Ontology based Multi Label Text Classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Student
Saba Sabrin

Supervisor
M.Sc. Wael Alkhatib

Prof. Dr.-Ing. Ralf Steinmetz
KOM - Multimedia Communications Lab

Agenda

Background

Motivation

Related Work

Proposed Method

Result and Evaluation

Conclusion

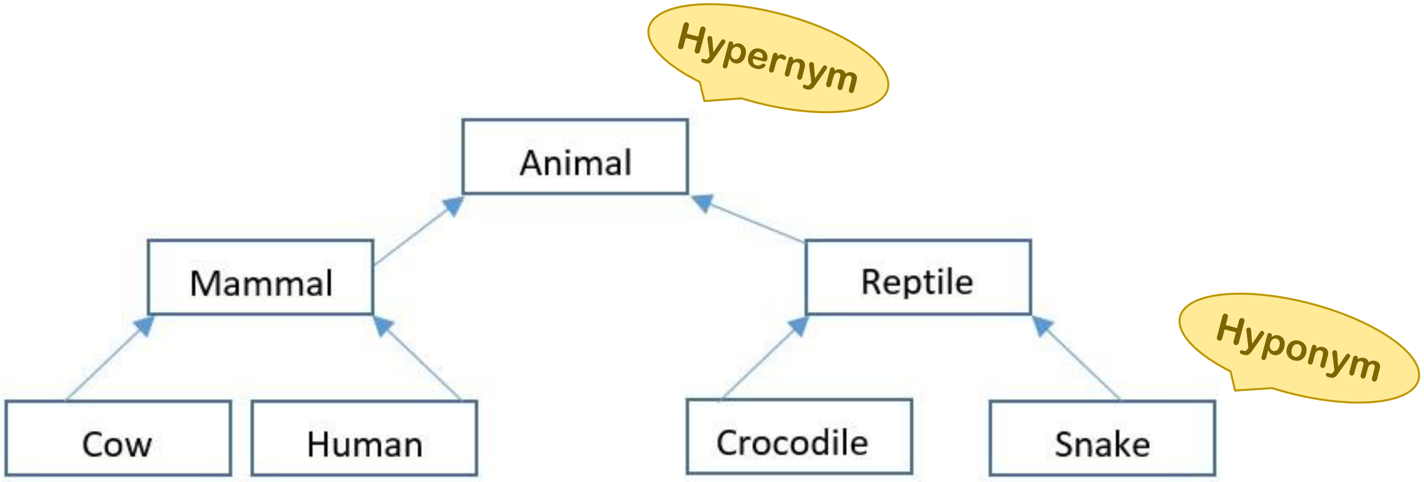
Future Work

Background

TECHNISCHE UNIVERSITÄT DARMSTADT

Ontology

- Nature of being, Domain knowledge
- is-a, part-of type relational data



Background

Multi Label Classification (MLC)

- Assignment of multiple labels or tags on a piece of data i.e. text, image, audio.
- Our focus: Document classification

News Paper

Problem transformation methods

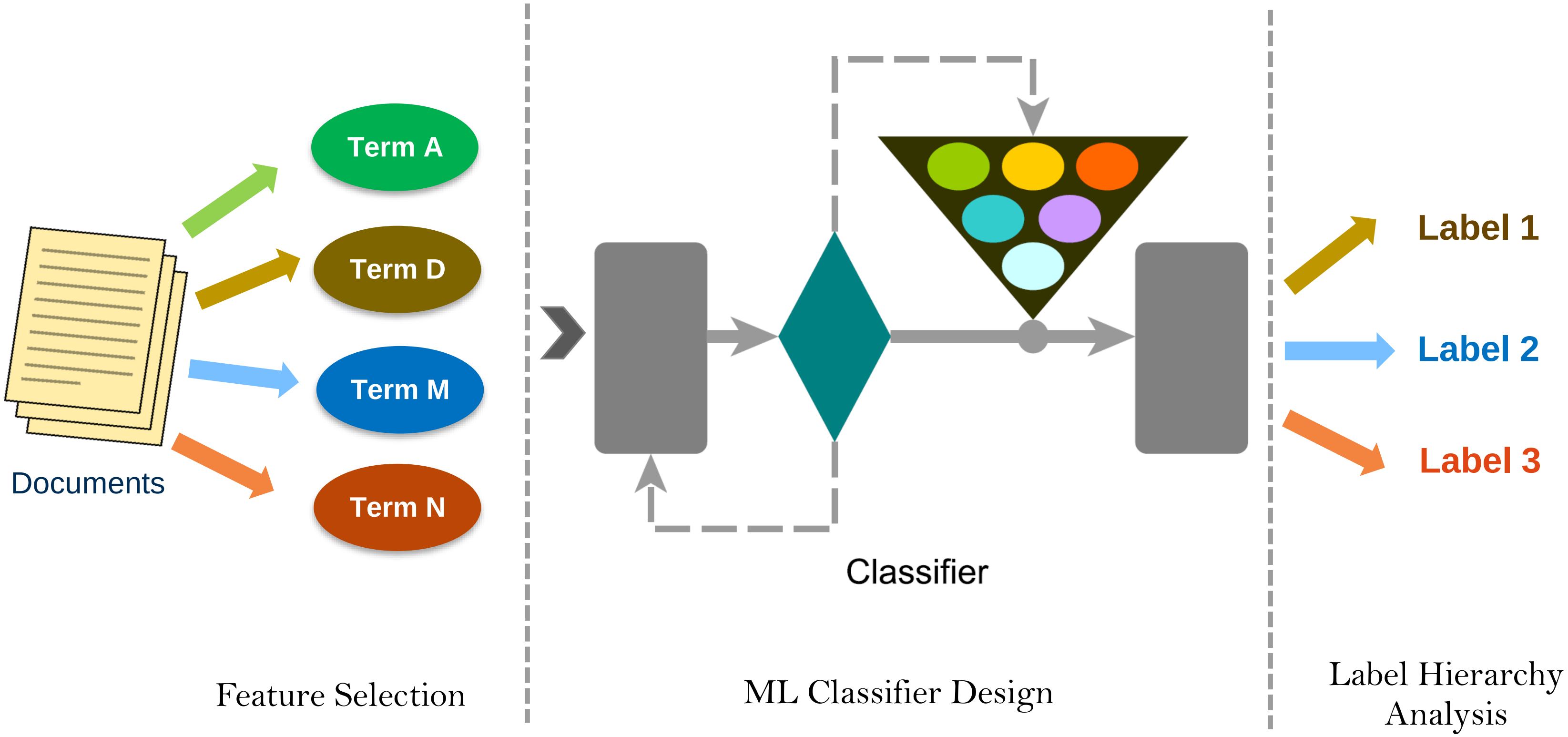
- Binary relevance
- Ranking
- Hierarchy based

Algorithm adaptation methods

- AdaBoost
- k-nearest neighborhood (kNN)
- Decision tree
- Neural Network (BP-MLL)

http://www.opentext.com/what-we-do/products/business-network/capture-and-recognition/opentext-capture-center

