Extreme Multi-Label Legal Text Classification: A case study in EU Legislation

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras and Ion Androutsopoulos
Extreme Multi-label Text Classification (XMTC)

Product Categorization

YANIS VAROUFAKIS
Adults in the Room

MY BATTLE WITH EUROPE’S DEEP ESTABLISHMENT

‘One of the greatest political memoirs of all time’
Guardian
Adults In The Room: My Battle With Europe’s Deep Establishment (paperback)

What happens when you take on the establishment? In this blistering, personal account, world-famous economist Yanis Varoufakis blows the lid on Europe’s hidden agenda and exposes what actually goes on in its corridors of power.

Varoufakis sparked one of the most spectacular and controversial battles in recent political history when, as finance minister of Greece, he attempted to re-negotiate his country’s relationship with the EU. Despite the mass support of the Greek people and the simple logic of his arguments, […]

As is now clear, the same policies that required the tragic and brutal suppression of Greece’s democratic uprising have led directly to authoritarianism, populist revolt and instability throughout the Western world.

Adults In The Room is an urgent wake-up call to renew European democracy before it is too late.
Adults In The Room: My Battle With Europe’s Deep Establishment (paperback)

What happens when you take on the establishment? In this blistering, personal account, world-famous economist Yanis Varoufakis blows the lid on Europe’s hidden agenda and exposes what actually goes on in its corridors of power.

Varoufakis sparked one of the most spectacular and controversial battles in recent political history when, as finance minister of Greece, he attempted to re-negotiate his country’s relationship with the EU. Despite the mass support of the Greek people and the simple logic of his arguments, […]

As is now clear, the same policies that required the tragic and brutal suppression of Greece’s democratic uprising have led directly to authoritarianism, populist revolt and instability throughout the Western world.

*Adults In The Room* is an urgent wake-up call to renew European democracy before it is too late.
Classifying EU legislation with EUROVOC concepts

- *EU legislation* (e.g., EU Directives, Regulations, Decisions, etc.) is published in EUR-Lex portal (https://eur-lex.europa.eu/).

- All documents have been labeled with *EUROVOC* concepts.

- **EUROVOC** (European Vocabulary)
  - multilingual thesaurus maintained by the *Publications Office of the European Union*.
  - more than 7,000 concepts referring to various activities of the EU (e.g., politics, economics, health-care, trade, etc.).
  - concepts are represented by concept descriptors, for example: (1309, import), (693, citrus fruit), (192, health control), (2511, agri-monetary policy), (863, Spain).
EURLEX57K Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Documents</th>
<th>EUROVOC concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>EURLEX (Mencia et al., 2007)</td>
<td>up to 2007</td>
<td>19,601</td>
<td>3,993</td>
</tr>
<tr>
<td>EURLEX57K (ours)</td>
<td>up-to-date</td>
<td>57,000</td>
<td>4,271</td>
</tr>
</tbody>
</table>

- Subsets: **training** (45,000 documents), **development** (6,000), and **test** (6,000)
- Avg. document length = **720 words**
- Avg. labels per documents = **5 labels**
COUNCIL DECISION of 31 July 1972 setting up a Standing Committee for Agricultural Statistics (72/279/EEC)

Having regard to the Treaty establishing the European Economic Community;
Whereas, for the purpose of facilitating their implementation, provision is made in acts adopted by the Council relating to agricultural statistics [...] 

HAS DECIDED AS FOLLOWS:

Article 1
A Standing Committee for Agricultural Statistics (hereinafter called the "Committee") is hereby set up; it shall consist of representatives of the Member States with a representative of the Commission as Chairman.

Article 2
The Committee shall carry out the duties assigned to it by the provisions adopted by the Council in the field of agricultural statistics in the cases and under the conditions provided for therein. It may, moreover, consider any other question arising in connection with such provisions and referred to it by the Chairman either on his own initiative or at the request of a Member State.
COUNCIL DECISION of 31 July 1972 setting up a Standing Committee for Agricultural Statistics (72/279/EEC)

THE COUNCIL OF THE EUROPEAN COMMUNITIES

Having regard to the Treaty establishing the European Economic Community;
Having regard to the draft decision submitted by the Commission;
Whereas, for the purpose of facilitating their implementation, provision is made in acts adopted by the Council relating to agricultural statistics for a procedure establishing close cooperation between Member States and the Commission; whereas, in order to achieve such cooperation, a Committee should be set up to carry out the duties assigned to it by such acts;
Whereas it is desirable that this Committee should consider any other question arising in connection with such acts.

