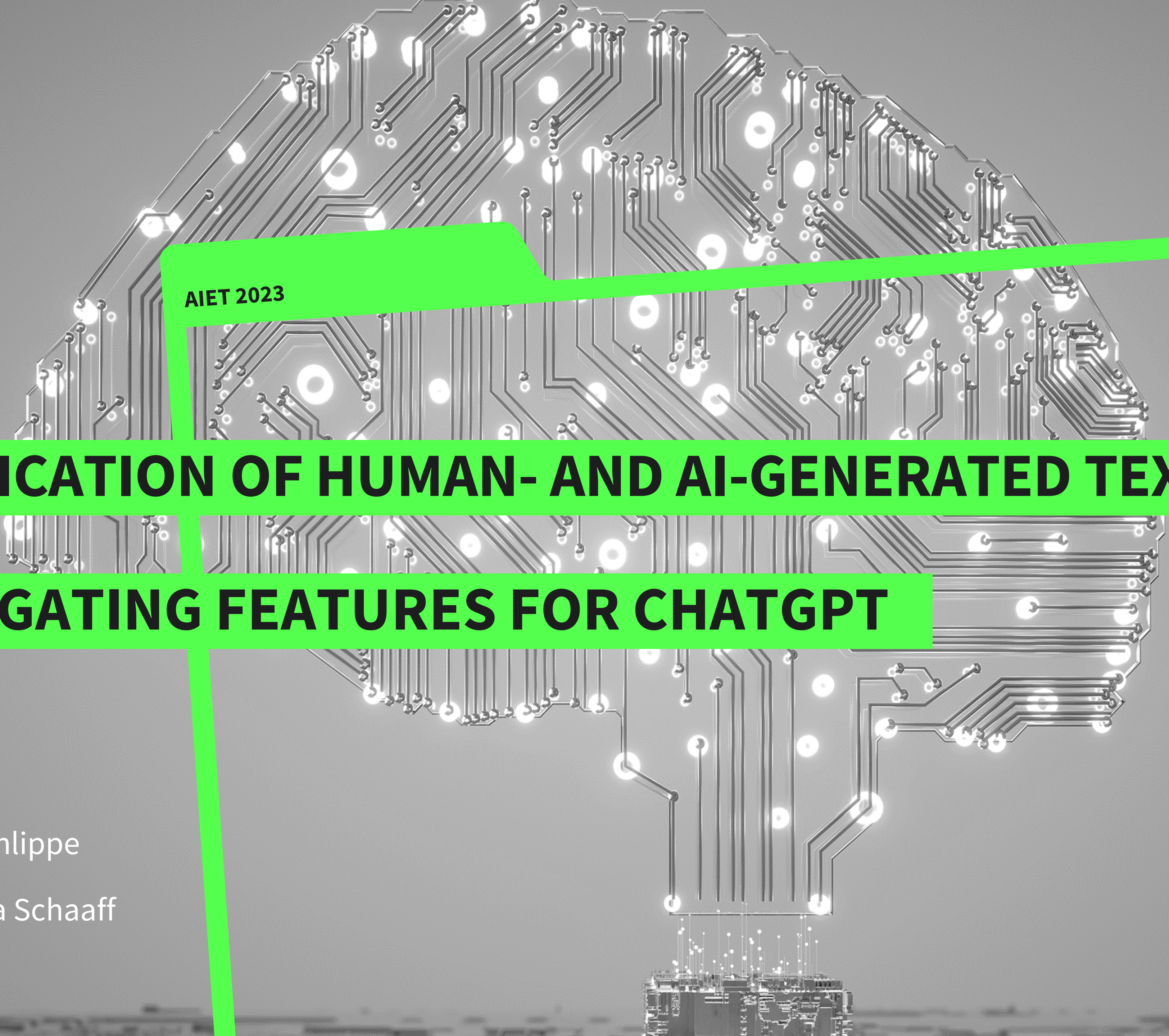


AIET 2023

CLASSIFICATION OF HUMAN- AND AI-GENERATED TEXTS: INVESTIGATING FEATURES FOR CHATGPT

Lorenz Mindner

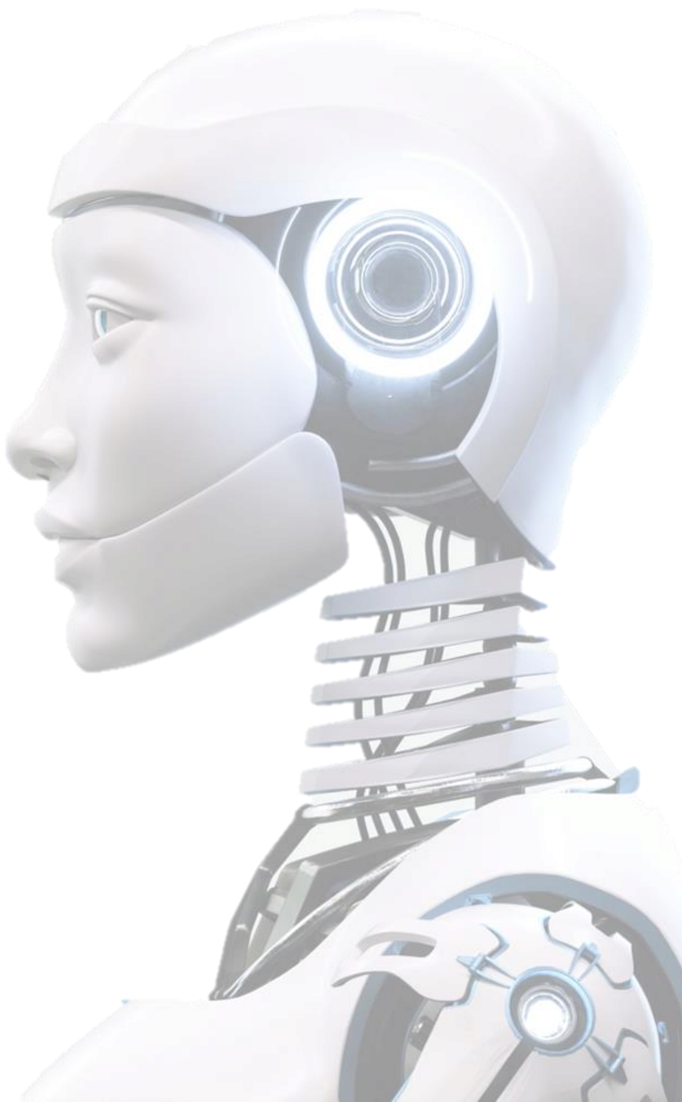
Prof. Dr. Tim Schlippe

Prof. Dr. Kristina Schaaff

01.07.2023

OUTLINE

- Motivation 1
- Related Research 2
- Human-AI-Generated Text Corpus 3
- Features 4
- Experiments and Results 5
- Conclusion & Future Work 6



MOTIVATION

Human-generated text

VS.

AI-generated text

01

PROVIDE NEW CORPUS

to contribute to the improvement of research

02

GAIN INSIGHTS INTO FEATURES

to distinguish human- and AI-generated texts

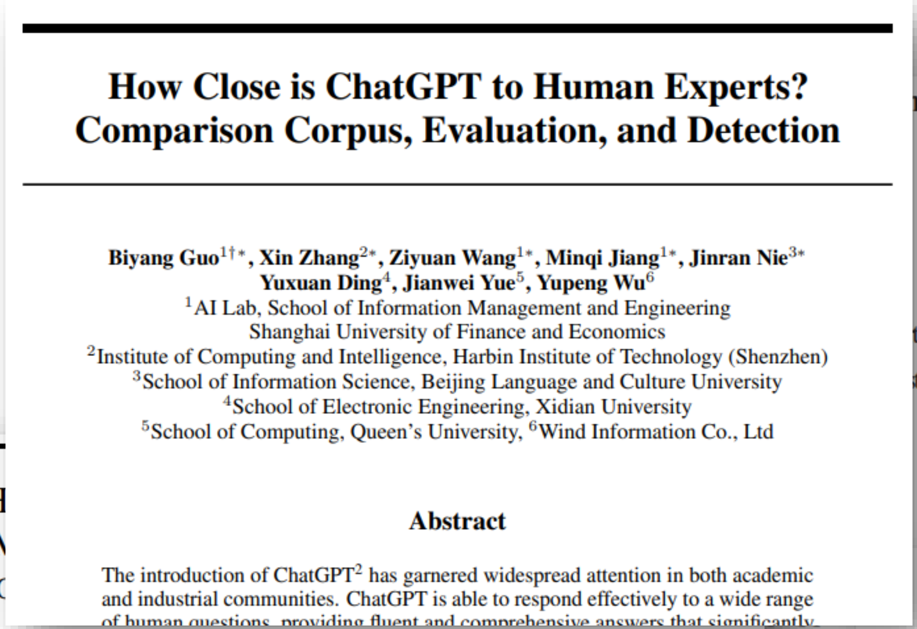
03

PROVIDE BENCHMARK

