# Detecting Vague Clauses in Italian Privacy Policies using Transformers, LLMs, and Cross-Lingual Techniques

**Giulia Grundler, Mariaceleste Musicco**,
Andrea Galassi, Francesca Lagioia, Ruta Liepina,
Giorgio Resta, Sara Roccu, Giovanni Sartor, Paolo Torroni

# Context

Growing **information asymmetry** between **digital service providers** and **consumers** about the process of their **personal data**


GDPR
(General Data Protection Regulation)

**Privacy Policies**
**should inform instead overwhelm**

# Context

Multilingualism ensures **clarity and accountability** towards consumers and Data Protection Authorities

Multilingual legal LLMs **challenges**:

- Nature of Legal Language
- Lack of annotated datasets for less resource languages

# Objective and Contribution

**Automatically detect insufficiently informative clauses in PPs across multiple European languages**

- **New corpus of Italian** privacy policies
- Experiments with BERT models and LMM for detecting insufficiently informative clauses **on Italian PP**
- Experiments with multilingual techniques to **transfer knowledge** between Italian and English

# Hierarchical annotation method

Detecting clauses regarding **the kind of personal data processed**

- **Sufficiently informative**

The required information are comprehensively specified and not vague

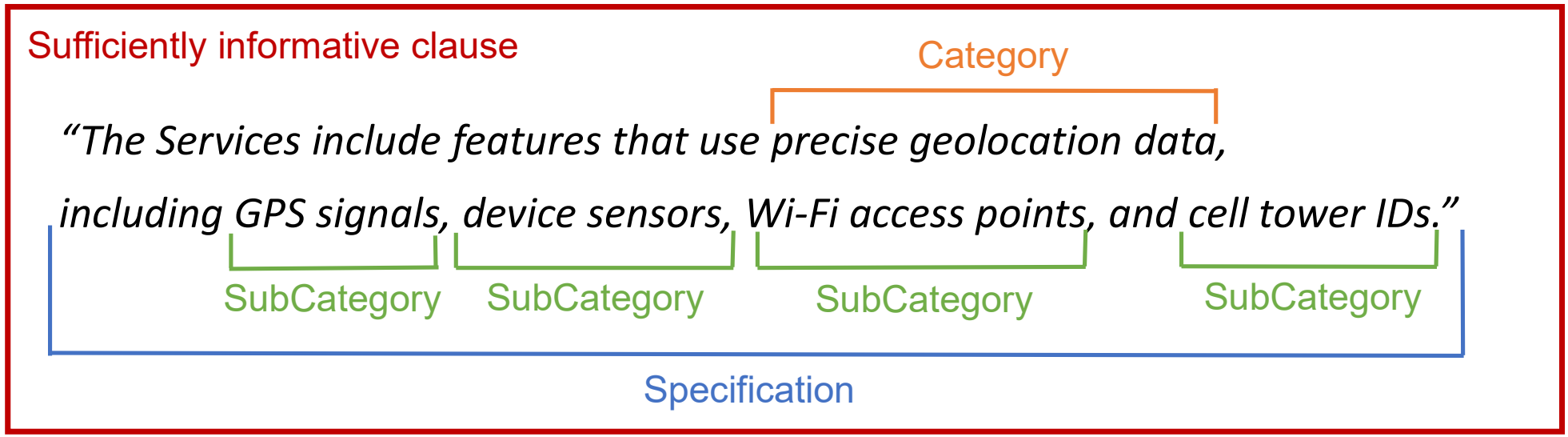
Example: *"We collect your **payment information** such as, name, surname, credit card number and expiration date"* [PayPal]

- **Insufficiently informative**

The required information are missing or they are present but vague and imprecise

Example: "*We collect **information about your interactions with the Airbnb Platform**, like the bookings you have made"* [AirBnB]

