UFRGS@PAN2010: Detecting External Plagiarism

Lab Report for PAN at CLEF 2010

Rafael Corezola Pereira, Viviane P. Moreira, and Renata Galante

Universidade Federal do Rio Grande do Sul

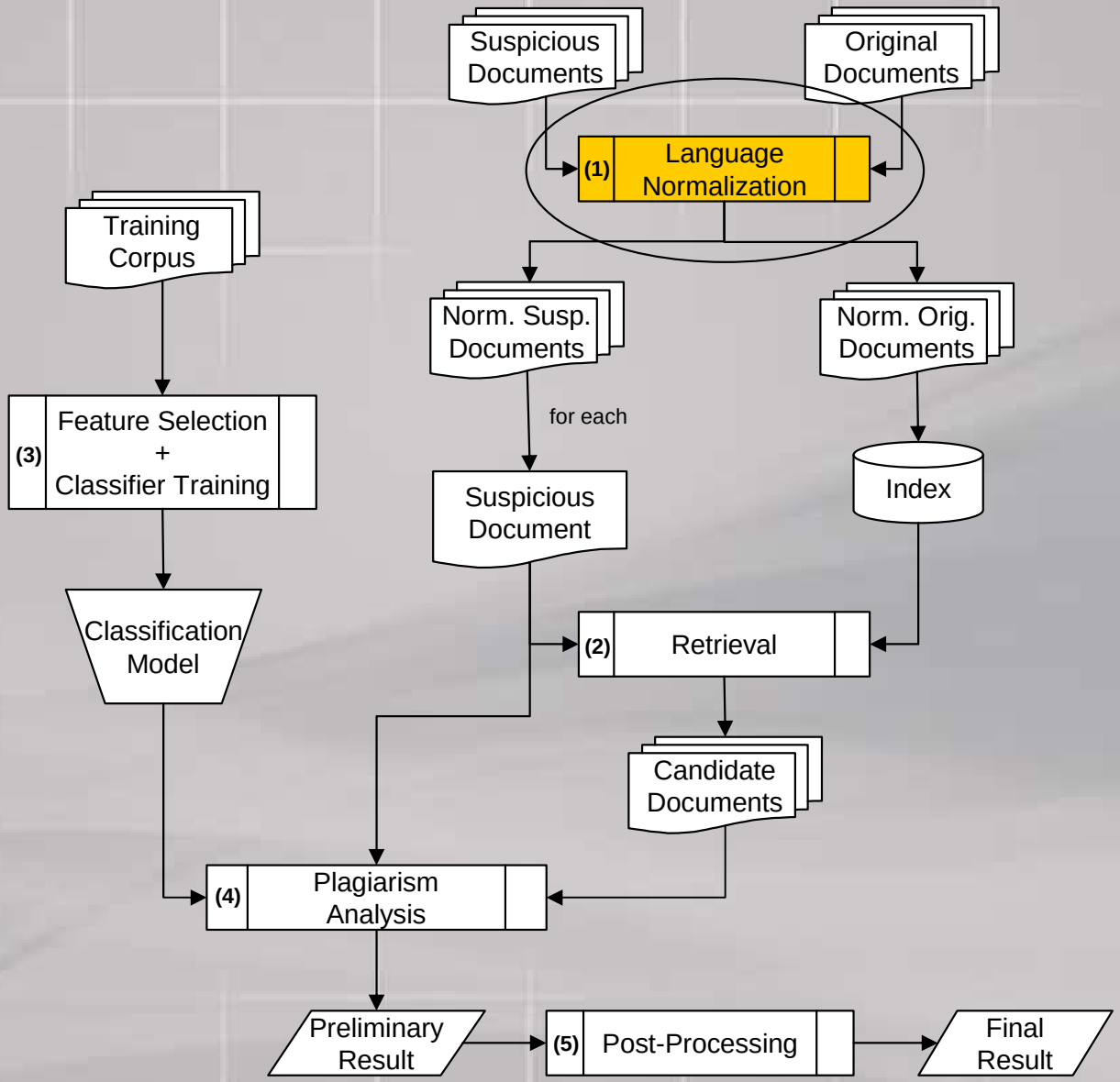
PAN 2010

The Task

- Detect the plagiarized passages in the suspicious documents and their corresponding text fragments in the source documents even if the documents are written in different languages

