Leveraging data science to understand and address music law issues

Using Logistic Regression and NER to classify public submissions and uncover key issues and actors

INTRO

- South Africa is undergoing significant reform in copyrights and performers act.
- Public submissions to bills like the Copyright Amendment Bill reflect valuable but underutilized stakeholder input
- These submissions are often buried in dense legal language making them hard to interpret or act upon
- The creative sector, particularly the music industry needs to bridge the gap between legal data and practical understanding
- This project uses NLP to analyze and structure these submissions
- The goal is to identify key legal issues and stakeholder actors
- The final tool allows users to upload legal text and receive searchable insights

METHODS

1. Combined labelled and unlabelled documents
2. Taxonomy mapping and labelling: Matched text to taxonomy categories using semantic similarity
3. Key actor identification using Named Entity Recognition
4. Trained and evaluated three models: LegalBert,Logistic Regression,Random Forest

5. Selected the highest performing model: Logistic Regression

RESULTS

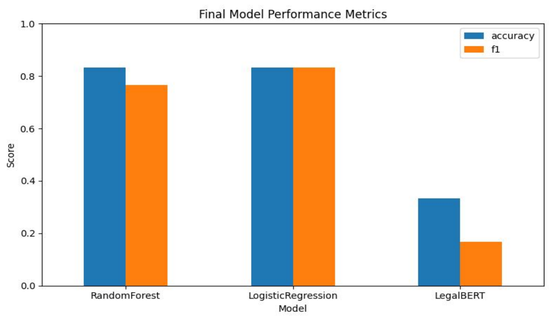

Figure 1: Model Performance Metrics

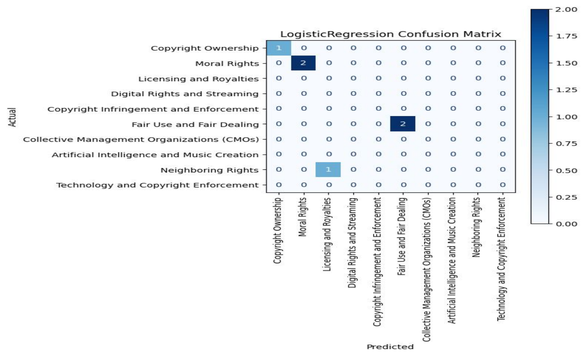

Figure 2: Logistic Regression Confusion Matrix

DISCUSSION

- We evaluated the three models on their ability to classify legal reform documents into taxonomy-defined categories
- Logistic Regression achieved the highest weighted F1 Score and accuracy of 0.83 for both metrics
- Categories like Moral Rights, Fair Use and Copyright Ownership were consistently classified correctly
- Feasibility of using NLP to classify legal reform content proven, however improvement is essential to maximize the tool's utility in real world legal settings

ADDITIONAL VISUALISATIONS

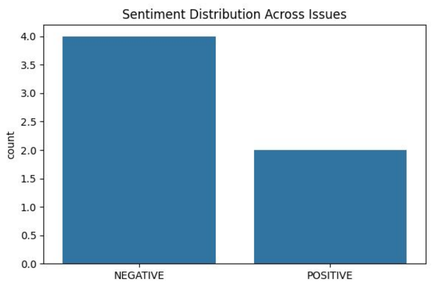

Figure 3: Sentiment Distribution Across Issues

Most documents carried negative sentiment suggesting stakeholder concern or criticism in their feedback which is an important insight for reform analysis.

Model	Accuracy	Precision (wtd)	Recall (wtd)	F1 (wtd)
RandomForest	0.833333333	0.722222222	0.833333333	0.766666667
LogisticRegression	0.833333333	0.833333333	0.833333333	0.833333333
LegalBERT	0.666666667	0.5	0.666666667	0.555555556

Figure 4: Detailed Model Performance Metrics

Logistic regression obtained the highest scores across different metrics making it the model option.

Figure 5: Training data distribution of legal documents

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me