for future classifiers

RELATED RESEARCH

Study characteristics of human- and AI-written expert texts

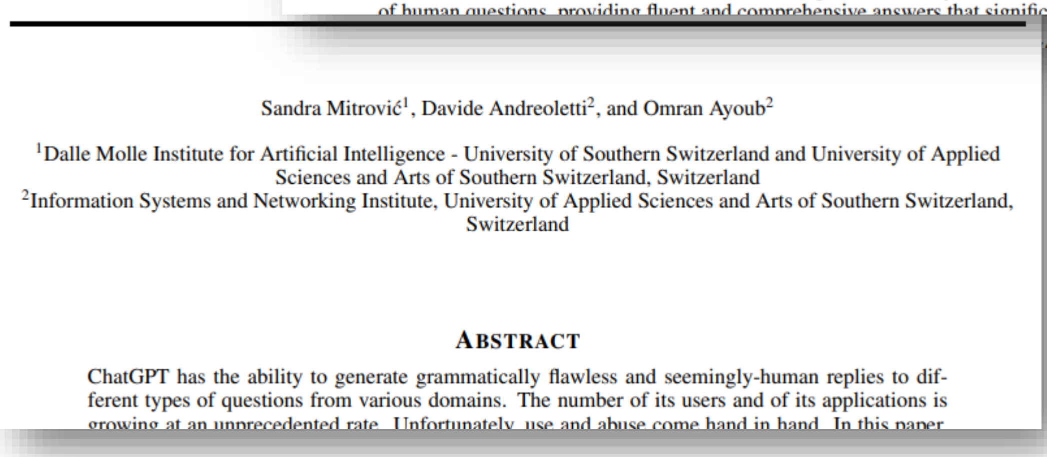


How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection

Biyang Guo[1†*], Xin Zhang[2*], Ziyuan Wang[1*], Minqi Jiang[1*], Jinran Nie[3*]
Yuxuan Ding[4], Jianwei Yue[5], Yupeng Wu[6]
[1]AI Lab, School of Information Management and Engineering
Shanghai University of Finance and Economics
[2]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen)
[3]School of Information Science, Beijing Language and Culture University
[4]School of Electronic Engineering, Xidian University
[5]School of Computing, Queen's University, [6]Wind Information Co., Ltd