- Known as **External** plagiarism analysis

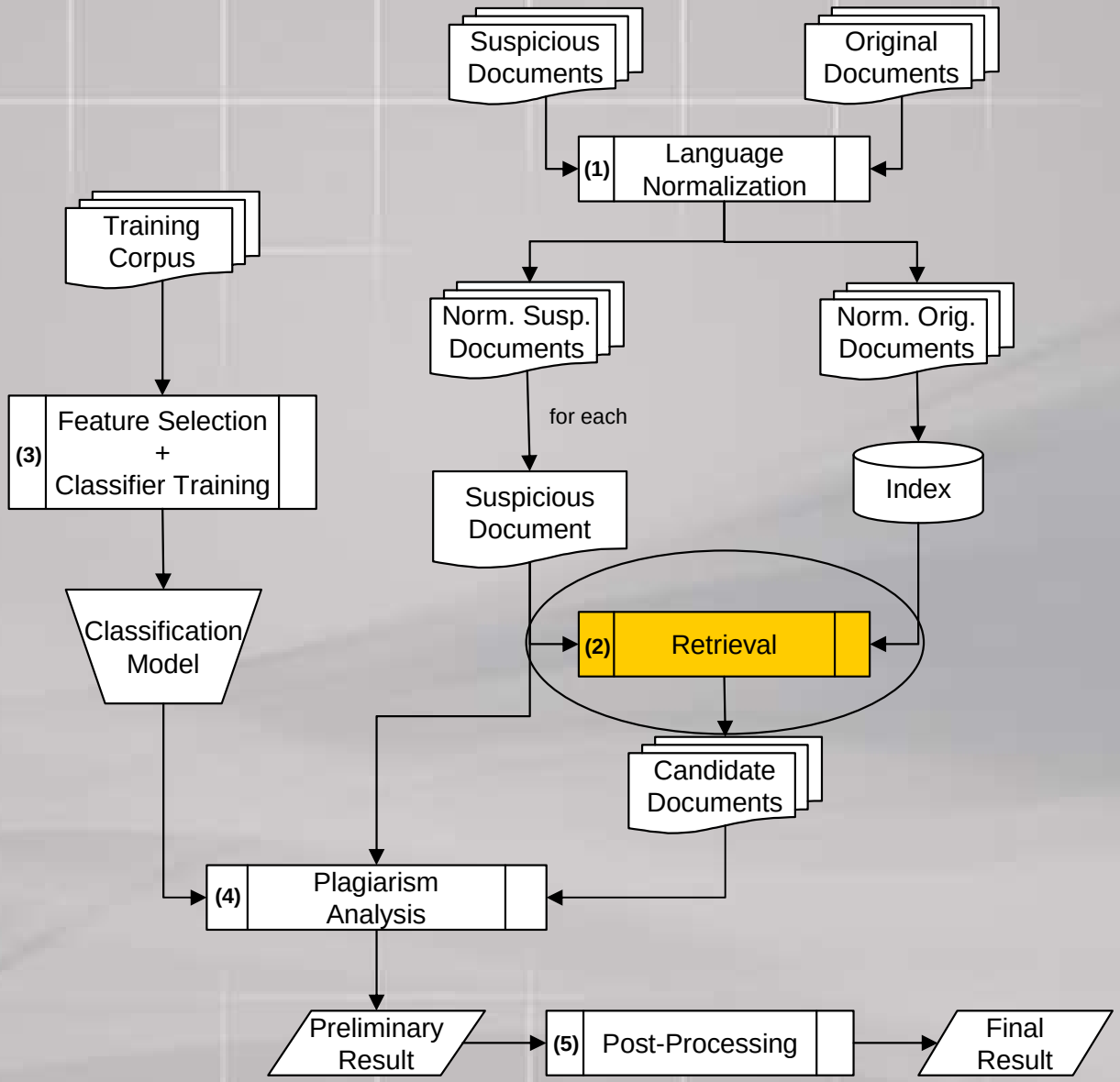
The Proposed Approach



(1) Language Normalization

- All documents are converted into a common language
- English was chosen
  - More translation resources
  - One of the easiest languages to translate into
- Used a language guesser and an automatic translation tool

