# Detecting Machine-Generated Text with Purely Linguistic Features

DetecTUM at the CLIN-33 Shared Task

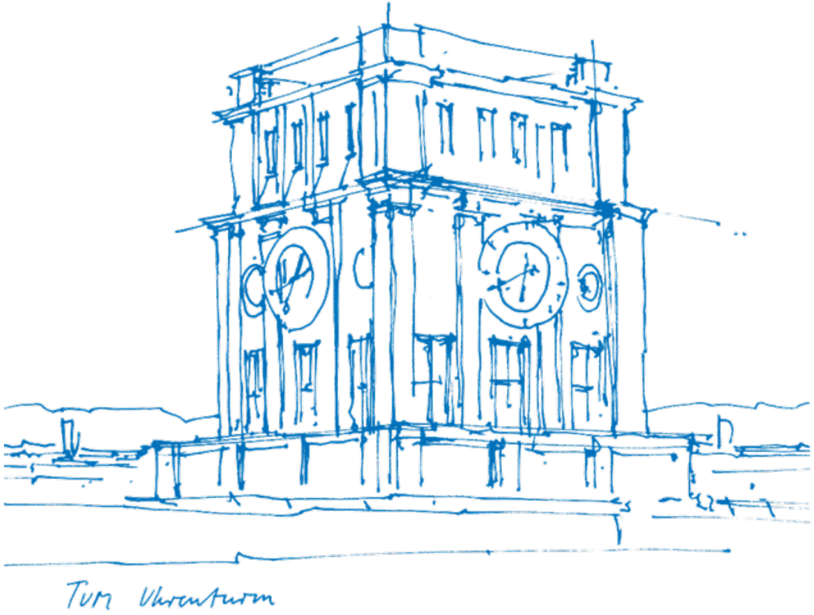

Wessel Poelman[†], Juraj Vladika[†], Esther Ploeger[§],

Florian Matthes[†] and Johannes Bjerva[§]

[†] *Technical University of Munich*

[§] *Aalborg University*


September 22, 2023



# Goals

Limited training data

# Goals

Limited training data → Cross-lingual features *transfer?*

# Goals

Limited training data → Cross-lingual features *transfer?*
Not all domains in provided data

# Goals

| | |
|---|---|
| Limited training data | $\rightarrow$ Cross-lingual features *transfer?* |
| Not all domains in provided data | $\rightarrow$ Cross-domain features *transfer?* |

# Goals

| | |
|---|---|
| Limited training data | $\rightarrow$ Cross-lingual features *transfer?* |
| Not all domains in provided data | $\rightarrow$ Cross-domain features *transfer?* |
| Explanation track | |

# Goals

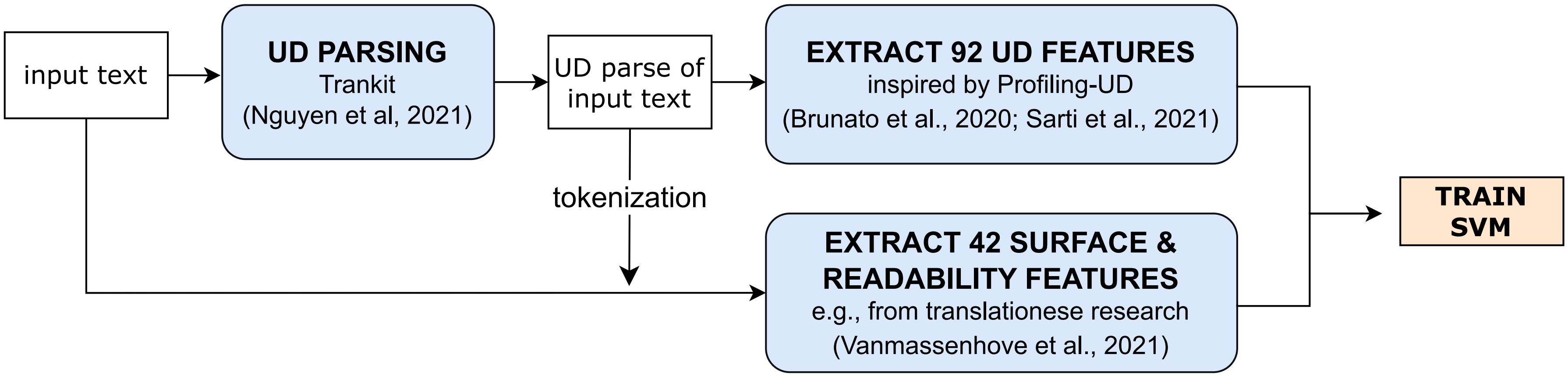
| | |
|---|---|
| Limited training data | $\rightarrow$ Cross-lingual features *transfer?* |
| Not all domains in provided data | $\rightarrow$ Cross-domain features *transfer?* |
| Explanation track | $\rightarrow$ Transparent and explainable features |