HAS DECIDED AS FOLLOWS:

Article 1
A Standing Committee for Agricultural Statistics (hereinafter called the "Committee") is hereby set up; it shall consist of representatives of the Member States with a representative of the Commission as Chairman.

Article 2
The Committee shall carry out the duties assigned to it by the provisions adopted by the Council in the field of agricultural statistics in the cases and under the conditions provided for therein. It may, moreover, consider any other question arising in connection with such provisions and referred to it by the Chairman either on his own initiative or at the request of a Member State.

Article 3
The Committee shall adopt its own rules of procedure.

Done at Brussels, 31 July 1972.
For the Council The President T. WESTERTERP
COUNCIL DECISION of 31 July 1972 setting up a Standing Committee for Agricultural Statistics (72/279/EEC)

THE COUNCIL OF THE EUROPEAN COMMUNITIES

Having regard to the Treaty establishing the European Economic Community;

Having regard to the draft decision submitted by the Commission;

Whereas, for the purpose of facilitating their implementation, provision is made in acts adopted by the Council relating to agricultural statistics for a procedure establishing close cooperation between Member States and the Commission; whereas, in order to achieve such cooperation, a Committee should be set up to carry out the duties assigned to it by such acts;

Whereas it is desirable that this Committee should consider any other question arising in connection with such acts.

HAS DECIDED AS FOLLOWS:

Article 1
A Standing Committee for Agricultural Statistics (hereinafter called the "Committee") is hereby set up; it shall consist of representatives of the Member States with a representative of the Commission as Chairman.

Article 2
The Committee shall carry out the duties assigned to it by the provisions adopted by the Council in the field of agricultural statistics in the cases and under the conditions provided for therein. It may, moreover, consider any other question arising in connection with such provisions and referred to it by the Chairman either on his own initiative or at the request of a Member State.

Article 3
The Committee shall adopt its own rules of procedure.

Done at Brussels, 31 July 1972.
For the Council The President T. WESTERTERP
COUNCIL DECISION of 31 July 1972 setting up a Standing Committee for Agricultural Statistics (72/279/EEC)

THE COUNCIL OF THE EUROPEAN COMMUNITIES

Having regard to the Treaty establishing the European Economic Community;

Having regard to the draft decision submitted by the Commission;

Whereas, for the purpose of facilitating their implementation, provision is made in acts adopted by the Council relating to agricultural statistics for a procedure establishing close cooperation between Member States and the Commission; whereas, in order to achieve such cooperation, a Committee should be set up to carry out the duties assigned to it by such acts;

Whereas it is desirable that this Committee should consider any other question arising in connection with such acts.

HAS DECIDED AS FOLLOWS:

Article 1
A Standing Committee for Agricultural Statistics (hereinafter called the "Committee") is hereby set up; it shall consist of representatives of the Member States with a representative of the Commission as Chairman.

Article 2
The Committee shall carry out the duties assigned to it by the provisions adopted by the Council in the field of agricultural statistics in the cases and under the conditions provided for therein. It may, moreover, consider any other question arising in connection with such provisions and referred to it by the Chairman either on his own initiative or at the request of a Member State.

Article 3
The Committee shall adopt its own rules of procedure.

Done at Brussels, 31 July 1972.
For the Council The President T. WESTERTERP
COUNCIL DECISION of 31 July 1972 setting up a Standing Committee for Agricultural Statistics (72/279/EEC)

THE COUNCIL OF THE EUROPEAN COMMUNITIES

Having regard to the Treaty establishing the European Economic Community;
Having regard to the draft decision submitted by the Commission;
Whereas, for the purpose of facilitating their implementation, provision is made in acts adopted by the Council relating to agricultural statistics for a procedure establishing close cooperation between Member States and the Commission; whereas, in order to achieve such cooperation, a Committee should be set up to carry out the duties assigned to it by such acts;
Whereas it is desirable that this Committee should consider any other question arising in connection with such acts.

HAS DECIDED AS FOLLOWS:

Article 1
A Standing Committee for Agricultural Statistics (hereinafter called the "Committee") is hereby set up; it shall consist of representatives of the Member States with a representative of the Commission as Chairman.

