2) À la suite de cette
gouvernement de la
(3) Il convient d'appro
DÉCIDE:

L'accord entre le gouve
approuvé au nom de l'U
Le texte de l'accord est j

Le président du Consei
européenne (1).

Text

22.6.2010 PL Dziennik Urzędowy Unii Europejskiej L 155/56

DECYZJA RADY

z dnia 17 listopada 2009 r.

dotycząca zawarcia Umowy między rządem Federacji Rosyjskiej a Unią Europejską w sprawie ochrony informacji niejawnych

(2010/348/WE)

RADA UNII EUROPEJS
uwzględniając Traktat o
uwzględniając zalecenie
a także mając na uwadze
(1) Na posiedzeniu w
Generalnego/Wysol
negocjacji z niektó
procedur bezpieczeń
(2) Po otrzymaniu upo
Federacją Rosyjską
(3) Umowa ta powinna
STANOWI, CO NASTĘ

Umowa między rządem
zatwierdzona w imieniu
Tekst umowy jest dołącz

Niniejszym upoważnia s
związania nią Unii Europ

Text

22.6.2010 EN Official Journal of the European Union L 155/56

COUNCIL DECISION

of 17 November 2009

concerning the conclusion of the Agreement between the Government of the Russian Federation and the European Union on the protection of classified information

(2010/348/EC)

THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on European Union, and in particular Articles 24 and 38 thereof,

Having regard to the recommendation from the Presidency,

Whereas:

(1) At its meeting on 27 and 28 November 2003, the Council decided to authorise the Presidency, assisted by the Secretary-General/High Representative ('SG/HR'), to open negotiations in accordance with Articles 24 and 38 of the Treaty on European Union with certain third States in order for the European Union to conclude with each of them an Agreement on security procedures for the exchange of classified information.

(2) Following this authorisation to open negotiations the Presidency, assisted by the SG/HR, negotiated an Agreement with the Government of the Russian Federation on the protection of classified information.

(3) The Agreement should be approved,

HAS DECIDED AS FOLLOWS:

Article 1

The Agreement between the Government of the Russian Federation and the European Union on the protection of classified information is hereby approved on behalf of the European Union.

The text of the Agreement is attached to this Decision.

Article 2

The President of the Council is hereby authorised to designate the person(s) empowered to sign the Agreement in order to bind the European Union (1).

Article 3

Manual assignment of labels/topics to documents

True labels

-Agreement (EU)
-Ratification of an agreement
-Public safety
-Access to information
-Data protection
-Russia
-Exchange of information
-Confidentiality

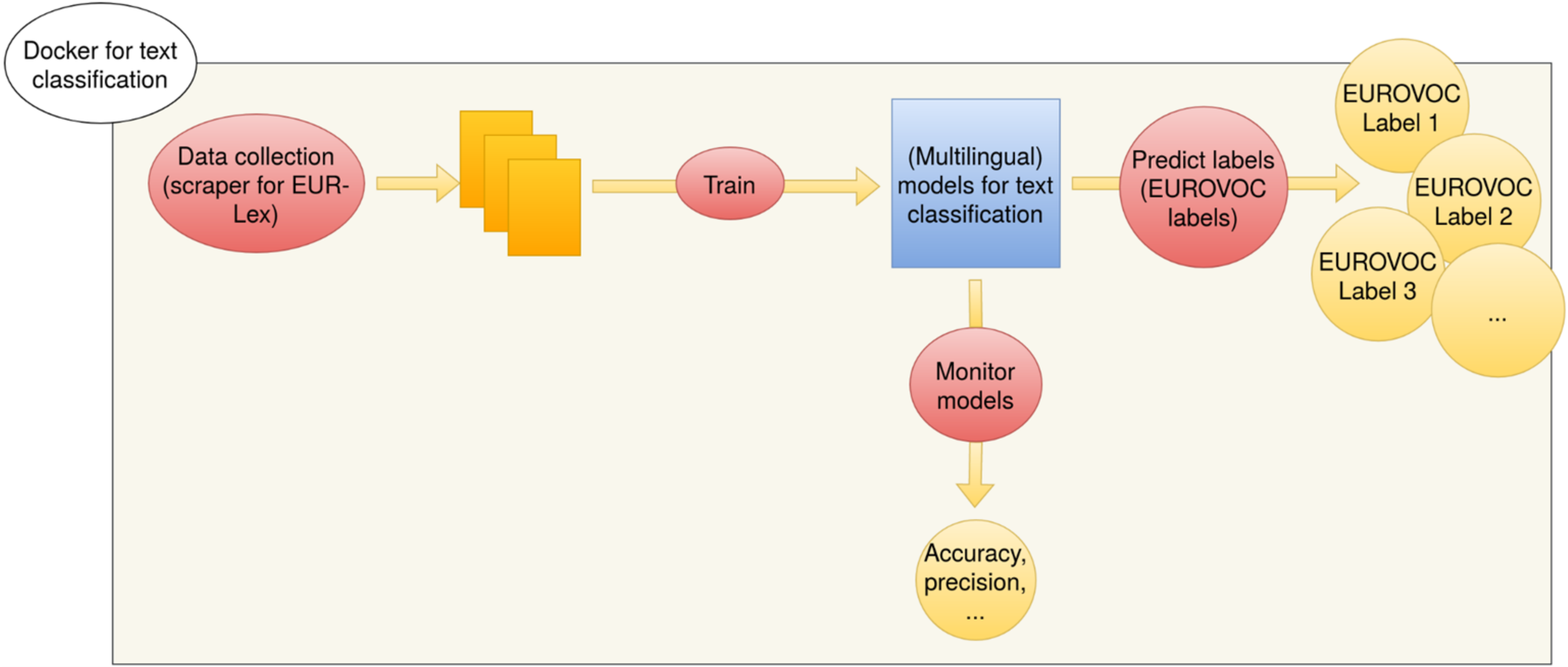
Correction automatically assigned labels

Predicted labels

-Exchange of information
-Ratification of an agreement
-Common foreign and security policy
-Data protection
-Agreement (EU)
-Russia
-Confidentiality
-Access to information

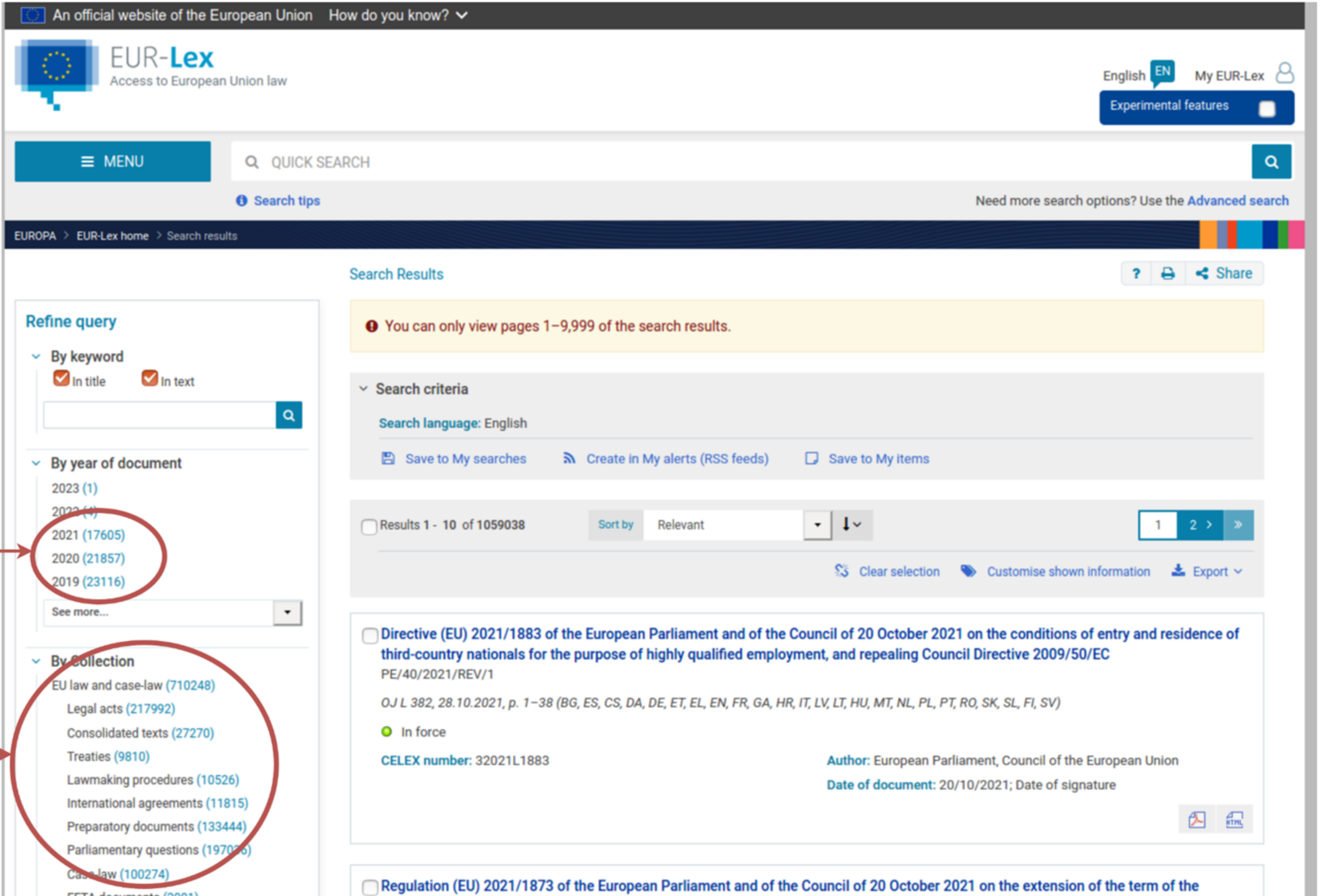
(Multilingual) text classifier/topic model

Automatic assignment of labels/topics to documents


Docker for text classification
Data collection (scraper for EUR-Lex)
Train
(Multilingual) models for text classification
Predict labels (EUROVOC labels)
EUROVOC Label 1
EUROVOC Label 2
EUROVOC Label 3
...
Monitor models
Accuracy, precision, ...