# Goals

| | |
|---|---|
| Limited training data | $\rightarrow$ Cross-lingual features *transfer?* |
| Not all domains in provided data | $\rightarrow$ Cross-domain features *transfer?* |
| Explanation track | $\rightarrow$ Transparent and explainable features |

**Universal Dependencies** and **Readability Metrics**

# Method



input text
UD PARSING
Trankit
(Nguyen et al, 2021)
UD parse of input text
tokenization
EXTRACT 92 UD FEATURES
inspired by Profiling-UD
(Brunato et al., 2020; Sarti et al., 2021)
EXTRACT 42 SURFACE & READABILITY FEATURES
e.g., from translationese research
(Vanmassenhove et al., 2021)
TRAIN SVM

# Experiments & Scrapped Ideas

Separate model per language $\rightarrow$ Mostly worse performance

# Experiments & Scrapped Ideas

| | |
|---|---|
| Separate model per language | $\rightarrow$ Mostly worse performance |
| `DeBERTa-v3-large` model | $\rightarrow$ Not in the spirit of explainable, cross-lingual & cross-domain |

# Experiments & Scrapped Ideas

| | |
|---|---|
| Separate model per language | $\rightarrow$ Mostly worse performance |
| `DeBERTa-v3-large` model | $\rightarrow$ Not in the spirit of explainable, cross-lingual & cross-domain |
| Domain splitting | $\rightarrow$ Performance not great and unknown domains in test set |

# Experiments & Scrapped Ideas

| | |
|---|---|
| Separate model per language | $\rightarrow$ Mostly worse performance |
| `DeBERTa-v3-large` model | $\rightarrow$ Not in the spirit of explainable, cross-lingual & cross-domain |
| Domain splitting | $\rightarrow$ Performance not great and unknown domains in test set |
| More data from similar shared task | $\rightarrow$ Only more for English, different setup |

# Results development set

| | Submission | Macro-accuracy | News | Twitter | Reviews | Poetry | Mystery genre | Open-source |
|---|---|---|---|---|---|---|---|---|
| English | DetecTUM_1 | 0.86 | 0.88 | 0.72 | 0.84 | 0.85 | 0.88 | 1.00 |
| | DetecTUM_2 | 0.83 | 0.90 | 0.68 | 0.88 | 0.70 | 0.94 | 0.90 |
| | Elsevier_2 | 0.82 | 0.99 | 0.72 | 0.69 | 0.50 | 1.00 | 1.00 |
| | Hans_van_Halteren_1 | 0.78 | 0.98 | 0.81 | 0.72 | 0.65 | 0.50 | 1.00 |
| | Elsevier_1 | 0.77 | 0.88 | 0.55 | 0.69 | 0.50 | 1.00 | 1.00 |
| | Hans_van_Halteren_2 | 0.67 | 0.92 | 0.71 | 0.55 | 0.35 | 0.50 | 1.00 |
| | NLP_MS_1 | 0.63 | 0.65 | 0.62 | 0.50 | 0.50 | 0.81 | 0.70 |
| Dutch | Elsevier_2 | 0.80 | 0.94 | 0.69 | 0.88 | 0.60 | 0.90 | - |
| | Hans_van_Halteren_1 | 0.79 | 0.99 | 0.69 | 0.76 | 0.65 | 0.85 | - |
| | Elsevier_1 | 0.79 | 0.94 | 0.64 | 0.88 | 0.60 | 0.90 | - |
| | Hans_van_Halteren_2 | 0.78 | 0.94 | 0.61 | 0.75 | 0.70 | 0.92 | - |
| | DetecTUM_1 | 0.73 | 0.96 | 0.70 | 0.82 | 0.55 | 0.62 | - |
| | DetecTUM_2 | 0.73 | 0.98 | 0.64 | 0.81 | 0.50 | 0.72 | - |
| | NLP_MS_1 | 0.53 | 0.59 | 0.50 | 0.49 | 0.55 | 0.50 | - |

DetecTUM_1: Single SVM trained on both languages and all domains. DetecTUM_2: Same, but model per language.

