AI-Generated Text Detection Using Ensemble and Combined Model Training

NLP M&S team

Hadi Mohammadi   Anastasia Giachanou   Ayoub Bagheri

Department of Methodology and Statistics,
Utrecht University,
The Netherlands.

November 28, 2024

Universiteit Utrecht

CLIN33 Shared task

The CLIN33 shared task addresses the crucial need to differentiate between human-written content and text generated by AI language models.

Universiteit Utrecht

Data Acquisition and Integration

- Read the primary development dataset.
- Integrate external data sources:
  - `'AuTexTification_train'`
  - `'AuTexTification_test'`

AUTEXTIFICATION 🤖👩🏻

AuTexTification: Automated Text Identification shared task

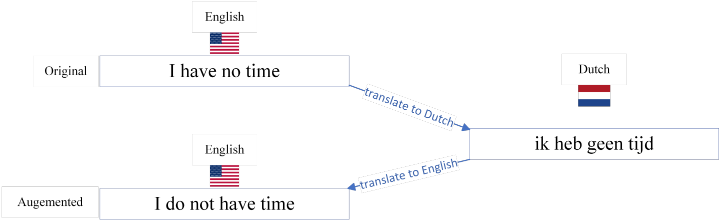
Data Pre-processing

- Convert text to lowercase.
- Remove URLs, special characters, and punctuation.
- Tokenize sentences.
- Lemmatization using `WordNetLemmatizer`.

Universiteit Utrecht

Data Augmentation Techniques

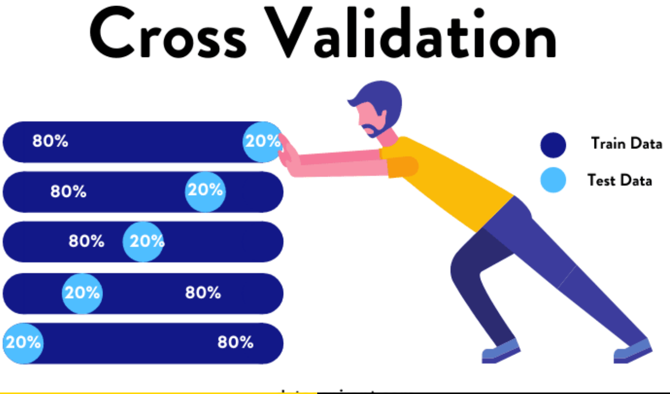
- Synonym augmentation: `aug_synonym.augment`
- Word swapping, insertion, substitution, deletion
- Introducing spelling variations
- Back translation techniques: English ¡-¿ Dutch
- ~~Paraphrasing by free generative AI models (such as GPT-2)~~



Universiteit Utrecht

Cross-Validation Data Preparation

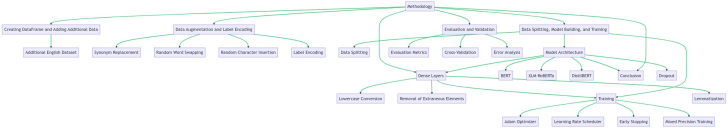
- Utilize `StratifiedKFold`
- Address class imbalance:
  - `RandomOverSampler`
  - `SMOTE`
- Compute class weights for balanced training.



Universiteit Utrecht

CustomBERT Model Architecture

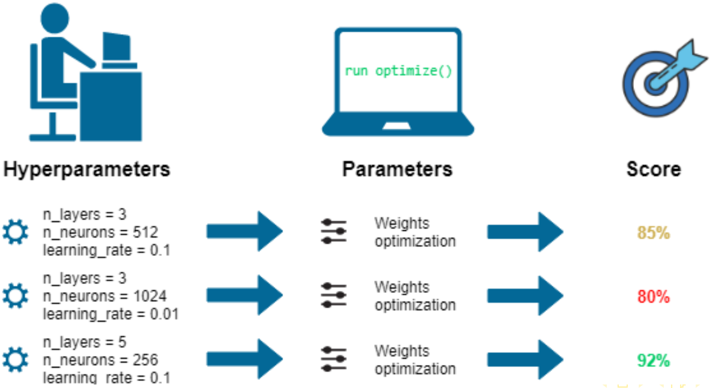
- Build and fine-tune sequence classification BERT model using:
  - bert-base-multilingual-cased
  - xlm-roberta-base
  - distilbert-base-multilingual-cased.
- Classification layer with `Dense` and `softmax`.
- Configuration: `Adam`, loss function, accuracy metric.



Universiteit Utrecht

Hyperparameter Optimization

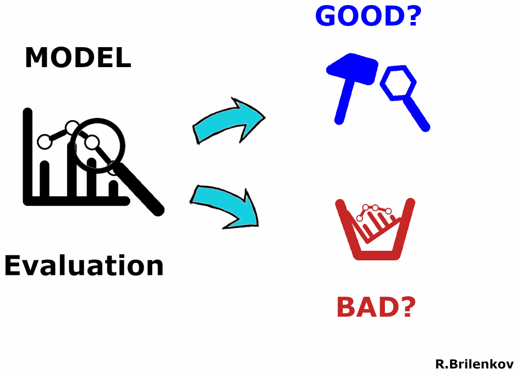
- Early stopping criteria.
- Learning rate scheduler setup.
- Fine-tuning and hyperparameter search space.
- Optimal hyperparameter identification.



Universiteit Utrecht

Evaluation on 'AuTexTification_test'

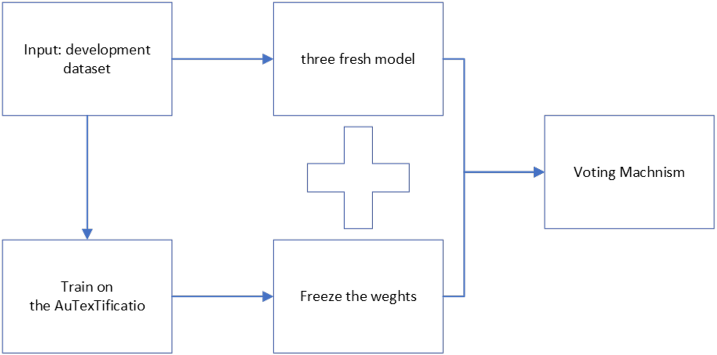
- Segment test data based on language and resource.
- Essential metrics: Accuracy, F1 Score, Recall, Precision.
- Average metrics across fold divisions.



Universiteit Utrecht

Combined Model

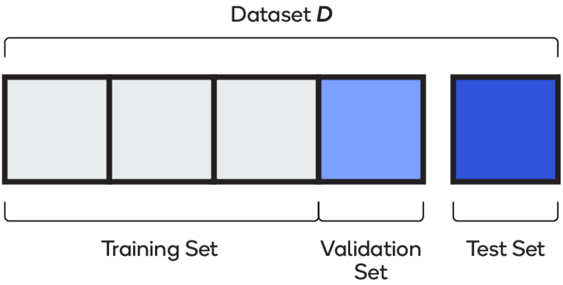
- Freeze weights of models trained on 'AuTexTification_train'.
- Integrate models: BERT, XLM-Roberta, DistilBERT.
- Employ a voting mechanism for the integrated models.



Universiteit Utrecht

Evaluation on Test Data

- Use untouched 10% of original dataset.
- Report performance metrics.
- Highlight: Achieved F1 score of 74%.



Universiteit Utrecht

Thank You!

Thank you for your attention!

For further questions or details, please contact:

h.mohammadi@uu.nl

https://hadimohammadi.info/

Universiteit Utrecht