Traditional Multi-Label Text Classification

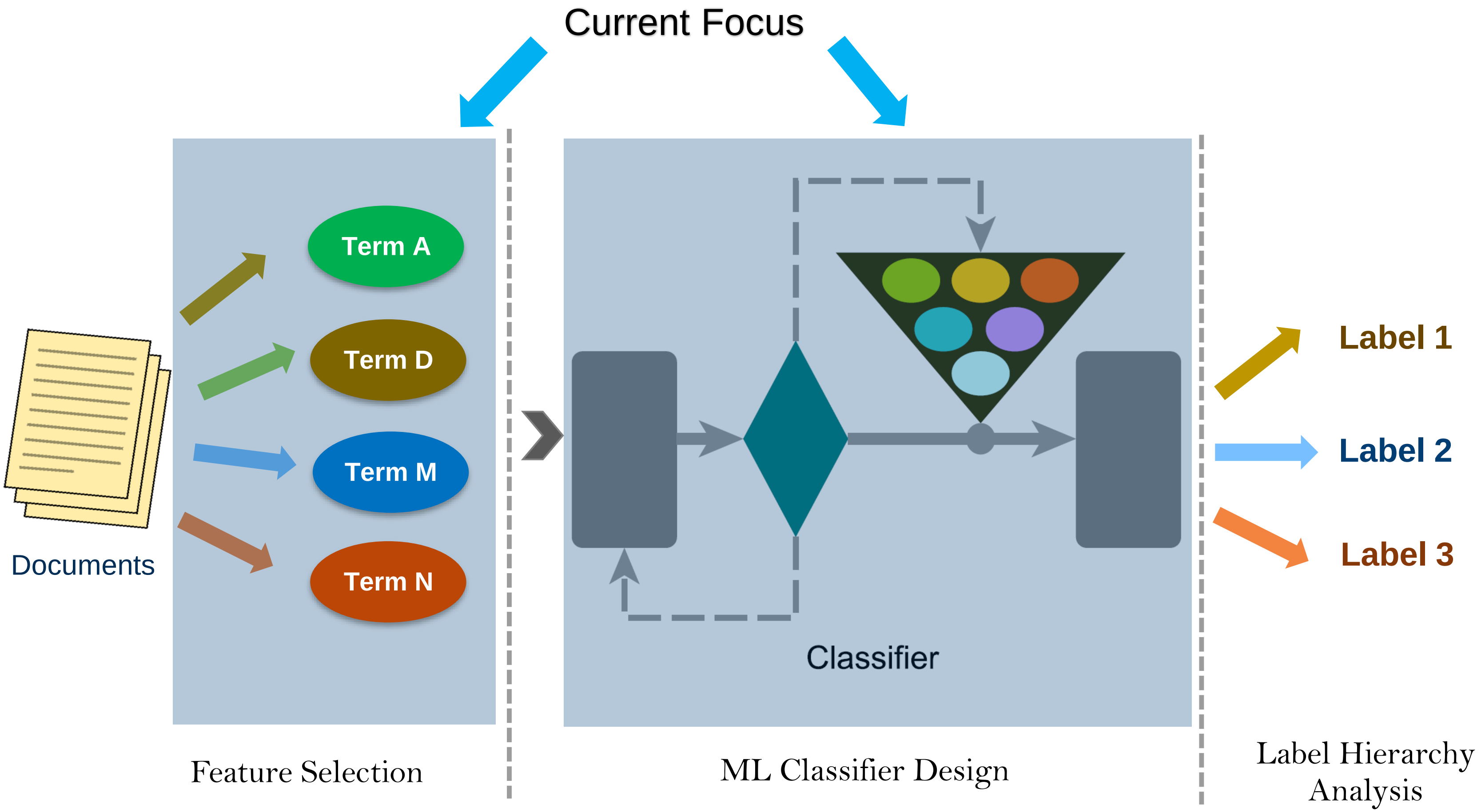


Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

Challenges	Traditional MLC	Ontology-based MLC
Higher number of feature extraction	✗	✓
Inter-Feature relationship consideration	✗	✓
Dependency between Document Features and Label Features	✗	✓
No Training for New Label	✗	✓
Reduced Training complexity	✗	✓
Improves performance on Low Frequent Label	✗	✓

Motivation

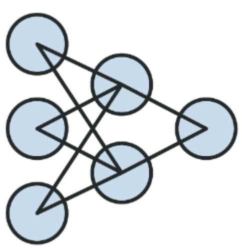


Current Focus
Term A
Term D
Term M
Term N
Documents
Classifier
Label 1
Label 2
Label 3
Feature Selection
ML Classifier Design
Label Hierarchy Analysis

Related Work

Ontology-Based Multi-Label Text Classification of Construction Regulatory Documents

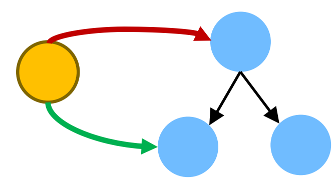
Manual process, Hierarchy based Domain Ontology



