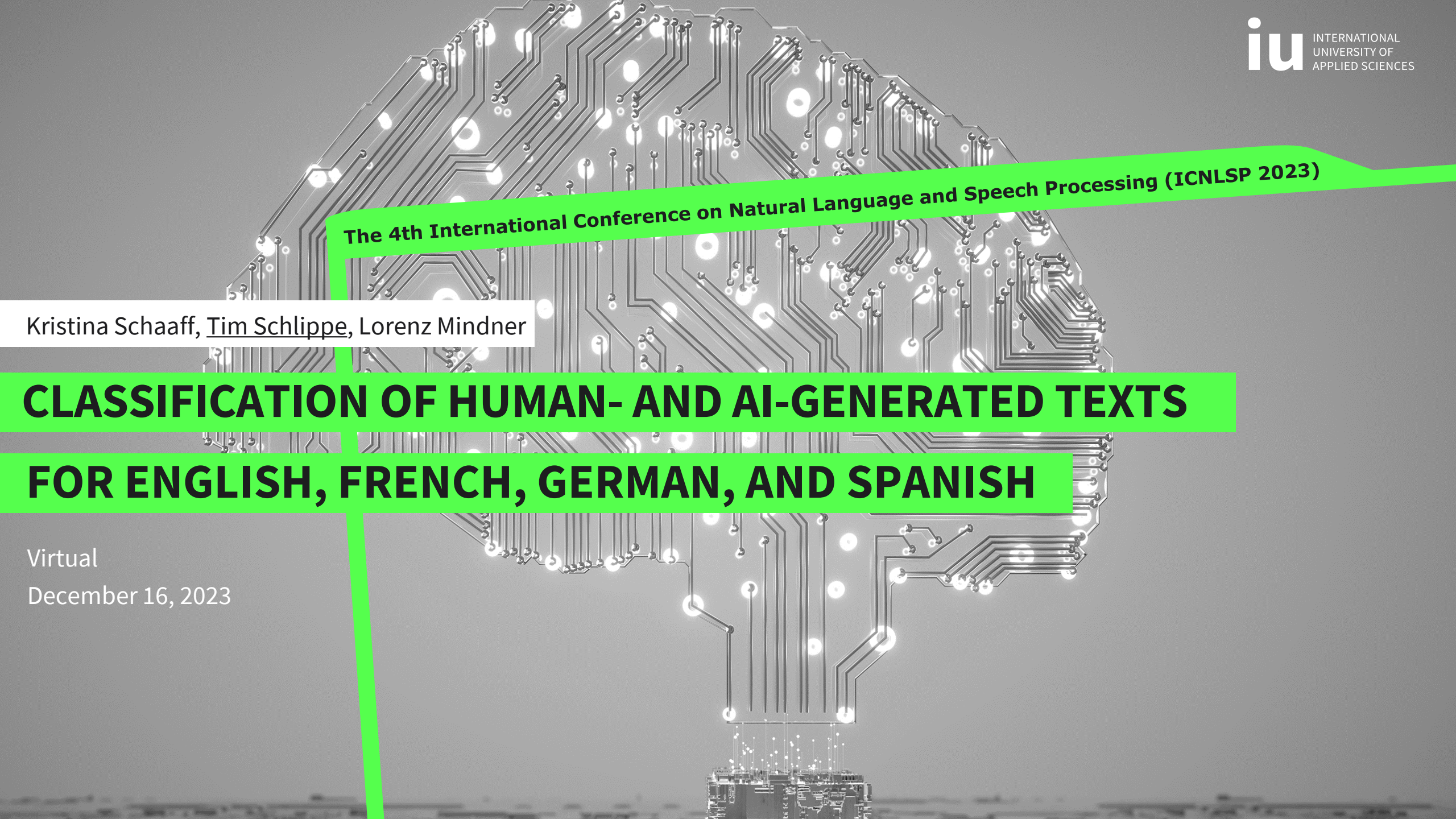


INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

The 4th International Conference on Natural Language and Speech Processing (ICNLSP 2023)