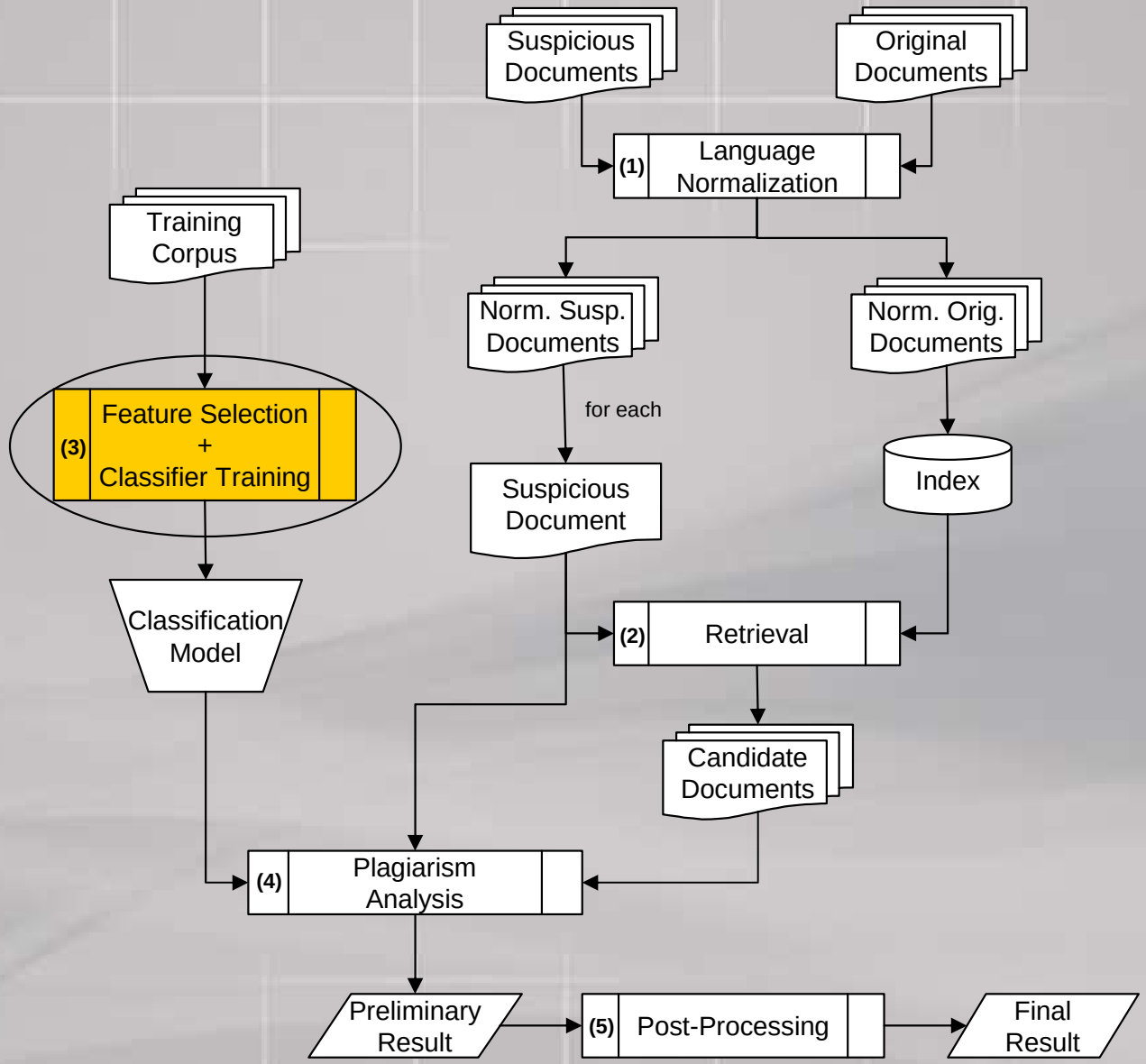
The Proposed Approach



(2) Retrieval of Candidate Documents

- Problem: It is not feasible to perform exhaustive comparisons
- Solution: Use the suspicious document as a query to be sent to an IR system
- Documents are divided into paragraphs (subdocuments)
- At the end of this phase, we have a list of at most ten candidate subdocuments for each passage in the suspicious document

