INTERNATIONAL HELLENIC UNIVERSITY

DEPARTMENT OF INFORMATION AND ELECTRONIC ENGINEERING / IHU

AN ENSEMBLE ARCHITECTURE OF CLASSIFIERS FOR PATENT CLASSIFICATION

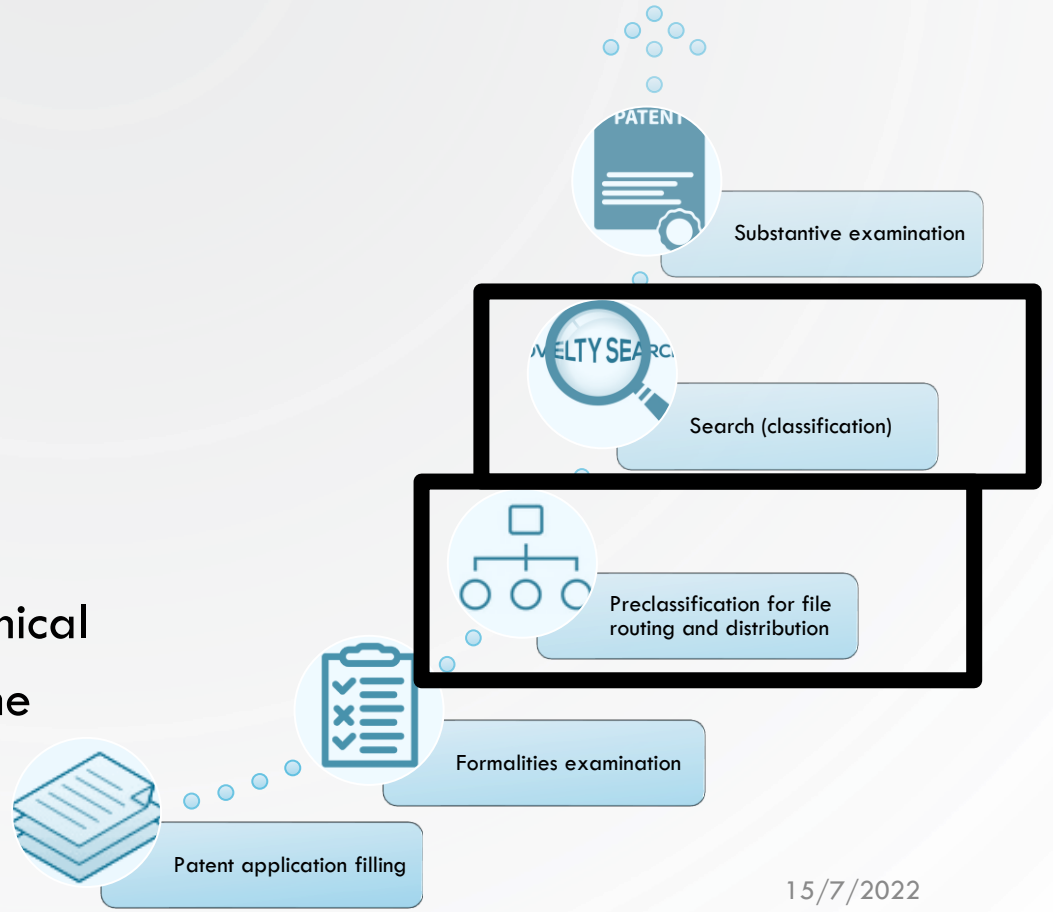
AUTHORS: ELENI KAMATERI AND MICHAIL SALAMPASIS
PRESENTER: ELENI KAMATERI

SIGIR 2022 Madrid

PATENTSEMTECH 2022

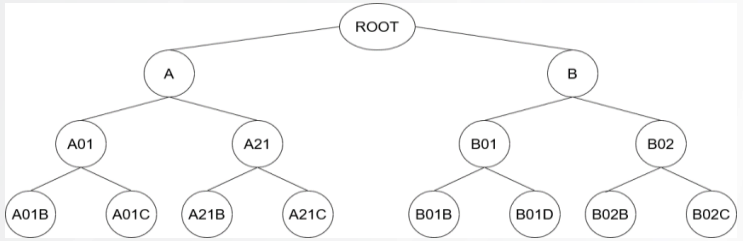
THE GLOBAL PATENT SYSTEM: HOW IT WORKS

- Preclassification: the assignment of one or more classification codes
- Quite important task:
  - it will enable correct routing and distribution to search division
  - it will enable patents with similar technical features to be grouped under the same classification code