Kristina Schaaff, Tim Schlippe, Lorenz Mindner

CLASSIFICATION OF HUMAN- AND AI-GENERATED TEXTS FOR ENGLISH, FRENCH, GERMAN, AND SPANISH

Virtual

December 16, 2023

AGENDA

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

1

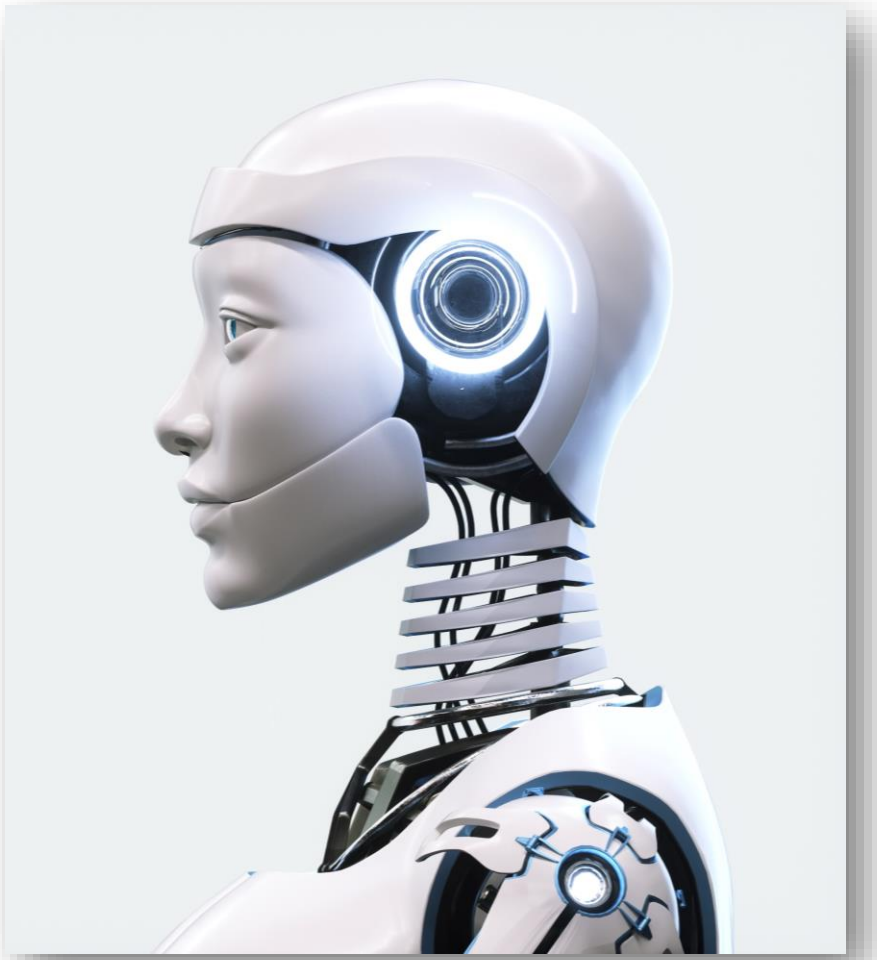
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

MOTIVATION

TYPICAL USE CASES FOR CHATBOTS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- Learning Support
- Chat Companion
- Medical Advice
- Customer Support



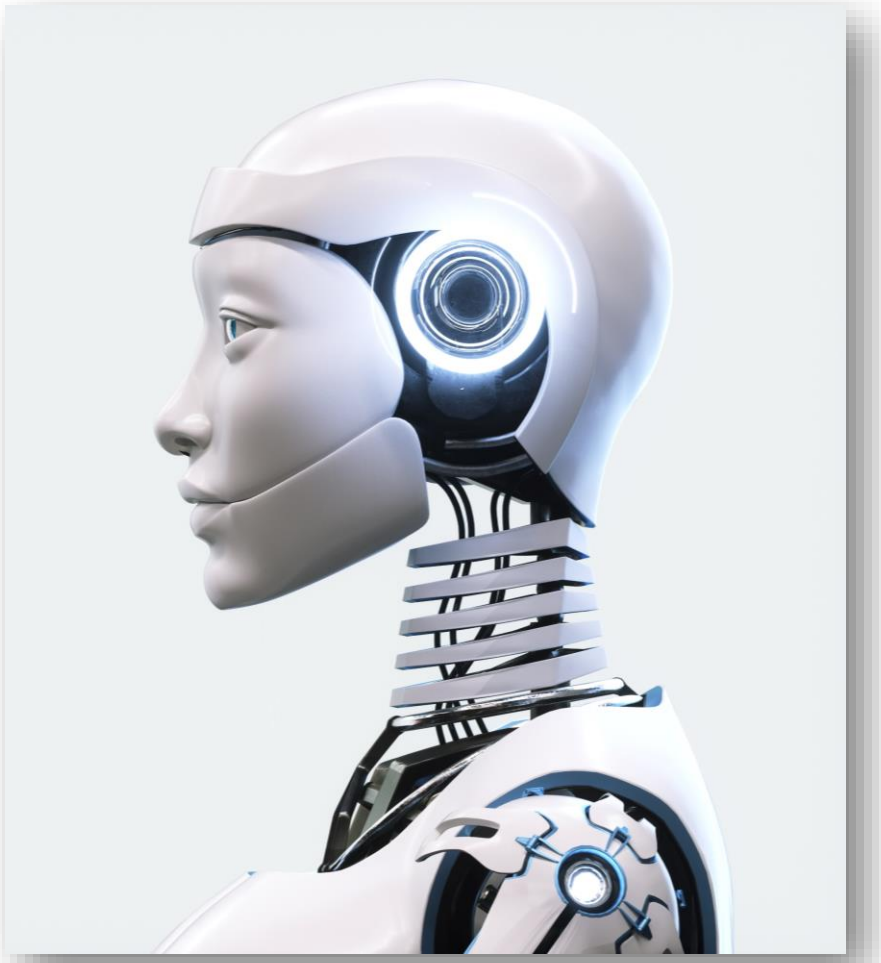
Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. Kristina Schaaff, Tim Schlippe, Lorenz Mindner. ICNLSP 2023

Image source: Microsoft Archive

DANGERS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- Fake News
- Spamming
- Plagiarism
- ...



Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. Kristina Schaaff, Tim Schlippe, Lorenz Mindner. ICNLSP 2023

Image source: Microsoft Archive

RESEARCH QUESTIONS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

RQ1

FEATURES

How effective are different features in distinguishing between human- and AI-generated texts from the educational domain?

RQ2

PROMPTS

What impact do we have if we specifically tell the AI to generate text as a human would do it?

RQ3

LANGUAGE

How do the classification results vary across different languages?

RQ4

TEXT DOMAIN

Do the features perform comparably well in the news domain?