Deep-Learning, Vector Similarity based Classification

Proposed Improvement

Automatic Ontology Extraction



Feature Graph Matching based Classification

Related Work

Training-less Ontology-based Text Categorization

RDF based static Ontology

External knowledge used for building inter-related Domain & Label Ontology

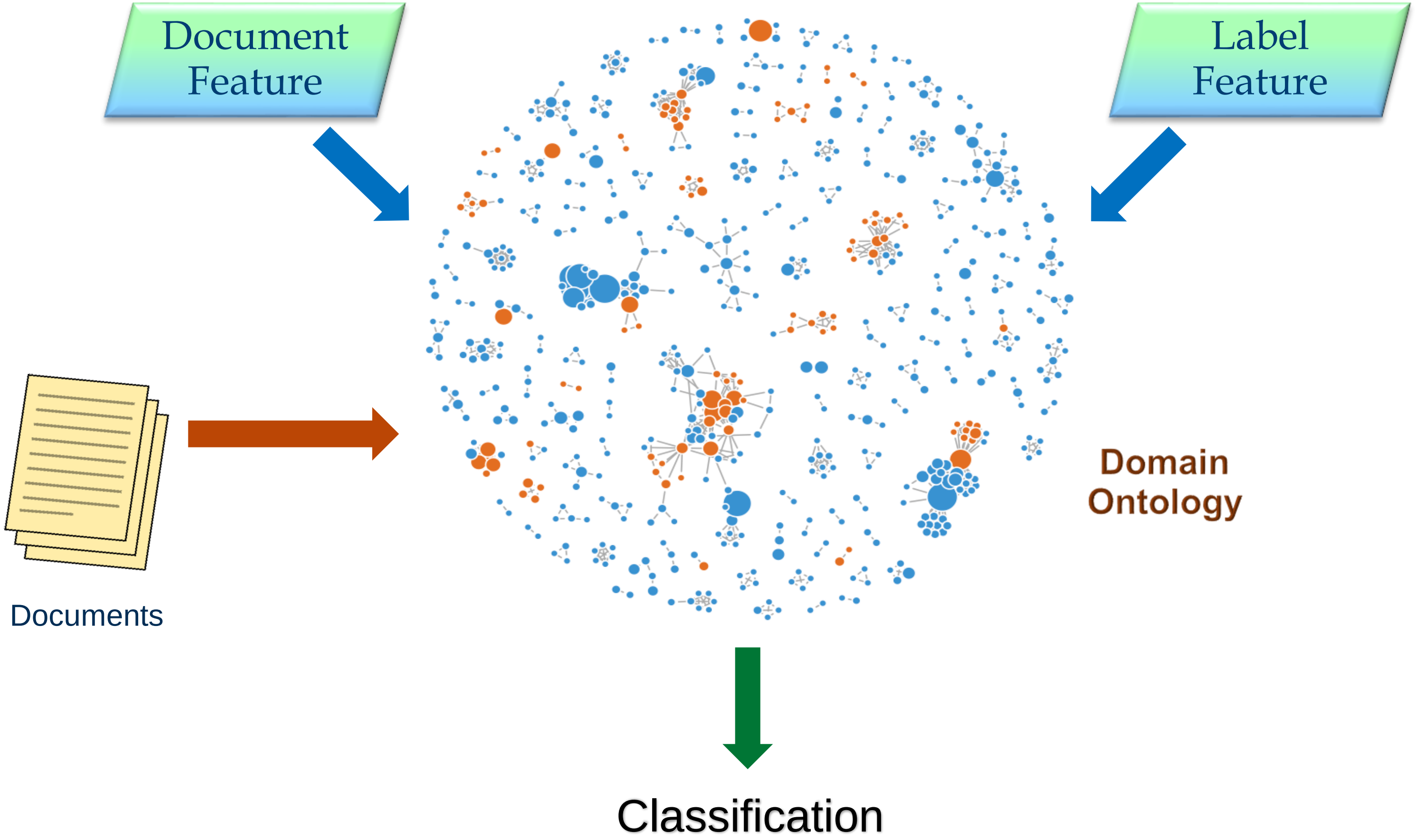
Proposed Improvement

Domain Ontology extraction using NLP techniques

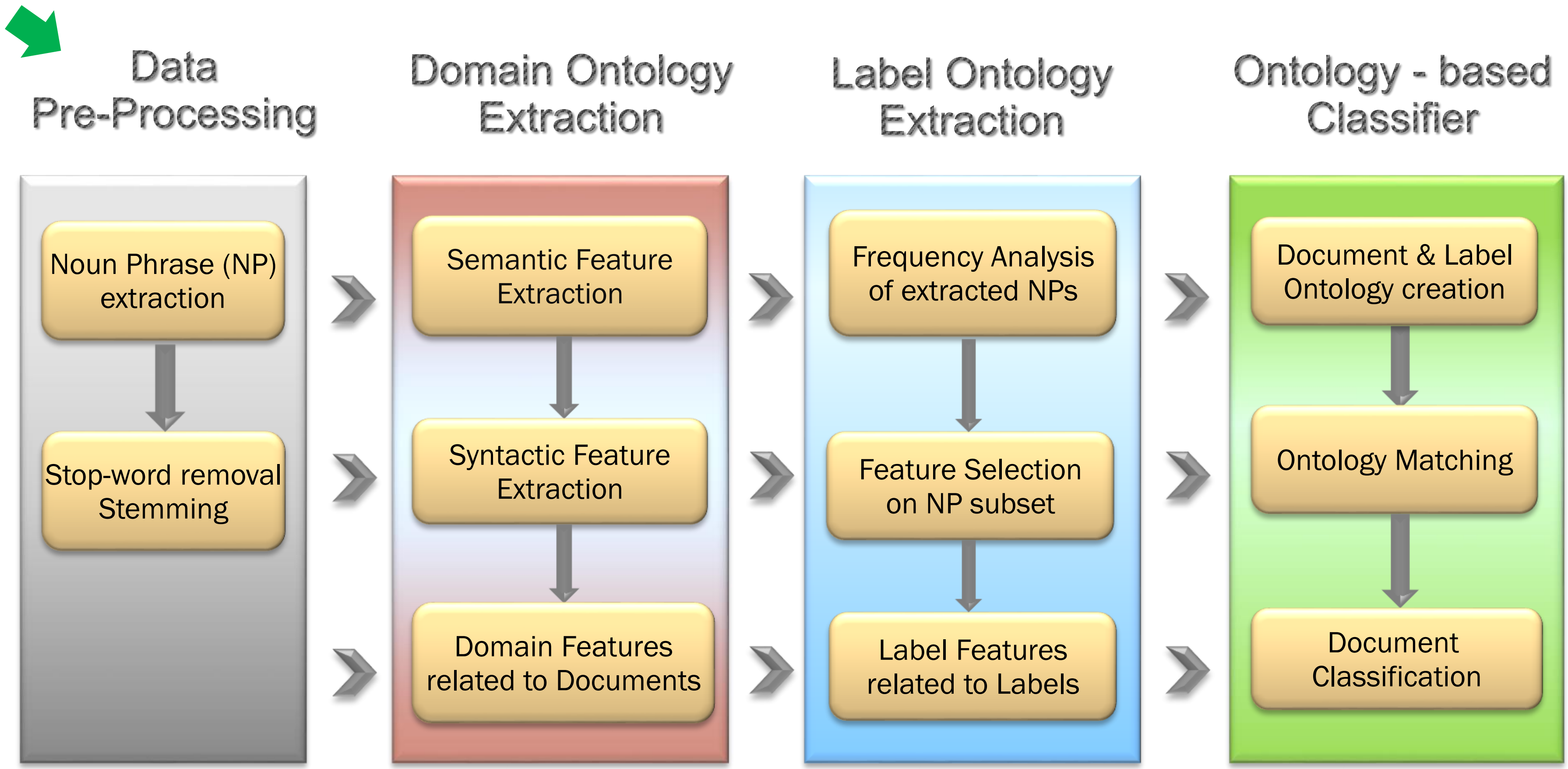
Scalability, Adaptability

Label Ontology extraction using statistical techniques

Proposed Architecture



Process Workflow



NP = Noun Phrase (*Example*: current social services)

Data Pre-Processing

TECHNISCHE UNIVERSITÄT DARMSTADT

Dataset

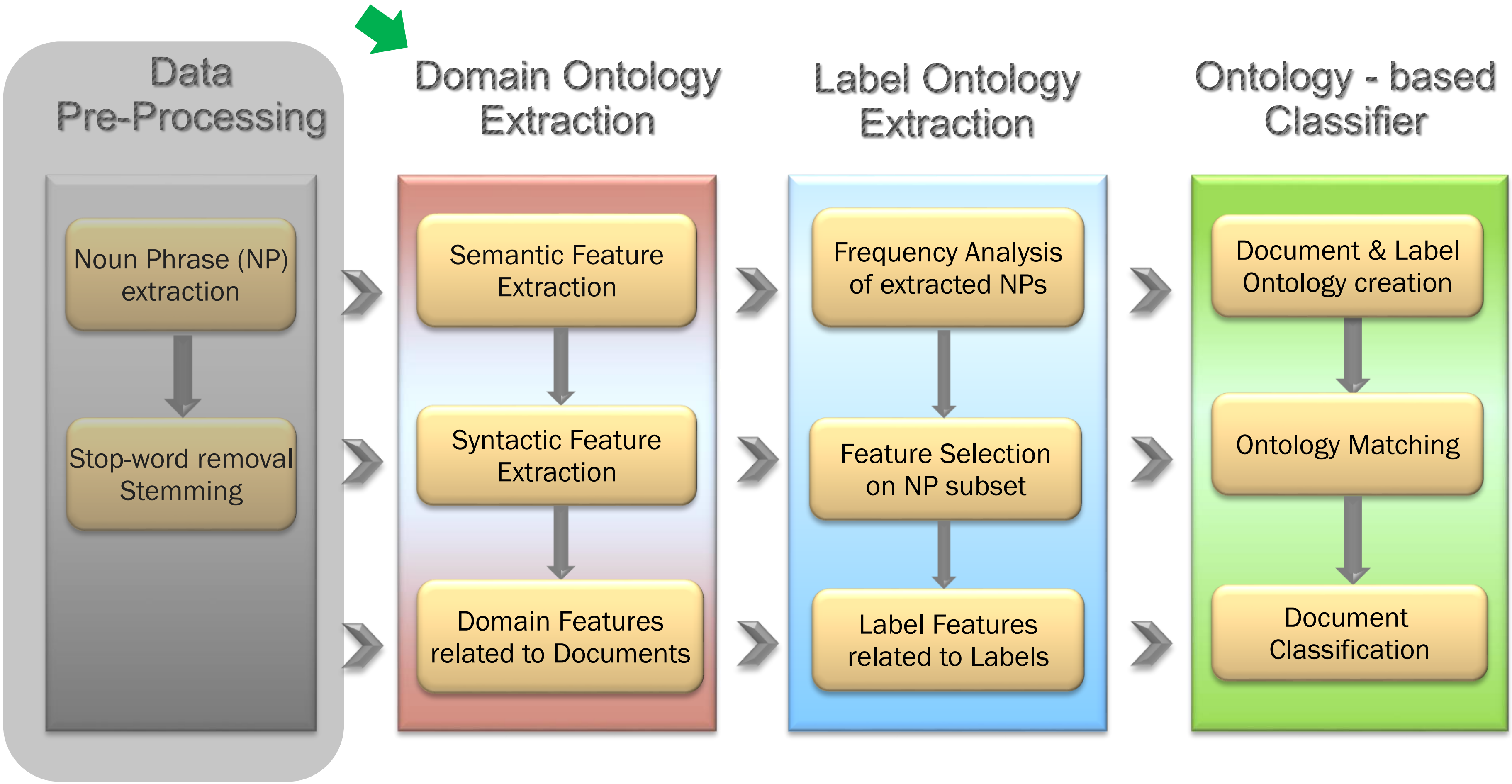
- EUR- Lex; Collection of documents about European Union law
- Numer of documents = 19,348
- Number of Labels = approx. ~ 4000

- Document Types:
  - Official journal, Treaties, International agreements, Legislation
  - Case law and parliamentary question

Stop-word Removal

- Remove frequent and non-discriminative words
  - Example: the, every, all, both
- Remove Domain related high frequent words such as
  - Example: behalf
- Remove Numbers, Dates, Symbols, One-letter words

Process Workflow



A1. Domain Ontology

Semantic Feature Extraction

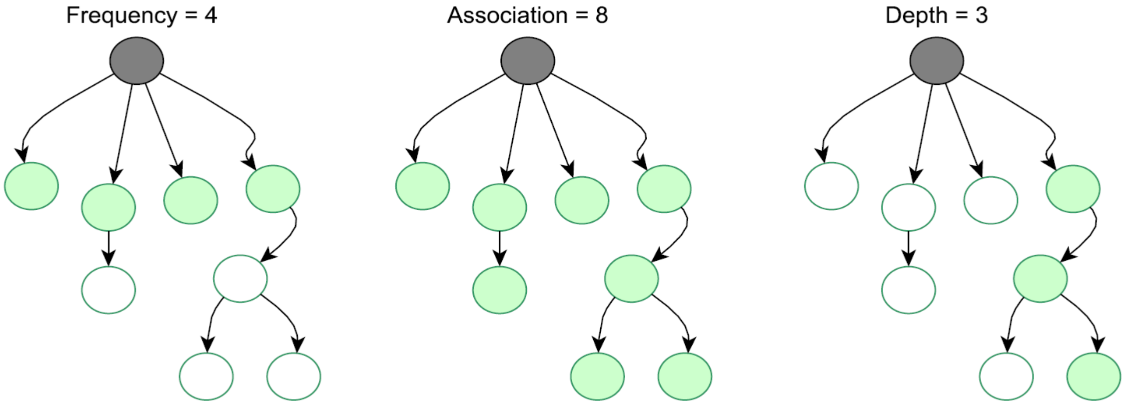
- Captures meaningful relations between concepts
- Lexico Syntactic Patterns, Hearst Patterns

(Adj | Noun) + Noun

NP such as NP, ∗ (or| and) NP

Feature representation

Community: (4, 8, 3)



- Extracted Semantic Relations = 26332
- Stanford NLP Parser is used to extract NPs