Overview of the classification task:

- EUR-Lex (text)
- EUROVOC taxonomy (labels)



An official website of the European Union How do you know?

EUR-Lex
Access to European Union law

English EN My EUR-Lex

Experimental features

MENU

QUICK SEARCH

Search tips

Need more search options? Use the Advanced search

EUROPA > EUR-Lex home > Search results

Search Results

? Share

Refine query

By keyword

☑ In title ☑ In text

By year of document

2023 (1)
2022 (1)
2021 (17605)
2020 (21857)
2019 (23116)

See more...

By collection

EU law and case-law (710248)
Legal acts (217992)
Consolidated texts (27270)
Treaties (9810)
Lawmaking procedures (10526)
International agreements (11815)
Preparatory documents (133444)
Parliamentary questions (197026)
Case-law (100274)
EFTA documents (2091)

You can only view pages 1–9,999 of the search results.

Search criteria

Search language: English

Save to My searches Create in My alerts (RSS feeds) Save to My items

□ Results 1 - 10 of 1059038 Sort by Relevant

1 2 > »

Clear selection Customise shown information Export

□ Directive (EU) 2021/1883 of the European Parliament and of the Council of 20 October 2021 on the conditions of entry and residence of third-country nationals for the purpose of highly qualified employment, and repealing Council Directive 2009/50/EC

PE/40/2021/REV/1

OJ L 382, 28.10.2021, p. 1–38 (BG, ES, CS, DA, DE, ET, EL, EN, FR, GA, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

● In force

CELEX number: 32021L1883

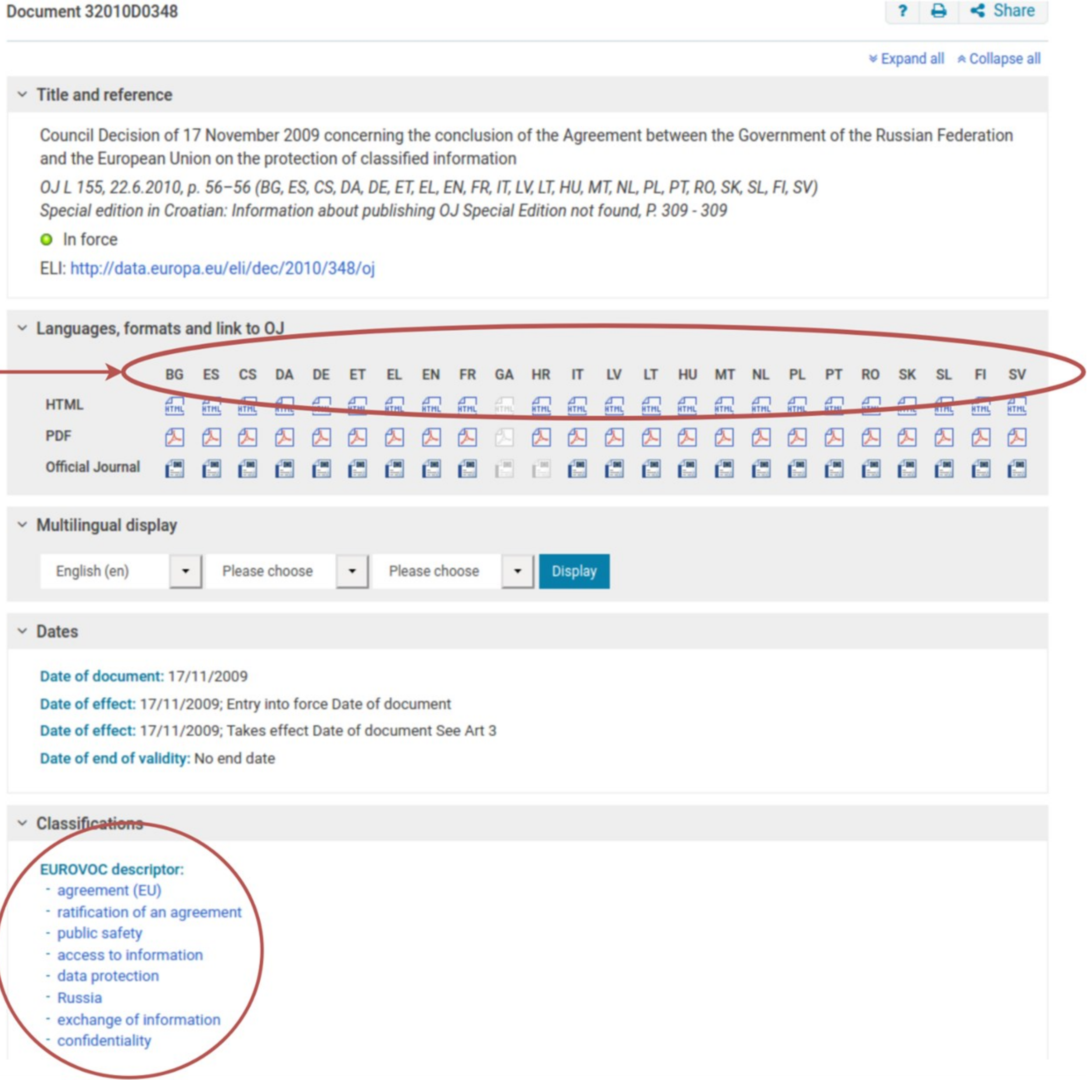
Author: European Parliament, Council of the European Union

Date of document: 20/10/2021; Date of signature

□ Regulation (EU) 2021/1873 of the European Parliament and of the Council of 20 October 2021 on the extension of the term of the

Around 20 000 EUR-Lex documents published each year.

Legal acts, Treaties, International agreements,...



Document 32010D0348

Title and reference

Council Decision of 17 November 2009 concerning the conclusion of the Agreement between the Government of the Russian Federation and the European Union on the protection of classified information

OJ L 155, 22.6.2010, p. 56–56 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)
Special edition in Croatian: Information about publishing OJ Special Edition not found, P. 309 - 309

In force

ELI: http://data.europa.eu/eli/dec/2010/348/oj

Languages, formats and link to OJ

BG ES CS DA DE ET EL EN FR GA HR IT LV LT HU MT NL PL PT RO SK SL FI SV

HTML

PDF

Official Journal

Multilingual display

Dates

Date of document: 17/11/2009

Date of effect: 17/11/2009; Entry into force Date of document

Date of effect: 17/11/2009; Takes effect Date of document See Art 3

Date of end of validity: No end date

Classifications

EUROVOC descriptor:
- agreement (EU)
- ratification of an agreement
- public safety
- access to information
- data protection
- Russia
- exchange of information
- confidentiality

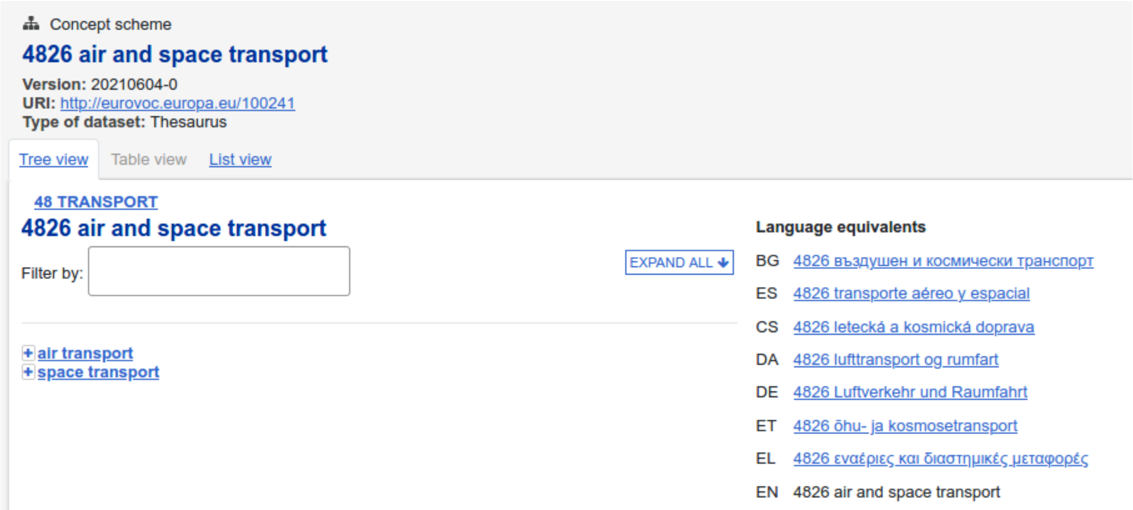
Most documents available in multiple languages.