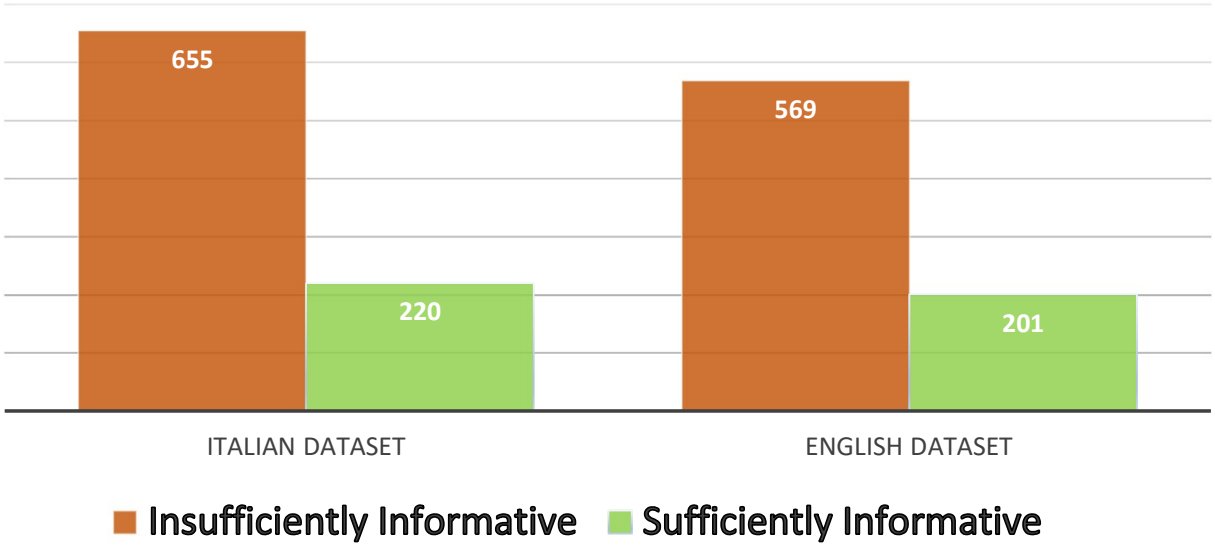
# Hierarchical annotation method




The assessment depends on a set of sub-elements and attributes:

- **Closed Terms**: concrete and exhaustive (e.g. Payment information)
- **Open Terms**: vague and ambiguous (e.g. Usage information)