# Results final set

| | Team | Macro-accuracy | News | Twitter | Reviews | Poetry | Mystery genre | Open-source |
|---|---|---|---|---|---|---|---|---|
| English | Hans_van_Halteren | 0.85 | 0.99 | 0.69 | 0.82 | 0.63 | 0.99 | 0.96 |
| | DetecTUM | 0.82 | 0.93 | 0.69 | 0.78 | 0.80 | 0.78 | 0.92 |
| | Elsevier | 0.81 | 0.98 | 0.65 | 0.75 | 0.50 | 0.99 | 0.98 |
| | NLP_MS | 0.74 | 0.87 | 0.63 | 0.63 | 0.65 | 0.85 | 0.82 |
| Dutch | Elsevier | 0.75 | 0.95 | 0.70 | 0.77 | 0.50 | 0.84 | - |
| | DetecTUM | 0.74 | 0.96 | 0.74 | 0.80 | 0.53 | 0.67 | - |
| | Hans_van_Halteren | 0.72 | 0.97 | 0.58 | 0.78 | 0.53 | 0.74 | - |
| | NLP_MS | 0.71 | 0.90 | 0.78 | 0.70 | 0.56 | 0.60 | - |

# Results final set

| | Team | Macro-accuracy | News | Twitter | Reviews | Poetry | Mystery genre | Open-source |
|---|---|---|---|---|---|---|---|---|
| English | Hans_van_Halteren | 0.85 | 0.99 | 0.69 | 0.82 | 0.63 | 0.99 | 0.96 |
| | DetecTUM | 0.82 | 0.93 | 0.69 | 0.78 | 0.80 | 0.78 | 0.92 |
| | Elsevier | 0.81 | 0.98 | 0.65 | 0.75 | 0.50 | 0.99 | 0.98 |
| | NLP_MS | 0.74 | 0.87 | 0.63 | 0.63 | 0.65 | 0.85 | 0.82 |
| Dutch | Elsevier | 0.75 | 0.95 | 0.70 | 0.77 | 0.50 | 0.84 | - |
| | DetecTUM | 0.74 | 0.96 | 0.74 | 0.80 | 0.53 | 0.67 | - |
| | Hans_van_Halteren | 0.72 | 0.97 | 0.58 | 0.78 | 0.53 | 0.74 | - |
| | NLP_MS | 0.71 | 0.90 | 0.78 | 0.70 | 0.56 | 0.60 | - |

A little disappointed. . .

# Insights

| Language | Domain | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|---|
| English | News | 0.66 (avg_syl) | 0.56 (dep_dist_det) | 0.56 (dep_dist_amod) |
| | Twitter | 0.54 (n_#) | 0.44 (n_}) | 0.35 (n_sentences) |
| | Reviews | 0.57 (n_,) | 0.39 (dep_dist_punct) | 0.38 (upos_dist_PUNCT) |
| | Poetry | 0.40 (n_') | 0.32 (n_,) | 0.21 (dep_dist_discourse) |
| | Columns | 0.84 (avg_syl) | 0.70 (dep_dist_det) | 0.68 (upos_dist_DET) |
| | Open-source | 0.64 (avg_syl) | 0.59 (char_per_tok) | 0.57 (dep_dist_amod) |
| Dutch | News | 0.66 (verbs_form_dist_Inf) | 0.63 (dep_dist_mark) | 0.63 (char_per_tok) |
| | Twitter | 0.55 (n_") | 0.44 (upos_dist_PUNCT) | 0.42 (dep_dist_punct) |
| | Reviews | 0.24 (dep_dist_det) | 0.24 (avg_syl) | 0.23 (n_}) |
| | Poetry | 0.32 (n_,) | 0.24 (upos_dist_ADJ) | 0.21 (lexical_density) |
| | Columns | 0.69 (lfp_b1) | 0.67 (dep_dist_mark) | 0.67 (dep_dist_cop) |

Top three most *positively* correlated* features per language and domain (out of 136 features).