* EPO, Guidelines for Examination in the European Patent Office, 2021

THE PROBLEM OF PATENT CLASSIFICATION



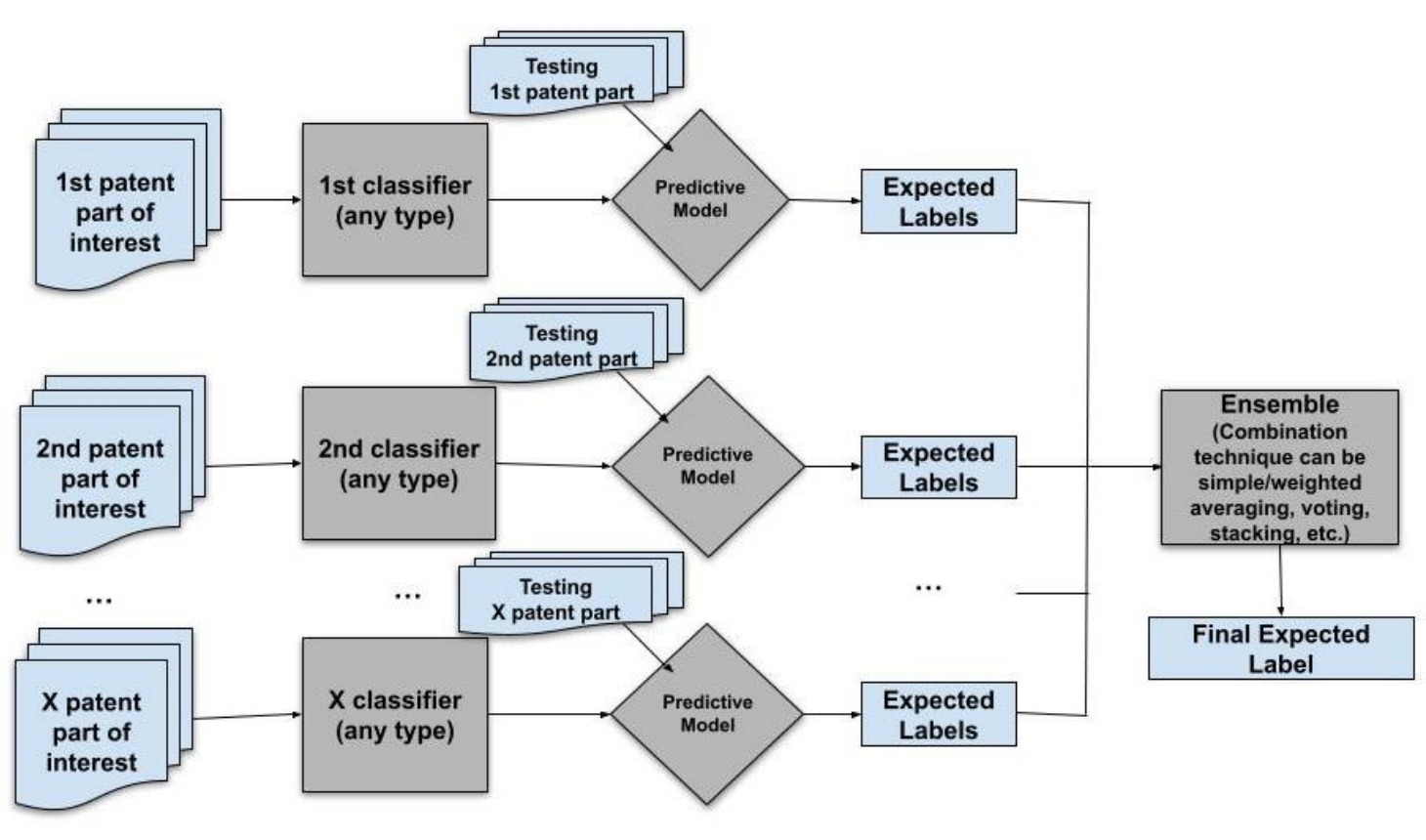
Level	IPC code	Description
Section	A	Human necessities
Class	A01	Agriculture; forestry; animal husbandry; hunting; trapping; fishing
Sub-class	A01B	Soil working in agriculture or forestry; parts, details, or accessories of agricultural machines or implements, in general