2

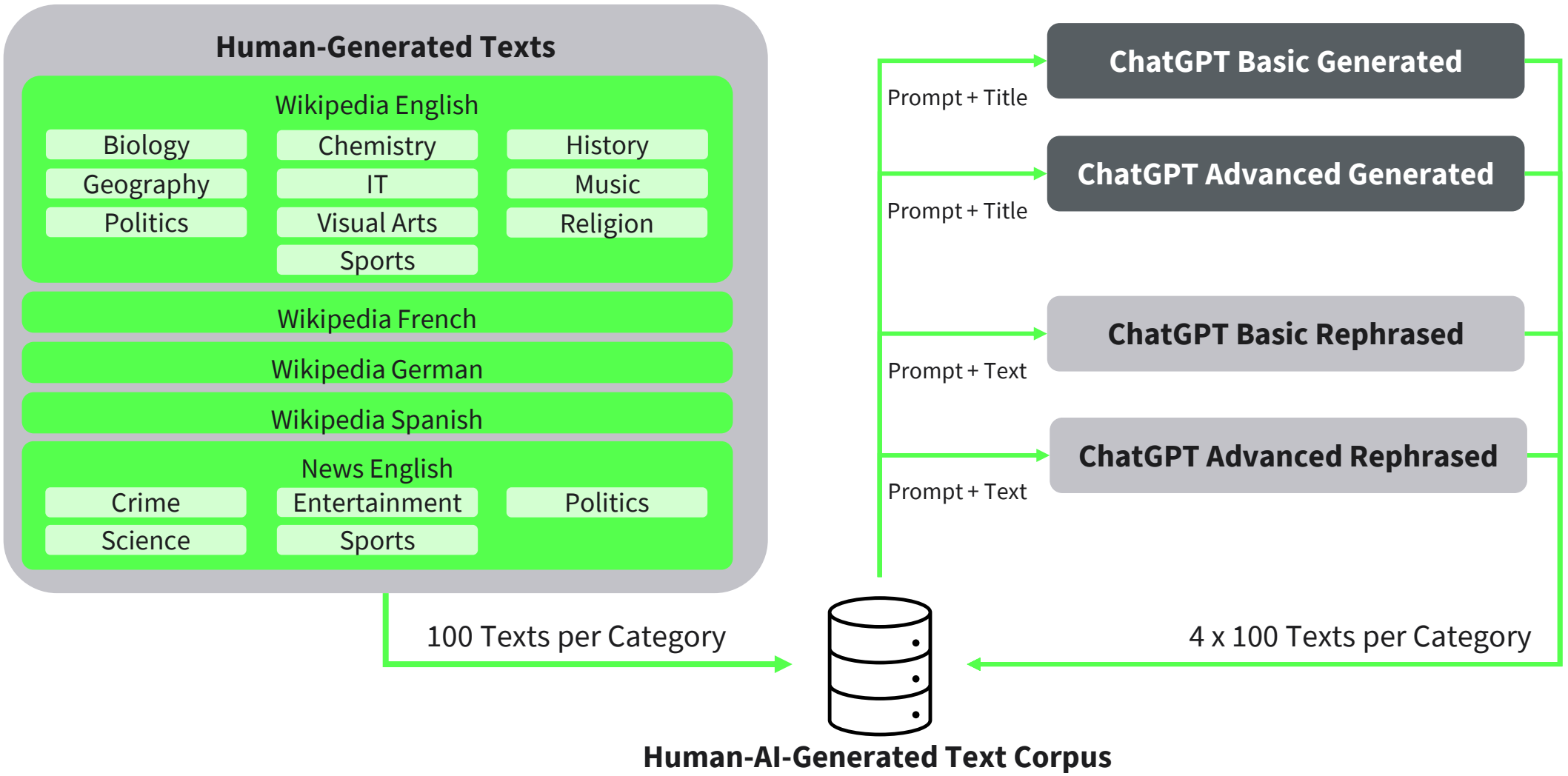
iu INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

HUMAN-AI-GENERATED TEXT CORPUS

https://github.com/LorenzM97/human-AIgeneratedTextCorpus

HUMAN-AI-GENERATED TEXT CORPUS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES



Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. Kristina Schaaff, Tim Schlippe, Lorenz Mindner. ICNLSP 2023

HUMAN-AI-GENERATED TEXT CORPUS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Basic Generated

Prompt

Generate a text on the following topic: Australia

Advanced Generated

Prompt

Generate a text on the following topic in a way a human would do it: Australia

Basic Rephrased

Prompt

Rephrase the following text: Australia, officially the Commonwealth of Australia, is a sovereign country […].

Advanced Rephrased

Prompt

Rephrase the following text in a way a human would do it: Australia, officially the Commonwealth of Australia, is a sovereign country […].

Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. Kristina Schaaff, Tim Schlippe, Lorenz Mindner. ICNLSP 2023

3

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

FEATURES FOR THE CLASSIFICATION OF HUMAN- AND AI-GENERATED TEXTS

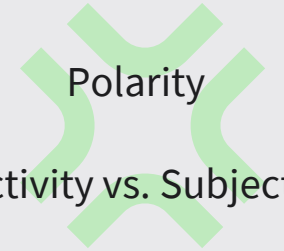
FEATURES

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Perplexity Features

Mean & max perplexity

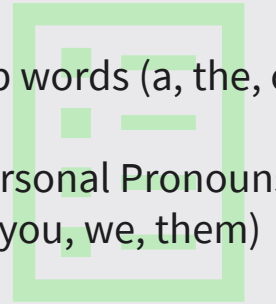
Semantic Features



Polarity

Objectivity vs. Subjectivity

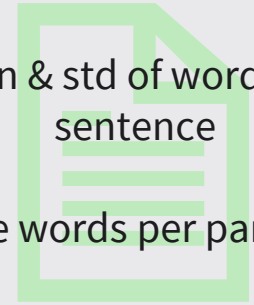
List Lookup Features



Stop words (a, the, of)