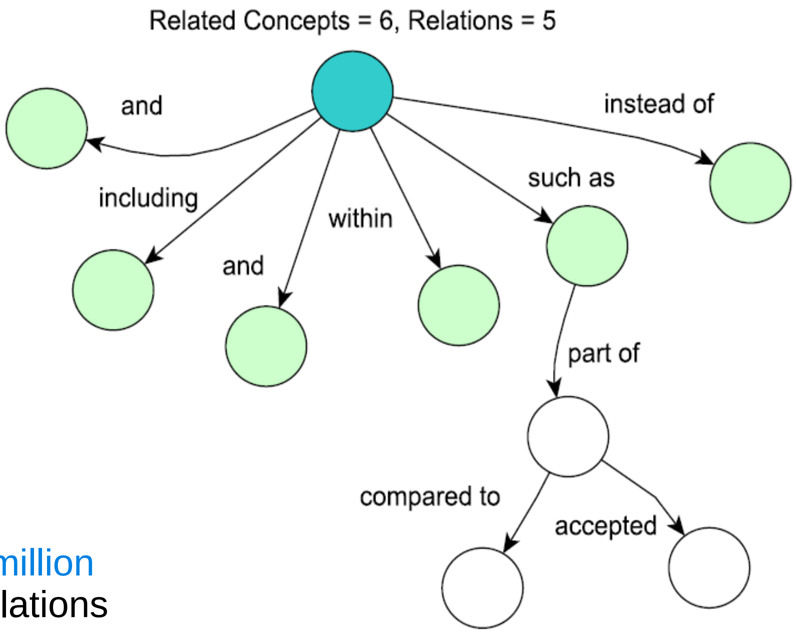
A2. Domain Ontology

Syntactic Feature Extraction

- Grammatical or lingustic specialized relations between document terms
- Syntactic Typed dependencies (Stanford Dependencies)

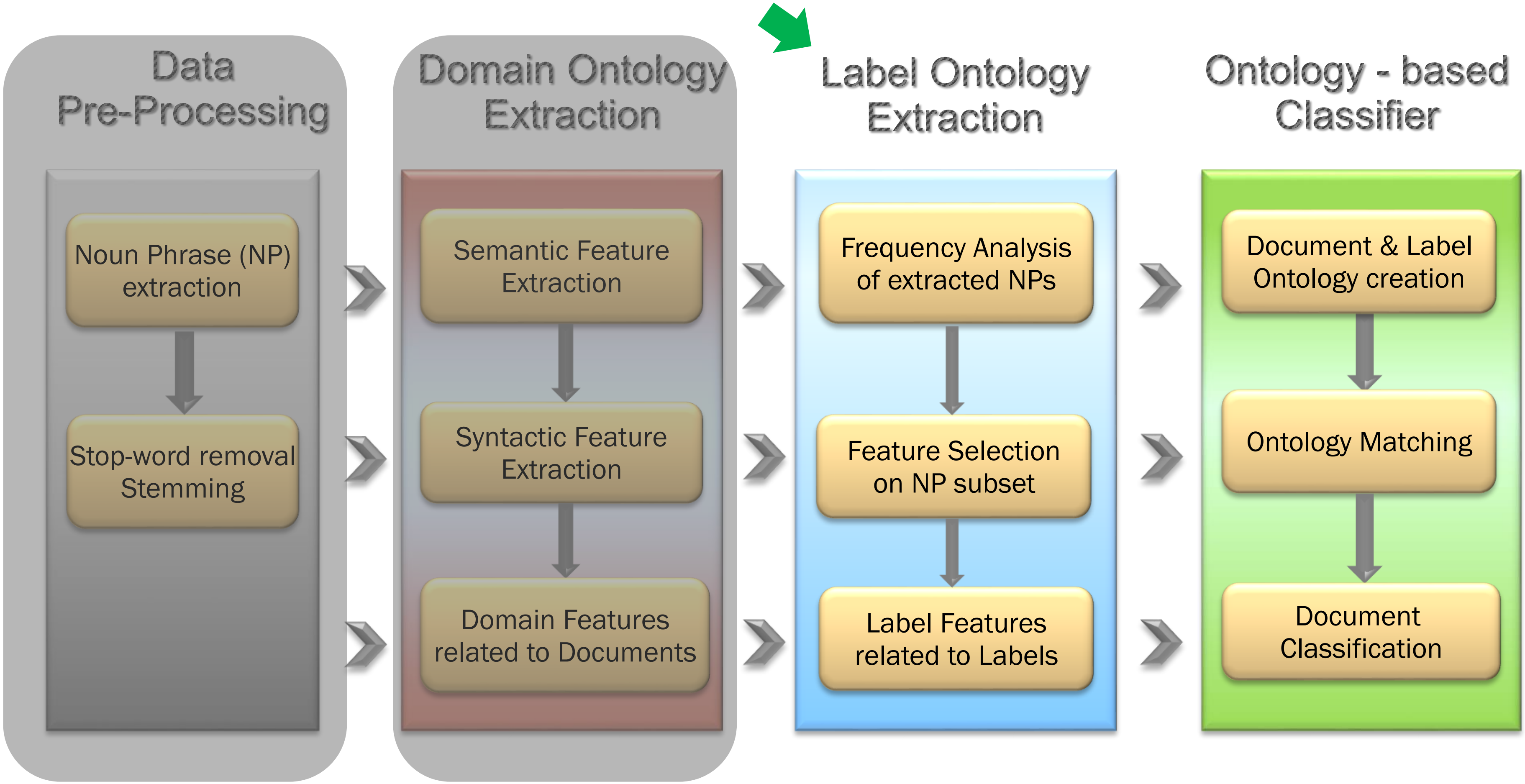
Feature representation

Commission : (6, 5, and = 2, including = 1, within = 1, such as = 5, instead of = 1)



- Extracted unique Syntactic Features near about 15 million
- Stanford Typed Dependencies used for extracting relations

Workflow



Data Pre-Processing
Noun Phrase (NP) extraction
Stop-word removal Stemming
Domain Ontology Extraction
Semantic Feature Extraction
Syntactic Feature Extraction
Domain Features related to Documents
Label Ontology Extraction
Frequency Analysis of extracted NPs
Feature Selection on NP subset
Label Features related to Labels
Ontology - based Classifier
Document & Label Ontology creation
Ontology Matching
Document Classification

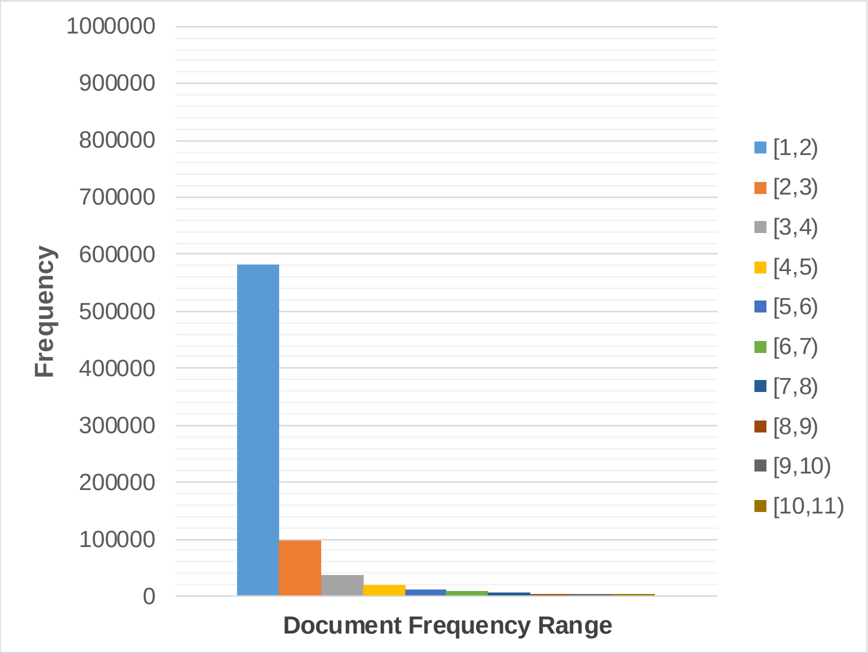
Label Ontology

- Frequency Analysis performed on all the extracted Noun Phrases (features)