- Challenges:
  - In the patent document:
    - Numerous, lengthy, full of technical terminologies, complexity of invention
  - In the classification scheme:
    - Complicated hierarchical structure, aproximetely 78,000 IPC/250,000 CPC individual codes, unbalanced distribution of patents among codes (80% of all patent documents are classified in about 20% of the codes)
  - Manual task so far => Need to be supported or fully automated by classification systems

STATE-OF-THE-ART AND OUR CONTRIBUTION

- Current research efforts in patent classification:
  - They combine NLP and ML/DL techniques [1-4]
  - They apply several simplifications:
    - Work with well-represented codes
    - Work at higher level of the classification hierarchy
- Current research efforts in patent classification with respect to ensemble techniques:
  - They apply ML methods
  - They work at higher level of the classification hierarchy
- **Our contribution:** We introduce a new ensemble architecture for automated patent classification at multiple levels (an extension of a previous work presented in [8])

ENSEMBLE ARCHITECTURE



DATA COLLECTION: CLEF-IP 2011

Patent documents that have English text

Patent documents that have main classification category

Patent documents that have all required metadata (title, abstract, description, claims, applicant, inventors)

1st data collection/pool: Patent text from title

2nd data collection/pool: Patents text from abstracts

3rd data collection/pool: Patents text from description

4th data collection/pool: Patent text from claims

5th data collection/pool: Patents text from applicants

6th data collection/pool: Patents text from inventors

```
<classification-ipc status="new">
  <main-classification status="new">C07C 323/60</main-classification>
  <further-classification status="new">A61K  31/16</further-classification>
  <further-classification status="new">C07D 209/48</further-classification>
  <further-classification status="new">C07D 213/75</further-classification>
  <further-classification status="new">C07K   5/06</further-classification>
  <further-classification status="new">A61K  38/05</further-classification>
  <further-classification status="new">C07D 295/12</further-classification>
</classification-ipc>
```

```
<abstract load-source="ep" status="new"
lang="EN">
<p>
An entertainment machine comprising a display
arranged to display a game, the display
comprising two or more zones 28, 30, 32, each
with an associated identifier 34, 36, 38. The
identifier may comprise for example a
colour ....
<img id="img-00000001" orientation="unknown"
wi="118" img-format="tif" img-
content="ad"file="00000001.tif" inline="no"
he="114"/>
</p>
</abstract>
```

Dataset used contains 541,131 patents and is available here: https://github.com/ekamater/CLEFIP2011_XML2MySQL

EXPERIMENTAL ORGANIZATION

DL model	Data utilization	Preprocessing	Language model	Ensemble combination	Task	Level
Bi-LSTM	Abstract, Description, Claims, Title, Applicants, Inventors	cleaning punctuation, symbols and numbers, and stop word removal	Domain-specific pre-trained embedding with 300 dimension [4]	Simple/Weighted averaging	Single	Subclass and Group