# Dataset

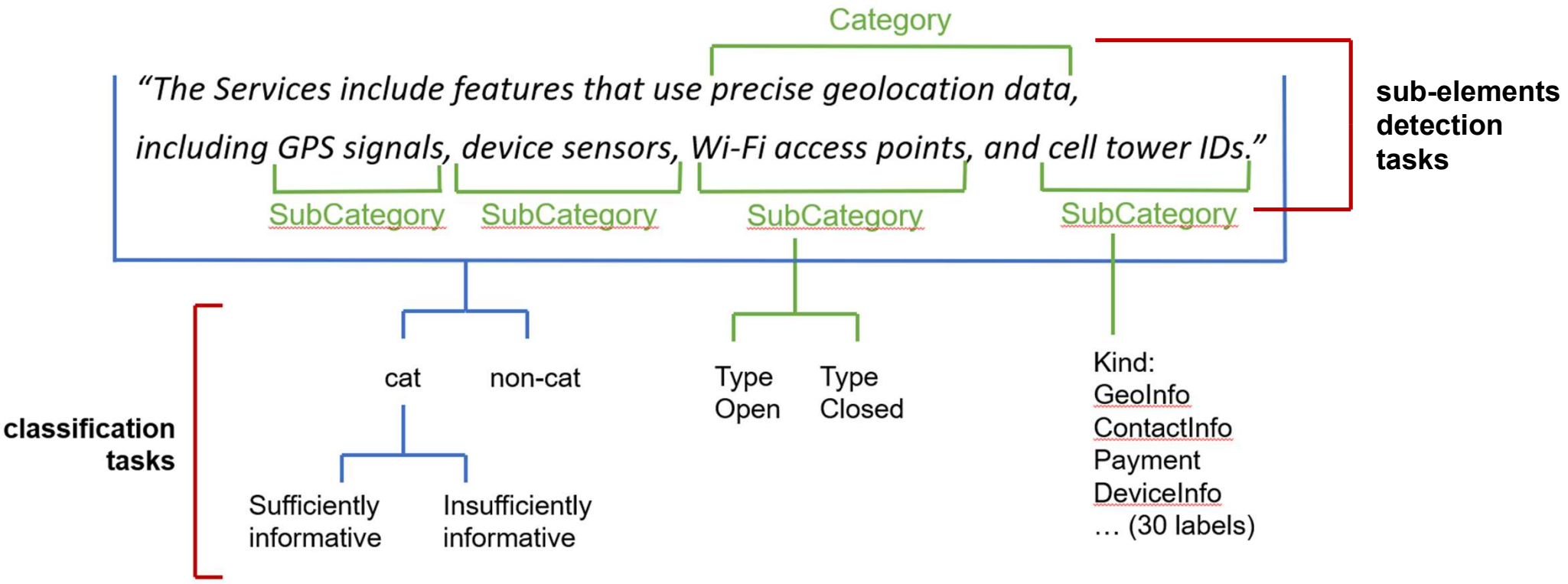



| Element | Ita | Eng |
|---|---|---|
| documents | 30 | 30 |
| sentences | 5862 | 6156 |
| cat | 875 | 770 |
| *Sufficiently inf.* | 220 | 201 |
| *Insufficiently inf.* | 655 | 569 |

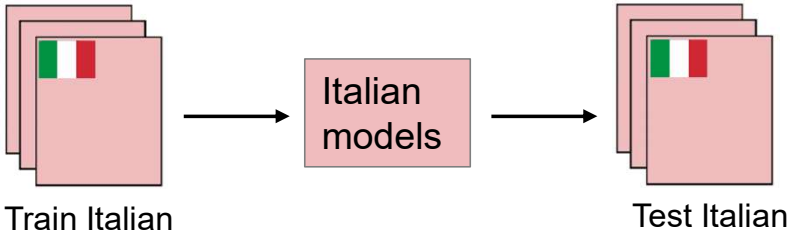
**NO PP WAS FULLY COMPLIANT**

- Italian PPs longer and more detailed
- **Selection criteria:** different market sectors, global relevance, users number
- **Manually annotated by legal expert**

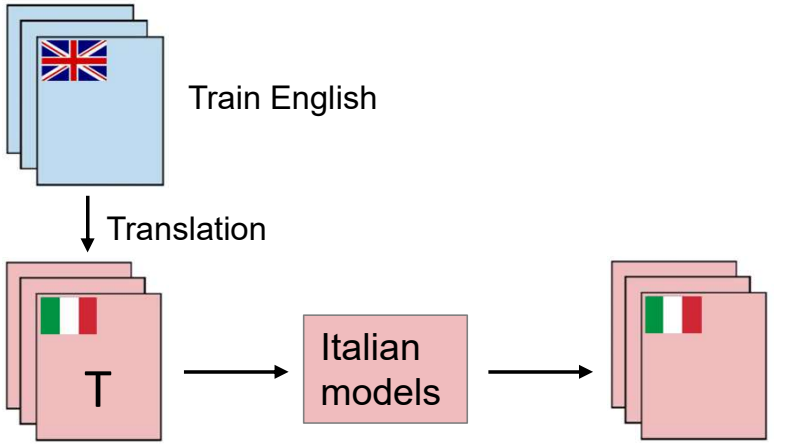
# Tasks


Category
"The Services include features that use precise geolocation data,
including GPS signals, device sensors, Wi-Fi access points, and cell tower IDs."
SubCategory
SubCategory
SubCategory
SubCategory
sub-elements detection tasks
classification tasks
cat
non-cat
Sufficiently informative
Insufficiently informative
Type Open
Type Closed
Kind:
GeoInfo
ContactInfo
Payment
DeviceInfo
… (30 labels)