Personal Pronouns (you, we, them)

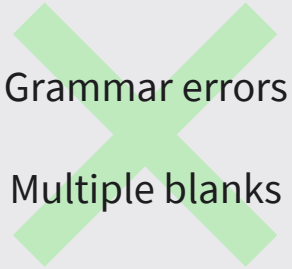
Document Features



Mean & std of words per sentence

Unique words per paragraph

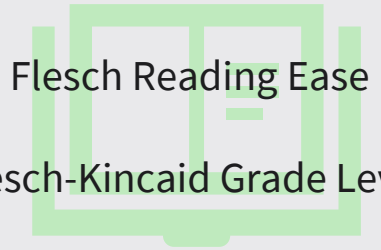
Error-Based Features



Grammar errors

Multiple blanks

Readability Features



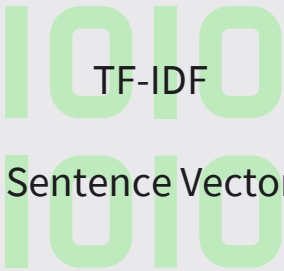
Flesch Reading Ease

Flesch-Kincaid Grade Level

AI Feedback Features

@ChatGPT:
"*Was this text generated by ChatGPT? [text]*"

Text Vector Features

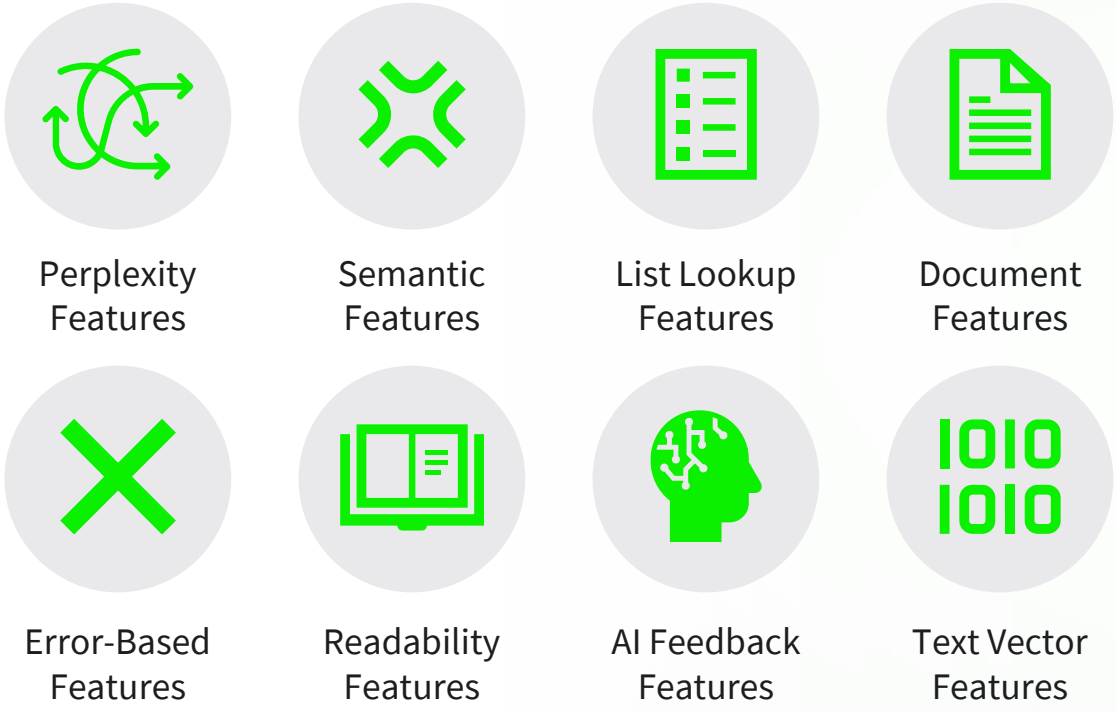


TF-IDF

Sentence Vector

Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. Kristina Schaaff, Tim Schlippe, Lorenz Mindner. ICNLSP 2023