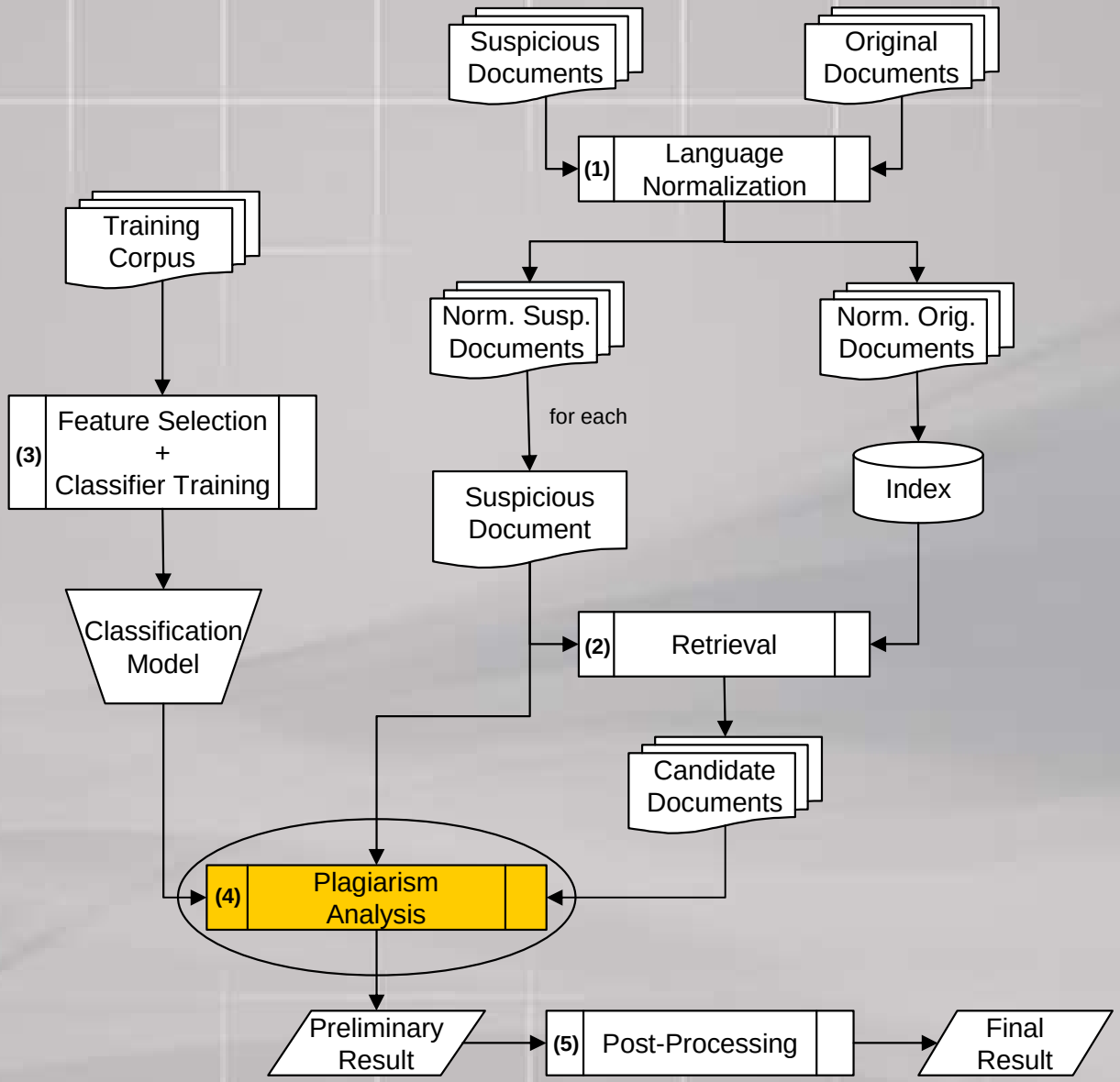
The Proposed Approach



(3) Feature Selection and Classifier Training

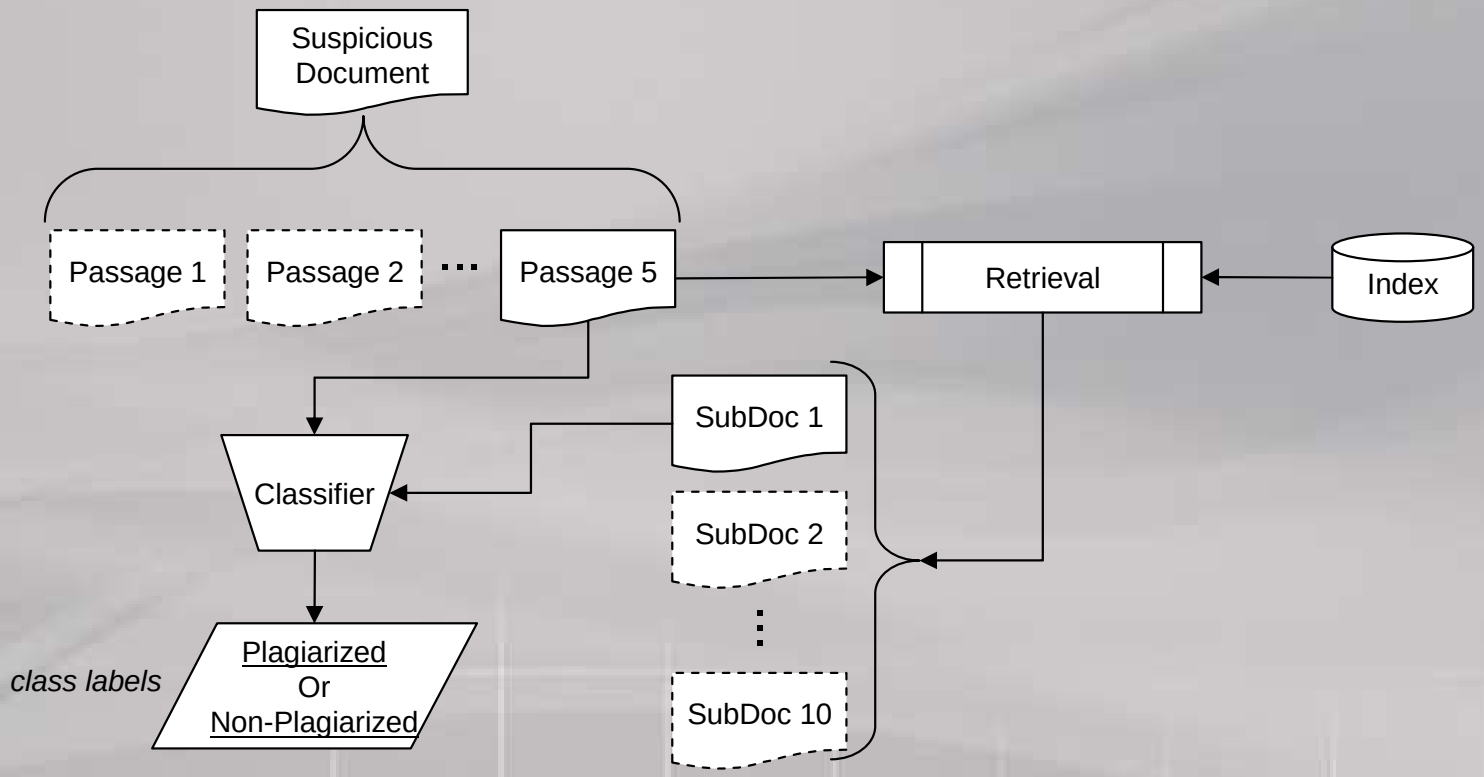
- The goal of the classifier is to decide whether a suspicious passage is plagiarized from a candidate subdocument
- Annotated synthetic examples used for training
- J48 classification algorithm
- Features
  - The cosine similarity between the suspicious passage and the candidate subdocument
  - The similarity score assigned by the IR system
  - The position of the candidate subdocument in the rank generated
  - The length (in characters) of the suspicious and the candidate subdocument