Abstract

The introduction of ChatGPT[2] has garnered widespread attention in both academic and industrial communities. ChatGPT is able to respond effectively to a wide range of human questions, providing fluent and comprehensive answers that significantly



ChatGPT or H
Decisions of M
Sho

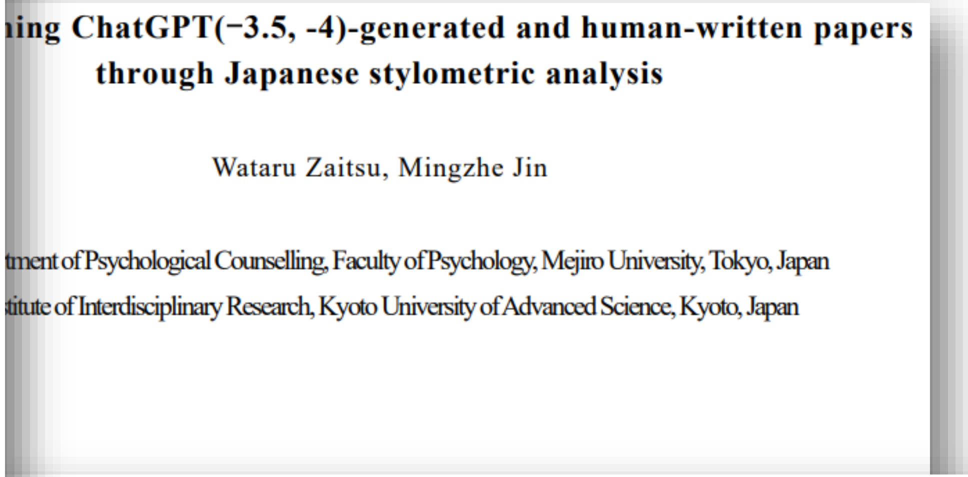
Sandra Mitrović[1], Davide Andreoletti[2], and Omran Ayoub[2]
[1]Dalle Molle Institute for Artificial Intelligence - University of Southern Switzerland and University of Applied Sciences and Arts of Southern Switzerland, Switzerland
[2]Information Systems and Networking Institute, University of Applied Sciences and Arts of Southern Switzerland, Switzerland

Abstract

ChatGPT has the ability to generate grammatically flawless and seemingly-human replies to different types of questions from various domains. The number of its users and of its applications is growing at an unprecedented rate. Unfortunately, use and abuse come hand in hand. In this paper

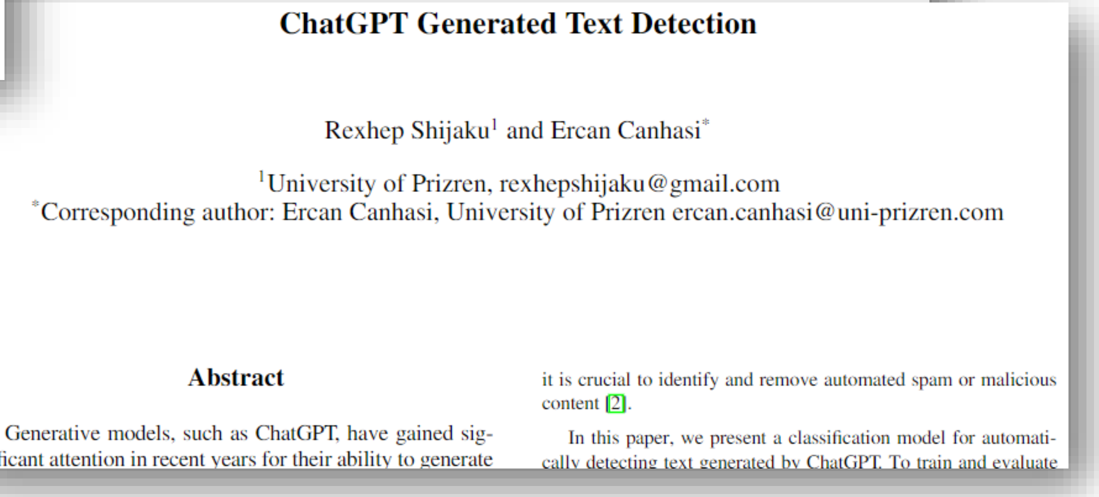
Detection of AI-written restaurant reviews

Focus on Japanese texts



ing ChatGPT(−3.5, -4)-generated and human-written papers through Japanese stylometric analysis

Wataru Zaitsu, Mingzhe Jin
tment of Psychological Counselling, Faculty of Psychology, Mejiro University, Tokyo, Japan
stitute of Interdisciplinary Research, Kyoto University of Advanced Science, Kyoto, Japan



ChatGPT Generated Text Detection

Rexhep Shijaku[1] and Ercan Canhasi[*]
[1]University of Prizren, rexhepshijaku@gmail.com
[*]Corresponding author: Ercan Canhasi, University of Prizren ercan.canhasi@uni-prizren.com