FEATURES



8 feature categories, 37 features

Category	Feature	Description	Reference
Perplexity	PPL_{mean}	mean PPL	[21][14][19]
	PPL_{max}	maximum PPL	[21][14][19]
Semantic	$sentiment_{polarity}$	degree of positivity/negativity [-1,+1]	[14][19]
	$sentiment_{subjectivity}$	degree of subjectivity [0,+1]	new
List Lookup	$stopWord_{count}$	number of stop words	[17]
	$specialChar_{count}$	number of special characters	[28]
	$discourseMarker_{count}$	number of discourse markers	new
	$titleRepetition_{count}$	absolute repetitions of title	new
	$titleRepetition_{relative}$	relative repetitions of title	new
Document	$wordsPerParagraph_{mean}$	ø number of words per paragraph	[28]
	$wordsPerParagraph_{stdev}$	stdev of $wordsPerParagraph$	[28]
	$sentencesPerParagraph_{mean}$	ø number of sentences per paragraph	[28]
	$sentencesPerParagraph_{stdev}$	stdev of $sentencesPerParagraph$	[28]
	$wordsPerSentence_{mean}$	ø number of words per sentence	[28]
	$wordsPerSentence_{stdev}$	stdev of $wordsPerSentence$	[28]
	$uniqWordsPerSentence_{mean}$	ø number of unique words per sentence	[17]
	$uniqWordsPerSentence_{stdev}$	stdev of $uniqWordsPerSentence$	new
	$words_{count}$	number of running words	[19][17][28]
	$uniqWords_{count}$	number of unique words	[28]
	$uniqWords_{relative}$	relative number of unique words	[28]
	$paragraph_{count}$	number of paragraphs	[28]
	$sentence_{count}$	number of sentences	[28]
	$punctuation_{count}$	number of punctuation marks	[28]
	$quotation_{count}$	number of quotation marks	new
	$character_{count}$	number of characters	[28]
	$uppercaseWords_{relative}$	relative number of words in uppercase	[17]
	$personalPronoun_{count}$	absolute number of personal pronouns	[14]
	$personalPronoun_{relative}$	relative number of personal pronouns	[14]
	$POSPerSentence_{mean}$	ø number of unique POS-tags/sentence	[19][28][18]
Error-Based	$grammarError_{count}$	number of spelling/grammar errors	new
	$multiBlank_{count}$	number of multiple blanks	new
Readability	$fleschReadingEase$	Flesch Reading Ease score [0-100]	[17][29]
	$fleschKincaidGradeLevel$	Readability as U.S. grade level [0-100]	[17][30]
AI Feedback	$AIFeedback$	Ask AI if text was generated by AI	new
Text Vector	*TF-IDF*	500-dim TF-IDF vector of 1-/2-grams	[17][31]
	Sentence-BERT	ø Sentence-BERT vector	[32]
	Sentence-BERT-dist	ø distance of Sentence-BERT vectors	new

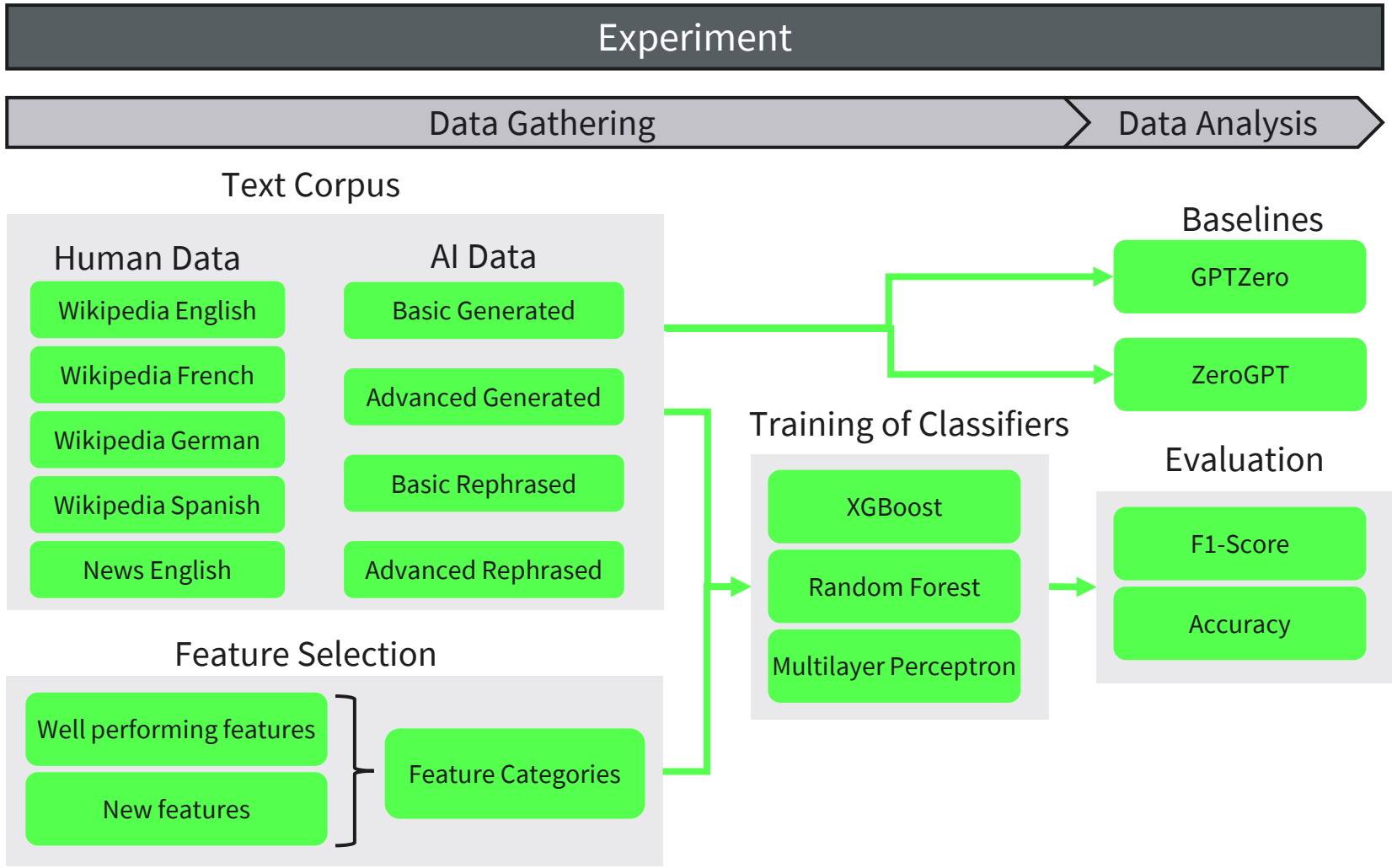
Table 3: Summary of our Features for the Classification of Generated Texts.