# Cross-lingual setting
## with BERT models

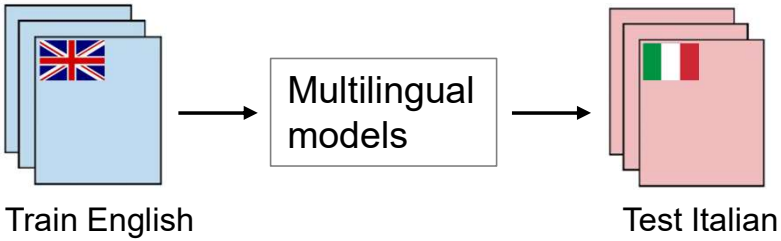



*Learning on Italian data*

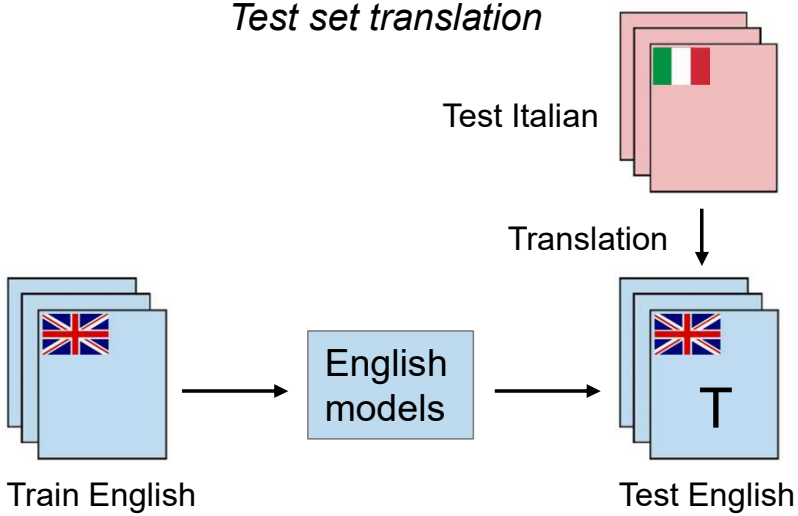



*Training set translation*

*Multilingual models*




*Test set translation*



# Cross-lingual results
## (macro F1 score)

| Method and Model | cat | Inform. | Type | Kind |
|---|---|---|---|---|
| *Learning on Italian data* | | | | |
| Italian BERT | **0.86** | 0.79 | **0.80** | 0.44 |
| *Multilingual model* | | | | |
| BERT multilingual | 0.68 | 0.65 | 0.73 | 0.30 |
| *Training set translation* | | | | |
| Italian BERT | 0.85 | 0.78 | 0.77 | 0.45 |
| UmBERTo | 0.85 | **0.80** | 0.73 | 0.32 |
| *Test set translation* | | | | |
| DeBERTa | 0.85 | **0.80** | 0.79 | 0.43 |

- Comparable results in all settings, except for the multilingual model
- Top score for INFORMATIVENESS when training in English and translating the test set at inference time (most convenient setting in term of resources)

# LLM experimental setting

- **Zero-shot**: definitions from the annotation guidelines
- **Static few-shot**: 15 random examples
  - balancing the number of examples per class
- **Dynamic few-shot**: 15 most similar examples to the given query
  - k-NN algorithm

- **Models**: Gemini 2.0 Flash, Llama-3.1-8B-Instruct

# Classification results
## (macro F1 score)

| Method and Model | cat | Inform. | Type | Kind |
|---|---|---|---|---|
| *No learning* | | | | |
| gemini zero-shot | 0.85 | 0.64 | 0.67 | **0.56** |
| llama zero-shot | 0.77 | 0.64 | 0.69 | 0.44 |
| *Learning on Italian data* | | | | |
| gemini static few-shot | 0.83 | 0.70 | 0.65 | 0.54 |
| gemini dynamic knn few-shot | **0.86** | 0.72 | 0.76 | 0.54 |
| llama static few-shot | 0.73 | 0.67 | 0.63 | 0.45 |
| llama dynamic knn few-shot | 0.77 | 0.71 | 0.73 | 0.51 |
| Italian BERT | **0.86** | **0.79** | **0.80** | 0.44 |
| DeBERTa | 0.85 | **0.80** | 0.79 | 0.43 |