Abstract

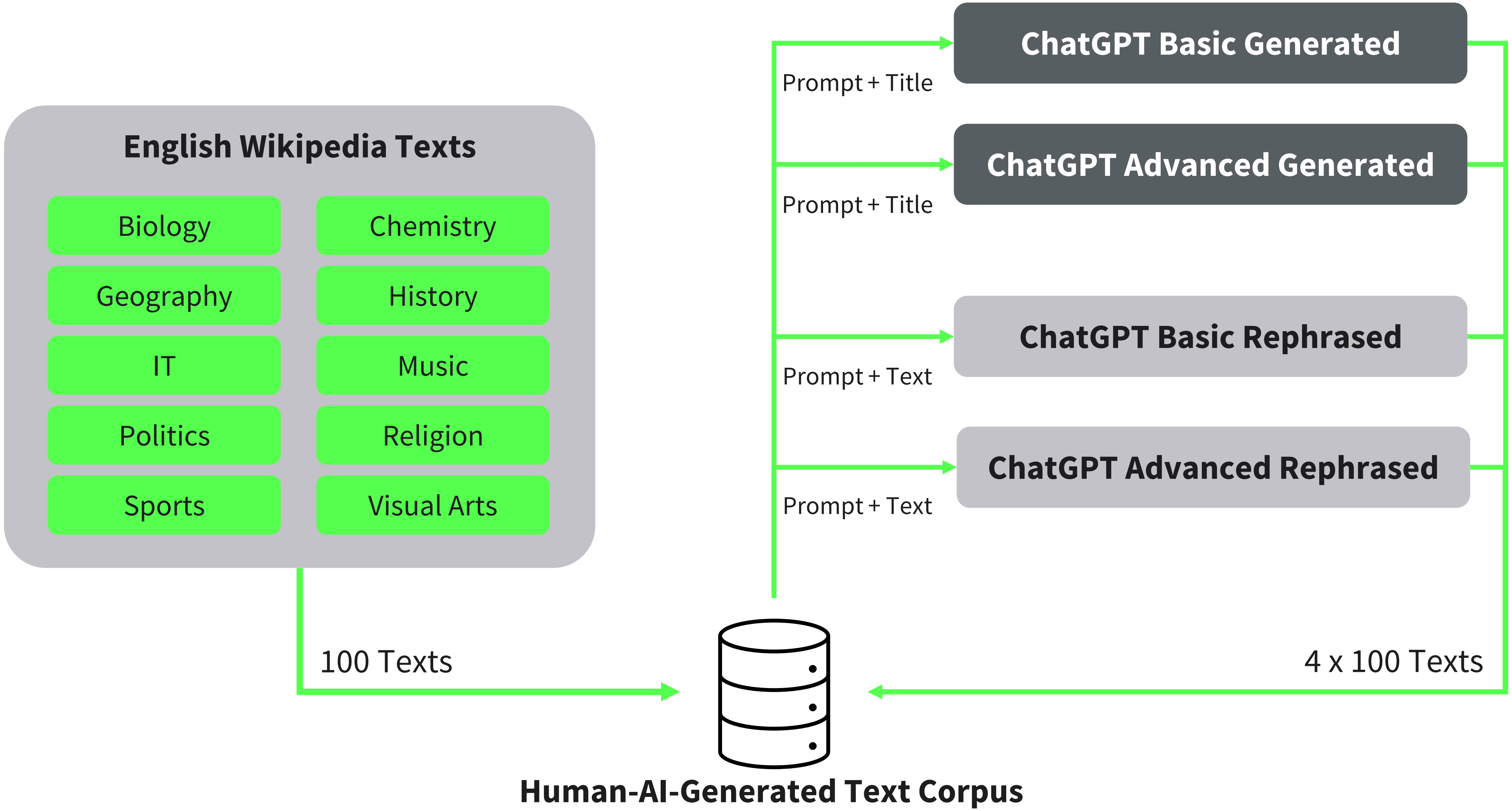
Generative models, such as ChatGPT, have gained significant attention in recent years for their ability to generate

it is crucial to identify and remove automated spam or malicious content [2].

In this paper, we present a classification model for automatically detecting text generated by ChatGPT. To train and evaluate

98% accuracy for detecting AI-generated essays

HUMAN-AI-GENERATED TEXT CORPUS



Corpus available on GitHub:
https://github.com/LorenzM97/human-AI-generatedTextCorpus

HUMAN-AI-GENERATED TEXT CORPUS

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 […].