Manual assignment of EUROVOC descriptors to (most) documents.

EUROVOC Taxonomy:

- Contains 7390 concepts.
- 8 levels of concepts.
- If a document is assigned a concept, ancestors and descendants of that concept are typically not assigned to the same document.

Example:



Concept scheme

4826 air and space transport

Version: 20210604-0
URI: http://eurovoc.europa.eu/100241
Type of dataset: Thesaurus

Tree view Table view List view

48 TRANSPORT

4826 air and space transport

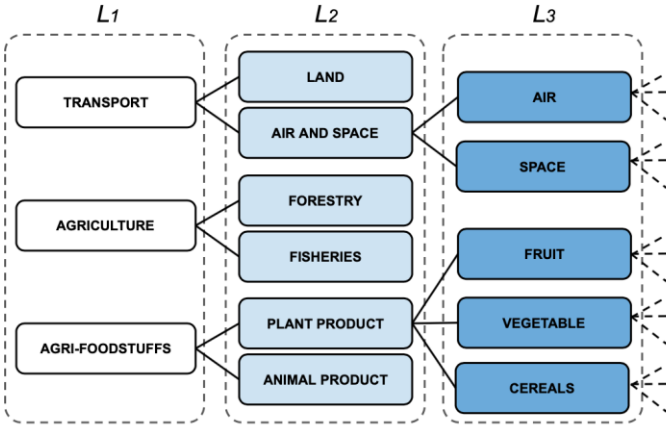
Filter by:

EXPAND ALL

+ air transport
+ space transport

Language equivalents

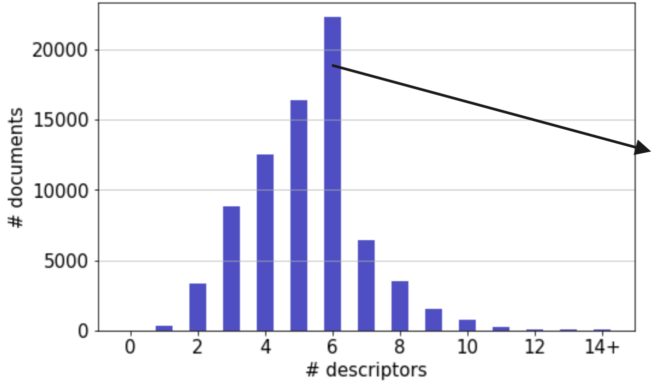
BG	4826 въздушен и космически транспорт
ES	4826 transporte aéreo y espacial
CS	4826 letecká a kosmická doprava
DA	4826 lufttransport og rumfart
DE	4826 Luftverkehr und Raumfahrt
ET	4826 õhu- ja kosmosetransport
EL	4826 εναέριες και διαστημικές μεταφορές
EN	4826 air and space transport



Example from levels L1, L2 and L3 from the EUROVOC taxonomy. Image taken from Chalkidis, I., Fergadiotis, E., Malakasiotis, P. and Androutsopoulos, I. (2021). MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross lingual transfer.

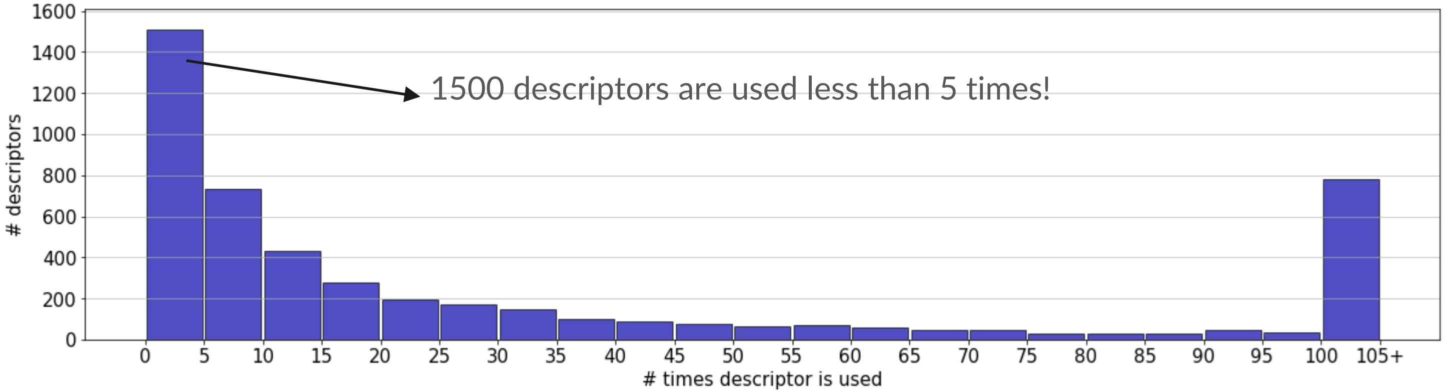
EUROVOC Taxonomy, data distribution:

- Number of descriptors per document (calculated on subset of EUR-Lex corpus, 76k documents):



Average number of EUROVOC descriptors per document is around 5-6.

- Histogram of number of times a descriptor is used:



1500 descriptors are used less than 5 times!

Classification task:

- Manual assignment of EUROVOC labels to EUR-Lex documents is **time consuming**.
- Manual assignment of EUROVOC labels is **complex** (>7K EUROVOC descriptors) and requires domain knowledge.
- Can we build a text classification model that can predict EUROVOC labels given an EUR-Lex document with **good accuracy?**

Pipeline for building a model for Text Classification:

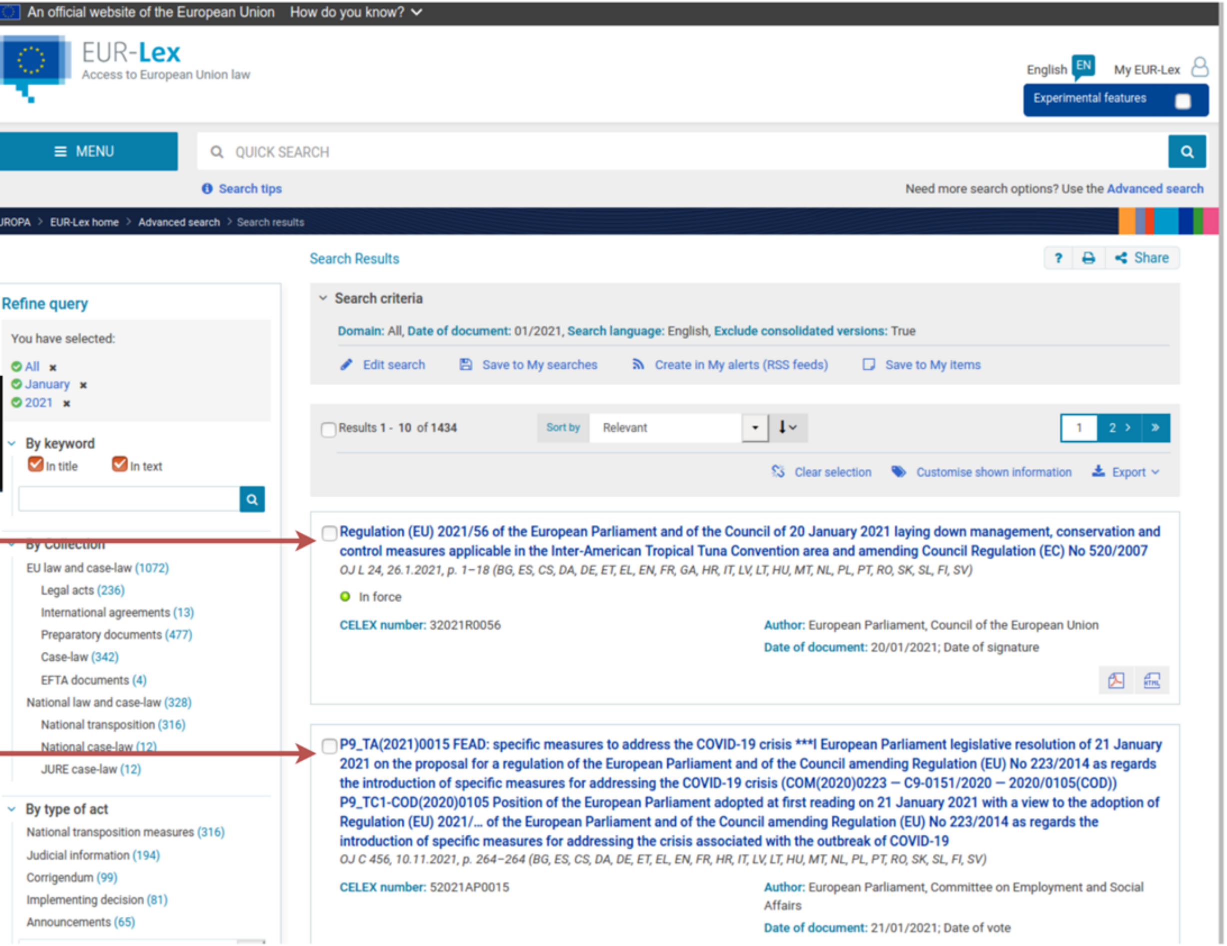
- Data collection (scraping)
- Preprocessing (extraction of text)
- Training the model