4

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

EXPERIMENTS AND RESULTS

EXPERIMENTAL SETUP

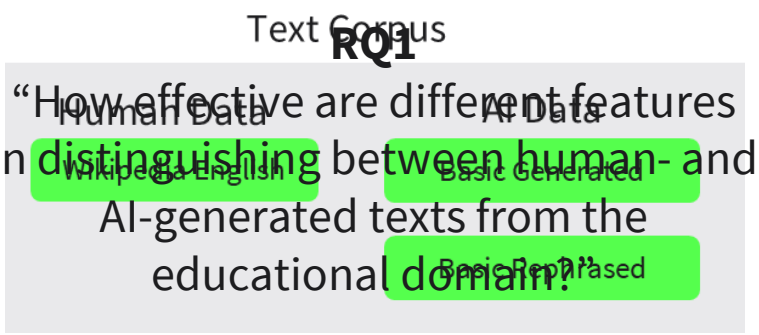


Training for:

- Human vs. basic AI-generated
- Human vs. advanced AI-generated
- Human vs. basic AI-rephrased
- Human vs. advanced AI-rephrased

Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. Kristina Schaaff, Tim Schlippe, Lorenz Mindner. ICNLSP 2023

RESULTS
RQ1: FEATURES



RQ1

“How effective are different features in distinguishing between human- and AI-generated texts from the educational domain?”

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Comparison of Feature Categories

F1-Scores Basic Generated

1	94.9% $\text{TextVector}_{\text{traditional+new}}$	$\text{Readability}_{\text{traditional}}$ 59.3%	1
1	94.9% $\text{Document}_{\text{traditional}}$	$\text{ErrorBased}_{\text{new}}$ 63.9%	2
3	85.3% $\text{Perplexity}_{\text{traditional}}$	$\text{AIFeedback}_{\text{new}}$ 68.1%	3

F1-Scores Basic Rephrased

1	79.7% $\text{Document}_{\text{traditional}}$	$\text{AIFeedback}_{\text{new}}$ 50.9%	1
2	78.2% $\text{TextVector}_{\text{traditional+new}}$	$\text{Readability}_{\text{traditional}}$ 51.1%	2
3	73.9% $\text{ListLookup}_{\text{traditional+new}}$	$\text{Semantic}_{\text{traditional+new}}$ 64.4%	3

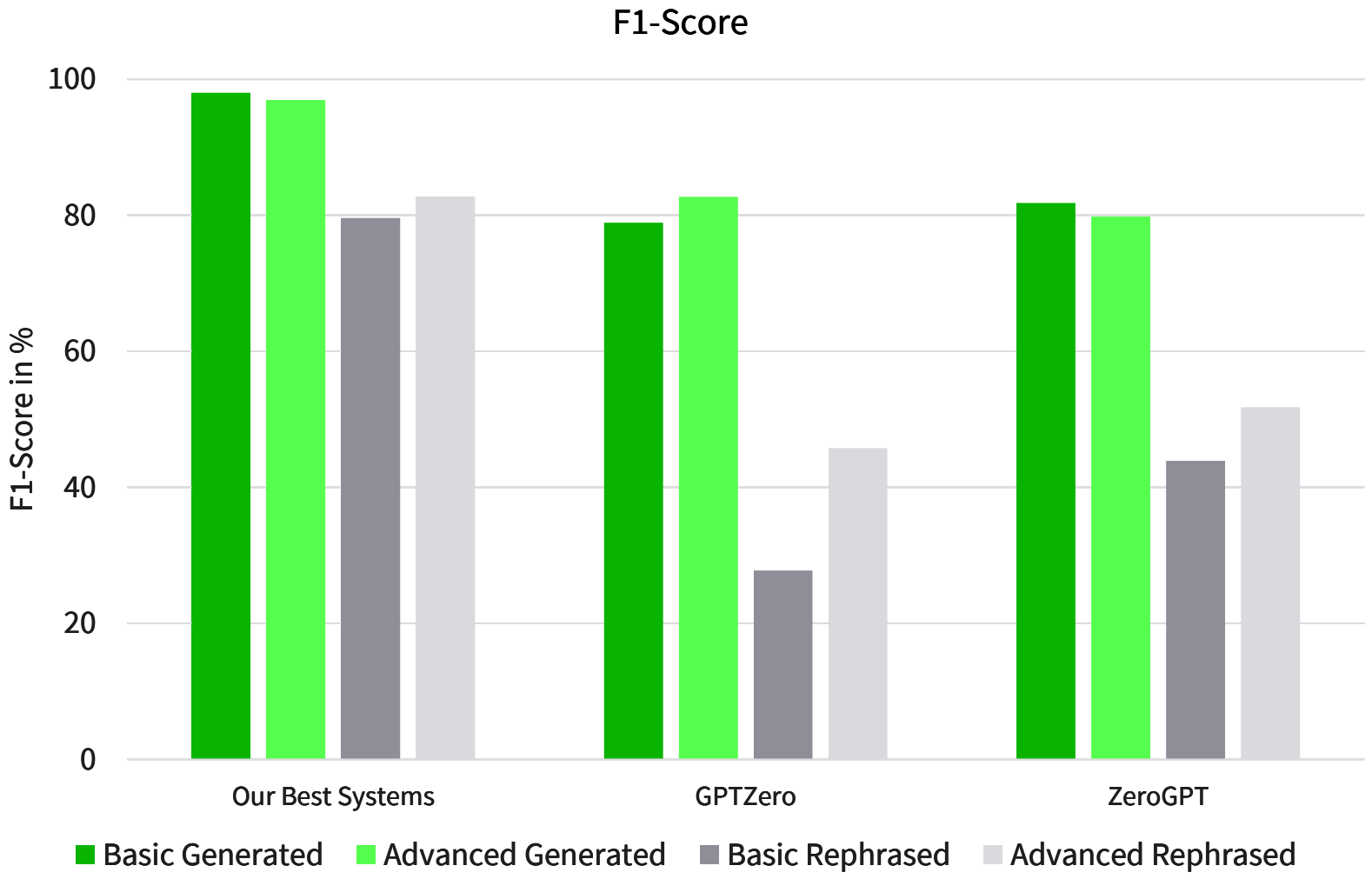
RESULTS
RQ2: PROMPT

RQ2

"What impact do we have if we specifically tell the AI to generate text as a human would do it?"