Article 2
The Committee shall carry out the duties assigned to it by the provisions adopted by the Council in the field of agricultural statistics in the cases and under the conditions provided for therein. It may, moreover, consider any other question arising in connection with such provisions and referred to it by the Chairman either on his own initiative or at the request of a Member State.

Article 3
The Committee shall adopt its own rules of procedure.

Done at Brussels, 31 July 1972.
For the Council The President T. WESTERTERP
EURLEX57K Dataset (Labels distribution)

While EUROVOC includes over 7,000 concepts (labels):

- only 4,271 (59.31%) of them are present in EURLEX57K.
- only 2,049 (47.97%) have been assigned >10 documents.

<table>
<thead>
<tr>
<th>Label Group</th>
<th>#Labels</th>
<th>Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequent</td>
<td>746</td>
<td>$D_{train} &gt; 50$</td>
</tr>
<tr>
<td>few-shot</td>
<td>3,362</td>
<td>$1 &gt; D_{train} \geq 50$</td>
</tr>
<tr>
<td>zero-shot</td>
<td>163</td>
<td>$D_{train} = 0$</td>
</tr>
</tbody>
</table>

12
Methods Considered (Baselines)

We experimented with two baselines:

- **Exact Match** (e.g., “We suggest the change of price for *citrus fruit* in Europe.”)
- **Logistic Regression classifier with TF-IDF scores**
Methods Considered (Neural Approaches)

BIGRUs with self-attention (BIGRU-ATT) - (Xu et al., 2015)

\[
\alpha_t = \frac{\exp(h_f^T u)}{\sum_j \exp(h_j^T u)}
\]

\[
d = \frac{1}{T} \sum_{t-1}^T \alpha_t h_t
\]
Methods Considered (Neural Approaches)

BIGRUs with self-attention (BIGRU-ATT) - (Xu et al., 2015)

![Diagram of BIGRU-ATT]

\[ a_t = \text{softmax}(h_t^\top u) \]

\[ d = \frac{1}{T} \sum_{t=1}^{T} a_t h_t \]

Trainable components \( u \)
Methods Considered (Neural Approaches)

BIGNUs with self-attention (BIGNU-ATT) - (Xu et al., 2015)

$$a_t = \text{softmax}(h_t^T u)$$
$$d = \frac{1}{T} \sum_{t=1}^{T} a_t h_t$$
$$p = \text{sigmoid}(W_o d + b_o)$$

Trainable components
$$u, W_o, b_o$$
Methods Considered (Neural Approaches)

Hierarchical Attention Network (HAN) - (Yang et al., 2016)

\[ v_{it} = \tanh(W^{(s)} h_{it} + b^{(s)}) \]
\[ a_{it}^{(s)} = \text{softmax}(v_{it}^T u^{(s)}) \]
\[ c_i = \frac{1}{T_i} \sum_{t=1}^{T_i} a_{it}^{(s)} h_{it} \]

Trainable components
\[ W^{(s)}, b^{(s)}, u^{(s)} \]
Methods Considered (Neural Approaches)

Hierarchical Attention Network (HAN) - (Yang et al., 2016)

\[ v_i = \tanh(W^{(d)} c_i + b^{(d)}) \]

\[ a_i^{(d)} = \text{softmax}(v_i^T u^{(d)}) \]

\[ d = \frac{1}{S} \sum_{i=1}^{S} a_i^{(d)} c_i \]

Trainable components

\[ W^{(d)}, b^{(d)}, u^{(d)} \]
Methods Considered (Neural Approaches)

Hierarchical Attention Network (HAN) - (Yang et al., 2016)

\[ v_i = \tanh(W^{(d)} c_i + b^{(d)}) \]
\[ a_i^{(d)} = \text{softmax}(v_i^T u^{(d)}) \]
\[ d = \frac{1}{S} \sum_{i=1}^{S} a_i^{(d)} c_i \]
\[ p = \text{sigmoid}(W_o d + b_o) \]

Trainable components

\[ W^{(d)}, b^{(d)}, u^{(d)}, W_o, b_o \]
Methods Considered (Neural Approaches)

MaxPooling over Hierarchical Section Scorers (MAX-HSS)

\[
c_i = \frac{1}{T_i} \sum_{t=1}^{T_i} a_{i,t}^{(s)} h_{it}
\]

\[
p_i^{(s)} = \text{sigmoid}(W^{(m)} c_i + b^{(m)})
\]

\[
p^{(d)} = \text{maxpool}(p_1^{(s)}, \ldots, p_S^{(s)})
\]