- Feature Selection Approach
  - Correlation Coefficient
  - Gain Ratio

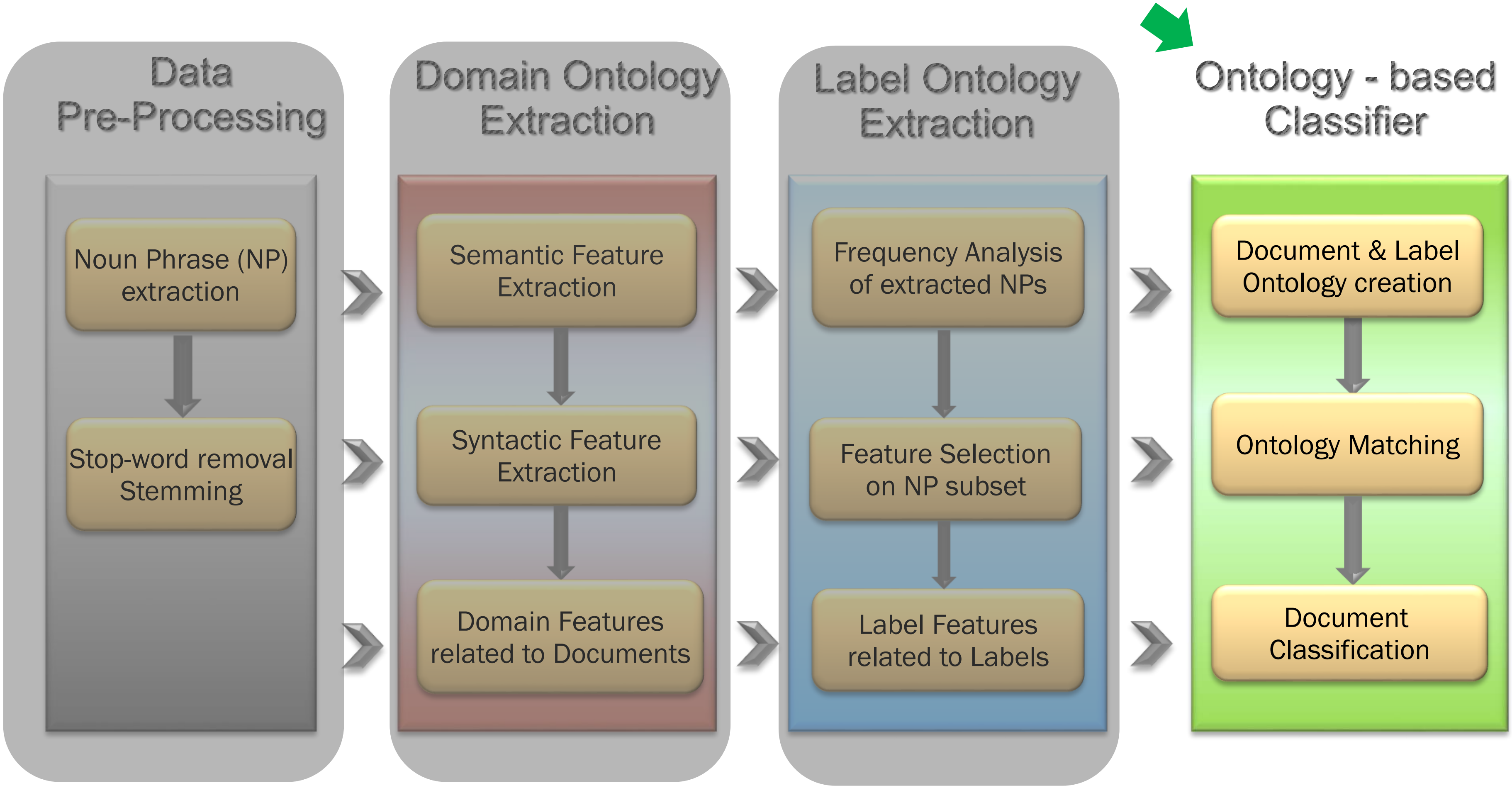
- Top features are selected based on Ranking

- Weka Feature Selection Tool is used for implementation.



Document Frequency of NPs

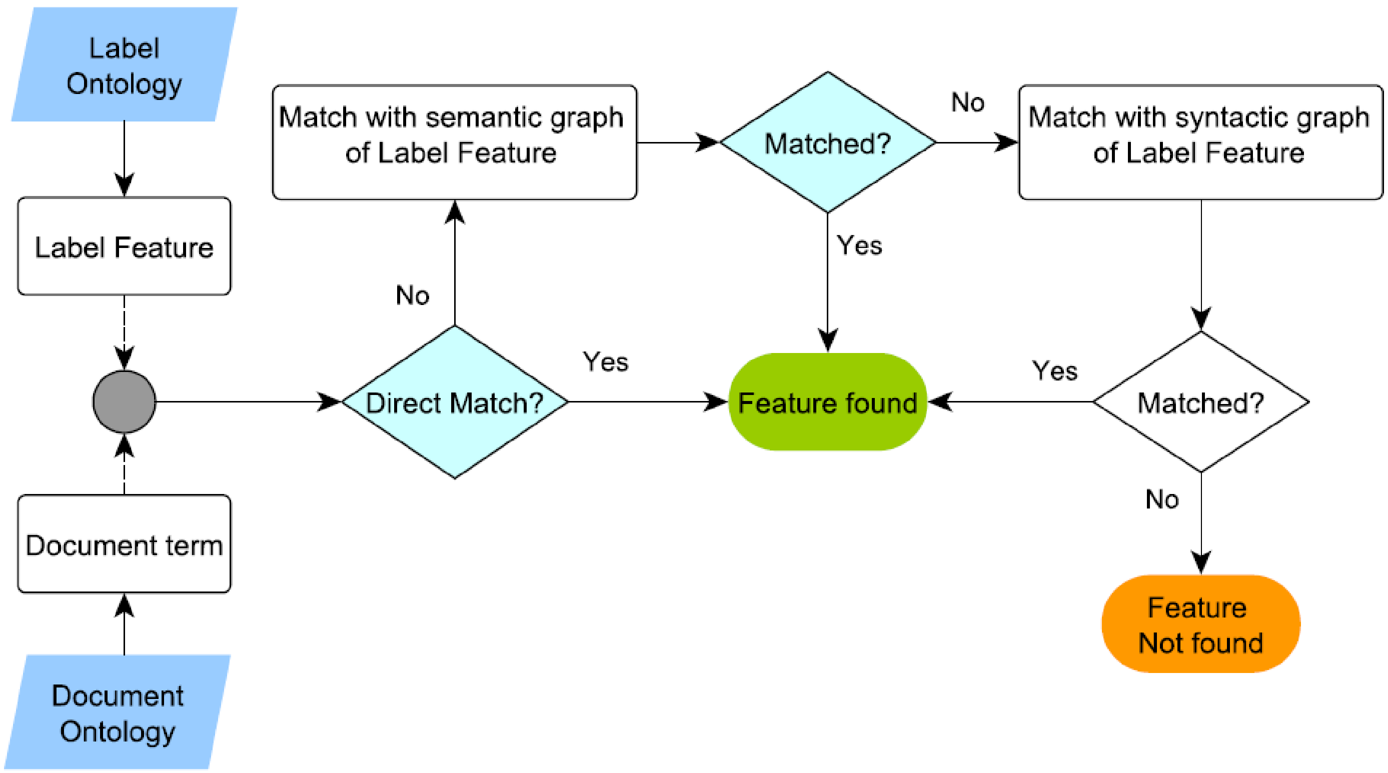
Workflow



Data Pre-Processing
Noun Phrase (NP) extraction
Stop-word removal Stemming
Domain Ontology Extraction
Semantic Feature Extraction
Syntactic Feature Extraction
Domain Features related to Documents
Label Ontology Extraction
Frequency Analysis of extracted NPs
Feature Selection on NP subset
Label Features related to Labels
Ontology - based Classifier
Document & Label Ontology creation
Ontology Matching
Document Classification