FEATURES

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

FEATURES

Perplexity Features

Semantic Features

List Lookup Features

Document Features

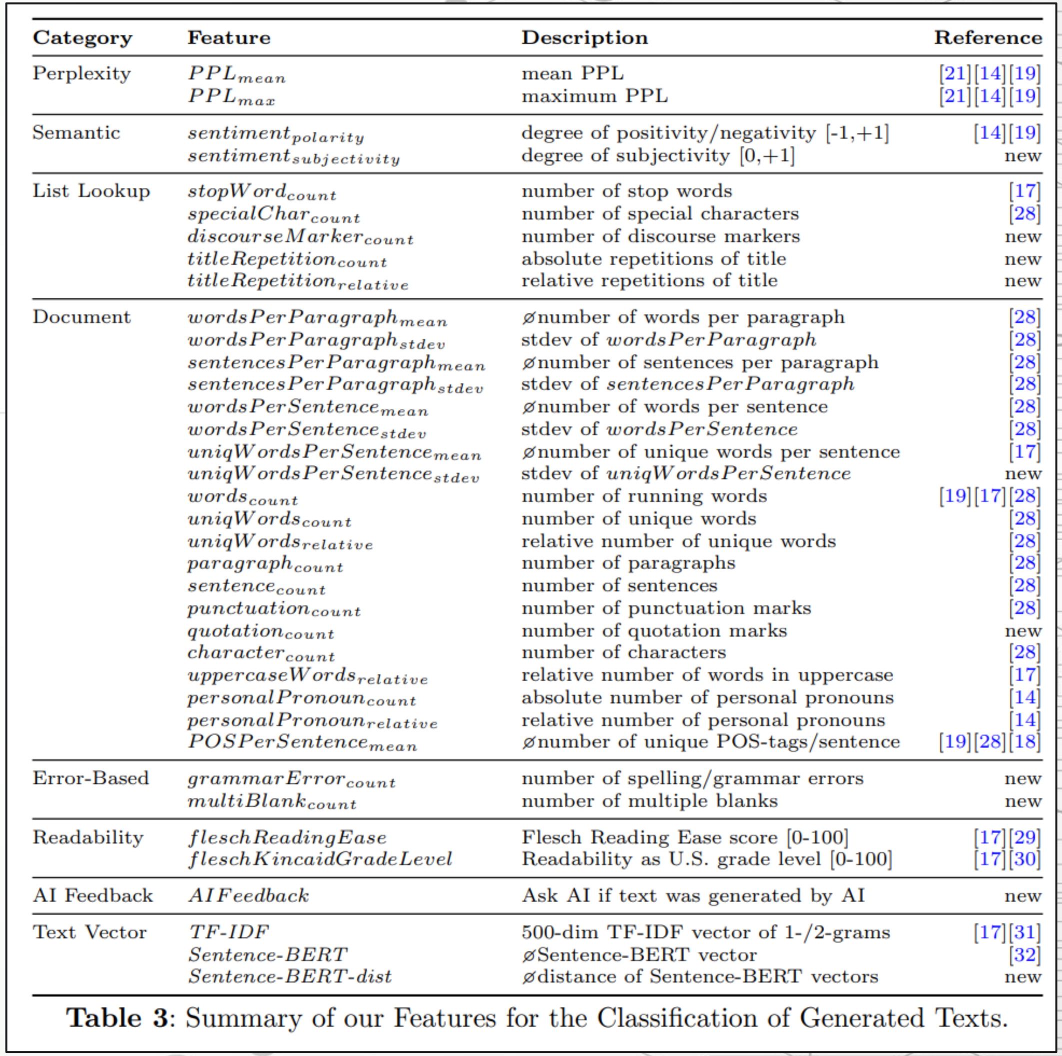
Error-Based Features

Readability Features

AI Feedback Features

Text Vector Features

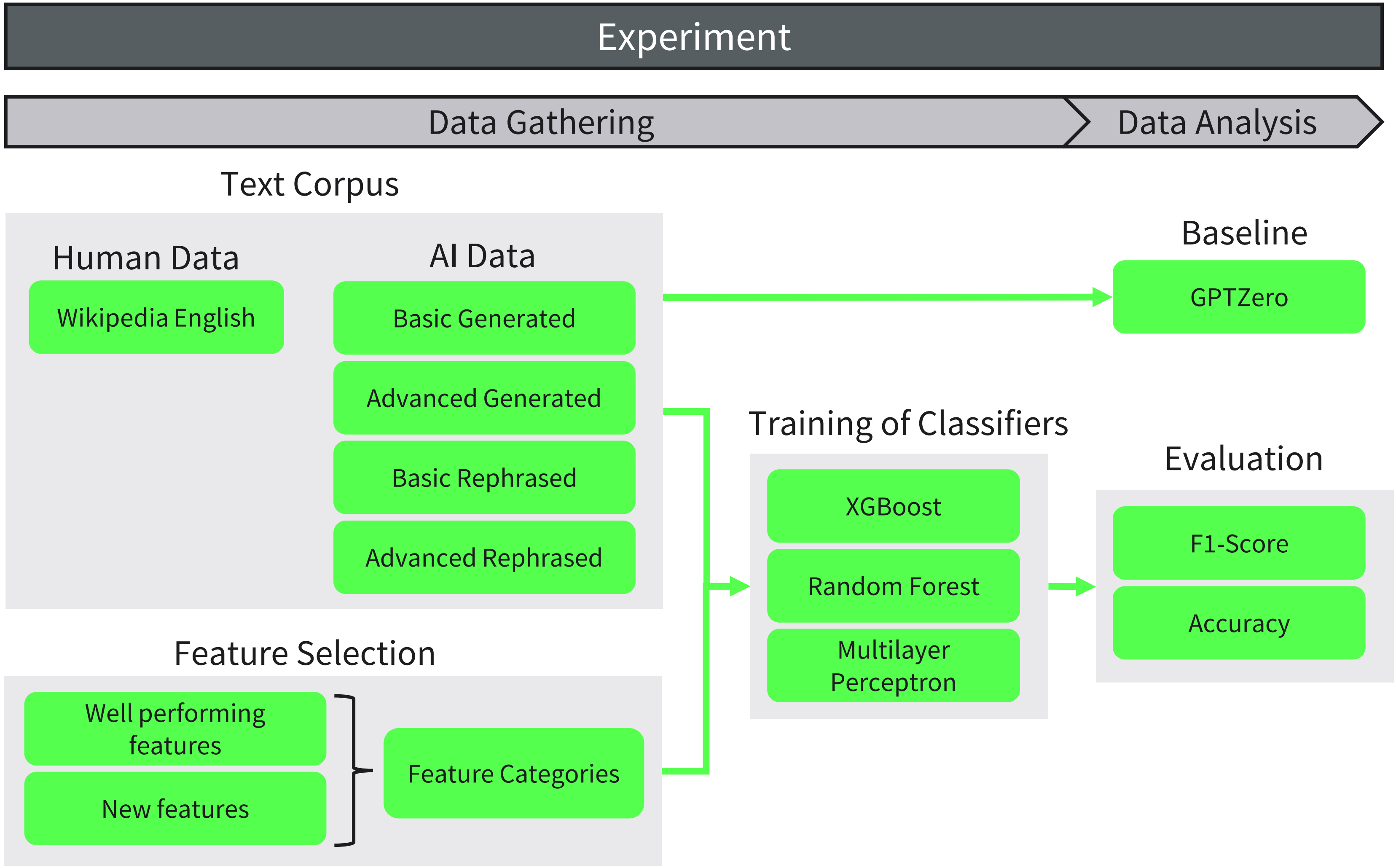
8 feature categories, 37 features



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	$wordsPerParagraph_{mean}$	⌀number of words per paragraph	[28]
	$wordsPerParagraph_{stdev}$	stdev of $wordsPerParagraph$	[28]
	$sentencesPerParagraph_{mean}$	⌀number of sentences per paragraph	[28]
	$sentencesPerParagraph_{stdev}$	stdev of $sentencesPerParagraph$	[28]