Trainable components

\[W^{(m)}, b^{(m)}\]
Methods Considered (Neural Approaches)

Label-Wise Attention Network (LWAN) - (Mullenbach et al., 2018)

\[
a_{lt} = \text{softmax}(h_t^\top u_l) \\
d_l = \sum_t a_{lt} h_t \ (l = 1, \ldots, L)
\]

Trainable components

\[
u_l (l = 1, \ldots, L)
\]
Methods Considered (Neural Approaches)

Label-Wise Attention Network (LWAN) - (Mullenbach et al., 2018)

\[
a_{lt} = \text{softmax}(h_l^\top u_l)
\]

\[
d_l = \sum_t a_{lt} h_t \ (l = 1, \ldots, L)
\]

\[
\varphi_l = \text{sigmoid}(W_{lo} d_l + b_{lo})
\]

Trainable components

\[
W_{lo}, b_{lo} \ (l = 1, \ldots, L)
\]
Methods Considered (Neural Approaches)

Zero Label-Wise Attention Network (Z-LWAN) - (Rios and Kavuluru, 2018)

\[ w_l = \frac{1}{E} \sum_{e=1}^{E} w_{le} \]

**Example**

\[ w_{l1} = E(\text{agricultural}) \]
\[ w_{l2} = E(\text{statistics}) \]

Trainable components

\[ \times \]
Methods Considered (Neural Approaches)

Zero Label-Wise Attention Network (Z-LWAN) - (Rios and Kavuluru, 2018)

\[
\mathbf{u}_l = \frac{1}{E} \sum_{e=1}^{E} \mathbf{w}_{le}
\]

\[
a_{lt} = \text{softmax}(h_t^\top \mathbf{u}_l)
\]

\[
d_l = \sum_{t} a_{lt} h_t (l = 1, \ldots, L)
\]

Trainable components
Methods Considered (Neural Approaches)

Zero Label-Wise Attention Network (Z-LWAN) - (Rios and Kavuluru, 2018)

\[ u_l = \frac{1}{E} \sum_{e=1}^{E} w_{le} \]

\[ a_{lt} = \text{softmax}(h_t^T u_l) \]

\[ d_l = \sum_t a_{lt} h_t \ (l = 1, \ldots, L) \]

\[ p_l = \text{sigmoid}(u_l^T d_l) \ (l = 1, \ldots, L) \]

Trainable components

\[ \text{X} \]
Evaluation Measures

- **Macro-averaged** measures are very sensitive to infrequent labels.
- We question the suitability of $\text{R@K}$ and $\text{P@K}$ for XMTC.
- $\text{R@K}$ penalizes methods when documents have $>K$ gold labels.
- $\text{P@K}$ penalizes methods when documents have $<K$ gold labels.

\[
\text{RP@K} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\text{Rel}(n,k)}{\min(K, R_n)}
\]

\[
\text{nDCG@K} = \frac{1}{N} \sum_{n=1}^{N} Z_{Kn} \sum_{k=1}^{K} \frac{2^{\text{Rel}(n,k)} - 1}{\log_2 (1 + k)}
\]
Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>ALL LABELS</th>
<th>FREQUENT</th>
<th>FEW</th>
<th>ZERO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$RP@5$</td>
<td>$nDCG@5$</td>
<td>$RP@5$</td>
<td>$nDCG@5$</td>
</tr>
<tr>
<td>Exact Match</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.097</td>
<td>0.099</td>
<td>0.120</td>
<td></td>
</tr>
<tr>
<td>BIGRU-ATT</td>
<td>0.758</td>
<td>0.789</td>
<td>0.689</td>
<td></td>
</tr>
<tr>
<td>HAN</td>
<td>0.746</td>
<td>0.778</td>
<td>0.680</td>
<td></td>
</tr>
<tr>
<td>CNN-LWAN</td>
<td>0.716</td>
<td>0.746</td>
<td>0.642</td>
<td></td>
</tr>
<tr>
<td>BIGRU-LWAN</td>
<td><strong>0.766</strong></td>
<td><strong>0.796</strong></td>
<td><strong>0.698</strong></td>
<td></td>
</tr>
<tr>
<td>Z-CNN-LWAN</td>
<td>0.684</td>
<td>0.717</td>
<td>0.618</td>
<td></td>
</tr>
<tr>
<td>Z-BIGRU-LWAN</td>
<td>0.718</td>
<td>0.752</td>
<td>0.652</td>
<td></td>
</tr>
<tr>
<td>ENSEMBLE-LWAN</td>
<td><strong>0.766</strong></td>
<td><strong>0.796</strong></td>
<td><strong>0.698</strong></td>
<td></td>
</tr>
<tr>
<td>MAX-HSS</td>
<td>0.737</td>
<td>0.773</td>
<td>0.671</td>
<td></td>
</tr>
<tr>
<td>LW-HAN</td>
<td>0.721</td>
<td>0.761</td>
<td>0.669</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results on EURLEX57K for all, frequent (> 50 training instances), few-shot (1 to 50 instances), and zero-shot labels. All the differences between the best (bold) and other methods are statistically significant ($p < 0.01$).
### Experimental Results