A1. Ontology Matching

TECHNISCHE UNIVERSITÄT DARMSTADT

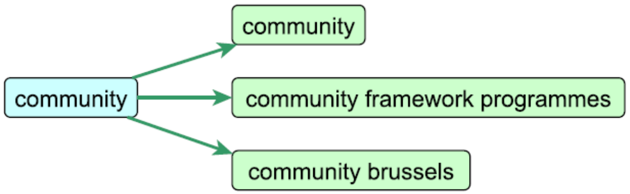


A2. Ontology Matching

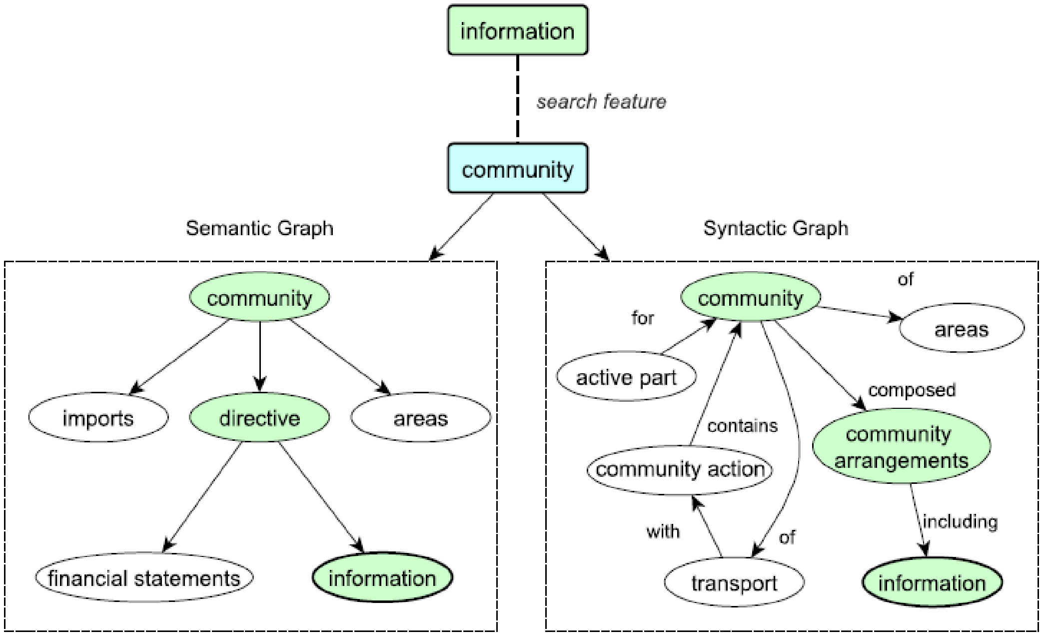
TECHNISCHE UNIVERSITÄT DARMSTADT

Direct Matching

Sub-Concept Matching

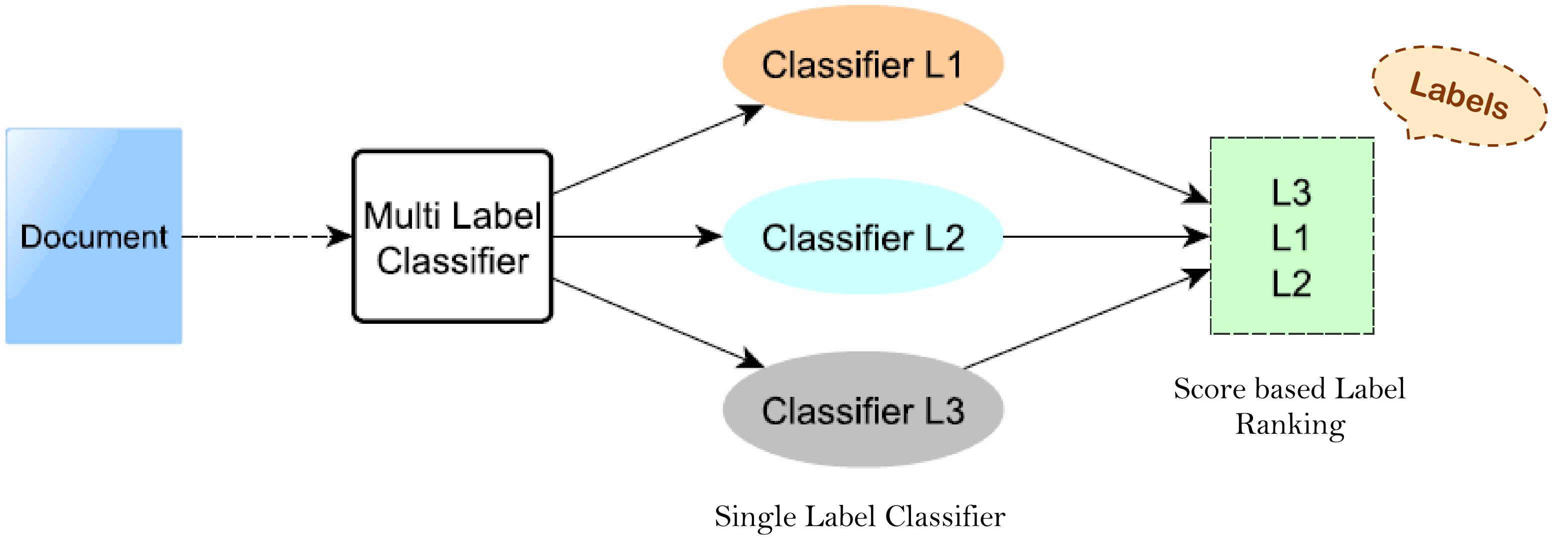


Feature Graph Matching



Ontology-based Multi-Label Classifier

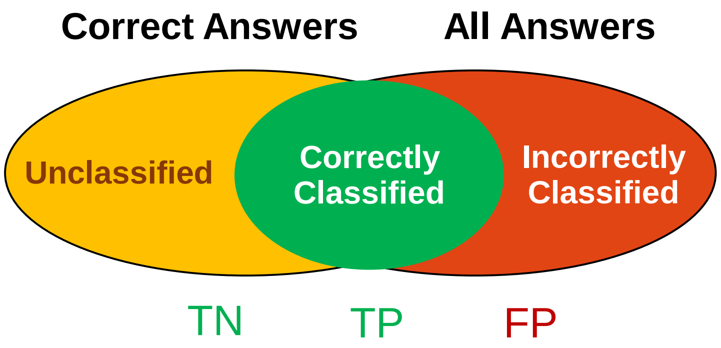
- **Classification approach:** Unsupervised Binary Relevance
- **Classification Algorithm**
  - Instance based grouping of Relevant and Irrelevant Document
  - Distance score measured using total document feature scores and label feature ranking scores



Evaluation of Classifier Performance

Evaluation Metrics

- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F – Measure = $2*\frac{Precision*Recall}{Precision+Recall}$
- Evaluation performed on all Documents



TP = True Positive

FP = False Positive

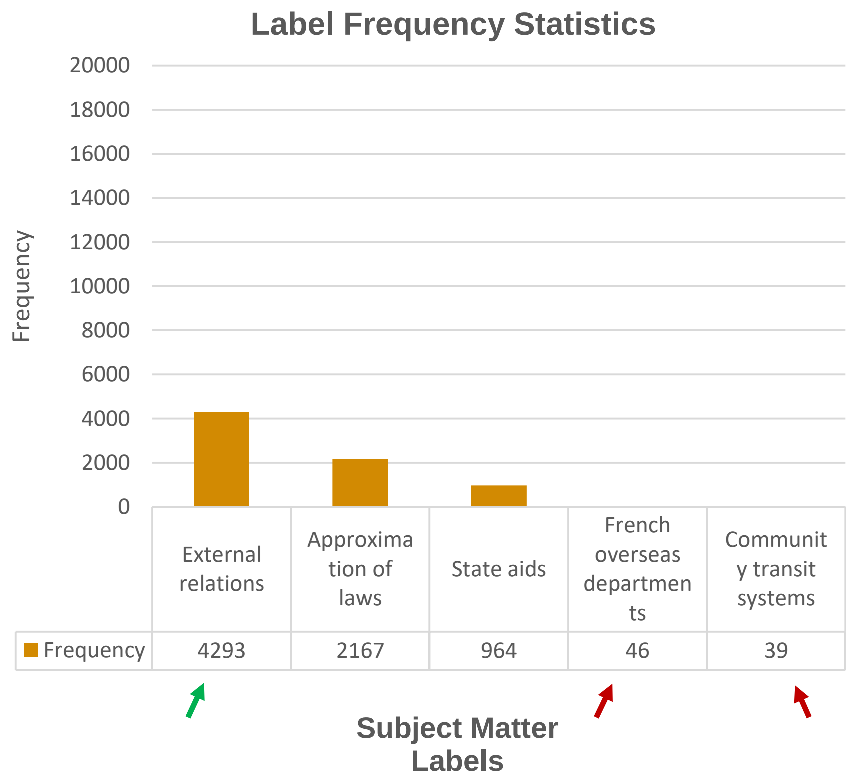
TN = True Negative

FN = False Negative

Evaluation Criteria

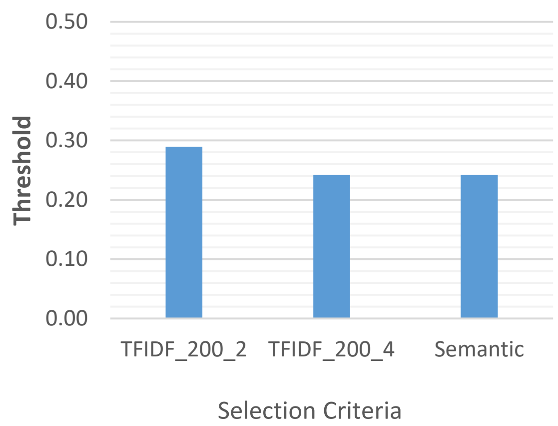
- Experimental Threshold Analysis conducted on the distance score
- Multiple evaluation criterias are considered in single sets
- Example: TFIDF_200_2 ➔ TFIDF based 200 document features are selected and matched with Label Feature‘s syntactic graph of depth 2

Selection Criteria	Range of Values
Gain Ratio	0.15, 0.20, 0.25 top = {200, 300}
Number of Document Features	100, 150, 200
Document Feature Selection Type	TFIDF, Semantic
Syntactic Graph Depth	2, 3, 4