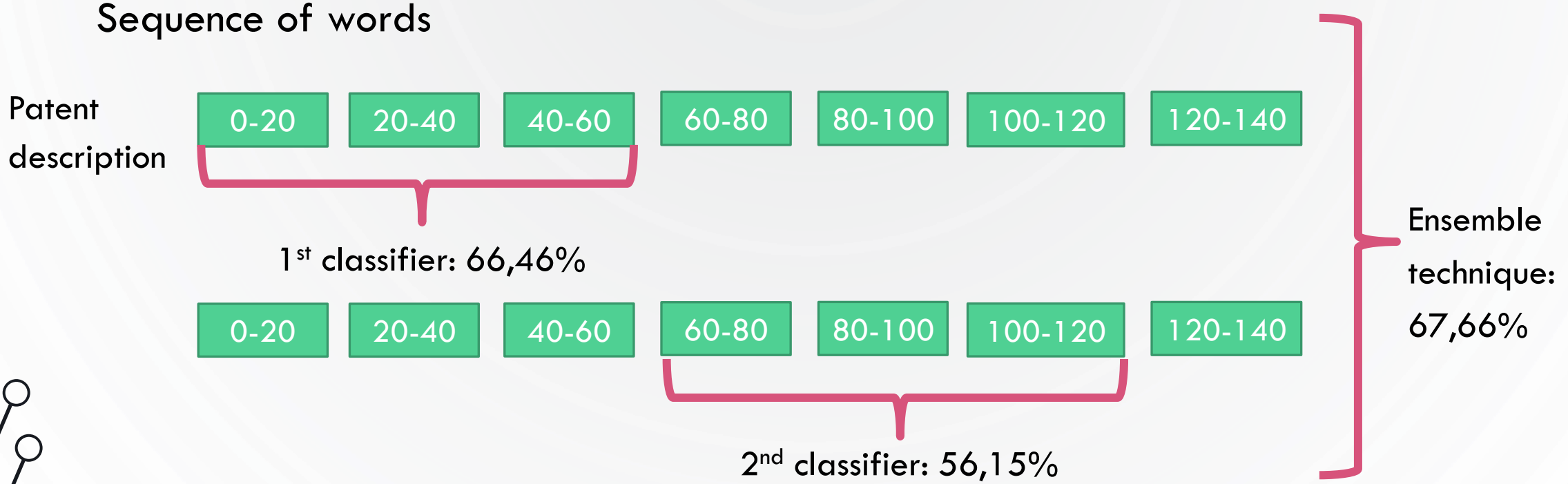
RESULTS

Subclass/Group

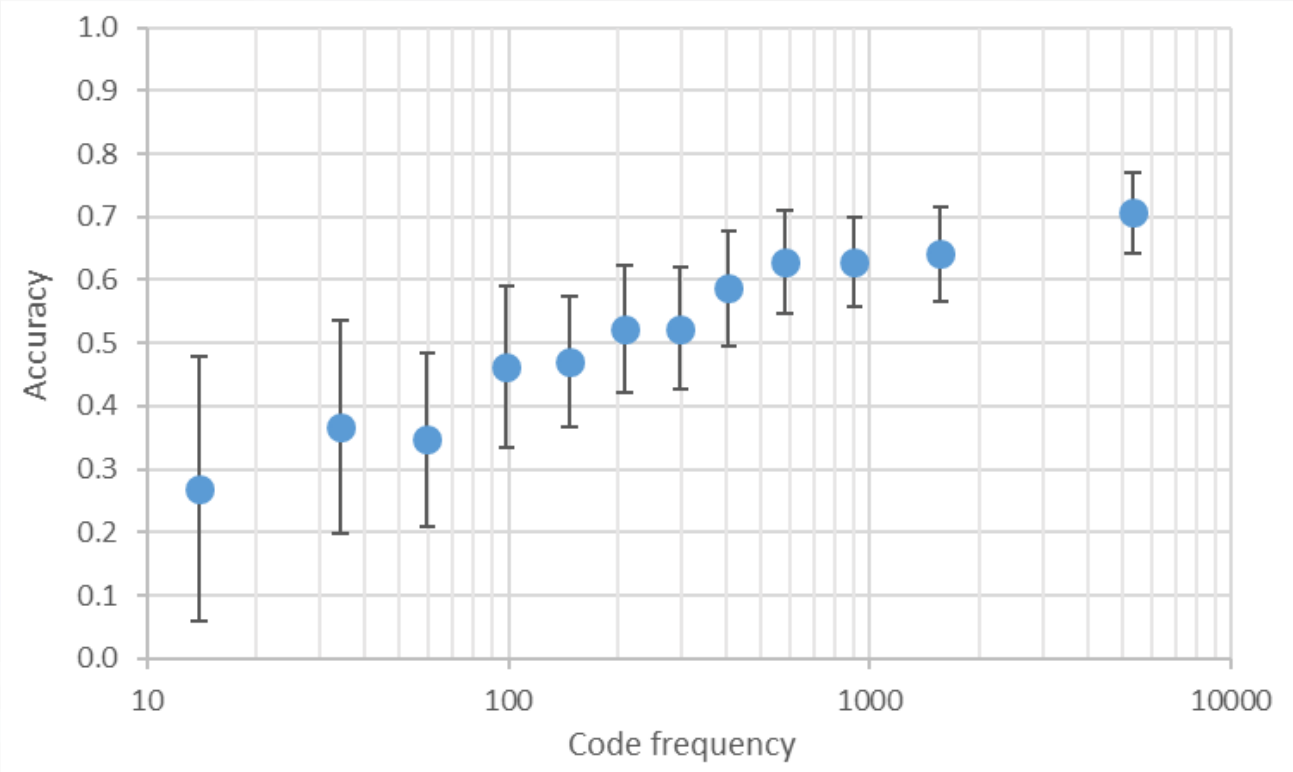
	1. Abstract	2. Description	3. Claims	4. Title	5. Applicants	6. Inventors
Individual Bi-LSTM classifier	63.76/44.68	66.46/47.23	64.56/45.10	59.58/40.74	24.32/12.93	11.52/6.01
Ensemble (simple averaging of classifiers 1-3)	70.39/52.52					
Ensemble (simple averaging of classifiers 1-6)	70.67/53.06					
Ensemble (weighted averaging of classifiers 1-6)	70.70/53.11					