	$wordsPerSentence_{mean}$	⌀number of words per sentence	[28]
	$wordsPerSentence_{stdev}$	stdev of $wordsPerSentence$	[28]
	$uniqWordsPerSentence_{mean}$	⌀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]
	$POSPerSentence_{mean}$	⌀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	$TF\text{-}IDF$	500-dim TF-IDF vector of 1-/2-grams	[17][31]
	$Sentence\text{-}BERT$	⌀Sentence-BERT vector	[32]
	$Sentence\text{-}BERT\text{-}dist$	⌀distance of Sentence-BERT vectors	new

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

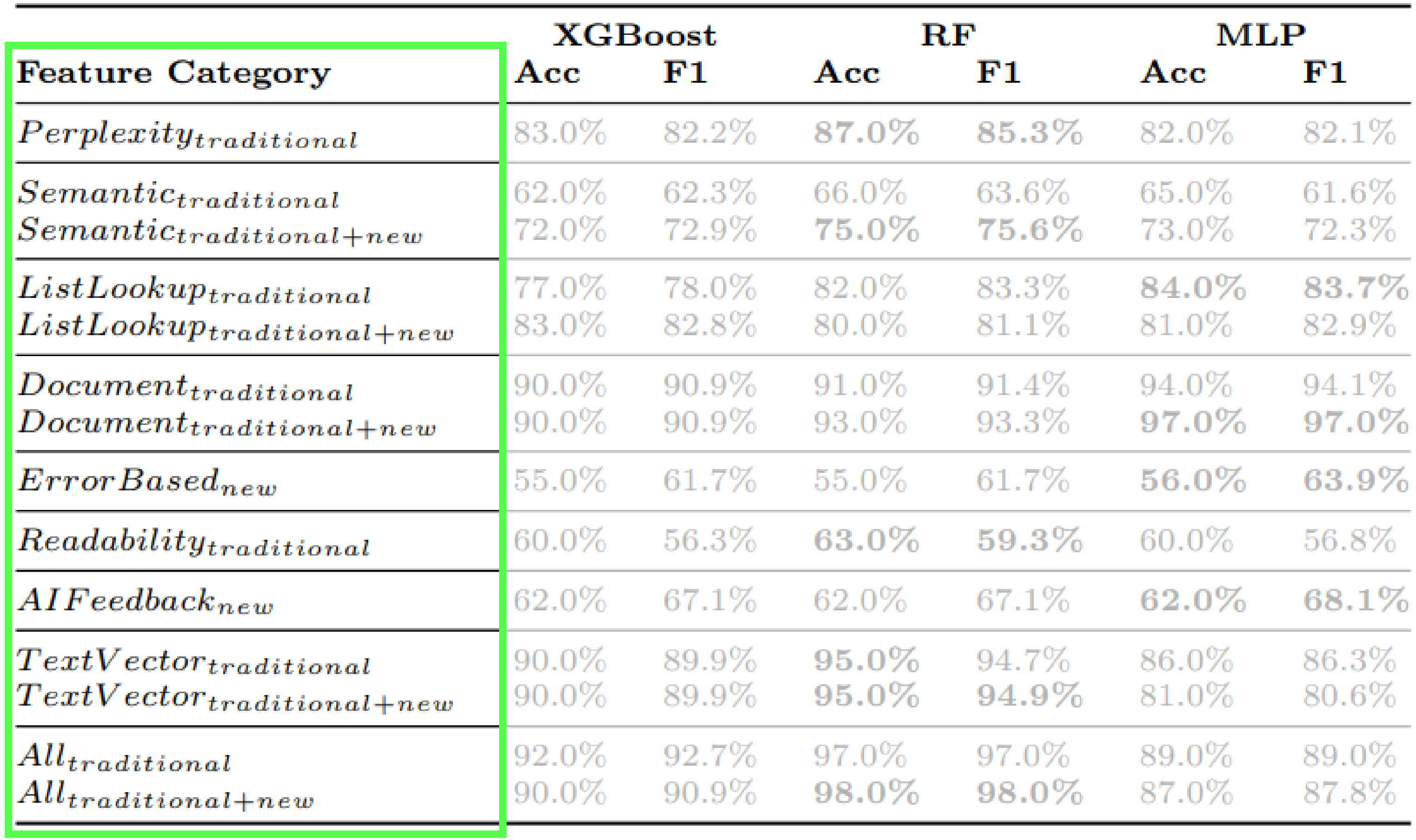
EXPERIMENTS AND RESULTS
OUR APPROACH



Training for:

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

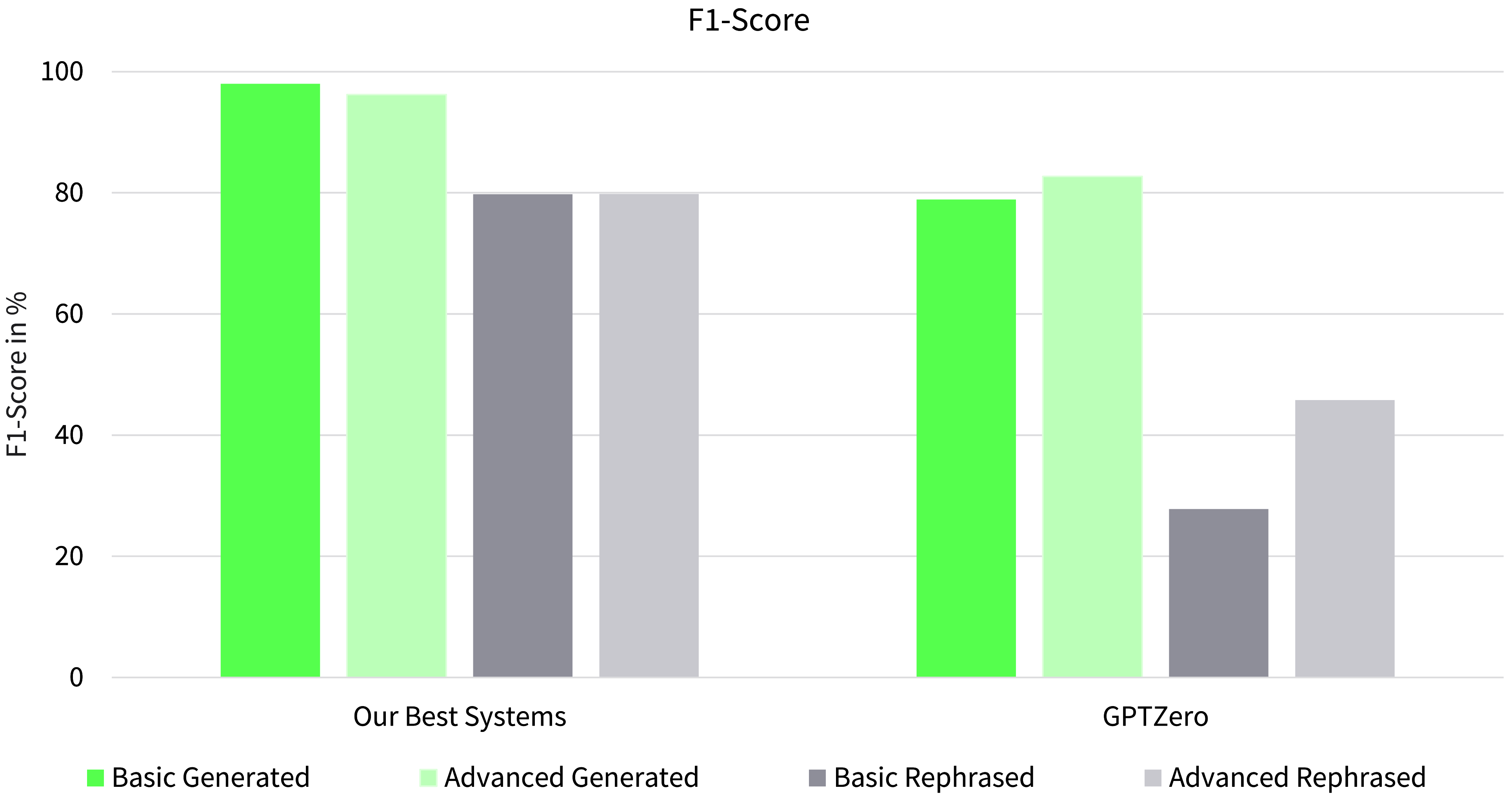
RESULTS



	XGBoost		RF		MLP	
Feature Category	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
$Perplexity_{traditional}$	83.0%	82.2%	**87.0%**	**85.3%**	82.0%	82.1%
$Semantic_{traditional}$	62.0%	62.3%	66.0%	63.6%	65.0%	61.6%
$Semantic_{traditional+new}$	72.0%	72.9%	**75.0%**	**75.6%**	73.0%	72.3%
$ListLookup_{traditional}$	77.0%	78.0%	82.0%	83.3%	**84.0%**	**83.7%**
$ListLookup_{traditional+new}$	83.0%	82.8%	80.0%	81.1%	81.0%	82.9%
$Document_{traditional}$	90.0%	90.9%	91.0%	91.4%	94.0%	94.1%
$Document_{traditional+new}$	90.0%	90.9%	93.0%	93.3%	**97.0%**	**97.0%**
$ErrorBased_{new}$	55.0%	61.7%	55.0%	61.7%	**56.0%**	**63.9%**
$Readability_{traditional}$	60.0%	56.3%	**63.0%**	**59.3%**	60.0%	56.8%
$AIFeedback_{new}$	62.0%	67.1%	62.0%	67.1%	**62.0%**	**68.1%**
$TextVector_{traditional}$	90.0%	89.9%	**95.0%**	94.7%	86.0%	86.3%
$TextVector_{traditional+new}$	90.0%	89.9%	**95.0%**	**94.9%**	81.0%	80.6%
$All_{traditional}$	92.0%	92.7%	97.0%	97.0%	89.0%	89.0%
$All_{traditional+new}$	90.0%	90.9%	**98.0%**	**98.0%**	87.0%	87.8%

Table 4: Results for Basic Text Generation: XGBoost vs. RF vs. MLP ($Acc_{GPTZero} = 76.0\%$, $F1_{GPTZero} = 78.9\%$).