Development of scrapers for EUR-Lex website (scraping of text and EUROVOC labels + extraction of text from html):

Development of 'scrapers' for data collection.

```
scrapy crawl eurlex \
-s FEED_URI='../../DATA/eurlex_012021.jsonl' \
-s FEED_FORMAT=jsonlines \
-a year='2021' \
-a month='01'
```

Scraper runs in docker container.


An official website of the European Union How do you know?
EUR-Lex
Access to European Union law
English EN
My EUR-Lex
Experimental features
MENU
QUICK SEARCH
Search tips
Need more search options? Use the Advanced search
EUROPA > EUR-Lex home > Advanced search > Search results
Search Results
? Share
Edit search
Save to My searches
Create in My alerts (RSS feeds)
Save to My items
Sort by Relevant
1 2 >
Clear selection
Customise shown information
Export

Refine query

You have selected:

- All ✖
- January ✖
- 2021 ✖

By keyword

☑ In title ☑ In text

By Collection

- EU law and case-law (1072)
  - Legal acts (236)
  - International agreements (13)
  - Preparatory documents (477)
  - Case-law (342)
  - EFTA documents (4)
- National law and case-law (328)
  - National transposition (316)
  - National case-law (12)
  - JURE case-law (12)

By type of act

- National transposition measures (316)
- Judicial information (194)
- Corrigendum (99)
- Implementing decision (81)
- Announcements (65)

Search criteria

Domain: All, Date of document: 01/2021, Search language: English, Exclude consolidated versions: True

□ Results 1 - 10 of 1434

□ Regulation (EU) 2021/56 of the European Parliament and of the Council of 20 January 2021 laying down management, conservation and control measures applicable in the Inter-American Tropical Tuna Convention area and amending Council Regulation (EC) No 520/2007

OJ L 24, 26.1.2021, p. 1–18 (BG, ES, CS, DA, DE, ET, EL, EN, FR, GA, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

● In force

CELEX number: 32021R0056

Author: European Parliament, Council of the European Union

Date of document: 20/01/2021; Date of signature

□ P9_TA(2021)0015 FEAD: specific measures to address the COVID-19 crisis ***I European Parliament legislative resolution of 21 January 2021 on the proposal for a regulation of the European Parliament and of the Council amending Regulation (EU) No 223/2014 as regards the introduction of specific measures for addressing the COVID-19 crisis (COM(2020)0223 — C9-0151/2020 — 2020/0105(COD)) P9_TC1-COD(2020)0105 Position of the European Parliament adopted at first reading on 21 January 2021 with a view to the adoption of Regulation (EU) 2021/... of the European Parliament and of the Council amending Regulation (EU) No 223/2014 as regards the introduction of specific measures for addressing the crisis associated with the outbreak of COVID-19

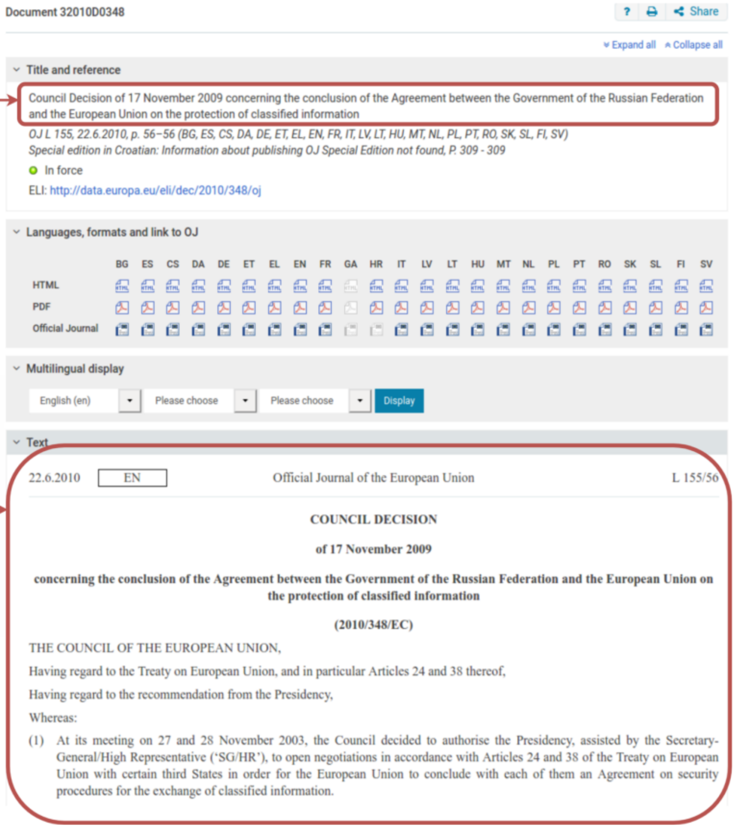
OJ C 456, 10.11.2021, p. 264–264 (BG, ES, CS, DA, DE, ET, EL, EN, FR, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

CELEX number: 52021AP0015

Author: European Parliament, Committee on Employment and Social Affairs

Date of document: 21/01/2021; Date of vote

Development of data preprocessing scripts for extraction and cleaning of text:



Extraction of the title

Extraction of the text

Document 32010D0348

Title and reference

Council Decision of 17 November 2009 concerning the conclusion of the Agreement between the Government of the Russian Federation and the European Union on the protection of classified information

OJ L 155, 22.6.2010, p. 56–56 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

Special edition in Croatian: Information about publishing OJ Special Edition not found, P. 309 - 309

● In force

ELI: http://data.europa.eu/eli/dec/2010/348/oj

Languages, formats and link to OJ

	BG	ES	CS	DA	DE	ET	EL	EN	FR	GA	HR	IT	LV	LT	HU	MT	NL	PL	PT	RO	SK	SL	FI	SV
HTML																								
PDF																								
Official Journal																								

Multilingual display

Text

22.6.2010 EN Official Journal of the European Union L 155/56

COUNCIL DECISION

of 17 November 2009

concerning the conclusion of the Agreement between the Government of the Russian Federation and the European Union on the protection of classified information

(2010/348/EC)

THE COUNCIL OF THE EUROPEAN UNION,

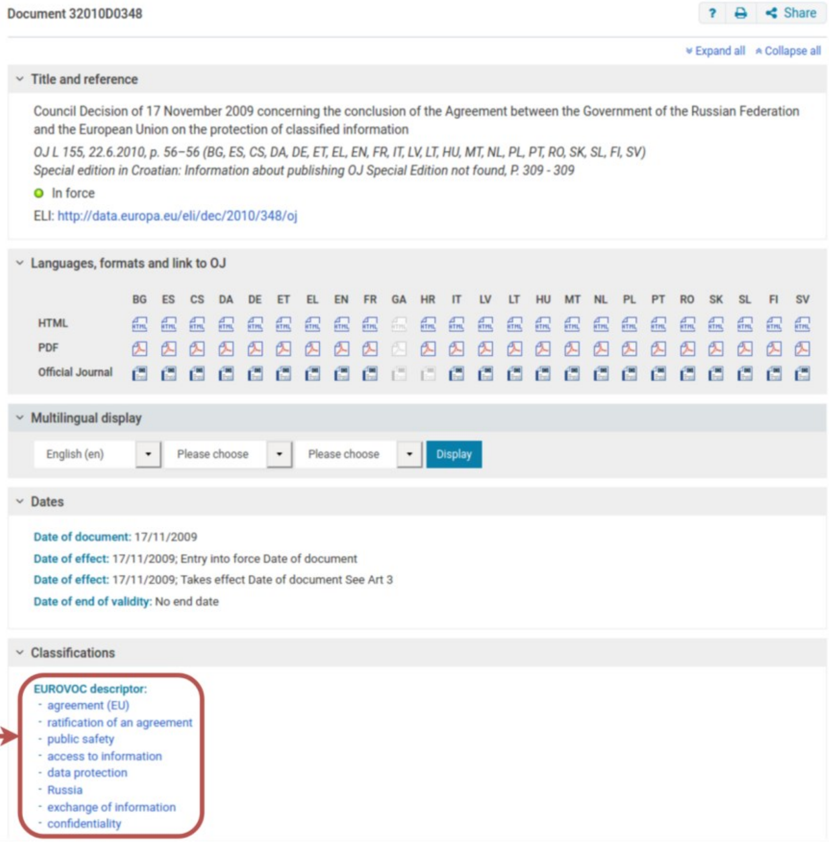
Having regard to the Treaty on European Union, and in particular Articles 24 and 38 thereof,

Having regard to the recommendation from the Presidency,

Whereas:

(1) At its meeting on 27 and 28 November 2003, the Council decided to authorise the Presidency, assisted by the Secretary-General/High Representative ('SG/HR'), to open negotiations in accordance with Articles 24 and 38 of the Treaty on European Union with certain third States in order for the European Union to conclude with each of them an Agreement on security procedures for the exchange of classified information.