INTERESTING OBSERVATION

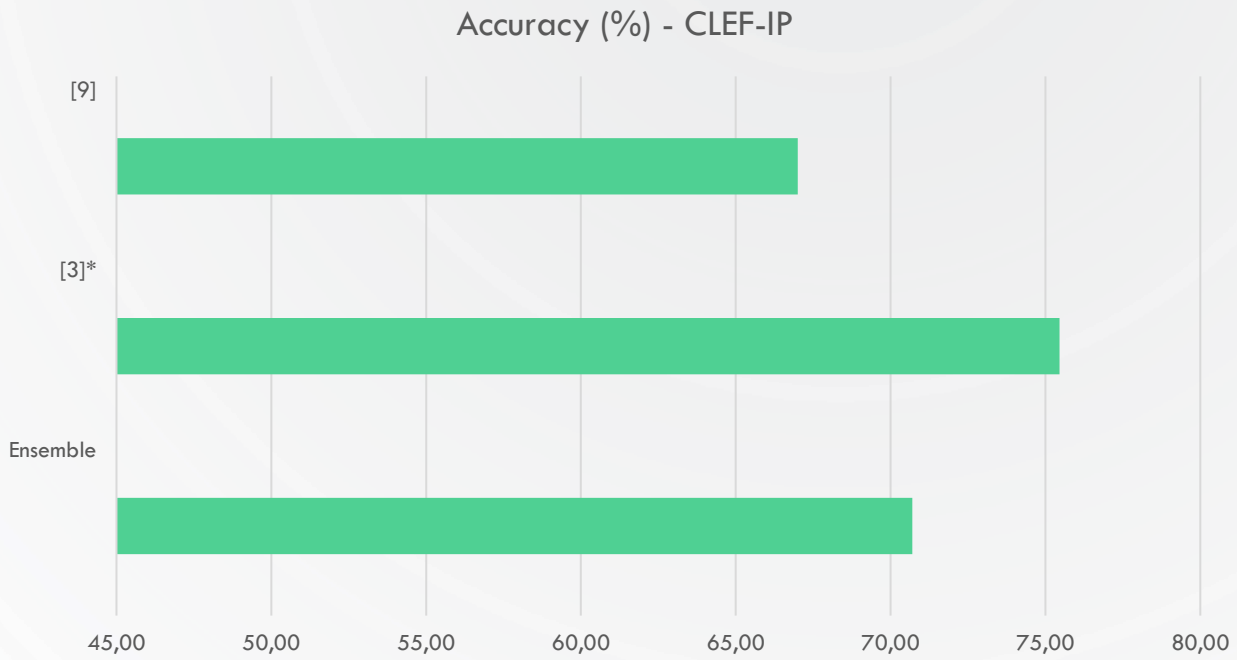
Sequence of words



INTERESTING OBSERVATION – IMPACT OF IMBALANCE DATASET

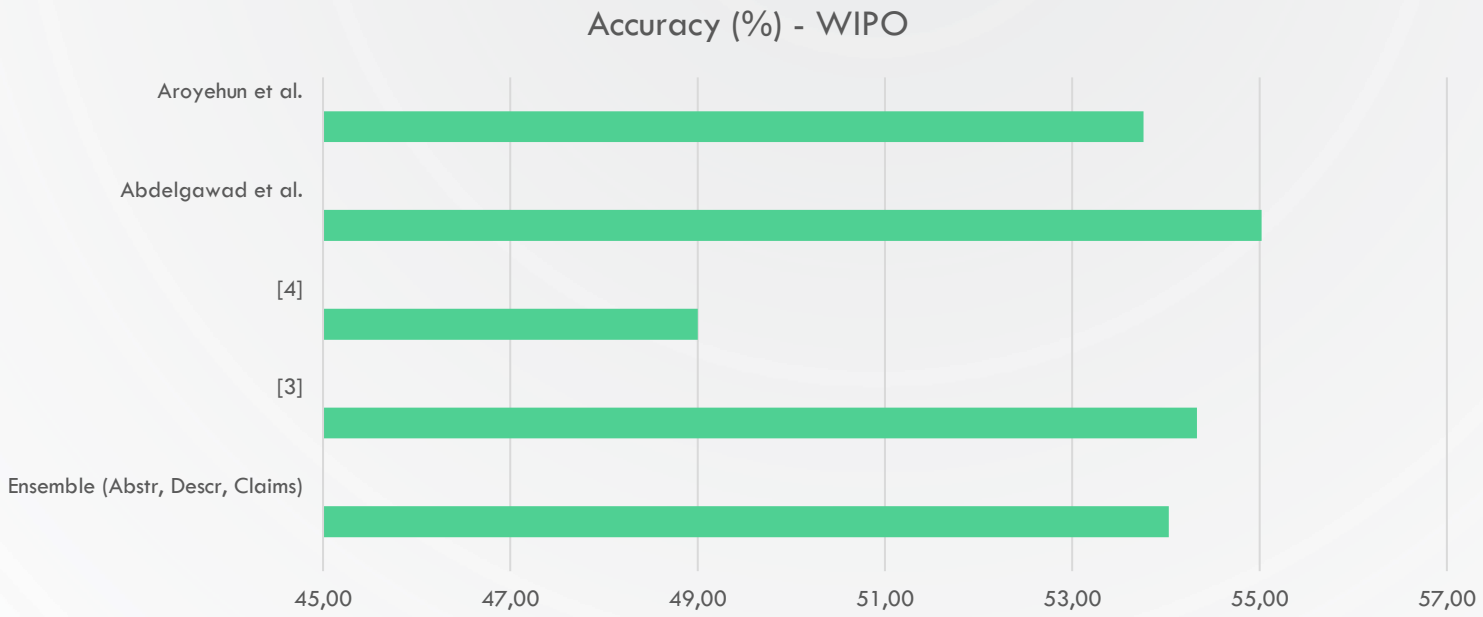


COMPARISON WITH RELATED WORK

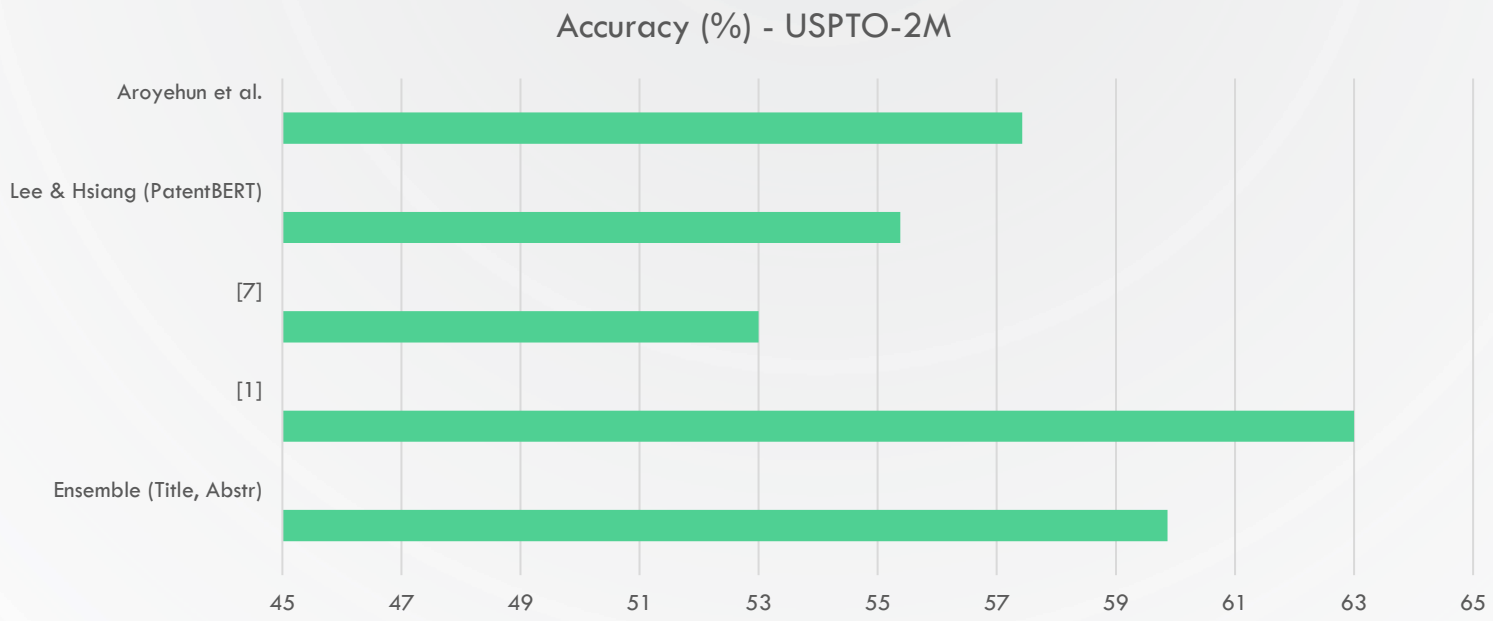


* Recall (top 4)

ONGOING WORK…



ONGOING WORK



CONCLUSIONS

- An ensemble architecture for automated patent classification was proposed.
- The ensemble architecture was instantiated in the single-label classification task at the subclass and group level category of the IPC 5+level hierarchy.
- The combination of classifiers outperform the same classifiers when used as standalone solutions and performs well compared with current state of the art (considering that we used a simple DL model without further fine-tuning).
- Combinations of different patent parts should be carefully explored (future work)