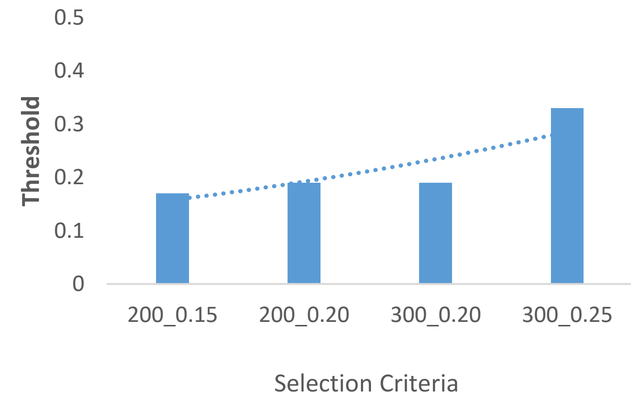
Most Frequent Label Performance

Label Feature Selection = Correlation

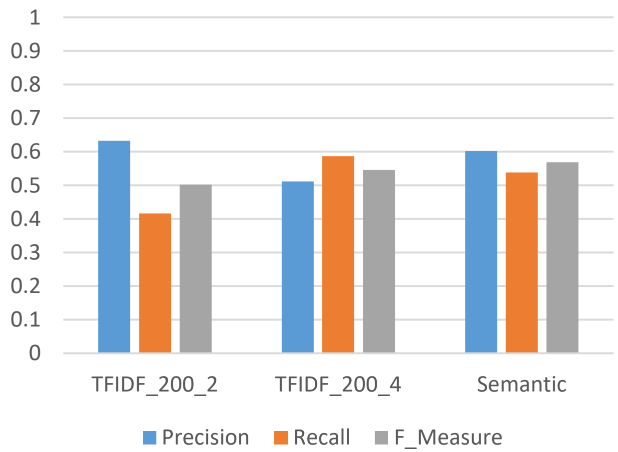


Label Feature Selection = Gain Ratio

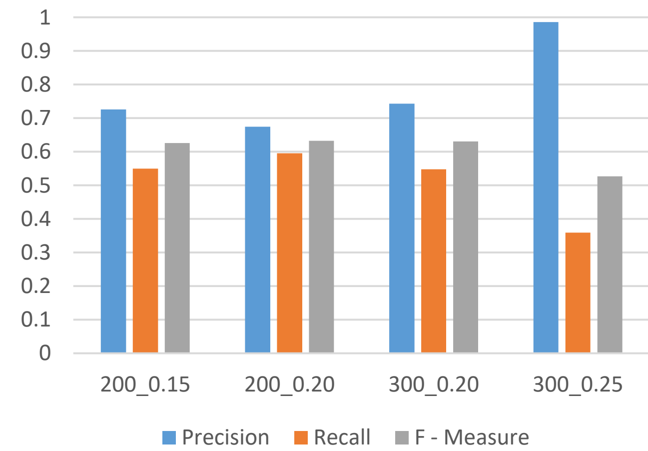
Document Feature Selection = TFIDF



Threshold Analysis

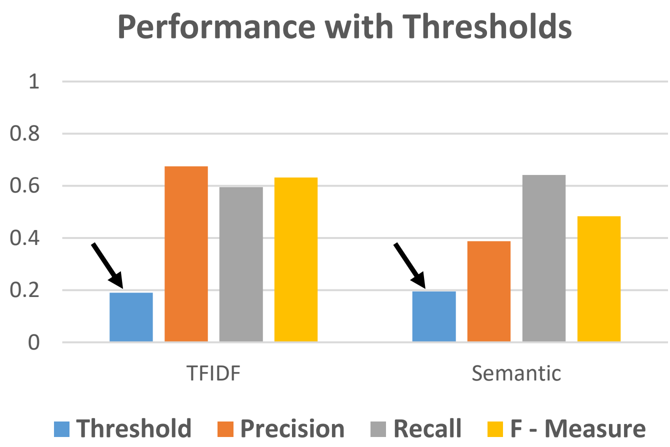


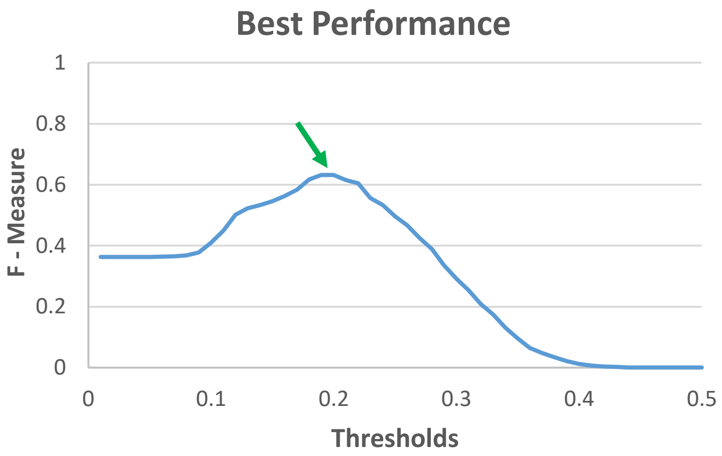
Performance

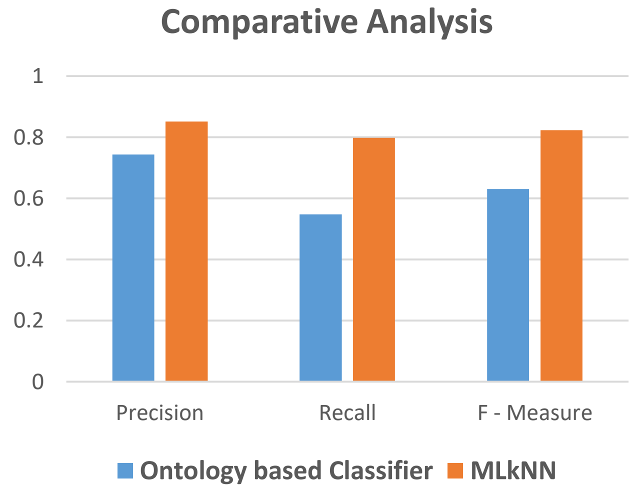


Most Frequent Label Performance

Label Feature Selection = Gain Ratio



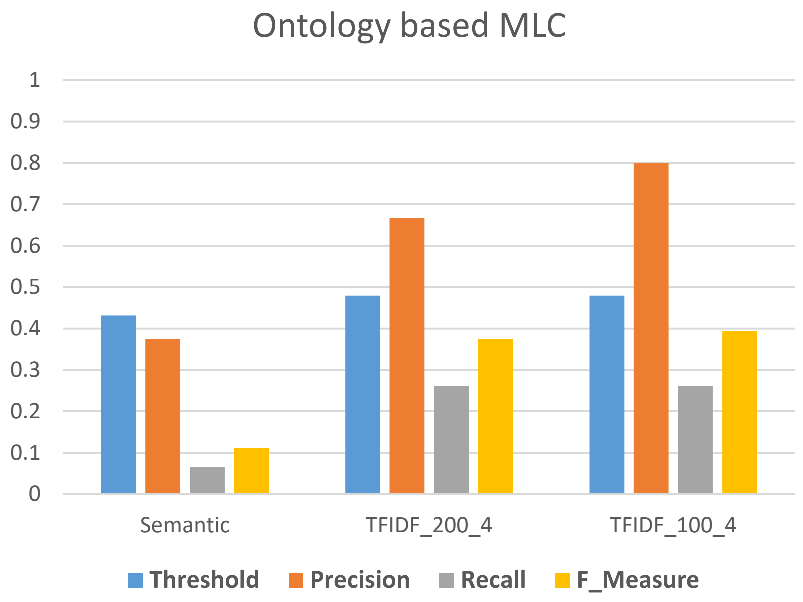




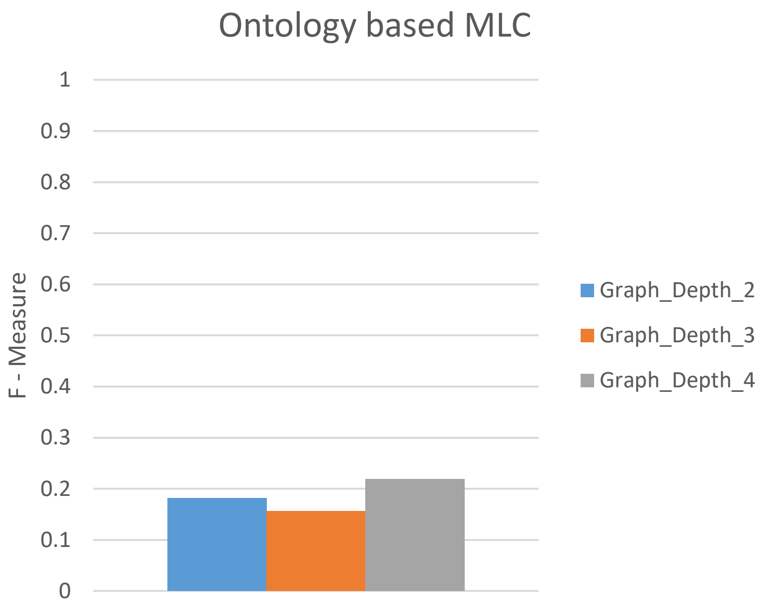
A1. Least Frequent Label Performance

Performance comparison criteria

- Label Feature Selection = Correlation
- Document Feature Selection = TFIDF, Semantic