<table>
<thead>
<tr>
<th></th>
<th><strong>ALL LABELS</strong></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>RP@5</strong></td>
<td><strong>nDCG@5</strong></td>
<td><strong>Micro-F1</strong></td>
<td><strong>RP@5</strong></td>
<td><strong>nDCG@5</strong></td>
<td><strong>RP@5</strong></td>
<td><strong>nDCG@5</strong></td>
<td><strong>RP@5</strong></td>
</tr>
<tr>
<td><strong>Exact Match</strong></td>
<td>0.097</td>
<td>0.099</td>
<td>0.120</td>
<td>0.219</td>
<td>0.201</td>
<td>0.111</td>
<td>0.074</td>
<td>0.194</td>
</tr>
<tr>
<td><strong>Logistic Regression</strong></td>
<td>0.710</td>
<td>0.741</td>
<td>0.539</td>
<td>0.767</td>
<td>0.781</td>
<td>0.508</td>
<td>0.470</td>
<td>0.011</td>
</tr>
<tr>
<td><strong>BIGRU-ATT</strong></td>
<td>0.758</td>
<td>0.789</td>
<td>0.689</td>
<td>0.799</td>
<td>0.813</td>
<td>0.631</td>
<td>0.580</td>
<td>0.040</td>
</tr>
<tr>
<td><strong>HAN</strong></td>
<td>0.746</td>
<td>0.778</td>
<td>0.680</td>
<td>0.789</td>
<td>0.805</td>
<td>0.597</td>
<td>0.544</td>
<td>0.051</td>
</tr>
<tr>
<td><strong>CNN-LWAN</strong></td>
<td>0.716</td>
<td>0.746</td>
<td>0.642</td>
<td>0.761</td>
<td>0.772</td>
<td>0.613</td>
<td>0.557</td>
<td>0.036</td>
</tr>
<tr>
<td><strong>BIGRU-LWAN</strong></td>
<td><strong>0.766</strong></td>
<td><strong>0.796</strong></td>
<td><strong>0.698</strong></td>
<td><strong>0.805</strong></td>
<td><strong>0.819</strong></td>
<td><strong>0.662</strong></td>
<td><strong>0.618</strong></td>
<td><strong>0.029</strong></td>
</tr>
<tr>
<td><strong>Z-CNN-LWAN</strong></td>
<td>0.684</td>
<td>0.717</td>
<td>0.618</td>
<td>0.730</td>
<td>0.745</td>
<td>0.495</td>
<td>0.454</td>
<td>0.321</td>
</tr>
<tr>
<td><strong>Z-BIGRU-LWAN</strong></td>
<td>0.718</td>
<td>0.752</td>
<td>0.652</td>
<td>0.764</td>
<td>0.780</td>
<td>0.561</td>
<td>0.510</td>
<td><strong>0.438</strong></td>
</tr>
<tr>
<td><strong>ENSEMBLE-LWAN</strong></td>
<td><strong>0.766</strong></td>
<td><strong>0.796</strong></td>
<td><strong>0.698</strong></td>
<td><strong>0.805</strong></td>
<td><strong>0.819</strong></td>
<td><strong>0.662</strong></td>
<td><strong>0.618</strong></td>
<td><strong>0.438</strong></td>
</tr>
<tr>
<td><strong>MAX-HSS</strong></td>
<td>0.737</td>
<td>0.773</td>
<td>0.671</td>
<td>0.784</td>
<td>0.803</td>
<td>0.463</td>
<td>0.443</td>
<td>0.039</td>
</tr>
<tr>
<td><strong>LW-HAN</strong></td>
<td>0.721</td>
<td>0.761</td>
<td>0.669</td>
<td>0.766</td>
<td>0.790</td>
<td>0.412</td>
<td>0.402</td>
<td>0.039</td>
</tr>
</tbody>
</table>