- Simplifications used in literature such as working with well-represented codes should be carefully explored (future work)

REFERENCES

[1] Grawe, M. F., Martins, C. A., & Bonfante, A. G. (2017). Automated patent classification using word embedding. *In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 408-411).

[2] Xiao, L., Wang, G., & Zuo, Y. (2018). Research on patent text classification based on word2vec and LSTM. *In 2018 11th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 1, pp. 71-74).

[3] Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117(2).

[4] Risch J. & Krestel, R. (2019). Domain-specific word embeddings for patent classification. *Data Technologies and Applications*

[5] Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137, 239-263.

[6] Mathiassen, H., & Ortiz-Arroyo, D. (2006). Automatic categorization of patent applications using classifier combinations. *In International Conference on Intelligent Data Engineering and Automated Learning* (pp. 1039-1047).

[7] Benites, F., Malmasi, S., & Zampieri, M. (2018). Classifying patent applications with ensemble methods. *arXiv preprint arXiv:1811.04695*.

[8] Kamateri, E., Stamatis, V., Diamantaras, K., & Salampasis, M. (2022). Automated Single-Label Patent Classification using Ensemble Classifiers. *ICMLC 2022*.

[9] Sofean, M. (2021). Deep learning based pipeline with multichannel inputs for patent classification. *World Patent Information*, 66, 102060.

[10] Tikk, D., Biró, G., & Törcsvári, A. (2008). A hierarchical online classifier for patent categorization. *In Emerging technologies of text mining: Techniques and applications* (pp. 244-267).

INTERNATIONAL HELLENIC UNIVERSITY

DEPARTMENT OF INFORMATION AND ELECTRONIC ENGINEERING / IHU

THANK YOU!

SIGIR 2022
Madrid

PATENTSEMTECH 2022