Horizon 2020 Programme
DG CNECT
Next-generation Internet

Isaac Newton Trust

NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

The Limits of Annotation in Machine Learning a Documents Hohfeldian Legal Entities

Ahmed Izzidien
Psychometrics Centre, Cambridge Judge Business School, University of Cambridge
ai297@cam.ac.uk

■ Introduction

▪ The Hohfeldian description of legal relations is an atomistic semiotic system, in which all legal relationships are described with only eight terms: right, duty, power, liability, immunity, disability, privilege, and no-right (Wenar 2005).

▪ To train a machine learning (ML) algorithm to recognize these terms, one requires an accurate classification of **all the agents** initiating these legal relations.

▪ Current software libraries can be limited by the scope of their named entity recognition (NER) when applied to legal documents, and to their capture of under-represented entities (Leitner, Rehm, and Moreno-Schneider 2020; Mehrabi et al. 2021).

▪ Our research question: How many labels are required to train a ML algorithm in order to identify all the agents behind Hohfeldian legal relations?

■ Methodology

BBC and Wikipedia articles were randomly collected from the web, and were hand annotated in accordance with the following rules:

▪ All possible parties to an agreement were hand labelled as Agents. Thus, a brand, e.g., BMW. The title of a person, e.g., the chief executive. A group of people, e.g., employees. The description of a group, e.g., the company board. The description of a functional entity, e.g., the court, the name of a country or city that can be party to an agreement even if the country does not use the same name today, e.g., Rhodesia. The acronym for any of the above, e.g., the SEC. The name of an entity that represents the actions of a person, people or organization, e.g., El Watan newspaper, all are labelled as Agents.

▪ Where Agents are mentioned with their title, for example, 'The President of the United States of America George Bush', each qualifying entity for the agent is hand labelled independently. Thus: 'President' is an Agent, 'United States of America' is an Agent, and 'George Bush' is an Agent.

▪ A ML algorithm was trained using Spacy 'en_core_web_en' (Explosion 2021) implemented in UBIAI, using n=2400 labels. The F1 scores are plot with the gradual addition of the labels to the training corpus (Fig.1) The plot is then extrapolated using a logarithmic trendline (Fig. 2).

▪ Hyperparameters: Iterations: 10, dropout 0.1, batch size 4, 80:20 training split.

■ Results

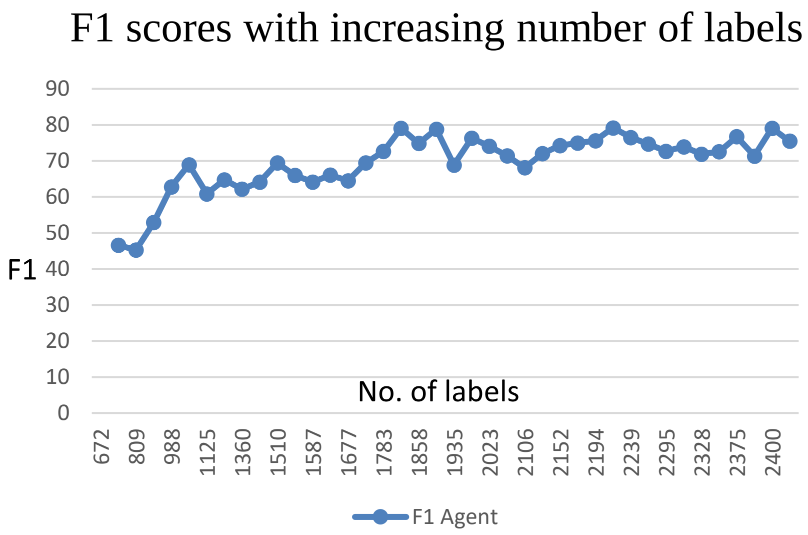

Fig. 1. Plot of F1 *vs.* number of legal entities labeled

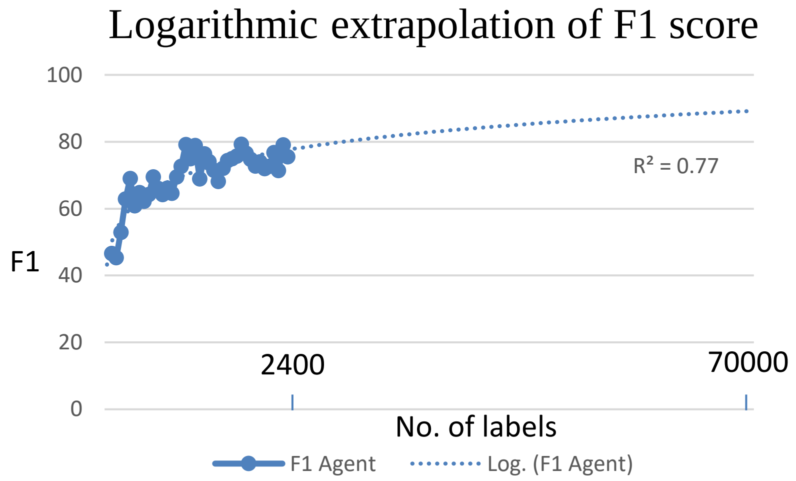

Fig 2. At 70,000 'Agent' F1 scores remain below 90.

Extrapolation: Reaches F1=100 with 10^6 labels.

■ Conclusion

The **inclusion of all legal entities**, particularly those that are under-represented in society, and by-extension, often under-represented in corpora, may require over 10^6 annotations. This being prohibitive, exploring the use of ontological abstractions is recommended for further work, whereby human agents are labeled based on hypernymy incorporating ontologies of *life*. While non-human agents, e.g., companies, are detected based on ontologies of *ownership* by *human* agents.

■ References

Explosion. 2021. Spacy: Industrial-Strength Natural Language Processing in Python (version 3.1.3). https://spacy.io.
Leitner, Elena, Georg Rehm, and Julián Moreno-Schneider. 2020. "A Dataset of German Legal Documents for Named Entity Recognition." ArXiv Preprint ArXiv:2003.13016.
Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." ACM Computing Surveys (CSUR) 54 (6): 1–35.
UBIAI "Easy to Use Text Annotation Tool" https://ubiai.tools/login.
Wenar, Leif. 2005. "The Nature of Rights." Philosophy & Public Affairs 33 (3): 223–52.