Table 2: Results on EURLEX57K for all, frequent (> 50 training instances), few-shot (1 to 50 instances), and zero-shot labels. All the differences between the best (bold) and other methods are statistically significant ($p < 0.01$).
# Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>ALL LABELS</th>
<th>FREQUENT</th>
<th>FEW</th>
<th>ZERO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$RP@5$</td>
<td>$nDCG@5$</td>
<td>$RP@5$</td>
<td>$nDCG@5$</td>
</tr>
<tr>
<td>Exact Match</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.097</td>
<td>0.099</td>
<td>0.120</td>
<td>0.219</td>
</tr>
<tr>
<td>BIGRU-ATT</td>
<td>0.710</td>
<td>0.741</td>
<td>0.539</td>
<td>0.767</td>
</tr>
<tr>
<td>HAN</td>
<td>0.746</td>
<td>0.788</td>
<td>0.680</td>
<td>0.799</td>
</tr>
<tr>
<td>CNN-LWAN</td>
<td>0.716</td>
<td>0.746</td>
<td>0.642</td>
<td>0.761</td>
</tr>
<tr>
<td>BIGRU-LWAN</td>
<td><strong>0.766</strong></td>
<td><strong>0.796</strong></td>
<td><strong>0.698</strong></td>
<td><strong>0.805</strong></td>
</tr>
<tr>
<td>Z-CNN-LWAN</td>
<td>0.684</td>
<td>0.717</td>
<td>0.618</td>
<td>0.730</td>
</tr>
<tr>
<td>Z-BIGRU-LWAN</td>
<td>0.718</td>
<td>0.752</td>
<td>0.652</td>
<td>0.764</td>
</tr>
<tr>
<td>ENSEMBLE-LWAN</td>
<td><strong>0.766</strong></td>
<td><strong>0.796</strong></td>
<td><strong>0.698</strong></td>
<td><strong>0.805</strong></td>
</tr>
<tr>
<td>MAX-HSS</td>
<td>0.737</td>
<td>0.773</td>
<td>0.671</td>
<td>0.784</td>
</tr>
<tr>
<td>LW-HAN</td>
<td>0.721</td>
<td>0.761</td>
<td>0.669</td>
<td>0.766</td>
</tr>
</tbody>
</table>

Table 2: Results on EURLEX57K for all, frequent (> 50 training instances), few-shot (1 to 50 instances), and zero-shot labels. All the differences between the best (bold) and other methods are statistically significant ($p < 0.01$).
## Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>ALL LABELS</th>
<th></th>
<th></th>
<th>FREQUENT</th>
<th></th>
<th></th>
<th>FEW</th>
<th></th>
<th></th>
<th>ZERO</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RP@5</td>
<td>nDCG@5</td>
<td>Micro-F1</td>
<td>RP@5</td>
<td>nDCG@5</td>
<td>RP@5</td>
<td>nDCG@5</td>
<td>RP@5</td>
<td>nDCG@5</td>
<td>RP@5</td>
<td>nDCG@5</td>
<td></td>
</tr>
<tr>
<td>Exact Match</td>
<td>0.710</td>
<td>0.741</td>
<td>0.539</td>
<td>0.767</td>
<td>0.781</td>
<td>0.508</td>
<td>0.470</td>
<td>0.011</td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.758</td>
<td>0.789</td>
<td>0.689</td>
<td>0.799</td>
<td>0.813</td>
<td>0.631</td>
<td>0.580</td>
<td>0.040</td>
<td>0.027</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIGRU-ATT</td>
<td>0.746</td>
<td>0.778</td>
<td>0.680</td>
<td>0.789</td>
<td>0.805</td>
<td>0.597</td>
<td>0.544</td>
<td>0.051</td>
<td>0.034</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAN</td>
<td>0.716</td>
<td>0.746</td>
<td>0.642</td>
<td>0.761</td>
<td>0.772</td>
<td>0.613</td>
<td>0.557</td>
<td>0.036</td>
<td>0.023</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN-LWAN</td>
<td>0.766</td>
<td>0.796</td>
<td>0.698</td>
<td>0.805</td>
<td>0.819</td>
<td>0.662</td>
<td>0.618</td>
<td>0.029</td>
<td>0.019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIGRU-LWAN</td>
<td>0.766</td>
<td>0.796</td>
<td>0.698</td>
<td>0.805</td>
<td>0.819</td>
<td>0.662</td>
<td>0.618</td>
<td>0.038</td>
<td>0.045</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-CNN-LWAN</td>
<td>0.321</td>
<td>0.264</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td>0.345</td>
<td>0.345</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-BIGRU-LWAN</td>
<td>0.438</td>
<td>0.345</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td>0.345</td>
<td>0.345</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSEMBLE-LWAN</td>
<td>0.766</td>
<td>0.796</td>
<td>0.698</td>
<td>0.805</td>
<td>0.819</td>
<td>0.662</td>
<td>0.618</td>
<td>0.038</td>
<td>0.045</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX-HSS</td>
<td>0.737</td>
<td>0.773</td>
<td>0.671</td>
<td>0.784</td>
<td>0.803</td>
<td>0.463</td>
<td>0.443</td>
<td>0.039</td>
<td>0.028</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LW-HAN</td>
<td>0.721</td>
<td>0.761</td>
<td>0.669</td>
<td>0.766</td>
<td>0.790</td>
<td>0.412</td>
<td>0.402</td>
<td>0.039</td>
<td>0.026</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results on EURLEX57K for all, frequent (> 50 training instances), few-shot (1 to 50 instances), and zero-shot labels. All the differences between the best (bold) and other methods are statistically significant ($p < 0.01$).