RESULTS


F1-Score
F1-Score in %
100
80
60
40
20
0
Our Best Systems
GPTZero
Basic Generated
Advanced Generated
Basic Rephrased
Advanced Rephrased

CONCLUSION & FUTURE WORK

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

Generated > Rephrased

Best F1-score for AI-generated texts: **98%**

Best F1-score for AI-rephrased texts: **78%**

GPTZero < Our Systems

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

Future Work

Improvement of text generation

Investigation of other domains & languages

THANK YOU

Human-AI-Generated Text Corpus

REFERENCES

[1] Pelau, C., Dabija, D.-C., Ene, I.: What Makes an AI Device Human-Like? The Role of Interaction Quality, Empathy and Perceived Psychological Anthropomorphic Characteristics in the Acceptance of Artificial Intelligence in the Service Industry. Computers in Human Behavior **122**, 106855 (2021)

[2] Adiwardana, D., Luong, M.-T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., Le, Q.V.: Towards a Human-Like Open-Domain Chatbot. ArXiv Preprint ArXiv:2001.09977 (2020)

[3] Dibitonto, M., Leszczynska, K., Tazzi, F., Medaglia, C.M.: Chatbot in a Campus Environment: Design of LiSA, a Virtual Assistant to Help Students in Their University Life. In: Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part III 20, pp. 103–116 (2018). Springer

[4] Arteaga, D., Arenas, J., Paz, F., Tupia, M., Bruzza, M.: Design of Information System Architecture for the Recommendation of Tourist Sites in the City of Manta, Ecuador through a Chatbot. In: 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6 (2019). IEEE

[5] Falala-Séchet, C., Antoine, L., Thiriez, I., Bungener, C.: OWLIE: A Chatbot that Provides Emotional Support for Coping With Psychological Difficulties. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, pp. 236–237 (2019)

[6] Taecharungroj, V.: "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data and Cognitive Computing **7**(1), 35 (2023)

[7] Baidoo-Anu, D., Owusu Ansah, L.: Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Available at SSRN 4337484 (2023)

[8] Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A.T., Topalis, J., Weber, T., Wesp, P., Sabel, B., Ricke, J., Ingrisch, M.: ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. ArXiv E-Prints (2022)

[9] Jiao, W., Wang, W., Huang, J.-t., Wang, X., Tu, Z.: Is ChatGPT a Good Translator? A Preliminary Study. ArXiv Preprint ArXiv:2301.08745 (2023)

[10] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. CoRR **abs/2005.14165** (2020)

[11] Kenton, J.D.M.-W.C., Toutanova, L.K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)

[12] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR **abs/1907.11692** (2019) 1907.11692

[13] Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P.J., Narang, S., Li, W., Zhou, Y.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Technical report, Google (2019)

[14] Mitrović, S., Andreoletti, D., Ayoub, O.: ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-Generated Text. arXiv preprint arXiv:2301.13852 (2023)

[15] Soni, M., Wade, V.: Comparing Abstractive Summaries Generated by ChatGPT to Real Summaries Through Blinded Reviewers and Text Classification Algorithms (2023)

[16] OpenAI: New AI Classifier for Indicating AI-written Text (2023). https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text Accessed 04/21/2023

[17] Shijaku, R., Canhasi, E.: ChatGPT Generated Text Detection (2023). https://doi.org/10.13140/RG.2.2.21317.52960

[18] Zaitsu, W., Jin, M.: Distinguishing ChatGPT(-3.5, -4)-generated and Human-Written Papers Through Japanese Stylometric Analysis (2023)

[19] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y.: How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection (2023)

[20] Vu, N.T., Schlippe, T., Kraus, F., Schultz, T.: Rapid Bootstrapping of Five Eastern European Languages Using the Rapid Language Adaptation Toolkit. In: Interspeech (2010)

[21] Gehrmann, S., Strobelt, H., Rush, A.: GLTR: Statistical Detection and Visualization of Generated Text. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 111–116. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-3019

[22] Bird, S., Loper, E.: NLTK: The Natural Language Toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 214–217. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/P04-3031

[23] Wankhade, M., Rao, A., Kulkarni, C.: A Survey on Sentiment Analysis Methods, Applications, and Challenges. Artificial Intelligence Review, 1–50 (2022) https://doi.org/10.1007/s10462-022-10144-1

[24] Rakhmanov, O., Schlippe, T.: Sentiment Analysis for Hausa: Classifying Students' Comments. In: SIGUL 2022, Marseille, France (2022)

[25] Mabokela, K.R., Schlippe, T.: AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa. In: SACAIR (2022)

[26] Natalie: What is ChatGPT? (2023). https://help.openai.com/en/articles/6783457-what-is-chatgpt Accessed 04/21/2023

[27] Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes **30**(1), 3–26 (2007) https://doi.org/10.1075/li.30.1.03nad

[28] Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., Liu, H.: Stylometric Detection of AI-Generated Text in Twitter Timelines (2023)