Extraction of EUROVOC labels:



Extraction of labels

Document 32010D0348

Title and reference

Council Decision of 17 November 2009 concerning the conclusion of the Agreement between the Government of the Russian Federation and the European Union on the protection of classified information

OJ L 155, 22.6.2010, p. 56–56 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

Special edition in Croatian: Information about publishing OJ Special Edition not found, P. 309 - 309

● In force

ELI: http://data.europa.eu/eli/dec/2010/348/oj

Languages, formats and link to OJ

Multilingual display

Dates

Date of document: 17/11/2009

Date of effect: 17/11/2009; Entry into force Date of document

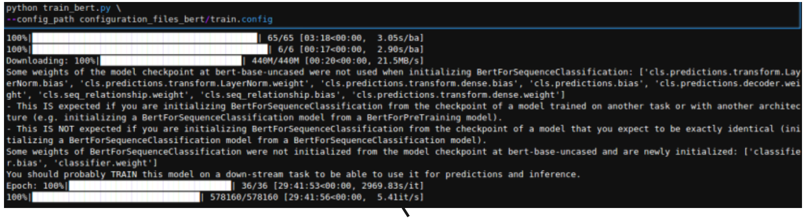
Date of effect: 17/11/2009; Takes effect Date of document See Art 3

Date of end of validity: No end date

Classifications

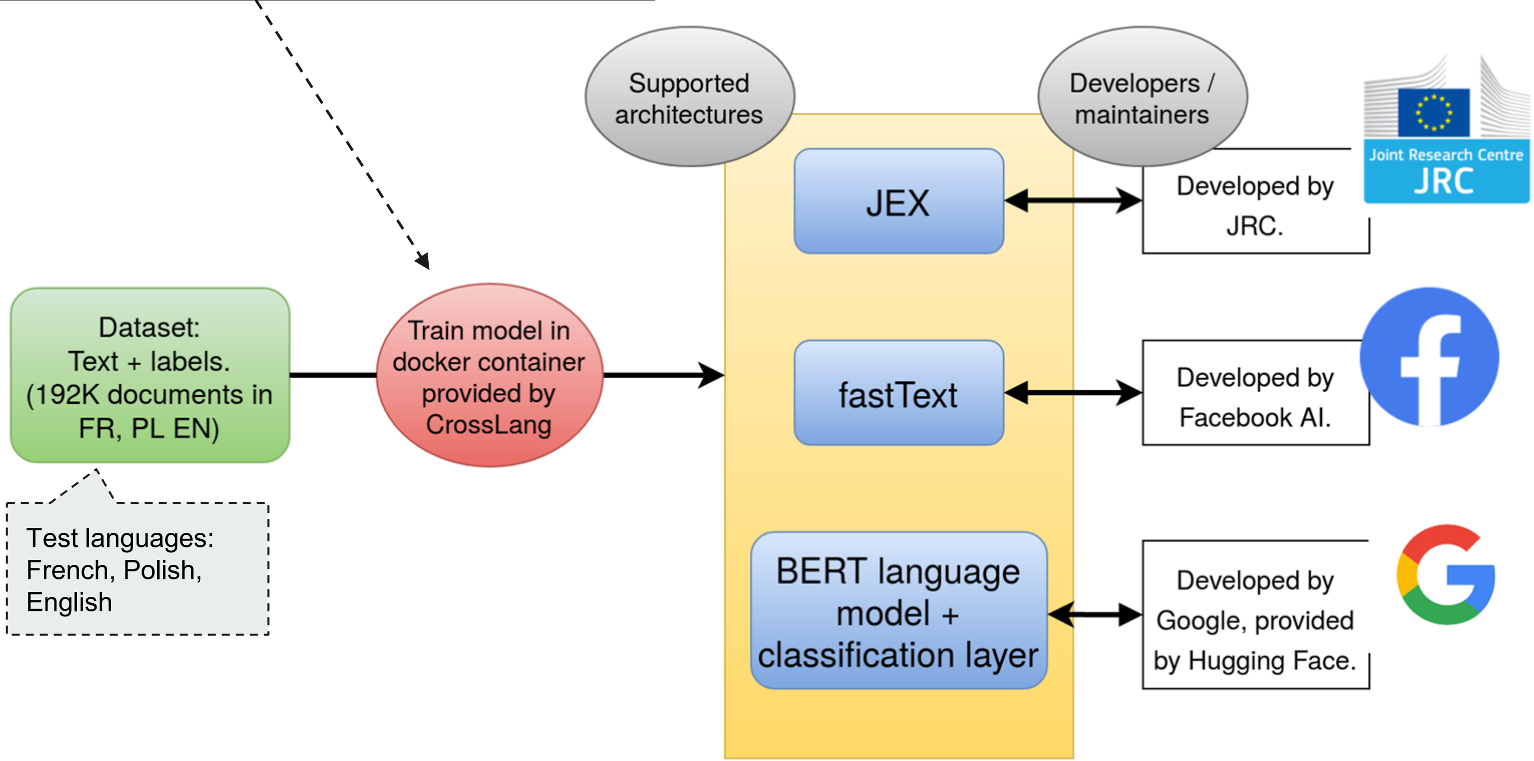
EUROVOC descriptor:
- agreement (EU)
- ratification of an agreement
- public safety
- access to information
- data protection
- Russia
- exchange of information
- confidentiality

Train the model for text classification using the docker provided by CrossLang:

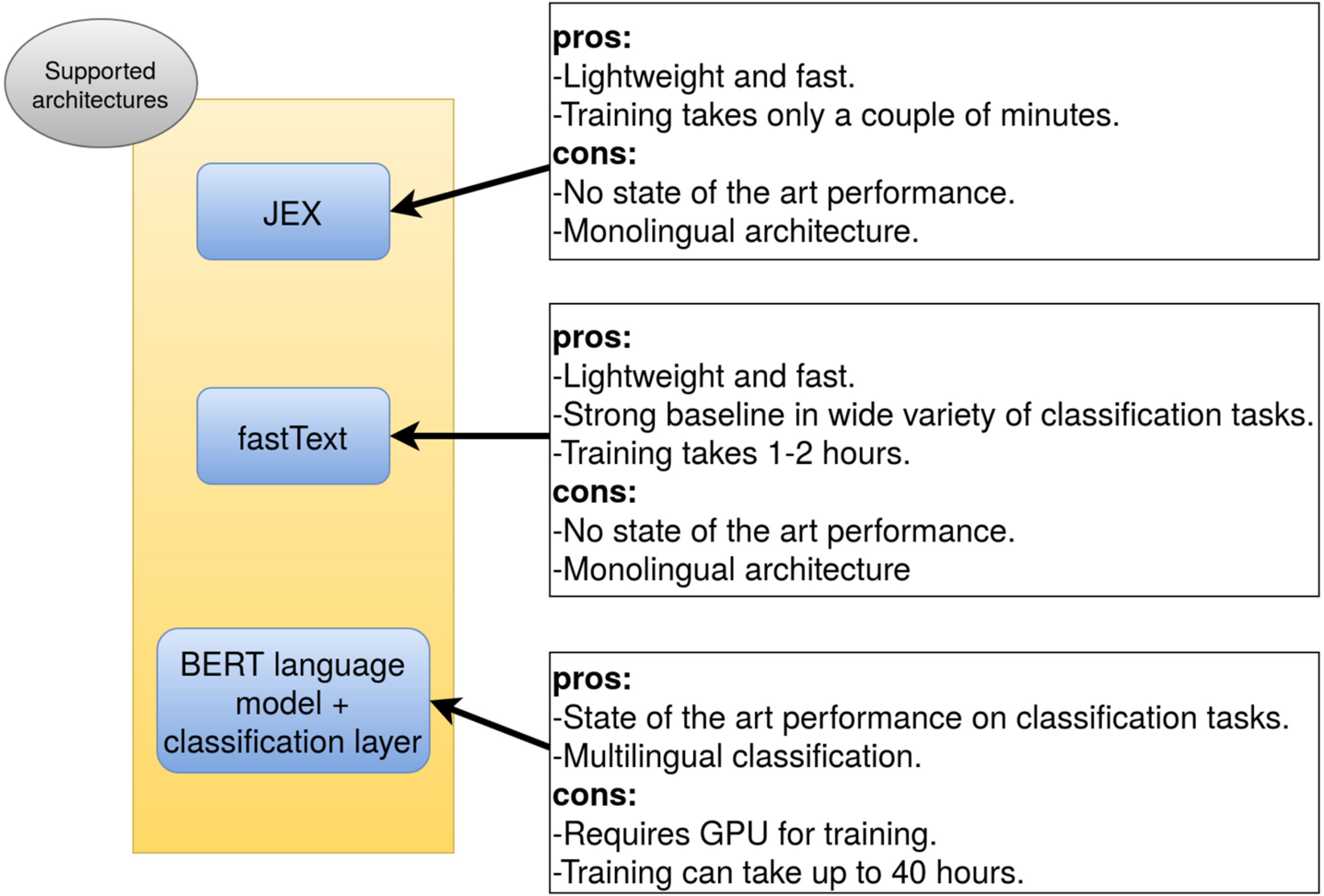


```
python train_bert.py \
--config_path configuration_files_bert/train.config

100%|██████████████████████████████████████| 65/65 [03:18<00:00,  3.05s/ba]
100%|████████████████████████████████████████| 6/6 [00:17<00:00,  2.90s/ba]
Downloading: 100%|████████████████████████| 440M/440M [00:20<00:00, 21.5MB/s]
Some weights of the model checkpoint at bert-base-uncased were not used when initializing BertForSequenceClassification: ['cls.predictions.transform.Lay
erNorm.bias', 'cls.predictions.transform.LayerNorm.weight', 'cls.predictions.transform.dense.bias', 'cls.predictions.bias', 'cls.predictions.decoder.wei
ght', 'cls.seq_relationship.weight', 'cls.seq_relationship.bias', 'cls.predictions.transform.dense.weight']
- This IS expected if you are initializing BertForSequenceClassification from the checkpoint of a model trained on another task or with another architec
ture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing BertForSequenceClassification from the checkpoint of a model that you expect to be exactly identical (ini
tializing a BertForSequenceClassification model from a BertForSequenceClassification model).
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at bert-base-uncased and are newly initialized: ['classifie
r.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Epoch: 100%|██████████████████████████████| 36/36 [29:41:53<00:00, 2969.83s/it]
100%|████████████████████████████████| 578160/578160 [29:41:56<00:00,  5.41it/s]
```



Supported architectures
Developers / maintainers
Joint Research Centre
JRC
Dataset:
Text + labels.
(192K documents in FR, PL EN)
Train model in docker container provided by CrossLang
JEX
fastText
BERT language model + classification layer
Developed by JRC.
Developed by Facebook AI.
Developed by Google, provided by Hugging Face.
Test languages:
French, Polish,
English



Supported architectures

JEX

pros:
-Lightweight and fast.
-Training takes only a couple of minutes.
cons:
-No state of the art performance.
-Monolingual architecture.

fastText

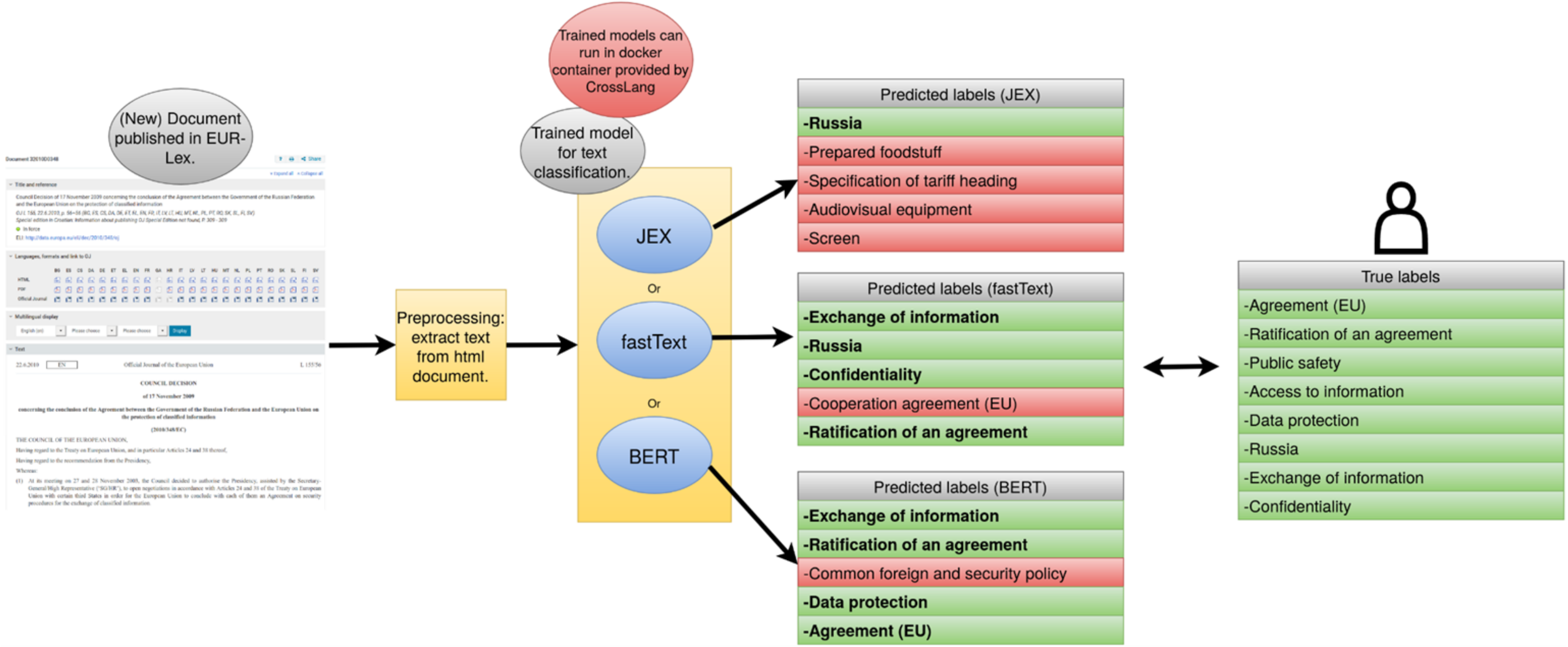
pros:
-Lightweight and fast.
-Strong baseline in wide variety of classification tasks.
-Training takes 1-2 hours.
cons:
-No state of the art performance.
-Monolingual architecture

BERT language model + classification layer

pros:
-State of the art performance on classification tasks.
-Multilingual classification.
cons:
-Requires GPU for training.
-Training can take up to 40 hours.

Pipeline for text classification (inference)

Pipeline for inference (given an html document, predict EUROVOC labels):



(New) Document published in EUR-Lex.
Preprocessing: extract text from html document.
Trained models can run in docker container provided by CrossLang
Trained model for text classification.
JEX
Or
fastText
Or
BERT

Document 32010D0348

?
Share
+ Expand all
+ Collapse all

Title and reference

Council Decision of 17 November 2009 concerning the conclusion of the Agreement between the Government of the Russian Federation and the European Union on the protection of classified information

OJ L 155, 22.6.2010, p. 56–56 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)
Special edition in Croatian: Information about publishing OJ Special Edition not found, P. 309 - 309

In force

ELI: http://data.europa.eu/eli/dec/2010/348/oj

Languages, formats and link to OJ

	BG	ES	CS	DA	DE	ET	EL	EN	FR	GA	HR	IT	LV	LT	HU	MT	NL	PL	PT	RO	SK	SL	FI	SV
HTML	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●
PDF	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●
Official Journal	●	●	●	●	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●	●

Multilingual display

Text

22.6.2010 EN Official Journal of the European Union L 155/56

COUNCIL DECISION

of 17 November 2009

concerning the conclusion of the Agreement between the Government of the Russian Federation and the European Union on the protection of classified information

(2010/348/EC)

THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on European Union, and in particular Articles 24 and 38 thereof,

Having regard to the recommendation from the Presidency,

Whereas:

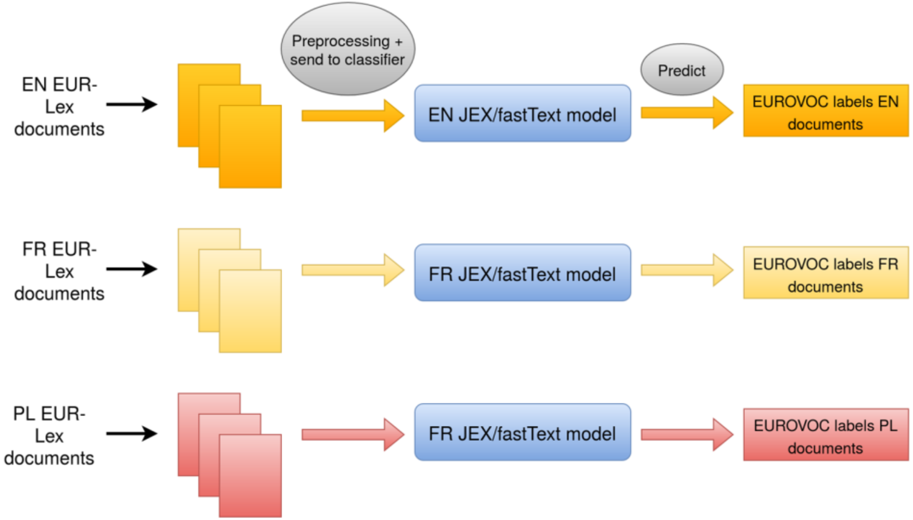
(1) At its meeting on 27 and 28 November 2005, the Council decided to authorise the Presidency, assisted by the Secretary-General/High Representative ('SG/HR'), to open negotiations in accordance with Articles 24 and 38 of the Treaty on European Union with certain third States in order for the European Union to conclude with each of them an Agreement on security procedures for the exchange of classified information.

Predicted labels (JEX)
-Russia
-Prepared foodstuff
-Specification of tariff heading
-Audiovisual equipment
-Screen

Predicted labels (fastText)
-Exchange of information
-Russia
-Confidentiality
-Cooperation agreement (EU)
-Ratification of an agreement

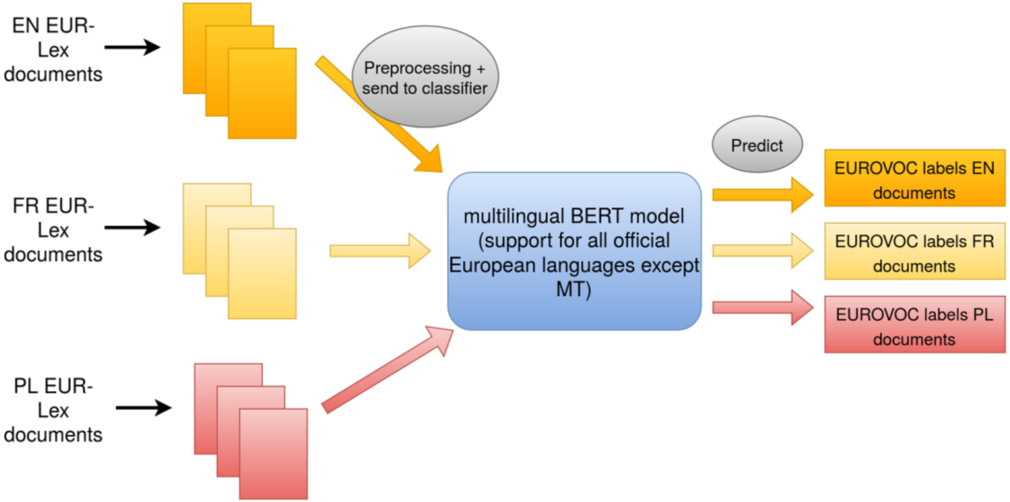
Predicted labels (BERT)
-Exchange of information
-Ratification of an agreement
-Common foreign and security policy
-Data protection
-Agreement (EU)

True labels
-Agreement (EU)
-Ratification of an agreement
-Public safety
-Access to information
-Data protection
-Russia
-Exchange of information
-Confidentiality



Monolingual architectures (JEX/fastText) require a separate classifier for each language....

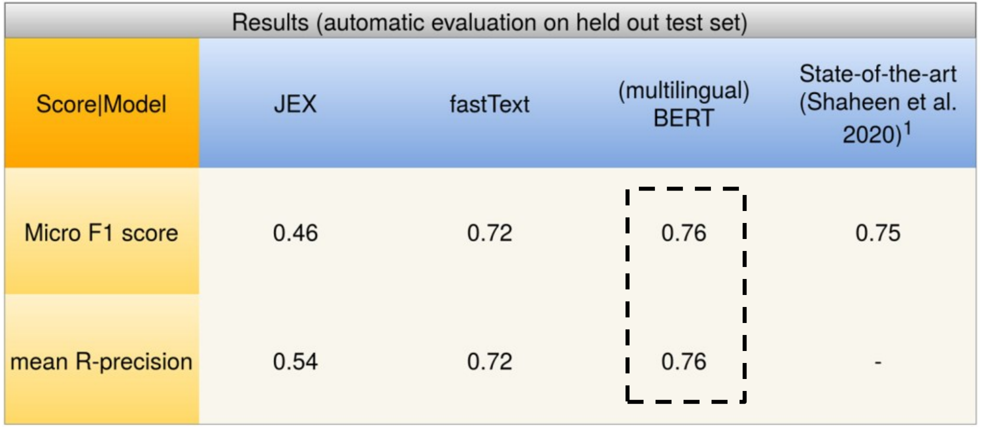
...while an architecture with multilingual support (multilingual BERT), only requires one model for all official European languages (except for Maltese).
This facilitates the deployment of the classification system.


EN EUR-Lex documents
Preprocessing + send to classifier
FR EUR-Lex documents
PL EUR-Lex documents
multilingual BERT model (support for all official European languages except MT)
Predict
EUROVOC labels EN documents
EUROVOC labels FR documents
EUROVOC labels PL documents

Evaluation of predicted labels:

- Automatic procedure
- Human assessment

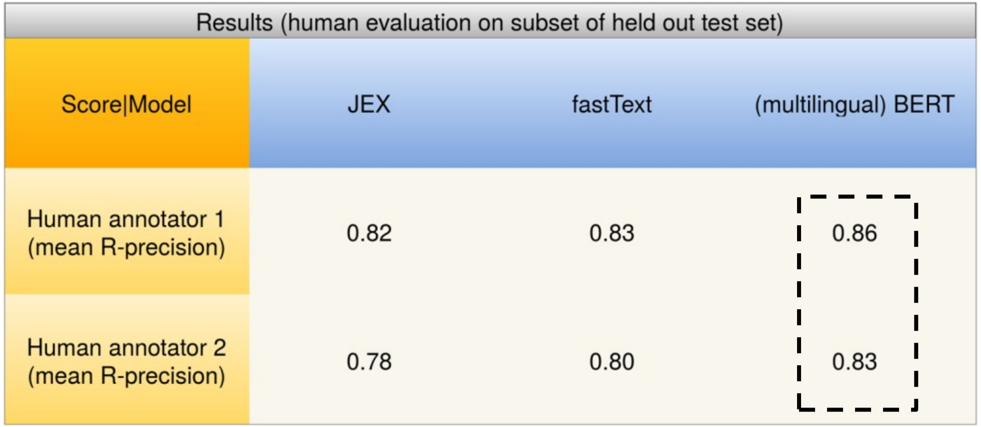
- **Automatic** evaluation: comparison of predicted labels with ground truth EUROVOC labels.



Results (automatic evaluation on held out test set)

Score|Model	JEX	fastText	(multilingual) BERT	State-of-the-art (Shaheen et al. 2020)[1]
Micro F1 score	0.46	0.72	0.76	0.75
mean R-precision	0.54	0.72	0.76	-

- **Human** assessment: evaluation of precision of predicted EUROVOC labels.



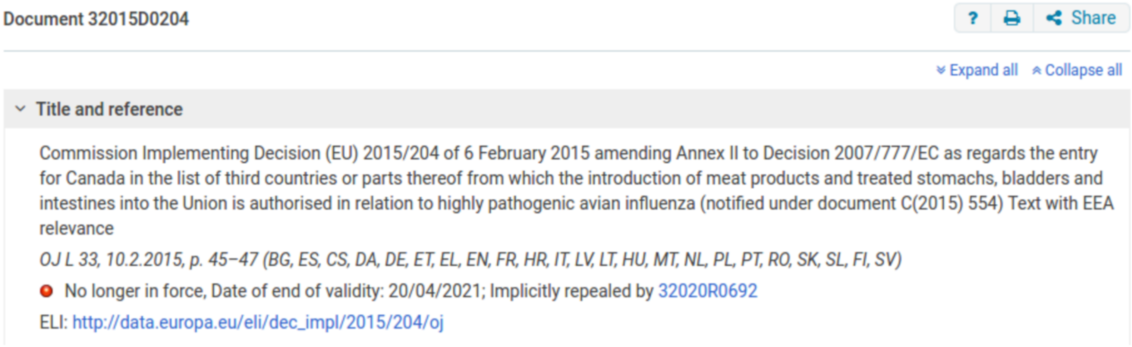
Results (human evaluation on subset of held out test set)

Score|Model	JEX	fastText	(multilingual) BERT
Human annotator 1 (mean R-precision)	0.82	0.83	0.86
Human annotator 2 (mean R-precision)	0.78	0.80	0.83

Inter annotator agreement of 84%

[1] Shaheen, Z., Wohlgenannt, G., and Filtz, E. (2020) Large Scale Legal Text Classification Using Transformer Models.