* *All are at least $p < 0.05$ significant, most are $p < 0.0001$ significant.*

# Insights

| Language | Domain | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|---|
| English | News | 0.66 (avg_syl) | 0.56 (dep_dist_det) | 0.56 (dep_dist_amod) |
| | Twitter | 0.54 (n_#) | 0.44 (n_}) | 0.35 (n_sentences) |
| | Reviews | 0.57 (n_,) | 0.39 (dep_dist_punct) | 0.38 (upos_dist_PUNCT) |
| | Poetry | 0.40 (n_') | 0.32 (n_,) | 0.21 (dep_dist_discourse) |
| | Columns | 0.84 (avg_syl) | 0.70 (dep_dist_det) | 0.68 (upos_dist_DET) |
| | Open-source | 0.64 (avg_syl) | 0.59 (char_per_tok) | 0.57 (dep_dist_amod) |
| Dutch | News | 0.66 (verbs_form_dist_Inf) | 0.63 (dep_dist_mark) | 0.63 (char_per_tok) |
| | Twitter | 0.55 (n_") | 0.44 (upos_dist_PUNCT) | 0.42 (dep_dist_punct) |
| | Reviews | 0.24 (dep_dist_det) | 0.24 (avg_syl) | 0.23 (n_}) |
| | Poetry | 0.32 (n_,) | 0.24 (upos_dist_ADJ) | 0.21 (lexical_density) |
| | Columns | 0.69 (lfp_b1) | 0.67 (dep_dist_mark) | 0.67 (dep_dist_cop) |

Top three most *positively* correlated* features per language and domain (out of 136 features).

** All are at least $p < 0.05$ significant, most are $p < 0.0001$ significant.*

# Insights

| Language | Domain | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|---|
| English | News | 0.66 (avg_syl) | 0.56 (dep_dist_det) | 0.56 (dep_dist_amod) |
| | Twitter | 0.54 (n_#) | 0.44 (n_}) | 0.35 (n_sentences) |
| | Reviews | 0.57 (n_,) | 0.39 (dep_dist_punct) | 0.38 (upos_dist_PUNCT) |
| | Poetry | 0.40 (n_') | 0.32 (n_,) | 0.21 (dep_dist_discourse) |
| | Columns | 0.84 (avg_syl) | 0.70 (dep_dist_det) | 0.68 (upos_dist_DET) |
| | Open-source | 0.64 (avg_syl) | 0.59 (char_per_tok) | 0.57 (dep_dist_amod) |
| Dutch | News | 0.66 (verbs_form_dist_Inf) | 0.63 (dep_dist_mark) | 0.63 (char_per_tok) |
| | Twitter | 0.55 (n_") | 0.44 (upos_dist_PUNCT) | 0.42 (dep_dist_punct) |
| | Reviews | 0.24 (dep_dist_det) | 0.24 (avg_syl) | 0.23 (n_}) |
| | Poetry | 0.32 (n_,) | 0.24 (upos_dist_ADJ) | 0.21 (lexical_density) |
| | Columns | 0.69 (lfp_b1) | 0.67 (dep_dist_mark) | 0.67 (dep_dist_cop) |

Top three most *positively* correlated* features per language and domain (out of 136 features).

* *All are at least $p < 0.05$ significant, most are $p < 0.0001$ significant.*

# Insights

| Language | Domain | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|---|
| English | News | -0.593 (upos_dist_PROPN) | -0.548 (verbs_tense_dist_Past) | -0.536 (dep_dist_nummod) |
| | Twitter | -0.284 (lexical_density) | -0.282 (tokens_per_sent) | -0.262 (dep_dist_flat) |
| | Reviews | -0.433 (lfp_b1) | -0.365 (char_per_tok) | -0.299 (n_() |
| | Poetry | -0.418 (n_;) | -0.365 (n_:) | -0.305 (upos_dist_NUM) |
| | Columns | -0.728 (dep_dist_ccomp) | -0.699 (dep_dist_obl:tmod) | -0.699 (yules_i) |
| | Open-source | -0.688 (verbs_tense_dist_Past) | -0.663 (verbs_form_dist_Fin) | -0.663 (verbs_mood_dist_Ind) |
| Dutch | News | -0.742 (ttr) | -0.727 (yules_i) | -0.705 (ttr_lemma_chunks_200) |
| | Twitter | -0.442 (char_per_tok) | -0.323 (ttr_lemma_chunks_200) | -0.320 (ttr) |
| | Reviews | -0.373 (upos_dist_ADV) | -0.327 (dep_dist_parataxis) | -0.322 (dep_dist_advmod) |
| | Poetry | -0.351 (n_:) | -0.336 (dep_dist_obl) | -0.294 (verbal_head_per_sent) |
| | Columns | -0.701 (flesch_mod) | -0.693 (gunning_fog) | -0.692 (flesch) |

Top three most *negatively* correlated* features per language and domain (out of 136 features).