- BERT > generative LLMs
  - Except for Kind
- Zero-shot > static few-shot
- Dynamic few-shot best among LLMs

- Example selection is crucial

# Detection results
## (F1 score)

| Model | Category-Subcategory | | | | | | Specification | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *B-C* | *I-C* | *B-S* | *I-S* | *O* | Avg. | *B-Sp* | *I-Sp* | *O* | Avg. |
| *BIO tagging* | | | | | | | | | | |
| gemini zero-shot | 0.48 | 0.51 | 0.71 | 0.72 | 0.85 | 0.65 | 0.64 | 0.51 | 0.80 | 0.65 |
| gemini few-shot | 0.66 | 0.60 | 0.81 | 0.77 | 0.88 | 0.75 | 0.51 | 0.69 | 0.82 | 0.67 |
| llama few-shot | 0.56 | 0.51 | 0.67 | 0.62 | 0.81 | 0.63 | 0.29 | 0.33 | 0.62 | 0.41 |
| Italian BERT | 0.74 | 0.68 | 0.81 | 0.80 | 0.90 | **0.79** | 0.74 | 0.91 | 0.93 | **0.86** |
| UmBERTo | 0.67 | 0.64 | 0.84 | 0.80 | 0.91 | 0.77 | 0.69 | 0.94 | 0.96 | **0.86** |
| *IO tagging* | | | | | | | | | | |
| gemini zero-shot | | 0.56 | | 0.76 | 0.85 | 0.72 | | 0.52 | 0.80 | 0.66 |
| gemini few-shot | | 0.63 | | 0.81 | 0.88 | 0.77 | | 0.70 | 0.82 | 0.76 |
| llama few-shot | | 0.54 | | 0.65 | 0.81 | 0.67 | | 0.34 | 0.62 | 0.48 |
| Italian BERT | | 0.70 | | 0.83 | 0.90 | **0.81** | | 0.89 | 0.92 | 0.91 |
| UmBERTo | | 0.68 | | 0.84 | 0.90 | **0.81** | | 0.92 | 0.94 | **0.93** |

- BERT > generative LLMs
- Few-shot > zero-shot

# Error Analysis

- We analyze the results of the **test set translation** approach
- Lexicon issues – differences in the domain-specific terminology in the two languages

*«Esaminiamo anche il tuo indirizzo IP per valutare…»*

*"process"* in original English

*"review"* in translated English

- Lack of context – when classifying the sub-elements

*«Deduciamo le vostre generalità (quali la fascia d'età e il genere)…»*

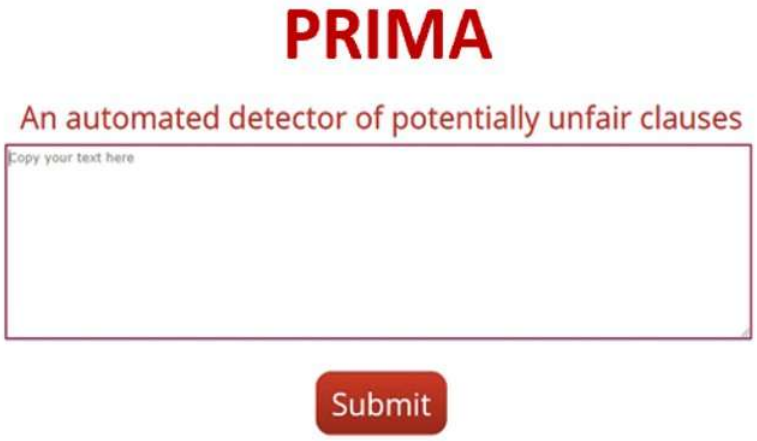
*"gender"* in original English

*"genre"* in translated English

# Conclusions and Future works

- BERT-based models yield better results than LLM
- English resources can be used to train a system for Italian with comparable results
  - Potentially generalizable to other languages
- New directions for LLM example selection:
  - Expert selected instead of random
  - Change number of examples
  - Dynamic selection with similarity+variety