Label Feature Selection = Gain Ratio



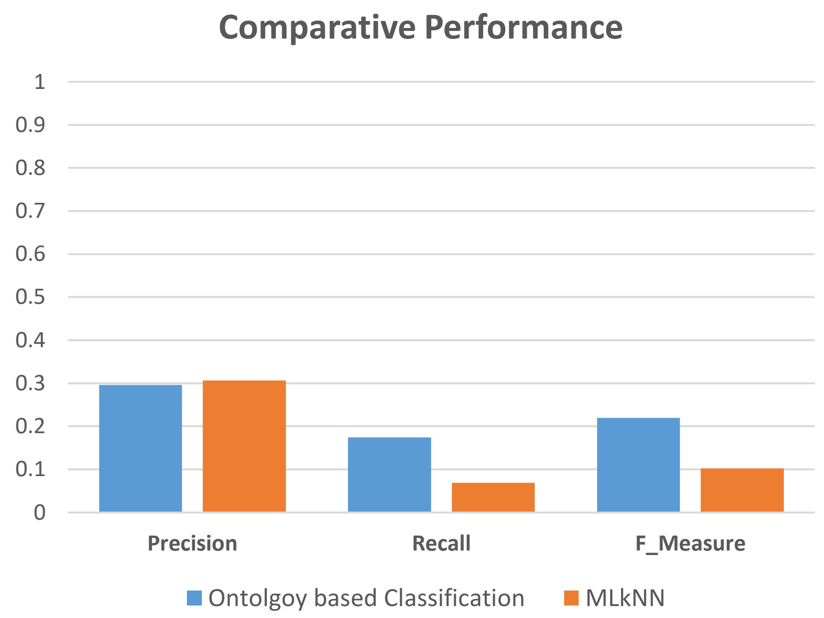
A2. Least Frequent Label Performance

Label Feature Selection = Gain Ratio

Document Feature Selection = TFIDF

Syntactic Graph Depth = 4

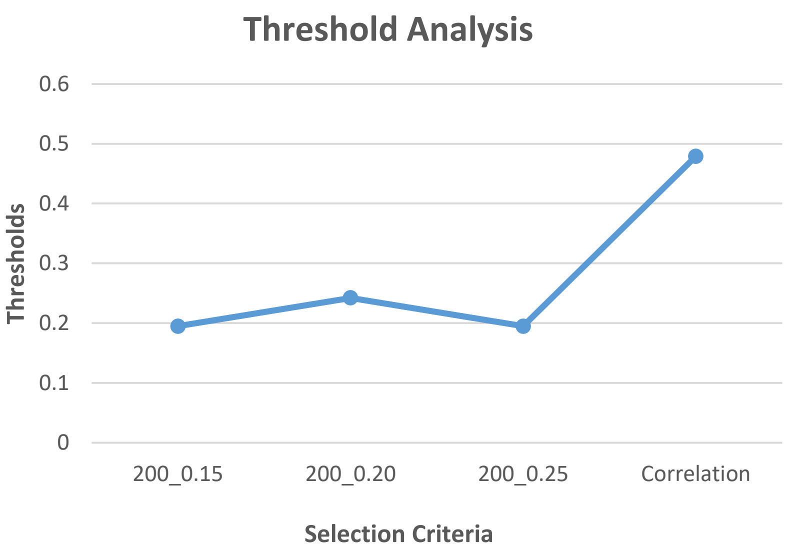
Number of Terms = 200



B1. Least Frequent Label Performance

Label Feature Selection = Gain Ratio {200_0.15, 200_0.20, 200_0.25}, Correlation

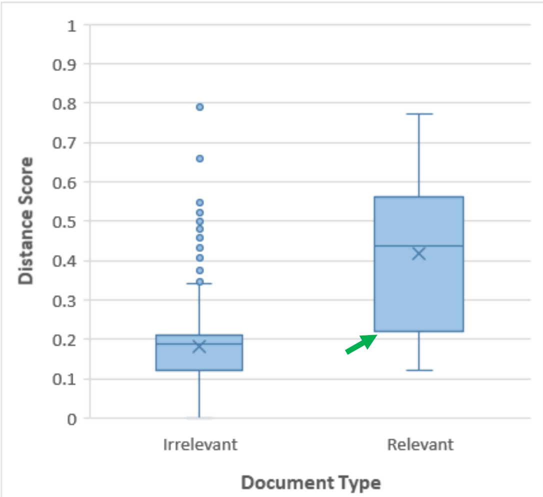
Syntactic Graph Depth = 4



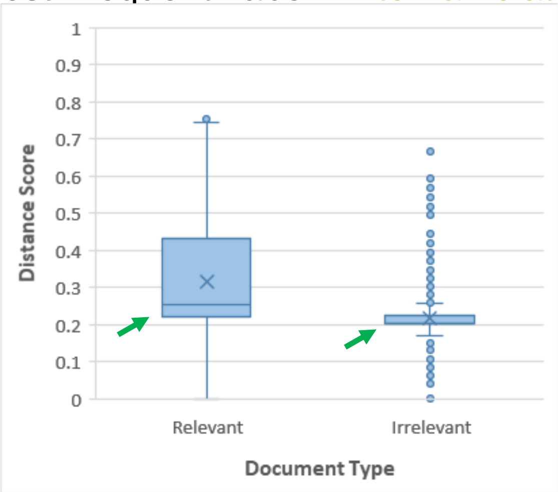
Threshold Evaluation

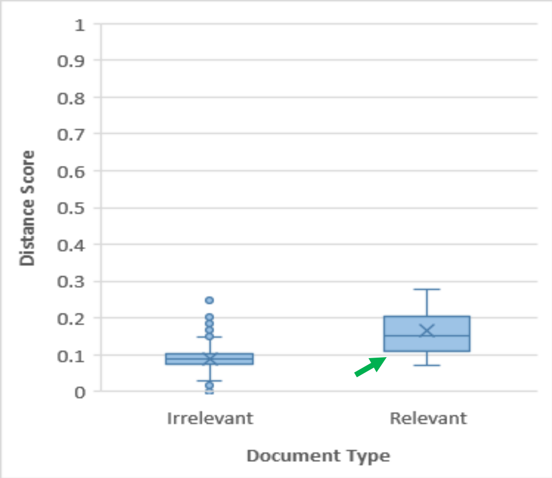
Least Frequent Label: Community transit systems

Most Frequent Label: External relations

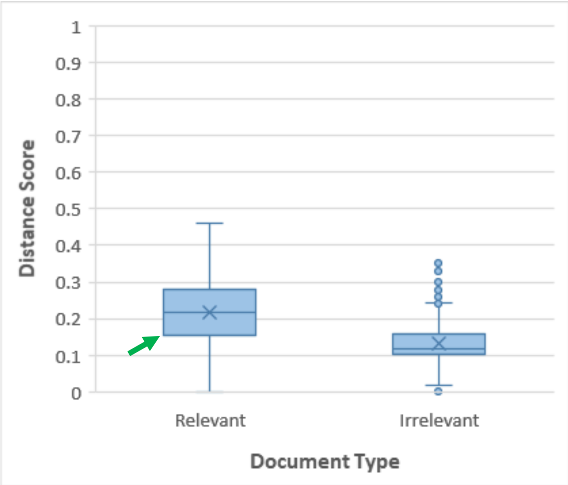


Correlation based Comparison





Gain-Ratio based Comparison



Conclusion

- The extraction of Domain Ontology solely based on NLP techniques shows good performance.
- The extraction of Label Ontology based on Statistical techniques is challenging due to the current tool time complexity but, performs well with the Domain Ontology matching process.
- The Ontology Matching process based on heuristical graph techniques shows more scope for improvement.
- Multi-Label classification is possible solely based on the ontological information and performs better on low frequent labels whereas, traditional Multi-Label classification approach failed.

Future Work

- Generalization of threshold for all Labels to develop Ontology based unsupervised MLC algorithm

- Adding Lexical Database knowledge to domain ontology for amplifying the ontology quality to improve classification performance

- Using more feature selection techniques for better label feure extraction

- Adding more NLP techniques to filter out unnecessary words and phrases.

Thank you for your attention! Questions?