Example:
EN document



Document 32015D0204

Title and reference

Commission Implementing Decision (EU) 2015/204 of 6 February 2015 amending Annex II to Decision 2007/777/EC as regards the entry for Canada in the list of third countries or parts thereof from which the introduction of meat products and treated stomachs, bladders and intestines into the Union is authorised in relation to highly pathogenic avian influenza (notified under document C(2015) 554) Text with EEA relevance

OJ L 33, 10.2.2015, p. 45–47 (BG, ES, CS, DA, DE, ET, EL, EN, FR, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

● No longer in force, Date of end of validity: 20/04/2021; Implicitly repealed by 32020R0692

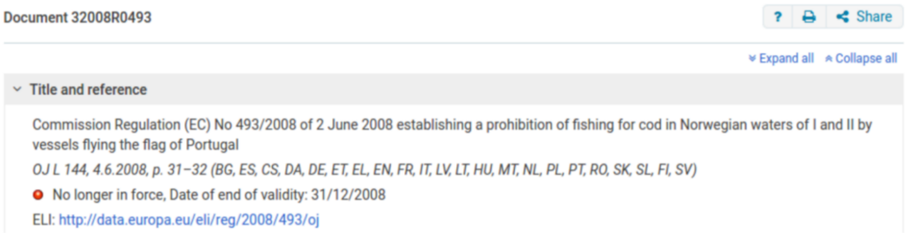
ELI: http://data.europa.eu/eli/dec_impl/2015/204/oj

Predicted labels (JEX)
-Avian influenza
-Poultry
-Veterinary inspection
-Health certificate
-Bird

Predicted labels (BERT)
-Veterinary inspection
-Import (EU)
-Poultry
-Animal product
-Avian influenza

True labels
-Veterinary inspection
-Animal product
-Import (EU)
-Import restriction
-Poultry
-Canada
-Avian influenza

Example:
FR document



Document 32008R0493

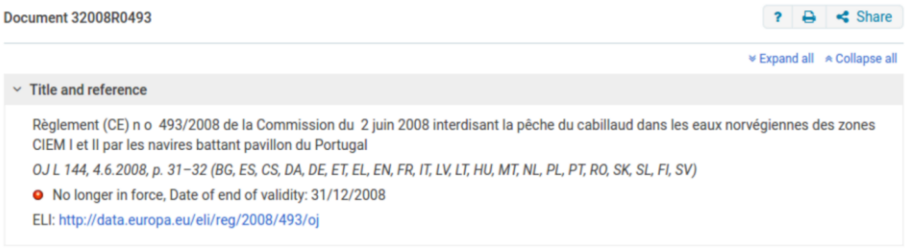
Title and reference

Commission Regulation (EC) No 493/2008 of 2 June 2008 establishing a prohibition of fishing for cod in Norwegian waters of I and II by vessels flying the flag of Portugal

OJ L 144, 4.6.2008, p. 31–32 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

● No longer in force, Date of end of validity: 31/12/2008

ELI: http://data.europa.eu/eli/reg/2008/493/oj



Document 32008R0493

Title and reference

Règlement (CE) n o 493/2008 de la Commission du 2 juin 2008 interdisant la pêche du cabillaud dans les eaux norvégiennes des zones CIEM I et II par les navires battant pavillon du Portugal

OJ L 144, 4.6.2008, p. 31–32 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

● No longer in force, Date of end of validity: 31/12/2008

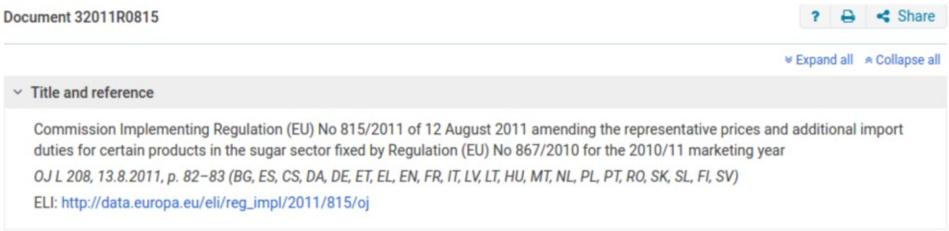
ELI: http://data.europa.eu/eli/reg/2008/493/oj

Predicted labels (JEX)
-Pavillion de navire (Ship's flag)
-Droit de pêche (Fishing rights)
-Quota de pêche (catch quota)
-Poisson de mer (Sea fish)
-Zone de pêche (Fishing area)

Predicted labels (BERT)
-Poisson de mer (Sea fish)
-Portugal
-Pavillion de navire (Ship's flag)
-Quota de pêche (catch quota)
-Norvège (Norway)

True labels
-Pavillion de navire (Ship's flag)
-Poisson de mer (Sea fish)
-Portugal
-Quota de pêche (catch quota)
-Réglementation de la pêche (fishing regulartions)
-Droit de pêche (Fishing rights)
-Eaux de l'UE (EU waters)

Example:
PL document



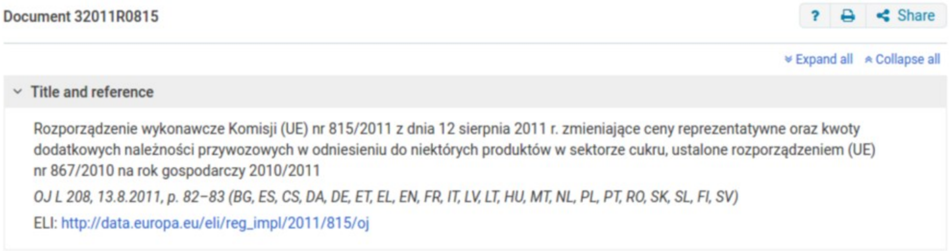
Document 32011R0815

Title and reference

Commission Implementing Regulation (EU) No 815/2011 of 12 August 2011 amending the representative prices and additional import duties for certain products in the sugar sector fixed by Regulation (EU) No 867/2010 for the 2010/11 marketing year

OJ L 208, 13.8.2011, p. 82–83 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

ELI: http://data.europa.eu/eli/reg_impl/2011/815/oj



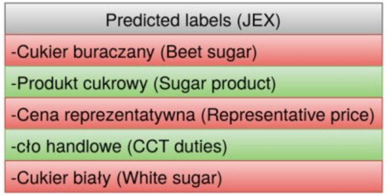
Document 32011R0815

Title and reference

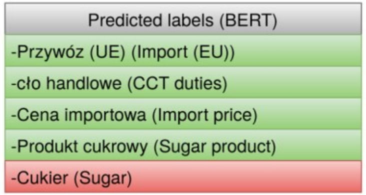
Rozporządzenie wykonawcze Komisji (UE) nr 815/2011 z dnia 12 sierpnia 2011 r. zmieniające ceny reprezentatywne oraz kwoty dodatkowych należności przywozowych w odniesieniu do niektórych produktów w sektorze cukru, ustalone rozporządzeniem (UE) nr 867/2010 na rok gospodarczy 2010/2011

OJ L 208, 13.8.2011, p. 82–83 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

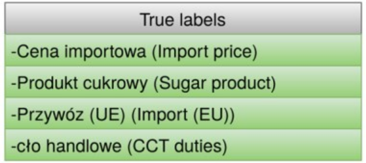
ELI: http://data.europa.eu/eli/reg_impl/2011/815/oj



Predicted labels (JEX)
-Cukier buraczany (Beet sugar)
-Produkt cukrowy (Sugar product)
-Cena reprezentatywna (Representative price)
-cło handlowe (CCT duties)
-Cukier biały (White sugar)



Predicted labels (BERT)
-Przywóz (UE) (Import (EU))
-cło handlowe (CCT duties)
-Cena importowa (Import price)
-Produkt cukrowy (Sugar product)
-Cukier (Sugar)



True labels
-Cena importowa (Import price)
-Produkt cukrowy (Sugar product)
-Przywóz (UE) (Import (EU))
-cło handlowe (CCT duties)

Takeaways:

- Labels predicted by JEX/fastText/BERT architectures can be considered relevant for the document in most cases (precision around +/-80%).

- (multilingual) BERT architecture learns to predict labels that are both **relevant** according to human evaluation and **consistent** with labeling efforts of human annotators.

- Only **one** (multilingual) BERT model needed for classification of documents in 22 official languages of the EU (MT not supported). Using JEX/fastText would require 22+1 separate models.

- Predicting using one multilingual model has the advantage that the result will be **more consistent** across languages, compared to the use of separate models for each language.

- Both data collection (scraping), (multilingual) model training and inference can be done inside the docker container provided by CrossLang.