**Zero-shot learning (Intact Prior Knowledge)**
## Experimental Results

| Method          | All Labels | | | Frequent | | | Few | | | Zero | | |
|-----------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
|                 | $RP@5$     | $nDCG@5$   | Micro-$F_1$| $RP@5$     | $nDCG@5$   | $RP@5$     | $nDCG@5$   | $RP@5$     | $nDCG@5$   | $RP@5$     | $nDCG@5$   |
| Exact Match     | 0.097      | 0.099      | 0.120      | 0.219      | 0.201      | 0.111      | 0.074      | 0.194      | 0.186      |            |
| Logistic Regression | 0.710  | 0.741      | 0.539      | 0.767      | 0.781      | 0.508      | 0.470      | 0.011      | 0.011      |            |
| BIGRU-ATT       | 0.758      | 0.789      | 0.689      | 0.799      | 0.813      | 0.631      | 0.580      | 0.040      | 0.027      |            |
| HAN             | 0.746      | 0.778      | 0.680      | 0.789      | 0.805      | 0.597      | 0.544      | 0.051      | 0.034      |            |
| CNN-LWAAN       | 0.716      | 0.746      | 0.642      | 0.761      | 0.772      | 0.613      | 0.557      | 0.036      | 0.023      |            |
| BIGRU-LWAAN     | 0.766      | 0.796      | 0.698      | 0.805      | 0.819      | 0.662      | 0.618      | 0.029      | 0.019      |            |
| Z-CNN-LWAAN     | 0.684      | 0.717      | 0.618      | 0.730      | 0.745      | 0.495      | 0.454      | 0.321      | 0.264      |            |
| Z-BIGRU-LWAAN   | 0.718      | 0.752      | 0.652      | 0.764      | 0.780      | 0.561      | 0.510      | 0.438      | 0.345      |            |
| ENSEMBLE-LWAAN  | 0.766      | 0.796      | 0.698      | 0.805      | 0.819      | 0.662      | 0.618      | 0.438      | 0.345      |            |
| MAX-HSS         | 0.737      | 0.773      | 0.671      | 0.784      | 0.803      | 0.463      | 0.443      | 0.039      | 0.028      |            |
| LW-HAN          | 0.721      | 0.761      | 0.669      | 0.766      | 0.790      | 0.412      | 0.402      | 0.039      | 0.026      |            |

Table 2: Results on EURLEX57K for all, frequent (> 50 training instances), few-shot (1 to 50 instances), and zero-shot labels. All the differences between the best (bold) and other methods are statistically significant ($p < 0.01$).
Providing Evidence through Attention (BIGRU-ATT)