The Proposed Approach

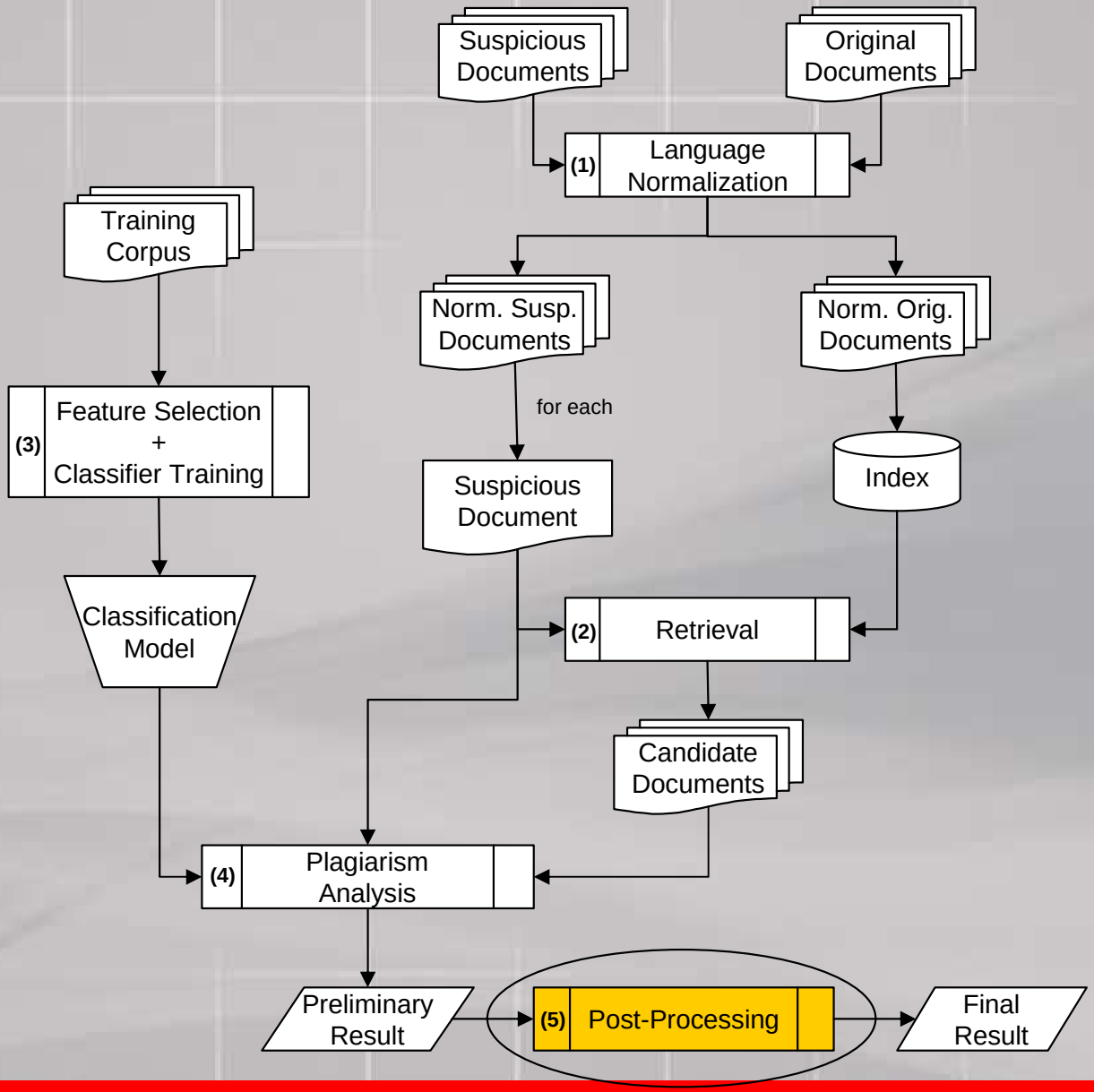


(4) Plagiarism Analysis

- Submit the test instances to the trained classifier and let it decide whether the suspicious passage is, in fact, plagiarized from one of the candidate subdocuments



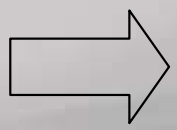
The Proposed Approach



(5) Result Post-Processing

- Join the contiguous plagiarized passages detected by the method in order to decrease its final granularity score

```
<?xml version="1.0" encoding="UTF-8"?>
<document reference="A.txt">
 <feature name="detected-plagiarism"
     this_offset="1000“
     this_length="500"
     source_reference="B.txt"
     source_offset="3000“
     source_length="500"
 />
 <feature name="detected-plagiarism"
     this_offset="1500“
     this_length="300"
     source_reference="B.txt"
     source_offset="3500“
     source_length="300"
 />
</document>
```



```
<?xml version="1.0" encoding="UTF-8"?>
<document reference="A.txt">
 <feature name="detected-plagiarism"
     this_offset="1000“
     this_length="800"
     source_reference="B.txt"
     source_offset="3000“
     source_length="800"
 />
</document>
```

Experiments

- Terrier IR System
  - Porter Stemmer