Working on AI-empowered tool to detect potentially unlawful clauses in PPs



# THANK YOU!

Detecting Vague Clauses in Italian Privacy Policies using Transformers, LLMs, and Cross-Lingual Techniques

**28th European Conference on Artificial Intelligence**

# Hierarchical Annotation System

**Element and Sub-elements:**

- **<Category>** is the genus of personal data processed [e.g. “your location”]

This element may include two optional sub-elements:

- **<Specification>** is the term that introduces a list of sub-categories [e.g., “such as”]

- **<SubCategory>** is the data denoting a species of the given genus [e.g., “GPS information”]

**Mandatory and Optional Attributes:**

**<Type>:** The precision of terms describing the data;

**<Kind>**: the terms describing the category of data and so identifying its content;

**<Ref>:** The link between the subelement and Category it refers to.

# Hierarchical Annotation System

| Sufficiently Informative (Level="1") | | Insufficiently Informative (Level="2") | |
|---|---|---|---|
| Rule 1 | Category Type = "Closed"<br>No Specification<br>No Subcategories | Rule 1 | Category Type = "Open"<br>No Specification<br>No Subcategories |
| Rule 2 | Category Type = "Closed"<br>Specification Type = "Closed"/"Open"<br>SubCategory Type = "Closed"/"Open" | Rule 2 | Category Type = "Open"<br>Specification Type = "Open"<br>Subcategory Type = "Closed"/"Open" |
| Rule 3 | Category Type = "Open"<br>Specification Type = "Closed"<br>SubCategory Type = "Closed" | Rule 3 | Category Type = "Open"<br>Specification Type = "Closed"<br>Subcategory Type = "Open" |