**True concepts:** chemical product | cosmetic product | toxic substance

**COMMISSION DIRECTIVE**

of 11 February 1982


THE COMMISSION OF THE EUROPEAN COMMUNITIES,

Having regard to the Treaty establishing the European Economic Community, Having regard to Council Directive 76/768/EEC of 27 July 1976 on the approximation of the laws of the Member States relating to cosmetic products (1), as last amended by Directive 79/661/EEC (2) and in particular Article 8 (2) thereof, Whereas according to the results of the most recent scientific and technical research the use of acetyl ethyl tetramethyl tetralin should be prohibited, account being taken of its neurotoxic effects harmful to health; Whereas the provisions of this Directive are in accordance with the opinion of the Committee on the Adaptation to Technical Progress of the Directives on the removal of technical barriers to trade in the cosmetic products sector,

**Article 1**

The following number is hereby added to Annex II to Council Directive 76/768/EEC:

'362 3'-ethyl-5',6',7',8'-tetrahydro-5',6',8',8'-tetramethyl-2'-ace phthone;

Syn.: 1,1,4,4-tetramethyl-6-ethyl-7-acetyl-1,2,3,4-tetrahydropaphene (acetyl ethyl tetramethyl tetralin, AETT)'.

**Article 2**

Member States shall bring into force the laws, regulations or administrative provisions necessary to comply with this Directive by 31 December 1982 at the latest and shall forthwith inform the Commission thereof.

**Article 3**

This Directive is addressed to the Member States.

**Predicted concepts:** cosmetic product | approximation of laws | chemical product | technological change | analytical chemistry
Providing Evidence through Attention (BIGRU-LWAN)

**True concepts:** chemical product | cosmetic product | toxic substance

**COMMISSION DIRECTIVE**

of 11 February 1982


THE COMMISSION OF THE EUROPEAN COMMUNITIES,

Having regard to the Treaty establishing the European Economic Community, Having regard to Council Directive 76/768/EEC of 27 July 1976 on the approximation of the laws of the Member States relating to cosmetic products (1), as last amended by Directive 79/661/EEC (2), and in particular Article 8 (2) thereof, Whereas according to the results of the most recent scientific and technical research the use of acetyl ethyl tetramethyl tetralin should be prohibited, account being taken of its neurotoxic effects harmful to health; Whereas the provisions of this Directive are in accordance with the opinion of the Committee on the Adaptation to Technical Progress of the Directives on the removal of technical barriers to trade in the cosmetic products sector, HAS ADOPTED THIS DIRECTIVE:

**Article 1**

The following number is hereby added to Annex II to Council Directive 76/768/EEC:

'362 3'-ethyl-5',6',7,8'-tetrahydro-5',6',8',8'-tetramethyl-2'-ace phthone;
Syn. : 1,1,4,4-tetramethyl-6-ethyl-7-acetyl-1,2,3,4-tetrahydroanaph e (acetyl ethyl tetramethyl tetralin, AETT)'.

**Article 2**

Member States shall bring into force the laws, regulations or administrative provisions necessary to comply with this Directive by 31 December 1982 at the latest and shall forthwith inform the Commission thereof.

**Article 3**

This Directive is addressed to the Member States.

**Predicted concepts:** cosmetic product | approximation of laws | chemical product | technological change | health risk
Conclusions

- Comparison of **various neural methods** on a new legal XMTC dataset, EURLEX57K.
- Investigation for **few-shot** and **zero-shot learning**.
- **BIGRU-ATT** is a **strong baseline** for this XMTC dataset, **outperforming** CNN-LWAN.
- **SOTA** by **replacing vanilla CNN** (BIGRU-LWAN) with BIGRU encoder.
- **Z-GRU-LWAN** - Rios and Kavuluru, (2018) SOTA on **zero-shot labels**.
- **HAN** (Yang et al., 2016) and other hierarchical methods (MAX-HSS, LW-HAN) are **weaker compared to the other neural methods**.
Future Work

- Analyze the impact of:
  - each zone (*header, recitals, main body, attachments*).
  - alternative word embeddings
    (e.g., *Law2Vec* (Chalkidis and Kampas, 2019), *ELMo embeddings* (Peters et al., 2018)).
- Release the beast! Try employing *BERT* (Devlin et al., 2019) over EURLEX57K…


*Code on GitHub* ([https://github.com/iliaschalkidis/lmtc-eurlex57k](https://github.com/iliaschalkidis/lmtc-eurlex57k))

- Alternative options (*Dilated CNNs, Transformers*) vs. computationally intensive BIGRUs.
- Experiment with *more datasets* from other domains.
Questions?

Collaboration ideas? Project proposals?
Contact us!