Text Corpus

Human Data	AI Data
Wikipedia English	Basic Generated
	Advanced Generated
	Basic Rephrased
	Advanced Rephrased



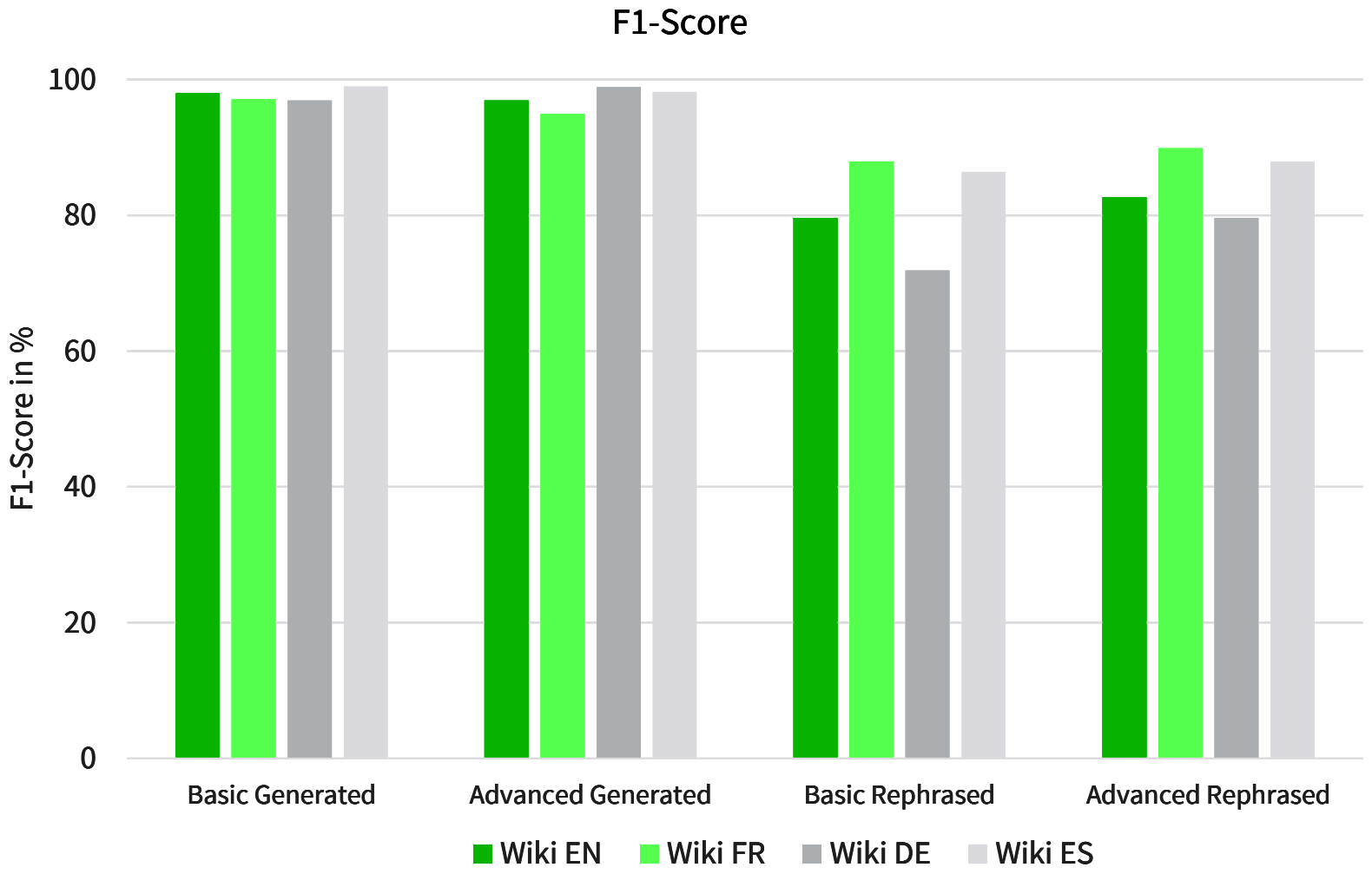
RESULTS
RQ3: LANGUAGE

RQ3

"How do the classification results vary across different languages?"