# Results extended

| Method and Model | cat | | | | Informativeness | | | | Type | | | | Kind | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | cat | non-cat | Avg. | $\sigma$ | suff. | insuff. | Avg. | $\sigma$ | open | closed | Avg. | $\sigma$ | Avg. | $\sigma$ |
| *No learning* | | | | | | | | | | | | | | |
| gemini zero-shot | 0.75 | 0.95 | 0.85 | - | 0.56 | 0.72 | 0.64 | - | 0.63 | 0.70 | 0.67 | - | **0.56** | - |
| llama zero-shot | 0.61 | 0.92 | 0.77 | - | 0.51 | 0.77 | 0.64 | - | 0.72 | 0.66 | 0.69 | - | 0.44 | - |
| *Learning on Italian data* | | | | | | | | | | | | | | |
| gemini static few-shot | 0.72 | 0.94 | 0.83 | - | 0.61 | 0.79 | 0.70 | - | 0.59 | 0.70 | 0.65 | - | 0.54 | - |
| gemini dynamic knn few-shot | **0.76** | 0.95 | **0.86** | - | 0.64 | 0.79 | 0.72 | - | 0.76 | 0.77 | 0.76 | - | 0.54 | - |
| llama static few-shot | 0.57 | 0.89 | 0.73 | - | 0.56 | 0.78 | 0.67 | - | 0.57 | 0.69 | 0.63 | - | 0.45 | - |
| llama dynamic knn few-shot | 0.64 | 0.91 | 0.77 | - | 0.62 | 0.79 | 0.71 | - | 0.71 | 0.74 | 0.73 | - | 0.51 | - |
| ITALIAN-LEGAL-BERT | 0.72 | 0.95 | 0.83 | 0.008 | 0.62 | 0.86 | 0.74 | 0.027 | 0.79 | 0.76 | 0.78 | 0.005 | 0.37 | 0.020 |
| Italian BERT | **0.76** | **0.96** | **0.86** | 0.011 | 0.69 | 0.88 | 0.79 | 0.016 | **0.82** | **0.78** | **0.80** | 0.014 | 0.44 | 0.019 |
| UmBERTo | **0.76** | **0.96** | **0.86** | 0.008 | 0.67 | **0.89** | 0.78 | 0.043 | 0.79 | 0.74 | 0.76 | 0.013 | 0.34 | 0.036 |
| BERT multilingual | 0.75 | **0.96** | 0.85 | 0.010 | 0.68 | 0.86 | 0.77 | 0.009 | 0.81 | 0.77 | 0.79 | 0.016 | 0.41 | 0.021 |
| *Multilingual model* | | | | | | | | | | | | | | |
| BERT multilingual | 0.41 | 0.94 | 0.68 | 0.015 | 0.45 | 0.85 | 0.65 | 0.109 | 0.75 | 0.72 | 0.73 | 0.011 | 0.30 | 0.024 |
| *Training set translation* | | | | | | | | | | | | | | |
| ITALIAN-LEGAL-BERT | 0.71 | 0.95 | 0.83 | 0.004 | 0.68 | 0.87 | 0.78 | 0.030 | 0.76 | 0.75 | 0.76 | 0.008 | 0.38 | 0.007 |
| Italian BERT | 0.73 | **0.96** | 0.85 | 0.005 | 0.68 | 0.88 | 0.78 | 0.056 | 0.76 | **0.78** | 0.77 | 0.020 | 0.45 | 0.026 |
| UmBERTo | 0.73 | **0.96** | 0.85 | 0.003 | **0.72** | 0.88 | **0.80** | 0.022 | 0.73 | 0.73 | 0.73 | 0.024 | 0.32 | 0.049 |
| BERT multilingual | 0.75 | **0.96** | 0.85 | 0.017 | 0.67 | 0.88 | 0.78 | 0.026 | 0.78 | 0.76 | 0.77 | 0.019 | 0.43 | 0.001 |
| *Test set translation* | | | | | | | | | | | | | | |
| LEGAL-BERT | 0.74 | **0.96** | 0.85 | 0.016 | 0.67 | 0.87 | 0.77 | 0.015 | 0.80 | **0.78** | 0.79 | 0.006 | 0.45 | 0.015 |
| DistilRoBERTa | 0.73 | **0.96** | 0.85 | 0.008 | 0.71 | 0.88 | **0.80** | 0.013 | 0.79 | 0.75 | 0.77 | 0.012 | 0.45 | 0.003 |
| DeBERTa | 0.74 | **0.96** | 0.85 | 0.017 | **0.72** | **0.89** | **0.80** | 0.005 | 0.80 | **0.78** | 0.79 | 0.003 | 0.43 | 0.009 |

| Method and Model | cat | | | | Informativeness | | | | Type | | | | Kind | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | cat | non-cat | Avg. | $\sigma$ | suff. | insuff. | Avg. | $\sigma$ | open | closed | Avg. | $\sigma$ | Avg. | $\sigma$ |
| *Learning on English data* | | | | | | | | | | | | | | |
| LEGAL-BERT | 0.75 | **0.96** | **0.86** | 0.004 | **0.77** | **0.92** | **0.84** | 0.029 | **0.81** | **0.79** | **0.80** | 0.009 | **0.46** | 0.017 |
| *Training set translation* | | | | | | | | | | | | | | |
| LEGAL-BERT | **0.76** | **0.96** | **0.86** | 0.014 | 0.69 | 0.91 | 0.80 | 0.025 | 0.80 | 0.76 | 0.78 | 0.016 | 0.36 | 0.017 |
| *Test set translation* | | | | | | | | | | | | | | |
| Italian BERT | 0.74 | **0.96** | 0.85 | 0.012 | 0.76 | **0.92** | **0.84** | 0.022 | 0.80 | 0.75 | 0.77 | 0.031 | 0.35 | 0.013 |