** All are at least $p < 0.001$ significant, most are $p < 0.0001$ significant.*

# Insights

| Language | Domain | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|---|
| English | News | -0.593 (upos_dist_PROPN) | -0.548 (verbs_tense_dist_Past) | -0.536 (dep_dist_nummod) |
| | Twitter | -0.284 (lexical_density) | -0.282 (tokens_per_sent) | -0.262 (dep_dist_flat) |
| | Reviews | -0.433 (lfp_b1) | -0.365 (char_per_tok) | -0.299 (n_() |
| | Poetry | -0.418 (n_;) | -0.365 (n_:) | -0.305 (upos_dist_NUM) |
| | Columns | -0.728 (dep_dist_ccomp) | -0.699 (dep_dist_obl:tmod) | -0.699 (yules_i) |
| | Open-source | -0.688 (verbs_tense_dist_Past) | -0.663 (verbs_form_dist_Fin) | -0.663 (verbs_mood_dist_Ind) |
| Dutch | News | -0.742 (ttr) | -0.727 (yules_i) | -0.705 (ttr_lemma_chunks_200) |
| | Twitter | -0.442 (char_per_tok) | -0.323 (ttr_lemma_chunks_200) | -0.320 (ttr) |
| | Reviews | -0.373 (upos_dist_ADV) | -0.327 (dep_dist_parataxis) | -0.322 (dep_dist_advmod) |
| | Poetry | -0.351 (n_:) | -0.336 (dep_dist_obl) | -0.294 (verbal_head_per_sent) |
| | Columns | -0.701 (flesch_mod) | -0.693 (gunning_fog) | -0.692 (flesch) |

Top three most *negatively* correlated* features per language and domain (out of 136 features).

* *All are at least* $p < 0.001$ *significant, most are* $p < 0.0001$ *significant.*

# Insights

| Language | Domain | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|---|
| English | News | -0.593 (upos_dist_PROPN) | -0.548 (verbs_tense_dist_Past) | -0.536 (dep_dist_nummod) |
| | Twitter | -0.284 (lexical_density) | -0.282 (tokens_per_sent) | -0.262 (dep_dist_flat) |
| | Reviews | -0.433 (lfp_b1) | -0.365 (char_per_tok) | -0.299 (n_() |
| | Poetry | -0.418 (n_;) | -0.365 (n_:) | -0.305 (upos_dist_NUM) |
| | Columns | -0.728 (dep_dist_ccomp) | -0.699 (dep_dist_obl:tmod) | -0.699 (yules_i) |
| | Open-source | -0.688 (verbs_tense_dist_Past) | -0.663 (verbs_form_dist_Fin) | -0.663 (verbs_mood_dist_Ind) |
| Dutch | News | -0.742 (ttr) | -0.727 (yules_i) | -0.705 (ttr_lemma_chunks_200) |
| | Twitter | -0.442 (char_per_tok) | -0.323 (ttr_lemma_chunks_200) | -0.320 (ttr) |
| | Reviews | -0.373 (upos_dist_ADV) | -0.327 (dep_dist_parataxis) | -0.322 (dep_dist_advmod) |
| | Poetry | -0.351 (n_:) | -0.336 (dep_dist_obl) | -0.294 (verbal_head_per_sent) |
| | Columns | -0.701 (flesch_mod) | -0.693 (gunning_fog) | -0.692 (flesch) |

Top three most *negatively* correlated* features per language and domain (out of 136 features).

* *All are at least* $p < 0.001$ *significant, most are* $p < 0.0001$ *significant.*

# Next steps

Analyse results more $\rightarrow$ Domain influence, other detection datasets

# Next steps

| | |
|---|---|
| Analyse results more | $\rightarrow$ Domain influence, other detection datasets |
| Ensemble of approaches? | $\rightarrow$ Power of neural models with explainable linguistic features |

# Next steps

| | |
|---|---|
| Analyse results more | $\rightarrow$ Domain influence, other detection datasets |
| Ensemble of approaches? | $\rightarrow$ Power of neural models with explainable linguistic features |
| Try out adversarial samples | $\rightarrow$ Rephrasing, prompting techniques, ... |

# Next steps

| | |
|---|---|
| Analyse results more | $\rightarrow$ Domain influence, other detection datasets |
| Ensemble of approaches? | $\rightarrow$ Power of neural models with explainable linguistic features |
| Try out adversarial samples | $\rightarrow$ Rephrasing, prompting techniques, ... |
| CLIN paper | $\rightarrow$ Explain contributing features, inspiration from other fields, ... |

Wessel Poelman[†], Juraj Vladika[†], Esther Ploeger[§],

Florian Matthes[†] and Johannes Bjerva[§]

† *Technical University of Munich*

§ *Aalborg University*

September 22, 2023