Text Corpus

Human Data	AI Data
Wikipedia English	Basic Generated
Wikipedia French	Advanced Generated
Wikipedia German	Basic Rephrased
Wikipedia Spanish	Advanced Rephrased



Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. Kristina Schaaff, Tim Schlippe, Lorenz Mindner. ICNLSP 2023

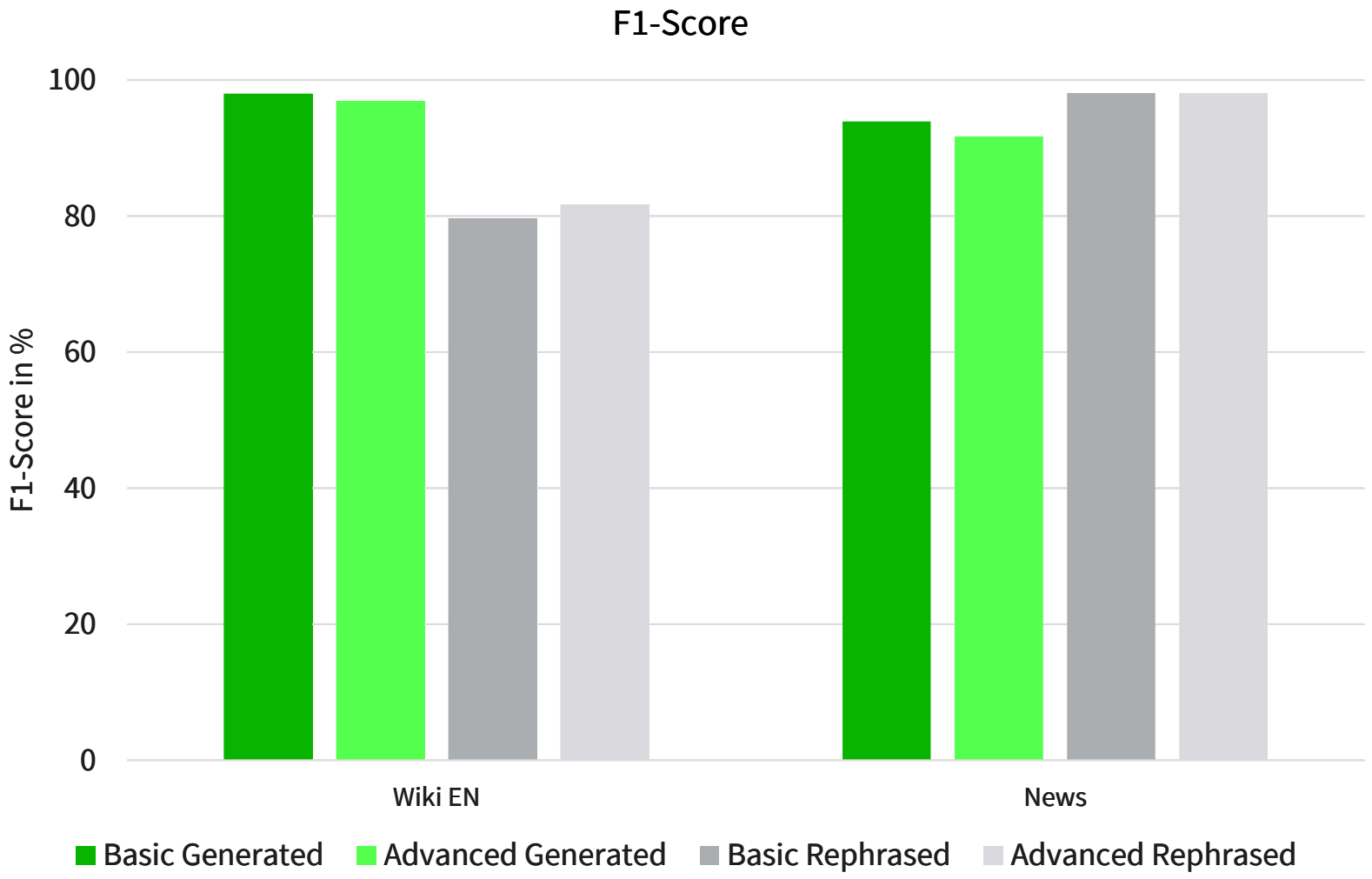
RESULTS
RQ4: TEXT DOMAIN

RQ4

"Do the features perform comparably well in the news domain?"

Text Corpus

Human Data	AI Data
Wikipedia English	Basic Generated
News English	Advanced Generated
	Basic Rephrased
	Advanced Rephrased



COMPARISON TO BASELINES
F1-SCORES OF BEST SYSTEMS

Wikipedia English

	GPTZero	ZeroGPT
Basic Generated	+24.2%	+19.8%
Basic Rephrased	+186.7%	+81.5%
Advanced Generated	+17.2%	+21.4%
Advanced Rephrased	+78.4%	+57.7%

News English

	GPTZero	ZeroGPT
Basic Generated	+25.9%	+21.0%
Basic Rephrased	+145.2%	+98.2%
Advanced Generated	+22.9%	+18.1%
Advanced Rephrased	+181.9%	+83.0%

	French	German	Spanish
	ZeroGPT	ZeroGPT	ZeroGPT
Basic Generated	+33.7%	+36.7%	+38.5%
Basic Rephrased	+30.4%	+45.3%	+35.6%
Advanced Generated	+28.9%	+51.2%	+37.3%
Advanced Rephrased	+37.7%	+61.5%	+40.6%

Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. Kristina Schaaff, Tim Schlippe, Lorenz Mindner. ICNLSP 2023

5

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

CONCLUSION AND FUTURE WORK

CONCLUSION & FUTURE WORK

iu INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Generated > Rephrased

Minor influence of prompt

Comparable Performance Across Languages

F1-scores for AI-generated texts between **97%** and **99%**

F1-scores for AI-rephrased texts between **78%** and **89%**

Educational Domain < News Domain

Comparable performance for AI-generated texts

Better results for AI-rephrased texts

Baselines < Our Systems

Our best basic text rephrasing detection system performs almost twice as good

Future Work

Improvement of text generation

Investigation of other prompts and text domains

Comparison to transformer-based models

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

THANK YOU

Tim Schlippe
tim.schlippe@iu.org