  - Stop-Word Removal (list of 733 words)
- Weka Data Mining Software
  - J48 classification algorithm
- Google Translator (as language guesser)
- LEC Power Translator

Summary

- Overall results (7th place) / No Obfuscation vs. Translated

---	Competition	Only External Cases
Recall	0.4036 (7th)	0.4966
Precision	0.7242 (11th)	0.7242
Granularity	1.0024 (1th)	1.0017
Overall Score	0.5175 (7th)	0.5881

	None	Translated	%
Precision	0.68	0.60 (4th)	88
Recall	0.51	0.43 (4th)	84
Granularity	1.00	1.01 (4th)	99

- The length of the plagiarized passage affects the results considerably
  - The larger the passage the easier the detection
- Low performance while detecting short plagiarized passages
  - Partially explained by our decision of indexing only the subdocuments with length greater than 250 characters

UFRGS@PAN2010: Detecting External Plagiarism

Lab Report for PAN at CLEF 2010

Questions?

Rafael Corezola Pereira, Viviane P. Moreira, and Renata Galante

Universidade Federal do Rio Grande do Sul

PAN 2010

Processing Time

- Notebook
  - Intel Core 2 Duo 1.6GHz
  - 2GB RAM
  - HD 5400 RPM

Total Analysis Time	~ 230 hours
Average Time / Suspicious Document	52 seconds
KB Analyzed / Minute	236KB